Quantized Overcomplete Expansions in $\mathbb{R}^N$: Analysis, Synthesis, and Algorithms

Vivek K Goyal, *Student Member, IEEE*, Martin Vetterli, *Fellow, IEEE*, and Nguyen T. Thao, *Member, IEEE*

***Abstract*—Coefficient quantization has peculiar qualitative effects on representations of vectors in $\mathbb{R}^N$ with respect to overcomplete sets of vectors. These effects are investigated in two settings: frame expansions (representations obtained by forming inner products with each element of the set) and matching pursuit expansions (approximations obtained by greedily forming linear combinations). In both cases, based on the concept of *consistency*, it is shown that traditional linear reconstruction methods are suboptimal, and better consistent reconstruction algorithms are given. The proposed consistent reconstruction algorithms were in each case implemented, and experimental results are included. For frame expansions, results are proven to bound distortion as a function of frame redundancy r and quantization step size for linear, consistent, and optimal reconstruction methods. Taken together, these suggest that optimal reconstruction methods will yield $O(1/r^2)$ mean-squared error (MSE), and that consistency is sufficient to insure this asymptotic behavior. A result on the asymptotic tightness of random frames is also proven. Applicability of quantized matching pursuit to lossy vector compression is explored. Experiments demonstrate the likelihood that a linear reconstruction is inconsistent, the MSE reduction obtained with a nonlinear (consistent) reconstruction algorithm, and generally competitive performance at low bit rates.**

***Index Terms*— Consistent reconstruction, frames, matching pursuit, MSE bounds, optimal reconstruction, overcomplete representations, quantization, source coding.**

I. Introduction

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 intricate when signal representations are with respect to redundant, or overcomplete, sets of vectors. This paper considers the effects of quantization in overcomplete finite linear expansions. Both fixed and adaptive basis methods are studied. Although it represents an input vector as a linear combination of elements from a representation set, the adaptive basis method is in fact a nonlinear mapping. While many other issues are explored, the unifying theme is that consistent reconstruction methods [1] give considerable improvement over linear reconstruction methods.

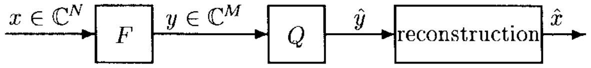


Fig. 1. Block diagram of reconstruction from quantized frame expansion.

Consider the expansion–quantization–reconstruction scenario depicted in Fig. 1. A vector $x \in \mathbb{C}^N$ is left-multiplied by a matrix $F \in \mathbb{C}^{M \times N}$ of rank N to get $y \in \mathbb{C}^M$. The transformed source vector y is scalar quantized, i.e., quantized with a quantizer which acts separately on each component of y, to get $\hat{y}$. As shown in Section II-A.2, this type of representation arises naturally in simple oversampled A/D conversion. In general, this sort of representation may be desirable when many coarse measurements can be made easily, but precise measurements are difficult to make. How can one best estimate x from $\hat{y}$? How does the quality of the estimate $\hat{x}$ depend on the properties of F, in particular its number of rows M? These are the fundamental questions addressed in Section II.

To put this in a solid framework, we review the basic properties of frames and prove a new result on the tightness of random frames. We then show that the quality of reconstruction can be improved by using deterministic properties of quantization (consistent reconstruction), as opposed to considering quantization to be the addition of noise that is independent in each dimension. The relationship between the redundancy of the frame and the minimum possible reconstruction error is explored.

Without sophisticated coding, a nonadaptive overcomplete expansion can be a very inefficient representation. In the context of Fig. 1, coding $\hat{y}$ may be an inefficient way to represent x. But could we get a good representation if we could choose a few components of $\hat{y}$ *a posteriori* which best describe x? This question is related to adaptive basis techniques described in Section III.

In Section III, the use of a greedy successive approximation algorithm for finding sparse linear representations with respect to an overcomplete set is studied. This algorithm, called matching pursuit (MP) [2], has recently been applied to image coding [3], [4] and video coding [5], [6], which inherently

Manuscript received March 15, 1996; revised May 15, 1997. This work was supported in part by the U.S. Advanced Research Projects Agency through a National Defense Science and Engineering Graduate Fellowship, the U.S. National Science Foundation under Grant MIP-93-21302, and the Swiss National Science Foundation under Grant 2100-043136.95. The material in this paper was presented in part at the IEEE Data Compression Conference, Snowbird, UT, March 28–30, 1995, and at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, May 7–10, 1996.

V. K. Goyal is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA.

M. Vetterli is with the Laboratoire de Communications Audiovisuelles, Département d'Électricité, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, and the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA.

N. T. Thao is with the Department of Electrical and Electronic Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.

Publisher Item Identifier S 0018-9448(98)00080-7.

require coarse coefficient quantization. However, to the best of our knowledge, the present work is the first to describe the qualitative effects of coefficient quantization in matching pursuit. In particular, as in Section II, we will find that reconstruction can be improved by consistent reconstruction techniques.

Except where noted, we consider vectors in a finite dimensional Hilbert space $H = \mathbb{R}^N$ or $\mathbb{C}^N$. For $x, y \in H$, we use the inner product $\langle x, y \rangle = x^T \overline{y}$ and the norm derived from the inner product through $\|x\| = \langle x, x \rangle^{1/2}$. $\mathcal{N}(\mu, \Lambda)$ is used to denote the Normal distribution with mean μ and covariance matrix Λ. The term squared error (SE) is used for the square of the norm of the difference between a vector and an estimate of the vector. The term mean-squared error (MSE) is reserved for the ensemble average of SE or expected SE.

II. Nonadaptive Expansions

This section describes frames, which provide a general framework for studying nonadaptive linear transforms. Frames were introduced by Duffin and Schaeffer [7] in the context of nonharmonic Fourier series. Recent interest in frames has been spurred by their utility in analyzing discrete wavelet transforms [8]–[10] and time–frequency decompositions [11]. We are motivated by a desire to understand quantization effects and efficient representations.

Section II-A begins with definitions and examples of frames. It concludes with a theorem on the tightness of random frames and a discussion of that result. Section II-B 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 [11] considered a particular type of frame and assumed the coefficients were subject to a stationary noise. This paper, on the other hand, is in the same spirit as [1], and [12]–[15] in that it utilizes the deterministic qualities of quantization.

A. Frames

1) Definitions and Basics: This subsection is largely adapted from [10, ch. 3]. Some definitions and notations have been simplified because we are limiting our attention to $H = \mathbb{R}^N$ or $\mathbb{C}^N$.

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

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

A and B are called the *frame bounds*. The cardinality of K is denoted by M. The lower bound in (1) is equivalent to requiring that Φ spans H. Thus a frame will always have $M \geq N$. Also notice that one can choose $B = \sum_{k \in K} \|\varphi_k\|^2$ whenever $M < \infty$. We will refer to $r = M/N$ as the *redundancy* of the frame. A frame Φ is called a *tight frame* if the frame bounds can be taken to be equal. It is easy to verify that if Φ is a tight frame with $\|\varphi_k\| = 1$ for all $k \in K$, then $A = r$.

Given a frame $\Phi = \{\varphi_k\}_{k \in K}$ in H, the associated *frame operator* F is the linear operator from H to $\mathbb{C}^M$ defined by

$$(Fx)_k = \langle x, \varphi_k \rangle. \tag{2}$$

Since H is finite-dimensional, this operation is a matrix multiplication where F is a matrix with kth row equal to φ_k^*. Using the frame operator, (1) can be rewritten as

$$AI_N \leq F^*F \leq BI_N \tag{3}$$

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, $F^*F = AI_N$ if and only if Φ is a tight frame. From (3) 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 (3) that F^*F is invertible because all of its eigenvalues are nonzero, and furthermore

$$B^{-1}I_N \leq (F^*F)^{-1} \leq A^{-1}I_N. \tag{4}$$

The *dual frame* of Φ is defined as $\tilde{\Phi} = \{\tilde{\varphi}_k\}_{k \in K}$, where

$$\tilde{\varphi}_k = (F^*F)^{-1}\varphi_k, \qquad \forall\, k \in K. \tag{5}$$

$\tilde{\Phi}$ is itself a frame with frame bounds B^{-1} and A^{-1}.

Since $\operatorname{Span}(\Phi) = H$, any vector $x \in H$ can be written as

$$x = \sum_{k \in K} \alpha_k \varphi_k \tag{6}$$

for some set of coefficients $\{\alpha_k\} \subset \mathbb{R}$. If $M > N$, $\{\alpha_k\}$ may not be unique. We refer to (6) as a *redundant representation* even though it is not necessary that more than N of the α_k's be nonzero.

2) Example: 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.[1] So in some sense, it is easier to find a frame than to give an example of a set of vectors which do not form a frame. In this section we give a single example of a structured family of frames. We will prove certain properties of these frames in Section II-B4.

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

[1] 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.

conversion [1]. Let $x = [x_1\ x_2\ \cdots\ x_N]^T \in \mathbb{R}^N$, with N odd. Define a corresponding continuous-time signal by

$$x_c(t) = x_1 + \sum_{k=1}^{W} \left[x_{2k}\sqrt{2}\cos\frac{2\pi kt}{T} + x_{2k+1}\sqrt{2}\sin\frac{2\pi kt}{T} \right] \tag{7}$$

where $W = (N-1)/2$. 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 $x_d[m] = x_c(mT/M)$ and let

$$y = [x_d[0]\ x_d[1]\ \cdots\ x_d[M-1]]^T.$$

Then we have $y = Fx$, where

$$F = [\varphi_1\quad \varphi_2\quad \cdots\quad \varphi_M]^T$$

with

$$\varphi_k = \left[1\quad \sqrt{2}\cos\frac{2\pi k}{M}\quad \sqrt{2}\sin\frac{2\pi k}{M}\right.$$
$$\left.\cdots\quad \sqrt{2}\cos\frac{2\pi Wk}{M}\quad \sqrt{2}\sin\frac{2\pi Wk}{M}\right]^T. \tag{8}$$

Using the orthogonality properties of sine and cosine, it is easy to verify that $F^*F = MI_N$, so F is an operator associated with a tight frame. Pairing terms and using the identity $\cos^2\theta + \sin^2\theta = 1$, 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.

3) Tightness of Random Frames: Tight frames constitute an important class of frames. As we will see in Section II-B1, since a tight frame is self-dual, it has some desirable reconstruction properties. These extend smoothly to nearly tight frames, i.e., frames with B/A close to one. Also, for a tight frame (1) 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 Section II-B4 that some properties of "typical" frame operators depend only on the redundancy. This motivates our interest in the following theorem.

Theorem 1. Tightness of Random Frames: Let $\{\Phi_M\}_{M=N}^{\infty}$ be a sequence of frames in $\mathbb{R}^N$ such that Φ_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 Φ_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 Appendix I-A. □

Theorem 1 shows that a sequence of random frames with increasing redundancy will approach a tight frame. Note that although the proof in Appendix I-A uses an unrelated strategy, the constant $1/N$ is intuitive: If Φ_M is a tight frame with normalized elements, then we have $F_M^*F_M = (M/N)I_N$ 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 Fig. 2 are averaged results for 1000 sequences of frames in $\mathbb{R}^4$. Fig. 2 shows that A/M and B/M converge to $1/N$ and that B/A converges to one.

