Quantization of Overcomplete Expansions

Vivek K Goyal, Martin Vetterli*
Dept. of Electrical Engineering
Univ. of California at Berkeley
{vkgoyal,martin}@eecs.berkeley.edu

Nguyen T. Thao
Dept. of Electrical & Electronic Eng.
Hong Kong Univ. of Science & Tech.
eethao@ee.ust.hk

Abstract

The use of overcomplete sets of vectors (redundant bases or frames) together with quantization is explored as an alternative to transform coding for signal compression. The goal is to retain the computational simplicity of transform coding while adding flexibility like adaptation to signal statistics. We show results using both fixed quantization in frames and greedy quantization using matching pursuit. An MSE slope of -6 dB/octave of frame redundancy is shown for a particular tight frame and is verified experimentally for another frame.

1 Introduction

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. In particular, transform coding is ubiquitous in image compression.

For fine quantization of a Gaussian signal with known statistics, the Karhunen-Loeve transform (KLT) is optimal for transform coding [2]. 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. 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 [5].

All varieties of transform coding represent a signal vector as a linear combination of orthogonal basis vectors. In this paper, we present a method that represents a signal with respect to an overcomplete set of vectors which we call a dictionary. The representation is generated through greedy successive approximation. Much

*Work supported in part by the National Science Foundation under grant MIP-93-21302

as the KLT finds the best representation "on average," this method finds a good representation for the particular vector being coded. The overhead in using this method is that the indices of the dictionary elements used must be coded. Hence 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.

For a signal with correlated samples, we expect certain dictionary elements to be chosen much more often than others. Thus entropy coding of the indices greatly reduces the overhead in this representation. In particular, this method can be used in a one-pass quantization system where the only adaptive component is a lossless coder. We do not adapt the dictionary, which would be computationally expensive. Note that one step of our algorithm is related to gain-shape vector quantization, and the overall scheme could be seen as a cascade form.

We begin in Section 2 with background material on frame representations and methods of generating frames. In Section 3 we discuss quantization in tight frames with no distributional assumptions and no adaptation to signal properties. Finally, in Section 4 we describe our quantization method based on matching pursuit. An example illustrates the flexibility of this approach. Experimental results based on a simple design which employs no distributional assumptions are also presented.

Throughout we will limit our attention to quantization of vectors from a finite dimensional Hilbert space $H = \mathbb{C}^N$ (or $\mathbb{R}^N$). We denote the inner product of $x, y \in H$ by $< x, y >$, and denote the norm of x by $||x|| =< x, x >^{1/2}$.

2 Redundant Representations and Frames

Let $\{\varphi_k\}_{k=1}^M \subset H$, where $M > N$. If $Span(\{\varphi_k\}_{k=1}^M) = H$, there exist $0 < A \leq B < \infty$ so that, $\forall f \in H$,

$$A||f||^2 \leq \sum_{k=1}^{M} | < f, \varphi_k > |^2 \leq B||f||^2. \tag{1}$$

We say that $\{\varphi_k\}_{k=1}^M$ is a *frame* or an *overcomplete* set of vectors with *redundancy ratio* $R = \frac{M}{N}$ [1]. Furthermore, if (1) holds for some $A = B$, we call the frame a *tight frame*. If $\{\varphi_k\}_{k=1}^M$ is a tight frame such that $||\varphi_k|| = 1 \; \forall \; k$, then $A = R$.

Since $Span(\{\varphi_k\}_{k=1}^M) = H$, any vector $f \in H$ can be written as

$$f = \sum_{k=1}^{M} \alpha_k \varphi_k \tag{2}$$

for some set of coefficients $\{\alpha_k\} \subset \mathbb{R}$ which are not unique. We refer to (2) as a *redundant representation*, although it may be the case that only N of the α_k's are non-zero. Define the *frame operator* F associated with $\{\varphi_k\}_{k=1}^M$ to be the linear operator from H to $\mathbb{C}^M$ given by

$$(Ff)_k =< f, \varphi_k > . \tag{3}$$

Note that 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

$$AI_N \leq F^*F \leq BI_N, \tag{4}$$

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 particular, $F^*F = AI_N$ shows that $\{\varphi_k\}_{k=1}^M$ is a tight frame.