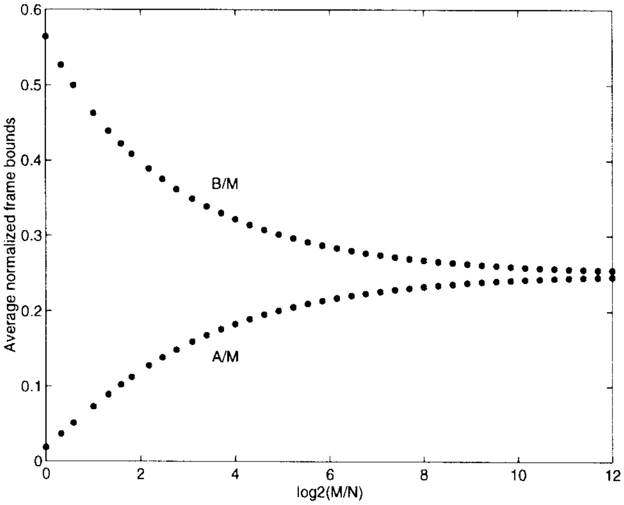


Fig. 2. Normalized frame bounds for random frames in $\mathbb{R}^4$.

B. Reconstruction from Frame Coefficients

One cannot rightly call a frame expansion a "signal representation" without considering the viability of reconstructing the original signal. This is the problem that we address presently.

In Section II-B1, we review the basic properties of reconstructing from (unquantized) frame coefficients. This material is adapted from [10]. 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 [1]. Consistent reconstruction methods yield smaller reconstruction errors than classical methods.

1) Unquantized Case: Let Φ be a frame and assume the notation of Section II-A1. In this subsection, we consider the problem of recovering x from $\{\langle x, \varphi_k\rangle\}_{k\in K}$. Let $\tilde{F}$: $H \rightarrow \mathbb{C}^M$ be the frame operator associated with $\tilde{\Phi}$. It can be shown [10, Proposition 3.2.3] that $\tilde{F}^*F = I_N$. Thus a possible reconstruction formula is given by

$$x = \tilde{F}^*Fx = \sum_{k\in K}\langle x, \varphi_k\rangle\tilde{\varphi}_k.$$

This formula is reminiscent of reconstruction from a discrete Fourier transform (DFT) representation, in which case

$$\varphi_k = \tilde{\varphi}_k = 1/\sqrt{N}[1\ e^{j2\pi k/N}\ \cdots\ e^{j2\pi k(N-1)/N}]^T.$$

In the DFT and inverse DFT, one set of vectors plays the roles of both Φ and $\tilde{\Phi}$ because it is a tight frame in $\mathbb{C}^N$. Other reconstruction formulas are possible; for details the reader is referred to [10, Sec. 3.2].

2) Classical Method: We now turn to the question of reconstructing when the frame coefficients $\{\langle x, \varphi_k\rangle\}_{k\in K}$ 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 x given $Fx+\beta$, where $\beta \in \mathbb{C}^M$ is a zero-mean noise, uncorrelated with x. The key to finding the best approximation is that $FH = \operatorname{Ran}(F)$ is an N-dimensional subspace of $\mathbb{C}^M$. Hence the component of β perpendicular to FH should not hinder our approximation, and the best approximation is the projection of $Fx+\beta$ onto $\operatorname{Ran}(F)$. By [10, Proposition 3.2.3], this approximation is given by

$$\hat{x} = \tilde{F}^*(Fx+\beta). \tag{9}$$

Furthermore, because the component of β orthogonal to $\operatorname{Ran}(F)$ does not contribute, we expect $\|x-\hat{x}\| = \|\tilde{F}^*\beta\|$ to be smaller than $\|\beta\|$. The following proposition makes this more precise.

Proposition 1. Noise Reduction in Linear Reconstruction: Let $\{\varphi_k\}_{k=1}^M$ 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 β_i's are independent random variables with mean zero and variance σ^2. Then the MSE of the classical reconstruction (9) satisfies

$$\frac{M\sigma^2}{B^2} \leq \text{MSE} \leq \frac{M\sigma^2}{A^2}.$$

Proof: See Appendix I-B. □

Corollary 1: If the frame in Proposition 1 is tight

$$\text{MSE} = \frac{N^2\sigma^2}{M} = \frac{N\sigma^2}{r}.$$

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.,

$$Q(y) = [q_1(y_1)\ q_2(y_2)\ \cdots\ q_M(y_M)]^T$$

where q_i: $\mathbb{R} \to \mathbb{R}$, $1 \leq i \leq M$, 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 $\hat{x} = \tilde{F}^*\hat{y}$ 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.

3) Consistent Reconstruction:

Definition 1. Consistency [1]: Let f: $X \to Y$. Let $x \in X$ and $y = f(x)$. If $f(\hat{x}) = y$ then $\hat{x} \in X$ is called a *consistent estimate of x from y.* An algorithm that produces consistent estimates is called a *consistent reconstruction* algorithm. An estimate that is not consistent is said to be *inconsistent.*

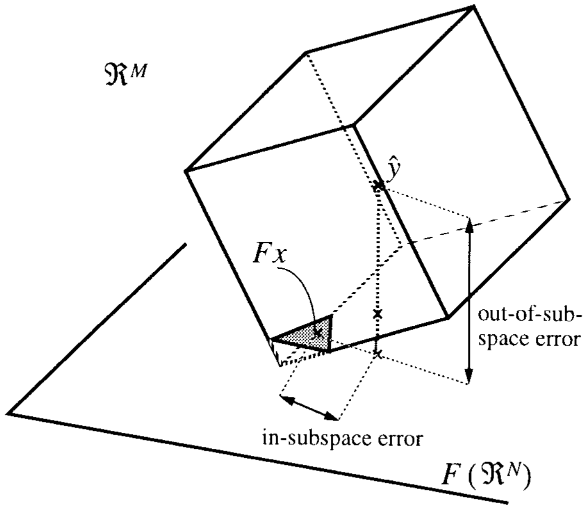


Fig. 3. Illustration of consistent reconstruction.

The essence of consistency is that $\hat{x}$ is a consistent estimate if it is compatible with the observed value of y, i.e., it is possible that $\hat{x}$ is exactly equal to x. In the case of quantized frame expansions $f = Q\circ F$, and one can give a geometric interpretation. Q induces a partitioning of $\mathbb{R}^M$, which in turn induces a partitioning of $\mathbb{R}^N$ through the inverse map of $Q\circ F$. A consistent estimate is simply one that falls in the same partition region as the original signal vector. These concepts are illustrated for $N = 2$ and $M = 3$ in Fig. 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. Further geometric interpretations of quantized frame expansions are given in Appendix II.

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 [16]. In this 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 and each component quantizer is uniform, a linear program can be used to find consistent estimates. For $i = 1, 2, \cdots, M$, denote the quantization stepsize in the ith component by Δ_i. For notational convenience, assume that the reproduction values lie halfway between decision levels. Then for each i, $|\hat{y}_i - y_i| \leq \Delta_i/2$. To obtain a consistent estimate, for each i we must have $|(F\hat{x})_i - \hat{y}_i| \leq \Delta_i/2$. Expanding the absolute value, we find the constraints

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

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

$$\begin{bmatrix} F \\ -F \end{bmatrix}\hat{x} \leq \begin{bmatrix} \frac{1}{2}\Delta + \hat{y} \\ \frac{1}{2}\Delta - \hat{y} \end{bmatrix}. \tag{10}$$

TABLE I
ALGORITHM FOR CONSISTENT RECONSTRUCTION FROM A QUANTIZED FRAME EXPANSION

1. Form
$$\overline{F} = \begin{bmatrix} F \\ -F \end{bmatrix} \quad \text{and} \quad \overline{y} = \begin{bmatrix} \frac{1}{2}\Delta + \hat{y} \\ \frac{1}{2}\Delta - \hat{y} \end{bmatrix}.$$
2. Pick an arbitrary cost function $c \in \mathbb{R}^N$.
3. Use a linear programming method to find $\hat{x}$ to minimize $c^T\hat{x}$ subject to $\overline{F}\hat{x} \le \overline{y}$.

The formulation (10) shows that $\hat{x}$ can be determined through linear programming [17]. The feasible set of the linear program is exactly the set of consistent estimates, so an arbitrary cost function can be used. This is summarized in Table I.

A linear program always returns a corner of the feasible set [17, Sec. 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 (9). 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.

4) Error Bounds for Consistent Reconstruction: In orthogonal representations, it is well understood that under very general conditions, the MSE is $O(\Delta^2)$ for small Δ. For frame expansions, how does the MSE depend on r, for large r, and how does it depend on the reconstruction method? The MSE obtained with any reconstruction method depends in general on the distribution of the source. The evidence suggests that any consistent reconstruction algorithm is essentially optimal, in a sense made clear by the following propositions, and gives $O(1/r^2)$ MSE.

Proposition 2. MSE Lower Bound: Let x be a random variable with probability density function $\boldsymbol{p}$ with support on a bounded subset $\mathcal{B}$ of $\mathbb{R}^N$. Consider any set of quantized frame expansions of x for which

$$\sup_M \max_{1 \le i \le M} (\|\varphi_i\|)/\Delta_i = d_0 < \infty.$$

Unless $\boldsymbol{p}$ is degenerate in a way which allows for exact reconstruction, any reconstruction algorithm will yield an MSE that can be lower-bounded by b/r^2, where b is a coefficient independent of r and a function of N, $\boldsymbol{p}$, the diameter D of $\mathcal{B}$, and the maximum density value d_0.

Proof: See Appendix I-C. □

Proposition 3. Squared-Error Upper Bound—DFT Case: Fix a quantization stepsize $\Delta \in \mathbb{R}^+$. For a sequence of quantized frame expansions of a fixed $x \in \mathbb{R}^N$ followed by consistent reconstruction, the squared error can be upper-bounded by an $O(1/r^2)$ expression under the following conditions:

i) $\underline{N \text{ odd}}$: The frame operators are as in (8) and
$$\|[x_2\, x_3\, \cdots\, x_N]^T\| > (N+1)\Delta/4$$
or

ii) $\underline{N \text{ even}}$: The frame operators are as in (8) with the first column removed and $\|x\| > (N+2)\Delta/4$.

Proof: See Appendix A-D. □

Conjecture 1. 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 sort of general upper bound to hold, some sort of nondegeneracy 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 such 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 (8). This is the case in which we have proven an $O(1/r^2)$ SE 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 [18].

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

Simulation results are given in Fig. 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 (10) 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 1, it may be useful to try to prove Conjecture 1 only for tight frames.

5) Rate-Distortion Tradeoffs: We have asserted that optimal reconstruction techniques give an MSE proportional to $1/r^2$, and the $O(\Delta^2)$ MSE for orthogonal representations extends to the frame case as well. Thus there are two ways to reduce the MSE by approximately a factor of four: double r or halve Δ. Our discussion has focused on expected distortion without concern for rate, and 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/N bits per sample. Doubling r gives $2M$ b-bit numbers, for a total rate of $2Mb/N$ bits per sample. On the other hand, halving Δ results in M $(b+1)$-bit numbers for a rate of only $M(b+1)/N$ bits per sample. This example suggests that

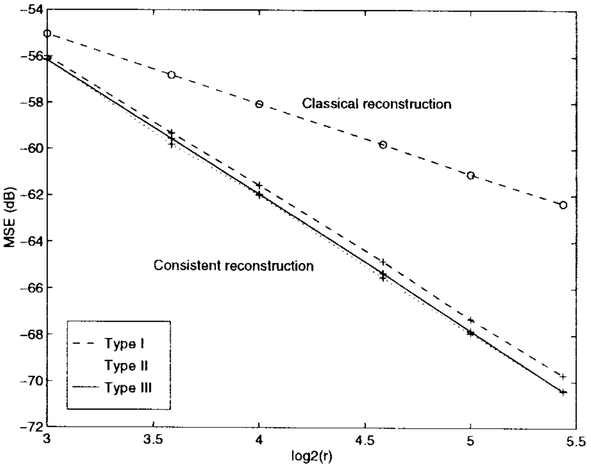


Fig. 4. Experimental results for reconstruction from quantized frame expansions. Shows $O(1/r^2)$ MSE for consistent reconstruction and $O(1/r)$ MSE for classical reconstruction.

halving Δ is always the better option, but a few comments are in order. One caveat is that in some situations, doubling r and halving Δ may have very different costs. For example, the much higher cost of halving Δ than of doubling r is a major motivating factor for oversampled A/D conversion. 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. This is discussed further in Appendix II.

III. Adaptive Expansions

Transform coding theory, as introduced in [19] and analyzed in detail in [20], is predicated on fine quantization approximations and assuming that signals are Gaussian. For most practical coding applications, these assumptions do not hold, so the wisdom of maximizing coding gain—which leads to the optimality of the Karhunen–Loève transform—has been questioned. More fundamentally, we can leave the usual framework of static orthogonal transform coding and consider the application of adaptive and nonlinear transforms.

The matching pursuit algorithm [2], described in Section III-A, has both adaptive and nonlinear aspects. Given a source vector x and a frame $\{\varphi_k\}_{k=1}^{M}$, it produces an approximate signal representation $x \approx \sum_{i=0}^{n-1} \alpha_i \varphi_{k_i}$. It is adaptive in the sense that the k_i's depend on x, yet it can be considered nonadaptive because it is time-invariant for transforming a sequence of source vectors. On the other hand, it has a linear nature because it produces a linear representation, but it is nonlinear because it does not satisfy additivity.[2]

The matching pursuit algorithm is a greedy algorithm for choosing a subset of the frame and finding a linear combination of that subset that approximates a given signal vector. The use of a greedy algorithm is justified by the computational intractability of finding the optimal subset of the original frame [21, ch. 2]. In our finite-dimensional setting, this is very similar to the problem of finding sparse approximate solutions to linear systems. In that context, this greedy heuristic is well-known and performance bounds have been proven [22].

Quantization of coefficients in matching pursuit leads to many interesting issues; some of these are discussed in Section III-B. Along with exploring general properties of matching pursuit, we are interested in its application to compressing data vectors in $\mathbb{R}^N$. A vector compression method based on matching pursuit is described in Section III-C.

A. Matching Pursuit

1) Algorithm: Let $\mathcal{D} = \{\varphi_k\}_{k=1}^{M} \subset H$ be a frame such that $\|\varphi_k\| = 1$ for all k. $\mathcal{D}$ is called the *dictionary*. Matching pursuit (MP) is an algorithm to represent $x \in H$ by a linear combination of elements of $\mathcal{D}$. In the first step of the algorithm, k_0 is selected such that $|\langle \varphi_{k_0}, x\rangle|$ is maximized. Then x can be written as its projection onto φ_{k_0} and a residue $R_1 x$

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

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

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

Identifying $R_0 x = x$, we can write

$$x = \sum_{i=0}^{n-1} \langle \varphi_{k_i}, R_i x\rangle \varphi_{k_i} + R_n x. \tag{11}$$

Hereafter, we will denote $\langle \varphi_{k_i}, R_i x\rangle$ by α_i. Note that the output of a matching pursuit expansion is not only the coefficients $(\alpha_0, \alpha_1, \cdots)$, but also the indices $(k_0, k_1, \cdots)$. For storage and transmission purposes, we will have to account for the indices.

Matching pursuit was introduced to the signal processing community in the context of time–frequency analysis by Mallat and Zhang [2]. Mallat and his coworkers have uncovered many of its properties [21], [23], [24].

2) Discussion: Since α_i is determined by projection, $\alpha_i \varphi_{k_i} \perp R_{i+1} x$. Thus we have the "energy conservation" equation

$$\|R_i x\|^2 = \|R_{i+1} x\|^2 + \alpha_i^2. \tag{12}$$

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 x is exactly represented.

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 algorithm which picks dictionary elements jointly would choose a basis from the dictionary and get an exact expansion in N steps. One way

[2] The usage of additivity is not obvious. Clearly if $x_1 \approx \sum_{i=0}^{n-1} \alpha_i \varphi_{k_i}$ and $x_2 \approx \sum_{i=0}^{n-1} \beta_i \varphi_{k_i}$, then

$$x_1 + x_2 \approx \sum_{i=0}^{n-1} (\alpha_i + \beta_i) \varphi_{k_i}.$$

But in general the expansions of x_1, x_2, and $x_1 + x_2$ would not use the same k_i's; for this reason the transform is nonlinear.

to speed convergence is to use an orthogonalized version of MP which at each step modifies the dictionary and chooses a dictionary element perpendicular to all previously chosen dictionary elements. Since orthogonalized matching pursuit does not converge significantly faster than the nonorthogonalized version for a small number of iterations [6], [21], [25], nonorthogonalized matching pursuit is not considered hereafter.

Matching pursuit has been found to be useful in source coding for two (related) reasons: The first reason—which was emphasized in the original Mallat and Zhang paper [2]—has been termed *flexibility*; the second is that the nonlinear approximation framework allows greater energy compaction than a linear transform.

MP is often said to have flexibility to differing signal structures. The archetypal illustration is that a Fourier basis provides a poor representation of functions well localized in time, while wavelet bases are not well suited to representing functions whose Fourier transforms have narrow, high-frequency support [2]. The implication is that MP, with a dictionary including a Fourier basis and a wavelet basis, would avoid these difficulties.

Looking at the energy compaction properties of MP gives a more extensive view of the potential of MP. Energy compaction refers to the fact that after an appropriately chosen transform, most of the energy of a signal can be captured by a small number of coefficients. In orthogonal transform coding, getting high-energy compaction is dependent on designing the transform based on knowledge of source statistics; for fine quantization of a stationary Gaussian source the Karhunen–Loève Transform is optimal [26]. Although both produce an approximation for a source vector which is a linear combination of basis elements, orthogonal transform coding contrasts sharply with MP in that the basis elements are chosen *a priori* and hence at best one can make the optimum basis choice *on average*. In MP, a subset of the dictionary is chosen in a *per vector* manner, so much more energy compaction is possible.

To illustrate the energy compaction property of MP, consider the following situation. A $\mathcal{N}(0, I_N)$ source is to be transform coded. Because the components of the source are uncorrelated, no orthogonal transform will give energy compaction; so in the linear coding case, k coefficients will capture k/N of the signal energy. A k-step MP expansion will capture much more of the energy. Fig. 5 shows the results of a simulation with $N = 8$. The plot shows the fraction of the signal energy in the residual when one- to four-term expansions are used. The dictionaries are generated randomly according to a uniform distribution on the unit sphere. For a corresponding number of terms, the energy compaction is much better with MP than with a linear transform. Notice in particular that this is true even if the dictionary is not overcomplete ($M = N$), in which case MP has no more "flexibility" than an orthogonal basis representation.

B. Quantized Matching Pursuit

Define *quantized matching pursuit* (QMP) to be a modified version of matching pursuit which incorporates coefficient quantization. In particular, the inner product $\alpha_i = \langle \varphi_{k_i}, R_i x \rangle$ is quantized to $\hat{\alpha}_i = q(\alpha_i)$ prior to the computation of the residual $R_{i+1}x$. The quantized value is used in the residual calculation: $R_{i+1}x = R_i x - \hat{\alpha}_i \varphi_{k_i}$. The use of the quantized value in the residual calculation reduces the propagation of the quantization error to subsequent iterations.

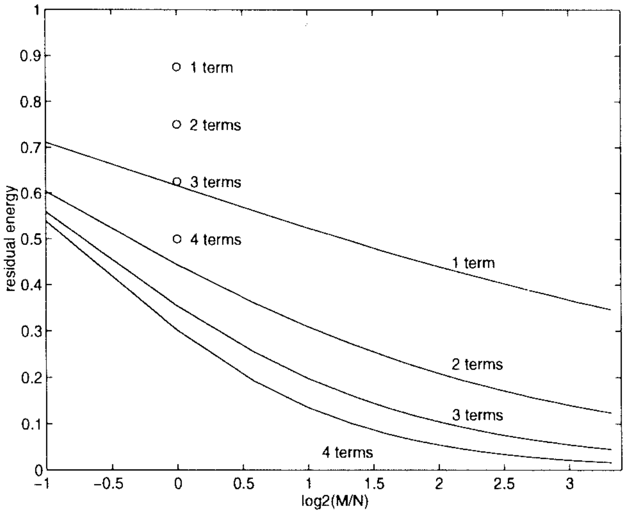


Fig. 5. Comparison of energy compaction properties for coding of a $\mathcal{N}(0, I_8)$ source. With a k-term orthogonal expansion, the residual has $(8 - k)/8$ of the energy (o's). The residual energy is much less with MP (solid curves).

Although QMP has been applied to low bit rate compression problems [5], [6], [25], which inherently require coarse coefficient quantization, little work has been done to understand the qualitative effects of coefficient quantization in matching pursuit. In this section we explore some of these effects. The relationship between quantized matching pursuit and other vector quantization (VQ) methods is discussed in Section III-B.1. The issue of consistency in these expansions is explored in Section III-B.2. 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. (Such an algorithm is presented in Section III-C.2.) In Section III-B.3, a detailed example on the application of QMP to quantization of an $\mathbb{R}^2$-valued source is presented. This serves to illustrate the concepts from Section III-B.2 and demonstrate the potential for improved reconstruction using consistency.