As an example, we will show that 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 [7]. Let $x = [X_1\ X_2\ \cdots\ X_N]^T \in \mathbb{R}^N$. We define a corresponding continuous-time signal by

$$x_c(t) = X_1 + \sum_{k=1}^{W} \left[X_{2k}\sqrt{2}\cos\frac{2\pi kt}{T} + X_{2k+1}\sqrt{2}\sin\frac{2\pi kt}{T} \right]. \tag{5}$$

(Any real-valued, T-period, bandlimited, continuous-time signal can be written in this form.) Define a sampled version of $x_c(t)$ by $x_d[m] = x_c(\frac{mT}{M})$ and let $y = [x_d(0)\ x_d(1)\ \cdots\ x_d(M-1)]^T$. Then we have $y = Fx$, where

$$F = \begin{bmatrix} 1 & \sqrt{2} & 0 & \cdots & \sqrt{2} & 0 \\ 1 & \sqrt{2}\cos\theta & \sqrt{2}\sin\theta & \cdots & \sqrt{2}\cos W\theta & \sqrt{2}\sin W\theta \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & \sqrt{2}\cos(M'\theta) & \sqrt{2}\sin(M'\theta) & \cdots & \sqrt{2}\cos(WM'\theta) & \sqrt{2}\sin(WM'\theta) \end{bmatrix}, \tag{6}$$

$M' = M - 1$, 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. Notice that multiplication by F can be done efficiently using an FFT-based algorithm. We will refer to generating a frame in $\mathbb{R}^N$ for odd N using (6) as "Method I".

A multitude of other families of frames can be found. For $N = 3$, 4, and 5, Hardin, Sloane and Smith have numerically found arrangements of up to 130 points on N-dimensional spheres that maximize the minimum Euclidean norm separation [3]. We refer to selecting one of these sets of points as "Method II".

A third method is to consider the corners of the hypercube $[-\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}]^N$. These form a set of 2^N symmetric points in $\mathbb{R}^N$. Taking the subset of points that have a positive first coordinate gives a frame of size 2^{N-1}. This frame has the feature that inner products can be computed by addition and subtraction without any multiplication.

3 Reconstruction from Frame Coefficients

Let $F \in \mathbb{R}^{M\times N}$ be a frame operator. Let $x \in \mathbb{R}^N$ and $y = Fx$. Suppose y is quantized as $\hat{y} = Q[y]$ according to some partition of $\mathbb{R}^M$. The uncertainty region associated

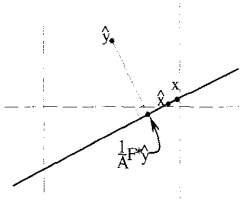


Figure 1: Illustration of consistent reconstruction

with $\hat{y}$ is obvious from the partition. This quantization also induces a partition of $\mathbb{R}^N$ insofar as a certain region of $\mathbb{R}^N$ would be mapped under $Q \circ F$ to $\hat{y}$. In general, $\hat{y}$ will not be in the image of $\mathbb{R}^N$ under F. Hence finding $\hat{x}$ such that $Q[F\hat{x}] = \hat{y}$ is not trivial.