1) Relationship to Other Vector Quantization Methods: A single iteration of matching pursuit is very similar to shape-gain VQ, which was introduced in [27]. 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 α_0, the only differences being that in MP one maximizes the *absolute value* of the correlation and thus the gain factor can be negative. Obtaining a good approximation in shape-gain VQ requires that $\mathcal{C}_s$ forms a dense subset of the unit sphere in $\mathbb{R}^N$. The area of the unit sphere increases exponentially with N, making it difficult to use shape-gain VQ in high-dimensional spaces. A multiple iteration application of matching pursuit can be seen as a cascade form of shape-gain VQ.

2) Consistency: We have thus far discussed only signal analysis (or encoding) using QMP and not synthesis (reconstruction) from a QMP representation. To the best of our knowledge, all previous work with QMP has used

$$\hat{x} = \sum_{i=0}^{p-1} \hat{\alpha}_i \varphi_{k_i} \tag{13}$$

which results from simply using quantized coefficients in (11) and setting the final residual to zero. Computing this reconstruction has very low complexity, but its shortcoming is that it disregards the effects of quantization; hence it can produce inconsistent estimates.

Suppose p iterations of QMP are performed with the dictionary $\mathcal{D}$ and denote the output by

$$\text{QMP}\,(x) = (k_0,\, \hat{\alpha}_0,\, k_1,\, \hat{\alpha}_1,\, \cdots,\, k_{p-1},\, \hat{\alpha}_{p-1}). \tag{14}$$

Denote the output of QMP (with the same dictionary and quantizers) applied to $\hat{x}$ by

$$(k_0',\, \hat{\alpha}_0',\, k_1',\, \hat{\alpha}_1',\, \cdots,\, k_{p-1}',\, \hat{\alpha}_{p-1}').$$

By the definition of consistency (Section II-B3), $\hat{x}$ is a consistent estimate of x if and only if $k_i = k_i'$ and $\hat{\alpha}_i = \hat{\alpha}_i'$ for $i = 0, 1, \cdots, p-1$.

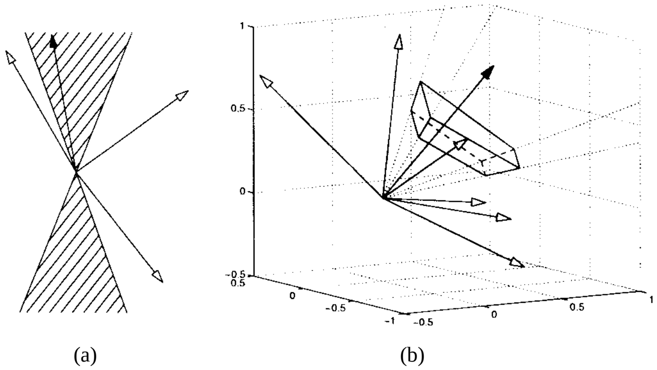
We now develop a description of the set of consistent estimates of x through simultaneous linear inequalities. For notational convenience, we assume uniform scalar quantization of the coefficients with stepsize Δ and midpoint reconstruction.[3] The selection of k_0 implies

$$|\langle \varphi_{k_0}, x\rangle| \geq |\langle \varphi, x\rangle|, \qquad \forall\, \varphi \in \mathcal{D}. \tag{15}$$

For each element of $\mathcal{D} \setminus \{\varphi_{k_0}\}$, (15) specifies a pair of half-space constraints with boundary planes passing through the origin. An example of such a constraint in $\mathbb{R}^2$ is shown in Fig. 6(a). If φ_{k_0} is the vector with the solid arrowhead (chosen from all of the marked vectors), the source vector must lie in the hatched area. For $N > 2$, the intersection of these constraints is two infinite convex polyhedral cones situated symmetrically with their apexes at the origin. The value of $\hat{\alpha}_0$ gives the constraint

$$\langle \varphi_{k_0}, x\rangle \in \left[\hat{\alpha}_0 - \frac{\Delta}{2},\, \hat{\alpha}_0 + \frac{\Delta}{2}\right]. \tag{16}$$

This specifies a pair of planes, perpendicular to φ_{k_0}, between which x must lie. Constraints (15) and (16) are illustrated in Fig. 6(b) for $\mathbb{R}^3$. The vector with the solid arrowhead was chosen among all the marked dictionary vectors as φ_{k_0}. Then the quantization of α_0 implies that the source vector lies in the volume shown.



(a) (b)

Fig. 6. (a) Illustration of consistency constraint (15) in $\mathbb{R}^2$. (b) Illustration of consistency constraints (15) and (16) in $\mathbb{R}^3$.

At the $(i-1)$st step, the selection of k_i gives the constraints

$$\left|\left\langle \varphi_{k_i},\, x - \sum_{\ell=0}^{i-1} \hat{\alpha}_\ell \varphi_{k_\ell}\right\rangle\right| \geq \left|\left\langle \varphi,\, x - \sum_{\ell=0}^{i-1} \hat{\alpha}_\ell \varphi_{k_\ell}\right\rangle\right|, \qquad \forall\, \varphi \in \mathcal{D}. \tag{17}$$

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

$$\left\langle \varphi_{k_i},\, x - \sum_{\ell=0}^{i-1} \hat{\alpha}_\ell \varphi_{k_\ell}\right\rangle \in \left[\hat{\alpha}_i - \frac{\Delta}{2},\, \hat{\alpha}_i + \frac{\Delta}{2}\right]. \tag{18}$$

This again specifies a pair of planes, now perpendicular to φ_{k_i}, between which x must lie.

By being explicit about the constraints as above, we see that, except in the case that $0 \in [\hat{\alpha}_i - \Delta/2,\, \hat{\alpha}_i + \Delta/2]$ for some i, the partition cell defined by (14) is convex.[4] Thus by using an appropriate projection operator, one can find a strictly consistent estimate from any initial estimate. The partition cells are intersections of cells of the form shown in Fig. 6(b).

Notice now that contrary to what would be surmised from (13), k_i gives some information on the signal even if $\hat{\alpha}_i = 0$. The experiments in Section III-C3 show that when $\hat{\alpha}_i = 0$, it tends to be inefficient in a rate-distortion sense to store or transmit k_i. If we know that $\hat{\alpha}_i = 0$ but do not know the value of k_i, then (17) and (18) reduce to

$$\left|\left\langle \varphi,\, x - \sum_{\ell=0}^{i-1} \hat{\alpha}_\ell \varphi_{k_\ell}\right\rangle\right| \leq \frac{\Delta}{2}, \qquad \forall\, \varphi \in \mathcal{D}. \tag{19}$$

Experiments were performed to demonstrate that (13) often gives inconsistent estimates and to assess how the probability of an inconsistent estimate depends on the dictionary size and the quantization. We present here results for an $\mathbb{R}^4$-valued source with the $\mathcal{N}(0, I)$ distribution. The consistency of reconstruction was checked for two iteration expansions with dictionaries generated randomly according to a uniform

[3] Ambiguities on partition cell boundaries due to arbitrary tie-breaking—in both dictionary element selection and nearest neighbor scalar quantization—are ignored.

[4] The "hourglass" cell that results from $0 \in [\hat{\alpha}_i - \Delta/2,\, \hat{\alpha}_i + \Delta/2]$ does not make consistent reconstruction more difficult, but is intuitively undesirable in a rate-distortion sense.

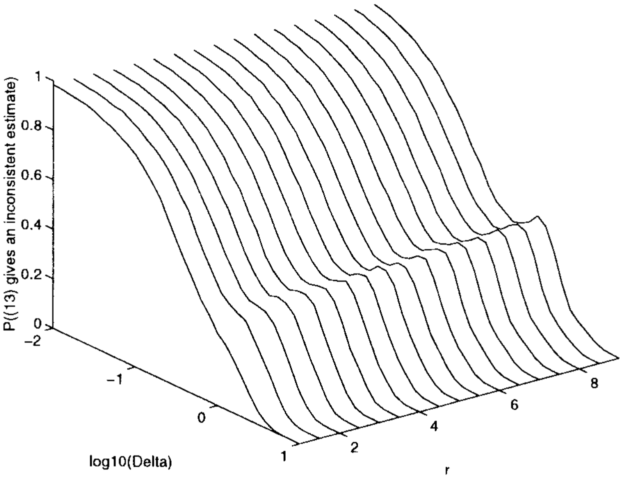


Fig. 7. Probability that (13) gives an inconsistent reconstruction for two iteration expansions of an $\mathbb{R}^4$-valued source.

distribution on the unit sphere. Dictionary sizes of $M = 4, 8, \cdots, 20$ were used. The quantization was uniform with reconstruction points $\{m\Delta\}_{m\in\mathbb{Z}}$. The results are shown in Fig. 7. The probability of inconsistency goes to zero for very coarse quantization and goes to one for fine quantization. The dependence on dictionary size and lack of monotonicity indicate complicated geometric factors. Similar experiments with different sources and dictionaries were reported in [28].

As noted earlier, the cells of the partition generated by QMP are convex or the union of two convex cells that share one point. This fact allows the computation of consistent estimates through the method of alternating projections [16]. One would normally start with an initial estimate given by (13). Given an estimate $\hat{x}$, the algorithm given in Table II performs the one "most needed" projection; namely, the first projection needed in enforcing (15)–(18). Among the possible projections in enforcing (17), the one corresponding to the largest deviation from consistency is performed. For notational convenience and concreteness, we assume again uniform quantization with

$$q_i(\alpha_i) = m\Delta \iff \alpha_i \in [(m - \frac{1}{2})\Delta,\, (m + \frac{1}{2})\Delta).$$

Steps 5 and 6 could easily be adjusted for a general quantizer.

In a broadly applicable special case, the inequalities (15)–(18) can be manipulated into a set of elementwise inequalities $Ax \le b$ suitable for reconstruction using linear or quadratic programming, where A and b are $2Mp \times N$ and $2Mp \times 1$, respectively, and A and b depend only on the QMP output. This formulation is possible when each QMP iteration either a) uses a quantizer with zero as a decision point; or b) uses a quantizer which maps a symmetric interval to zero, and the value of k_i is discarded when $\hat{\alpha}_i = 0$.

Consider first the case where q_i has zero as a decision point. For notational convenience, we will assume the decision points and reconstruction values are given by $\{m\Delta_i\}_{m\in\mathbb{Z}}$ and $\{(m + \frac{1}{2})\Delta_i\}_{m\in\mathbb{Z}}$, respectively, but all that is actually necessary is that the quantized coefficient $\hat{\alpha}_i$ reveals the sign of the unquantized coefficient α_i. Denote $\operatorname{sgn}(\alpha_i)$ by σ_i, and furthermore define the following $2(M-1) \times N$ matrices:

TABLE II
PROJECTION ALGORITHM FOR CONSISTENT RECONSTRUCTION FROM A QMP REPRESENTATION

1. Set $c = 0$. This is a counter of the number of steps of QMP that $\hat{x}$ is consistent with.
2. Let
$$\overline{x} = \hat{x} - \sum_{i=0}^{c-1} \hat{\alpha}_i \mathcal{D}_{k_i}$$
where it is understood that the summation is empty for $c = 0$
3. Find $\varphi \in \mathcal{D}$ that maximizes $|\langle \varphi, \overline{x}\rangle|$. If $\varphi = \varphi_{k_c}$, go to Step 5; else go to Step 4.
4. ($\hat{x}$ is not consistent with k_c.) Let
$$\tilde{\varphi}_{k_c} = \text{sgn}\,(\langle \varphi_{k_c}, \overline{x}\rangle)\varphi_{k_c}$$
and
$$\tilde{\varphi} = \text{sgn}\,(\langle \varphi, \overline{x}\rangle)\varphi.$$
Let
$$\hat{x} = \hat{x} - \langle \tilde{\varphi}_{k_c} - \tilde{\varphi}, \overline{x}\rangle(\tilde{\varphi}_{k_c} - \tilde{\varphi})$$
the orthogonal projection of $\hat{x}$ onto the set described by (17). Terminate
5. ($\hat{x}$ is consistent with k_c.) If
$$\langle \varphi_{k_c}, \overline{x}\rangle \in [\hat{\alpha}_c - \frac{1}{2}\Delta,\, \hat{\alpha}_c + \frac{1}{2}\Delta)$$
go to Step 7; else go to Step 6.
6. ($\hat{x}$ is not consistent with $\hat{\alpha}_c$.) Let
$$\begin{aligned}\beta &= \text{sgn}\,(\langle \varphi_{k_c}, \overline{x}\rangle - \hat{\alpha}_c)\\ &\quad \cdot \min\left\{\left|\langle \varphi_{k_c}, \overline{x}\rangle - \left(\hat{\alpha}_c + \frac{\Delta}{2}\right)\right|,\, \left|\langle \varphi_{k_c}, \overline{x}\rangle - \left(\hat{\alpha}_c - \frac{\Delta}{2}\right)\right|\right\}.\end{aligned}$$
Let $\hat{x} = \hat{x} - \beta\varphi_{k_c}$, the orthogonal projection of $\hat{x}$ onto the set described by (18). Terminate.
7. ($\hat{x}$ is consistent with $\hat{\alpha}_c$.) Increment c. If $c = p$, terminate ($\hat{x}$ is consistent); else go to Step 2.

$$\begin{aligned}F_i &= [\varphi_{k_i} \quad \varphi_{k_i} \quad \cdots \quad \varphi_{k_i}]^T\\ F_{\underline{i}} &= [\varphi_{k_1} \quad \cdots \quad \varphi_{k_{i-1}} \quad \varphi_{k_{i+1}} \quad \cdots \quad \varphi_{k_M}]^T.\end{aligned}$$

First, write (17) as

$$|\varphi_{k_i}^T(x - c)| \ge |\varphi^T(x - c)|, \qquad \forall\, \varphi \in \mathcal{D}$$

where c is shorthand for $\sum_{\ell=0}^{i-1} \hat{\alpha}_\ell \varphi_{k_\ell}$. Combining the $M - 1$ nontrivial inequalities gives

$$\sigma_i F_i(x - c) \ge |F_{\underline{i}}(x - c)|.$$

Expanding the absolute value one can obtain

$$\begin{bmatrix} F_{\underline{i}} - \sigma_i F_i \\ -F_{\underline{i}} - \sigma_i F_i \end{bmatrix} x \le \begin{bmatrix} (F_{\underline{i}} - \sigma_i F_i)c \\ (-F_{\underline{i}} - \sigma_i F_i)c \end{bmatrix}. \tag{20}$$

Writing (18) first as

$$|\varphi_{k_i}^T(x - c) - \hat{\alpha}_i| \le \frac{\Delta_i}{2}$$

one easily obtains

$$\begin{bmatrix} \varphi_{k_i}^T \\ -\varphi_{k_i}^T \end{bmatrix} x \le \begin{bmatrix} \dfrac{\Delta_i}{2} + \hat{\alpha}_i + \varphi_{k_i}^T c \\ \dfrac{\Delta_i}{2} - \hat{\alpha}_i - \varphi_{k_i}^T c \end{bmatrix}. \tag{21}$$

On the other hand, if q_i maps an interval $[-(\Delta_i/2), \Delta_i/2]$ to zero and k_i is not coded, then (19) leads similarly to the $2M$ inequalities

$$\begin{bmatrix} F \\ -F \end{bmatrix} x \leq \begin{bmatrix} Fc + \dfrac{\Delta_i}{2} \\ -Fc + \dfrac{\Delta_i}{2} \end{bmatrix}. \tag{22}$$

A formulation of the form $Ax \leq b$ is obtained by stacking inequalities (20)–(22) appropriately.

3) An Example in $\mathbb{R}^2$*:* Consider quantization of an $\mathbb{R}^2$-valued source. Assume that two iterations will be performed with the four-element dictionary

$$\mathcal{D} = \left\{ \left[\cos\frac{(2k-1)\pi}{8} \ \sin\frac{(2k-1)\pi}{8} \right]^T \right\}_{k=1}^{4}.$$

Even if the distribution of the source is known, it is difficult to find analytical expressions for optimal quantizers. (The issue of optimal quantizer design is considered for the case of a source with a uniform distribution on $[-1, 1]^2$ in [28, Sec. 3.3.2].) Since we wish to use fixed, untrained quantizers, we will use uniform quantizers for α_0 and α_1. It will generally be true that $\varphi_{k_0} \perp \varphi_{k_1}$, so it makes sense for the quantization step sizes for α_0 and α_1 to be equal.

The partitions generated by matching pursuit are very intricate. In Fig. 8, the heavy lines show the partitioning of the first quadrant when zero is a quantizer reconstruction value, i.e., the quantizer reconstruction points are $\{m\Delta\}_{m\in\mathbb{Z}}$ and decision points are $\{(m+\frac{1}{2})\Delta\}_{m\in\mathbb{Z}}$ for some quantization stepsize Δ.[5] The dotted lines show boundaries that are created by choice of k_0 (k_1) but, depending on the reconstruction method, might not be important because $\hat{\alpha}_0 = 0$ $(\hat{\alpha}_1 = 0)$. In this partition, most of the cells are squares, but there are also some smaller cells. The fraction of cells that are not square goes to zero as $\Delta \to 0$.

This quantization of $\mathbb{R}^2$ gives concrete examples of the inconsistency resulting from using (13). The linear reconstruction points are indicated in Fig. 8 by $\circ$'s. The light line segments connect these to the corresponding optimal[6] reconstruction points. Such a line segment crossing a cell boundary indicates a case of (13) giving an inconsistent estimate.

C. Lossy Vector Coding with Quantized Matching Pursuit

This section explores the efficacy of using QMP as an algorithm for lossy compression of vectors in $\mathbb{R}^N$. In order to reveal qualitative properties most clearly, very simple dictionaries and synthetic sources are used in the experiments. Experiments with other dictionaries and sources appear in [28]. We do not explore the design of a dictionary or scalar quantizers for a particular application. Dictionary structure has a great impact on the computational complexity of QMP as demonstrated, for example, in [29].

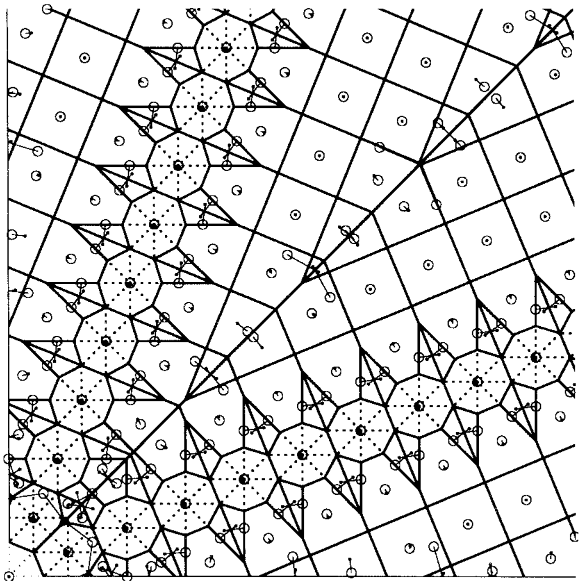


Fig. 8. Partitioning of first quadrant of $\mathbb{R}^2$ by matching pursuit with four-element dictionary (heavy lines). Linear reconstruction points ($\circ$'s) are connected to optimal reconstruction points ($\times$'s) by light line segments.

For simplicity, rate and distortion are measured by sample entropy and MSE per component, respectively. The sources used are multidimensional Gaussian with zero mean and independent components. The inner product quantization is uniform with midpoint reconstruction values at $\{m\Delta\}_{m\in\mathbb{Z}}$. Furthermore, the quantization stepsize Δ is constant across iterations. This is consistent with equal weighting of error in each direction.

1) Basic Experimental Results: In the first experiment, $N = 4$ and the dictionary was composed of $M = 11$ maximally spaced points on the unit sphere [18]. Rate was measured by summing the (scalar) sample entropies of $k_0, k_1, \cdots, k_{p-1}$ and $\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_{p-1}$, where p is the number of iterations. The results are shown in Fig. 9. The three dotted curves correspond to varying p from 1 to 3, while reconstructing according to (13). The points along each dotted curve are obtained by varying Δ. Notice that the number of iterations that minimizes the distortion depends on the available rate. The solid curve is the convex hull of these R–D operating points (converted to a decibel scale). In subsequent graphs, only this convex hull performance is shown.

2) Improved Reconstruction Using Consistency: Continuing the experiment above, the degree of improvement obtained by using a consistent reconstruction algorithm was ascertained. Using consistent reconstruction gives the performance shown by the dashed curve in Fig. 9. Notice that there is no improvement at low bit rates because consistency is not an issue for a single-iteration expansion. The improvement increases monotonically with the bit rate.

3) An Effective Stopping Criterion: Regardless of the reconstruction method, the coding results shown in Fig. 9 are far from satisfactory, especially at low bit rates. For a p-step expansion, the "baseline" coding method is to apply entropy codes (separately) to $k_0, \hat{\alpha}_0, k_1, \hat{\alpha}_1, \cdots, k_{p-1}, \hat{\alpha}_{p-1}$. This coding places a rather large penalty of roughly $\log_2 M$ bits

[5] The partition is somewhat different when the quantizer has different decision points, e.g., $\{(m + \frac{1}{2})\Delta\}_{m\in\mathbb{Z}}$ [28, Sec. 3.3.2]. The ensuing conclusions are qualitatively unchanged.

[6] Optimality is with respect to a uniform source distribution.

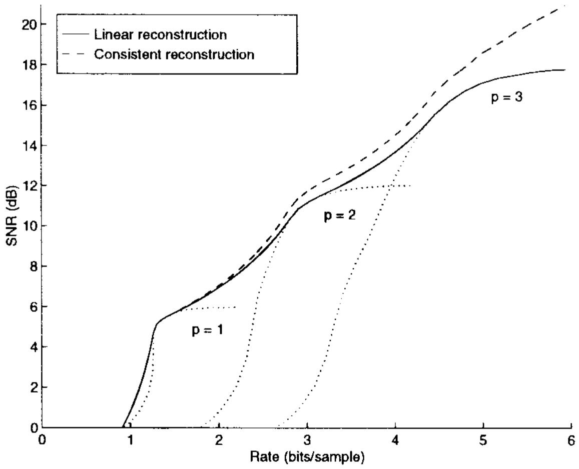


Fig. 9. Performance comparison between reconstruction based on (13) and consistent reconstruction. $N = 4$ and the dictionary is composed of $M = 11$ maximally spaced points on the unit sphere [18].

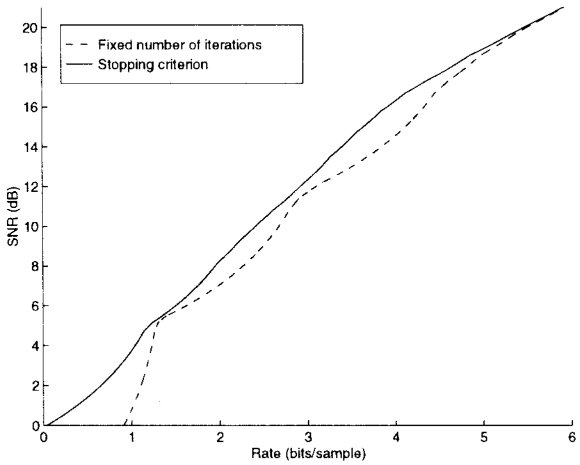


Fig. 10. Performance comparison between a fixed number of iterations and a simple stopping criterion. $N = 4$ and the dictionary is composed of $M = 11$ maximally spaced points on the unit sphere [18].

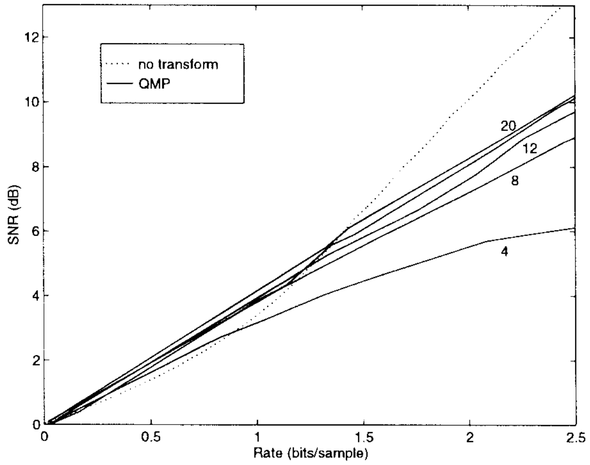


Fig. 11. Performance of QMP as the dictionary size is varied (solid curves, labeled by M) compared to the performance of independent uniform quantization of each sample (dotted curve).

on each iteration, i.e., this many bits must be spent in addition to the coding of the coefficient. In particular, the minimum achievable bit rate is about $(\log_2 M)/N$.

Assume that the same scalar quantization function is used at each iteration and that the quantizer maps a symmetric interval to zero. Based on a few simple observations, we can devise a simple alternative coding method which greatly reduces the rate. The first observation is that if $\hat{\alpha}_i = 0$, then $\hat{\alpha}_j = 0$ for all $j > i$ because the residual remains unchanged. Secondly, if $\hat{\alpha}_i = 0$, then k_i carries relatively little information. Thus we propose that a) $\hat{\alpha}_i = 0$ be used as a stopping criterion which causes a block to be terminated even if the maximum number of iterations has not been reached and b) k_i be considered conceptually to come after $\hat{\alpha}_i$, so k_i is not coded if $\hat{\alpha}_i = 0$.

Simulations were performed with the same source, dictionary, and quantizers as before to demonstrate the improvement due to the use of a stopping criterion. The results, shown in Fig. 10, indicate a sizable improvement at low bit rates.

4) Further Explorations: Having established the merits of consistent reconstruction and the stopping criterion of Section III-C3, we now explore the effects of varying the size of the dictionary. Again the source is independent and identically distributed (i.i.d.) Gaussian in blocks of $N = 4$ samples, and dictionaries generated randomly according to a uniform distribution on the unit sphere were used. Fig. 11 shows the performance of QMP with $M = 4, 8, \cdots, 20$ (solid curves); and of independent uniform scalar quantization followed by entropy coding (dotted curve). The performance of QMP improves as M is increased and exceeds that of independent uniform scalar quantization at low bit rates. This result highlights the advantage of a nonlinear transform, since no linear transform would give any coding gain for this source.[7]

In the final experimental investigation, we consider the lowest complexity instance of QMP. This occurs when the dictionary is an orthonormal set. In this case, QMP reduces to nothing more than a linear transform followed by sorting by absolute value and quantization. Here we code an i.i.d. Gaussian source with block sizes $N = 1, 2, \cdots, 8$.[8] The results shown in Fig. 12 indicate that even in this computationally simple case without a redundant dictionary, QMP performs well at low bit rates. An interesting phenomenon is revealed: $N = 1$ is best at high bit rates and $N = 2$ is best at low bit rates; no larger value of N is best at any bit rate.

5) A Few Possible Variations: The experiments of the previous subsections are the tip of the iceberg in terms of possible design choices. To conclude our discussion of source coding, 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.

There is no *a priori* reason to use the same dictionary at every iteration. Given a p iteration estimate, the entropy

[7] We are not advocating the use of random dictionaries. Slightly better performance is expected with an appropriately chosen fixed dictionary.

[8] Of course, $N = 1$ gives independent uniform scalar quantization of each sample.

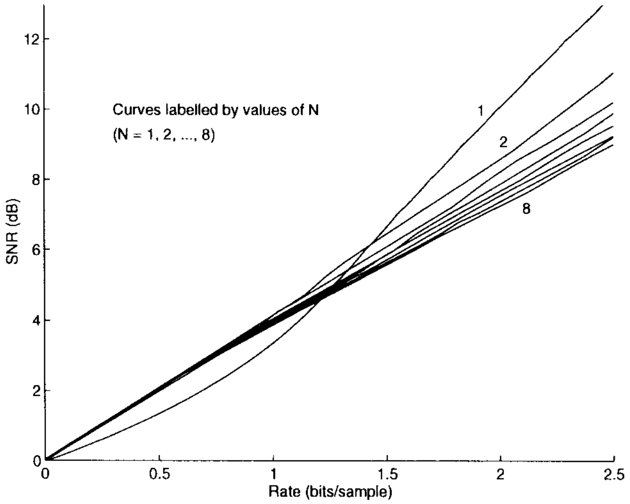


Fig. 12. Performance of QMP with an orthogonal basis dictionary as the block size N is varied.

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

The experimental results that have been presented are based on entropy coding each $\hat{\alpha}_i$ independently of the indices, which are in turn coded separately; there are other possibilities. Joint entropy coding of indices was explored in [28] and [30]. Also, conditional entropy coding could exploit the likelihood of consecutively chosen dictionary vectors being orthogonal or nearly orthogonal.

Finally, for a broad class of source distributions, the distributions of the α_i's will have some common properties because they are similar to order statistics. For example, the probability density of α_0 will be small near zero. This could be exploited in quantizer design.

IV. Conclusions

This paper has considered the effects of coefficient quantization in overcomplete expansions. Two classes of overcomplete expansions were considered: fixed (frame) expansions and expansions that are adapted to each particular source sample, as given by matching pursuit. In each case, the possible inconsistency of linear reconstruction was exhibited, computational methods for finding consistent estimates were given, and the distortion reduction due to consistent reconstruction was experimentally assessed.

For a quantized frame expansion with redundancy r, it was proven 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, it was conjectured that any reconstruction method that gives consistent estimates will have an MSE that can be upper-bounded by an $O(1/r^2)$ expression. Taken together, these suggest that optimal reconstruction methods will yield $O(1/r^2)$ MSE, and that consistency is sufficient to insure this asymptotic behavior.

Experiments on the application of quantized matching pursuit as a vector compression method demonstrated good low bit rate performance when an effective stopping criterion was used. Since it is a successive approximation method, matching pursuit may be useful in a multiresolution framework, and the inherent hierarchical nature of the representation is amenable to unequal error protection methods for transmission over noisy channels. Because of the dependencies between outputs of successive iterations, MP might also work well coupled with adaptive and/or universal lossless coding.

Appendix I
Proofs

A. Theorem 1

Let $\Phi_M = \{\varphi_k\}_{k=1}^M$. The corresponding frame operator is given by $F = [\varphi_1 \; \varphi_2 \; \cdots \; \varphi_M]^T$. Thus the (i, j)th element of $(1/M)F^*F$ is given by

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

where $(\varphi_k)_i$ is the ith component of φ_k.

First consider the diagonal elements $(i = j)$. Since for a fixed i the random variables $(\varphi_k)_i$, $1 \le k \le M$ are i.i.d., and have zero mean, we find that

$$\begin{aligned}E\left[\left(\frac{1}{M}F^*F\right)_{ii}\right] &= \mu_2\\ \mathrm{Var}\left[\left(\frac{1}{M}F^*F\right)_{ii}\right] &= \frac{1}{M}\left(\mu_4 - \frac{M-3}{M-1}\mu_2^2\right) \qquad (23)\end{aligned}$$

where $\mu_2 = E[(\varphi_k)_i^2]$ and $\mu_4 = E[(\varphi_k)_i^4]$ [31, Sec. 8-1]. For the off-diagonal elements $(i \neq j)$

$$E\left[\left(\frac{1}{M}F^*F\right)_{ij}\right] = 0 \qquad (24)$$

$$\mathrm{Var}\left[\left(\frac{1}{M}F^*F\right)_{ij}\right] = \frac{1}{M}E[(\varphi_k)_i^2(\varphi_k)_j^2]. \qquad (25)$$

Noting that μ_2 and μ_4 are independent of M, (23) shows that $\mathrm{Var}\,[(F^*F/M)_{ii}] \to 0$ as $M \to \infty$, so $(F^*F/M)_{ii} \to \mu_2$ in the mean-squared sense [31, Sec. 8-4]. Similarly, (24) and (25) show that for $i \neq j$, $(F^*F/M)_{ij} \to 0$ in the mean-squared sense. This completes the proof, provided $\mu_2 = 1/N$.

We now derive explicit formulas (depending on N) for μ_2, μ_4, and $E[(\varphi_k)_i^2(\varphi_k)_j^2]$. 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 $(\varphi_1, \varphi_2, \cdots, \varphi_N)$. Denote the N-dimensional unit sphere (centered at the origin) by S_N.

Since φ is uniformly distributed on S_N, the probability density function (p.d.f.) of φ is given by

$$f(\varphi) = \frac{1}{c_N}, \qquad \forall\, \varphi \in S_N \tag{26}$$

where c_N is the surface area of S_N. Using spherical coordinates, c_N is given by

$$c_N = \left(\int_0^{2\pi} d\theta\right)\left(\int_0^{\pi} \sin\,\omega_1\, d\omega_1\right)\left(\int_0^{\pi} \sin^2\,\omega_2\, d\omega_2\right)\cdots \left(\int_0^{\pi} \sin^{N-2}\,\omega_{N-2}\, d\omega_{N-2}\right). \tag{27}$$

Using (26), we can make the following calculation:

$$\begin{aligned}
\mu_2 &= E[\varphi_i^2] = E[\varphi_N^2] \\
&= \int_{S_N} \frac{\varphi_N^2}{c_N}\, dA \text{ where } dA \text{ is a differential area element} \\
&= \frac{1}{c_N}\left(\int_0^{2\pi} d\theta\right)\left(\int_0^{\pi} \sin\,\omega_1\, d\omega_1\right) \\
&\quad \cdot \left(\int_0^{\pi} \sin^2\,\omega_2\, d\omega_2\right)\cdots\left(\int_0^{\pi} \sin^{N-3}\,\omega_{N-3}\, d\omega_{N-3}\right) \\
&\quad \cdot \left(\int_0^{\pi} \cos^2\,\omega_{N-2}\, \sin^{N-2}\,\omega_{N-2}\, d\omega_{N-2}\right) \qquad (28)\\
&= \left(\int_0^{\pi} \sin^{N-2}\,\omega_{N-2}\, d\omega_{N-2}\right)^{-1} \\
&\quad \cdot \left(\int_0^{\pi} \cos^2\,\omega_{N-2}\, \sin^{N-2}\,\omega_{N-2}\, d\omega_{N-2}\right) \qquad (29)\\
&= \frac{1}{N}.
\end{aligned}$$

In this calculation, (28) results from using spherical coordinates and (29) follows from substituting (27) and canceling like terms. The final simplification is due to a standard integration formula [32, eq. (323)]. Similar calculations give

$$\mu_4 = E[\varphi_i^4] = 3/N(N+2)$$

and, for $i \neq j$,

$$E[\varphi_i^2 \varphi_j^2] = 1/N(N+2).$$

B. Proposition 1

Subtracting

$$\hat{x} = \sum_{k=1}^{M} (\langle x, \varphi_k \rangle + \beta_k)\tilde{\varphi}_k$$

from

$$x = \sum_{k=1}^{M} \langle x, \varphi_k \rangle \tilde{\varphi}_k$$

gives

$$x - \hat{x} = -\sum_{k=1}^{M} \beta_k \tilde{\varphi}_k.$$

Then we can calculate

$$\begin{aligned}
\text{MSE} &= E\|x - \hat{x}\|^2 = E\left\|\sum_{k=1}^{M} \beta_k \tilde{\varphi}_k\right\|^2 \\
&= E\left[\sum_{i=1}^{M}\sum_{k=1}^{M} \overline{\beta_i}\beta_k \tilde{\varphi}_i^* \tilde{\varphi}_k\right] = \sum_{i=1}^{M}\sum_{k=1}^{M} \delta_{ik}\sigma^2 \tilde{\varphi}_i^* \tilde{\varphi}_k \qquad (30)\\
&= \sigma^2 \sum_{k=1}^{M} \|\tilde{\varphi}_k\|^2 = \sigma^2 \sum_{k=1}^{M} \|(F^*F)^{-1}\varphi_k\|^2 \qquad (31)
\end{aligned}$$

where (30) results from evaluating expectations using the conditions on β, and (31) uses (5). From (4) we can derive

$$B^{-2}\|\varphi_k\|^2 \leq \|(F^*F)^{-1}\varphi_k\|^2 \leq A^{-2}\|\varphi_k\|^2$$

which simplifies to

$$B^{-2} \leq \|(F^*F)^{-1}\varphi_k\|^2 \leq A^{-2} \tag{32}$$

because of the normalization of the frame. Combining (31) and (32) completes the proof.

C. Proposition 2

Let us consider a given reconstruction algorithm. It maps every possible discrete vector $\hat{y}$ of $\mathbb{R}^M$ into a vector $\hat{x} = R(\hat{y})$ of $\mathbb{R}^N$. The reconstruction algorithm thus approximates any input vector $x \in \mathbb{R}^N$ by $\hat{x} = VQ(x)$ where $VQ(x) = R(Q(Fx))$. The reconstruction MSE is thus $E(\|VQ(x) - x\|^2)$. The mapping $x \mapsto VQ(x) = R(Q(Fx))$ is a vector quantizer of $\mathbb{R}^N$. For each discrete vector $\hat{y} \in Q(F\mathbb{R}^N)$, it maps all vectors of the subset $F^{-1}Q^{-1}(\hat{y})$ of $\mathbb{R}^N$ into the single vector $\hat{x} = R(\hat{y})$ of $\mathbb{R}^N$. According to the terminology in vector quantization [26], $F^{-1}Q^{-1}(\hat{y})$ is a cell of the partition defined by the vector quantizer VQ in $\mathbb{R}^N$, and $\hat{x} = R(\hat{y})$ is the corresponding code vector.

Let C be the number of cells that can be found in the region $\mathcal{B}$. It was proved by Zador [33], [34] (see also [35]) that there exists a coefficient $c(N, \boldsymbol{p})$ which only depends on N and $\boldsymbol{p}$, such that, for C large enough

$$E(\|VQ(x) - x\|^2) \geq \frac{c(N, \boldsymbol{p})}{C^{2/N}}. \tag{33}$$

To obtain a lower bound in terms of r, let us calculate an upper bound on C in terms of r. From (2) and the definition of Q, we have $Q(Fx) = [c_1(x)\; c_2(x)\; \cdots\; c_M(x)]^T$, where $c_i(x) = q_i(\langle x, \varphi_i \rangle)$. For each $i \in \{1, \cdots, M\}$, $x \mapsto c_i(x)$ is a mapping from $\mathbb{R}^N$ to $\mathbb{R}$. Because

$$Q(Fx) = \hat{y} \iff c_i(x) = \hat{y}_i \quad \forall\, i \in \{1, \cdots, M\}$$

we have

$$F^{-1}Q^{-1}(\hat{y}) = c_1^{-1}(\hat{y}_1) \cap c_2^{-1}(\hat{y}_2) \cap \cdots \cap c_M^{-1}(\hat{y}_M). \tag{34}$$

Consider a fixed $i \in \{1, \cdots, M\}$. Because q_i is a uniform scalar quantizer of step size Δ_i

$$\begin{aligned}
x \in c_i^{-1}(\hat{y}_i) &\iff q_i(\langle x, \varphi_i \rangle) = \hat{y}_i \\
&\iff \hat{y}_i - \frac{\Delta_i}{2} \leq \langle x, \varphi_i \rangle < \hat{y}_i + \frac{\Delta_i}{2}.
\end{aligned}$$

Thus $c_i^{-1}(\hat{y}_i)$ is the subset of $\mathbb{R}^N$ delimited by the two hyperplanes of equations $\langle x, \varphi_i \rangle = \hat{y}_i - \Delta_i/2$ and $\langle x, \varphi_i \rangle =$

$\hat{y}_i + \Delta_i/2$, respectively. These two parallel hyperplanes are perpendicular to the vector φ_i; the distance between them is $\ell_i = \Delta_i/\|\varphi_i\|$. Because $\hat{y}_i$ has its values on a discrete set of equidistant points separated by Δ_i, the set of all possible subsets $c_i^{-1}(\hat{y}_i)$ forms a partition of $\mathbb{R}^N$ whose cell boundaries are formed by parallel and equidistant hyperplanes. This type of partition was studied in [14] and is called a "hyperplane wave partition." The number $1/\ell_i$ represents the density of hyperplanes, or the number of hyperplanes per unit length in their orthogonal direction. The vector $d_i = (1/\ell_i)(\varphi_i/\|\varphi_i\|)$ is called the density vector of the hyperplane wave partition.

Thanks to (34), we see that the partition induced by VQ is obtained by intersecting M hyperplane wave partitions. It was shown in [14, Theorem A.7] that the number of cells induced by such a partition in a region of diameter D can be upper-bounded as

$$C \le \binom{M}{N}(\lceil Dd \rceil + 1)^N \tag{35}$$

where $d = \max_{1\le i\le M} \|d_i\|$. In our case, $\|d_i\| = 1/\ell_i$. Therefore,

$$d = \max_{1\le i\le M} \|\varphi_i\|/\Delta_i \le d_0.$$

Writing $M = rN$, we have

$$\binom{M}{N} \le \frac{M^N}{N!} = \frac{N^N}{N!} r^N. \tag{36}$$

By combining (33), (35), and (36), we obtain

$$E(\|VQ(x) - x\|^2) \ge b/r^2$$

where

$$b = \left(\frac{N!^{1/N}}{N(\lceil Dd_0 \rceil + 1)}\right)^2 c(N, \boldsymbol{p}).$$

D. Proposition 3

The proof is based on establishing the hypotheses of the following lemma:

Lemma 1: Assume $x_c(t)$ defined in (7) has at least $n = 2W+1$ quantization threshold crossings (QTC's) and consider sampling at a rate of M samples per period. Then there exist constants $c > 0$ and $r_0 \ge 1$ depending only on $x_c(t)$ such that for all $M = rn \ge r_0 n$, whenever $x_c(t)$ and $x'_c(t)$ have the same quantized sampled versions

$$1/T \int_T |x_c(t) - x'_c(t)|^2 \, dt < c/r^2.$$

Proof: This is a version of [1, Theorem 4.1] for real-valued signals. □

The following lemma gives a rough estimate which allows us to relate signal amplitude to signal power.[9]

Lemma 2: Among zero-mean, periodic signals with power P, the minimum possible peak-to-peak amplitude is $2\sqrt{P}$.

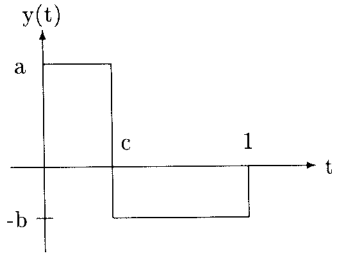


Fig. 13. One period of the signal used in the proof of Lemma 2.

Proof: We will construct a signal $y(t)$ with power P of minimum peak-to-peak amplitude. For convenience, let $T = 1$. Without loss of generality, we can assume that $\exists\, c$, $0 < c < 1$, such that $y(t) > 0$ for $t < c$ and $y(t) < 0$ for $t > c$. Then, to have minimum amplitude for given power, $y(t)$ must be piecewise-constant as in Fig. 13, with $a > 0$ and $b > 0$. The mean and power constraints can be combined to give $ab = P$. Under this constraint, the amplitude $a+b$ is uniquely minimized by $a = b = \sqrt{P}$. □

This final lemma relates the peak-to-peak amplitude of a continuous signal to its quantization threshold crossings:

Lemma 3: A continuous, periodic signal with peak-to-peak amplitude $> A$ which is subject to uniform quantization with stepsize Δ has at least $2\lfloor A/\Delta \rfloor$ quantization threshold crossings per period.

Proof: Consider first a signal $y(t)$ with peak-to-peak amplitude A. The "worst case" is for $A = k\Delta$ with $k \in \mathbb{Z}$, and for $\min y(t)$ and $\max y(t)$ to lie at quantization thresholds. In this case, the most we can guarantee is $k-1$ "increasing" QTC's and $k-1$ "decreasing" QTC's per period. If the peak-to-peak amplitude exceeds A, this worst case cannot happen, and we get at least $2\lfloor A/\Delta \rfloor$ QTC's. □

Proof of the Proposition:

i) N odd: Quantized frame expansion of x with M frame vectors is precisely equivalent to quantized sampling of $x_c(t)$ with M samples per period (see Section II-A2). Denote the quantized frame expansion of x and the corresponding continuous-time signal by x' and $x'_c(t)$, respectively. It is easy to verify that the average time-domain SE

$$1/T \int_T |x_c(t) - x'_c(t)|^2 \, dt$$

is the same as the vector SE $\|x - x'\|^2$. Let $\tilde{x}_c(t) = x_c(t) - x_1$. Then $\tilde{x}_c(t)$ is a zero-mean T-periodic signal with power $\|[x_2 \; x_3 \; \cdots \; x_N]^T\|^2$, which by hypothesis is greater than $((N+1)\Delta/4)^2$. Applying Lemma 2 we conclude that $\tilde{x}_c(t)$ has peak-to-peak amplitude greater than $(N+1)\Delta/2$. Since $x_c(t)$ has precisely the same peak-to-peak amplitude as $\tilde{x}_c(t)$, we can apply Lemma 3 to $x_c(t)$ to conclude that $x_c(t)$ has at least

$$2\lfloor ((N+1)\Delta/2)/\Delta \rfloor = 2\lfloor (N+1)/2 \rfloor = N+1$$

QTC's. Applying Lemma 1 with $n = N$ completes the proof.

[9] We use the standard notion of power; for $y(t)$ with period T:

$$1/T \int_T |y(t)|^2 \, dt.$$

ii) <u>N even</u>: We need only make slight adjustments from the previous case. Let

$$x_c(t) = \sum_{k=1}^{N/2} \left[x_{2k-1}\sqrt{2}\cos\frac{2\pi kt}{T} + x_{2k}\sqrt{2}\sin\frac{2\pi kt}{T} \right]$$

and define x' and $x'_c(t)$ correspondingly. Again the average time-domain SE

$$1/T \int_T |x_c(t) - x'_c(t)|^2 \, dt$$

is the same as the vector SE $\|x - x'\|^2$. The power of $x_c(t)$ equals $\|x\|^2 > ((N+2)\Delta/4)^2$. Applying Lemmas 2 and 3 implies that $x_c(t)$ has at least

$$2\lfloor((N+2)\Delta/2)/\Delta\rfloor = 2\lfloor(N+2)/2\rfloor = N+2$$

QTC's. We apply Lemma 1, this time with $n = N+1$ to match the form of (7), to complete the proof.

Note that the bounds in the hypotheses of the proposition are not tight. This is evidenced in particular by the fact that the bound in Lemma 2 is not attainable by bandlimited signals. For example, for $N = 2$ the minimum peak-to-peak amplitude is $\sqrt{2} \cdot 2\sqrt{P}$ and for $N = 4$ the minimum is $\approx 1.3657 \cdot 2\sqrt{P}$, compared to the bound of $2\sqrt{P}$. Because of Gibbs' phenomenon, the bound is not even asymptotically tight, but a more complicated lemma would serve no purpose here.

Appendix II
Frame Expansions and Hyperplane Wave Partitions

This appendix gives an interpretation of frame coefficients as measurements along different directions. Given a frame $\Phi = \{\varphi_k\}_{k=1}^M$, the kth component of $y = Fx$ is $y_k = \langle x, \varphi_k\rangle$. Thus y_k is a measurement of x *along* φ_k. We can thus interpret y as a vector of M "measurements" of x in directions specified by Φ. 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 us suppose y is scalar-quantized to give $\hat{y}$ by rounding each component to the nearest multiple of Δ. Since y_k specifies the measurement of a component parallel to φ_k, $\hat{y}_k = (i + \frac{1}{2})\Delta$ specifies an $(N-1)$-dimensional hyperplane perpendicular to φ_k. Thus quantization of y_k gives a set of parallel hyperplanes spaced by Δ, called a *hyperplane single wave*. The M hyperplane single waves give a partition with a particular structure called a *hyperplane wave partition* [14]. Examples of hyperplane wave partitions are shown in Fig. 14. In each diagram, a set of vectors comprising a frame in $\mathbb{R}^2$ is shown superimposed on the hyperplane wave partition induced by quantized frame expansion with that frame.

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 Δ^2. Section II-B4 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.

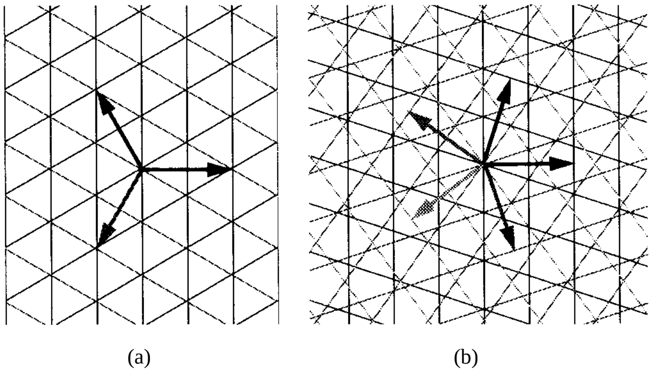


Fig. 14. Examples of hyperplane wave partitions in $\mathbb{R}^2$. (a) $M = 3$. (b) $M = 5$.

In Section II-B5 it was mentioned that coding each component of $\hat{y}$ separately is inefficient when $r \gg 1$. This can be explained by reference to Fig. 14. Specifying $\hat{y}_1$ and $\hat{y}_2$ defines a parallelogram within which x lies. Then there are a limited number of possibilities for $\hat{y}_3$. (In Fig. 14(a), there are exactly two possibilities. In Fig. 14(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.

Acknowledgment

The authors wish to thank Dr. D. Taubman for asking the first author whether consistency is an issue in matching pursuit and Dr. T. Kalker for providing some of the building blocks for the simulation software. An anonymous reviewer made several valuable comments.

References

[1] 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 Trans. Signal Processing*, vol. 42, pp. 200–203, Jan. 1994.

[2] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[3] F. Bergeaud and S. Mallat, "Matching pursuit of images," in *Proc. IEEE Int. Conf. on Image Processing* (Washington, DC, Oct. 1995), vol. I, pp. 53–56.

[4] V. K Goyal and M. Vetterli, "Dependent coding in quantized matching pursuit," in *Proc. SPIE Conf. on Visual Communication and Image Processing* (San Jose, CA, Feb. 1997), vol. 3024, pp. 2–14.

[5] R. Neff, A. Zakhor, and M. Vetterli, "Very low bit rate video coding using matching pursuits," in *Proc. SPIE Conf. on Visual Communication and Image Processing* (Chicago, IL, Sept. 1994), vol. 2308, pp. 47–60.

[6] M. Vetterli and T. Kalker, "Matching pursuit for compression and application to motion compensated video coding," in *Proc. IEEE Int. Conf. on Image Processing* (Austin, TX, Nov. 1994), vol. 1, pp. 725–729.

[7] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series," *Trans. Amer. Math. Soc.*, vol. 72, pp. 341–366, 1952.

[8] C. Heil and D. Walnut, "Continuous and discrete wavelet transforms," *SIAM Rev.*, vol. 31, pp. 628–666, 1989.

[9] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, Sept. 1990.

[10] ——, *Ten Lectures on Wavelets.* Philadelphia, PA: Soc. Industrial and Appl. Math., 1992.

[11] N. J. Munch, "Noise reduction in tight Weyl–Heisenberg frames," *IEEE Trans. Inform. Theory*, vol. 38, pp. 608–616, Mar. 1992.

[12] N. T. Thao, "Deterministic analysis of oversampled A/D conversion and sigma-delta modulation, and decoding improvements using consistent estimates," Ph.D. dissertation, Columbia Univ., New York, 1993.

[13] N. T. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. Signal Processing*, vol. 42, pp. 519–531, Mar. 1994.

[14] ——, "Lower bound on the mean-squared error in oversampled quantization of periodic signals using vector quantization analysis," *IEEE Trans. Inform. Theory*, vol. 42, pp. 469–479, Mar. 1996.

[15] Z. Cvetković and M. Vetterli, "Error analysis in oversampled A/D conversion and quantization of Weyl–Heisenberg frame expansions," submitted to *IEEE Trans. Inform. Theory*, 1996.

[16] D. C. Youla, "Mathematical theory of image restoration by the method of convex projections," in *Image Recovery: Theory and Application*, H. Stark, Ed. New York: Academic, 1987.

[17] G. Strang, *Introduction to Applied Mathematics.* Cambridge, MA: Wellesley-Cambridge, 1986.

[18] 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" [Online]. Availabe WWW: `http://www.research.att.com/~njas/packings`.

[19] H. P. Kramer and M. V. Mathews, "A linear coding for transmitting a set of correlated signals," *IRE Trans. Inform. Theory*, vol. IT-23, pp. 41–46, Sept. 1956.

[20] J.-Y. Huang and P. M. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun.*, vol. COM-11, pp. 289–296, Sept. 1963.

[21] G. Davis, "Adaptive nonlinear approximations," Ph.D. dissertation, New York Univ., Sept. 1994.

[22] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, Apr. 1995.

[23] Z. Zhang, "Matching pursuit," Ph.D. dissertation, New York Univ., 1993.

[24] G. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency approximations with matching pursuits," New York Univ., Tech. Rep. 657, Mar. 1994.

[25] T. Kalker and M. Vetterli, "Projection methods in motion estimation and compensation," in *Proc. IS & T/SPIE* (San Jose, CA, Feb. 1995), vol. 2419, pp. 164–175.

[26] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Boston, MA: Kluwer, 1992.

[27] A. Buzo, A. H. Gray, Jr., R. M. Gray, and J. D. Markel, "Speech coding based upon vector quantization," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-28, pp. 562–574, Oct. 1980.

[28] V. K. Goyal, "Quantized overcomplete expansions: Analysis, synthesis and algorithms," UC-Berkeley/ERL, Tech. Rep. M95/57, July 1995.

[29] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (Munich, Germany, Apr. 1997), pp. 2037–2040.

[30] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantization of overcomplete expansions," in *Proc. IEEE Data Compression Conf.*, J. A. Storer and M. Cohn, Eds. Snowbird, UT: IEEE Comp. Soc. Press, Mar. 1995, pp. 13–22.

[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1994.

[32] S. M. Selby, Ed., *Standard Mathematical Tables*, 18th ed. Boca Raton, FL: CRC, 1970.

[33] P. L. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1963.

[34] ——, "Asymptotic quantization error of continuous signals and the quantization dimension," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 139–148, Mar. 1982.

[35] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups.* New York: Springer-Verlag, 1988.