Consider the case where the quantization in $\mathbb{R}^M$ is scalar, *i.e.* independent in each dimension. For $i = 1, \ldots, M$, denote the quantization stepsize in the ith component by Δ_i. For convenience we assume that the reproduction values lie halfway between decision levels. Then for each i, $|\hat{y}_i - y_i| \leq \frac{\Delta}{2}$. Given $\hat{y}$, suppose we wish to find $\hat{x}$ such that $Q[F\hat{x}] = \hat{y}$. We refer to this as a *consistent reconstruction* [8]. Then for each i, we must have

$$|(F\hat{x})_i - \hat{y}_i| \leq \frac{\Delta_i}{2}. \tag{7}$$

Expanding the absolute value, we find the constraints

$$F\hat{x} \leq \frac{1}{2}\Delta + \hat{y} \text{ and } F\hat{x} \geq -\frac{1}{2}\Delta + \hat{y} \tag{8}$$

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

$$\begin{bmatrix} F \\ -F \end{bmatrix} \hat{x} \leq \begin{bmatrix} \frac{1}{2}\Delta + \hat{y} \\ \frac{1}{2}\Delta - \hat{y} \end{bmatrix}. \tag{9}$$

The formulation (9) shows that $\hat{x}$ can be determined through linear programming [6]. An arbitrary cost function can be used.

It is important to note that even if $F^*F = AI_N$, $\frac{1}{A}F^*\hat{y}$ is not, in general, a solution for $\hat{x}$. This is due to the fact that $Q[Fx]$ is generally not in the image of $\mathbb{R}^N$ under F. This is illustrated in Figure 1 for the case $N = 1$, $M = 2$. The position of the bold line gives the mapping $\mathbb{R}^N \to \mathbb{R}^M$. The labels on that line are with respect to $\mathbb{R}^N$. The quantized value of $y = Fx$ is $\hat{y}$. A "naive" reconstruction gives $\frac{1}{A}F^*y$, which is not consistent. A possible consistent reconstruction is $\hat{x}$.

A general theory relating the partition of $\mathbb{R}^N$ to the partition of $\mathbb{R}^M$ is beyond the scope of this paper. However, some results about the relationship between R and the MSE are known in the case of the frame operator F given in (6). Indeed, some partition properties were derived in [7], using specifically the sampling interpretation of F. It was shown that for a given $x \in \mathbb{R}^N$ with certain conditions, the size of the

partition cell in $\mathbb{R}^N$ diminishes with R as $O(1/R^2)$ in the MSE sense. The proof explicitly uses the fact that $\hat{y} = Q[Fx]$ gives the sequence obtained by oversampling and quantizing the continous-time signal $x_c(t)$ defined from x in (5). The condition on x, however, is that $x_c(t)$ must cross the thresholds of the quantizer at least N times in one period T. In the case $N = 3$ ($W = 1$), this is guaranteed for all vector $x = [X_1\ X_2\ X_3]^T$ such that $\sqrt{X_2^2 + X_3^2} \geq \Delta$. In the three dimensional case, a stronger result can in fact be shown. If we consider a given bounded region of $\mathbb{R}^3$ whose elements x satisfy $\sqrt{X_2^2 + X_3^2} \geq \Delta$, then all the cells of the partition within this region have a size which can be upper bounded by α/R^2 where α is a positive constant. We conjecture that this type of result is true for any finite dimensional tight frame of (6). Figure 2(a) shows a partition of $\mathbb{R}^2$ achieved with $R = \frac{5}{2}$.

Experimental results confirm the $O(1/R^2)$ MSE behavior for Method I and suggest that this behavior is a more general phenomenon. Simulations in which the quantization stepsize was fixed and the frame redundancy was varied were performed. Results for $N = 3$ are shown in Figure 2(b). Results for $N = 4$ and 5 were similar. Using the "naive" reconstruction value of $\frac{1}{A}F^*\hat{y}$ gives an MSE slope of -3 dB/octave of frame redundancy (dash-dot curve). Expanding using a frame generated by Method I or II and using consistent reconstruction results in an MSE slope of -6 dB/octave of frame redundancy (solid and dotted curves). A linear program always returns a corner of the consistent region. Since the consistent region is convex, we can get better performance by averaging the reconstructions found using two different cost functions. This is shown by the dashed curve.

4 Quantization Using Matching Pursuit

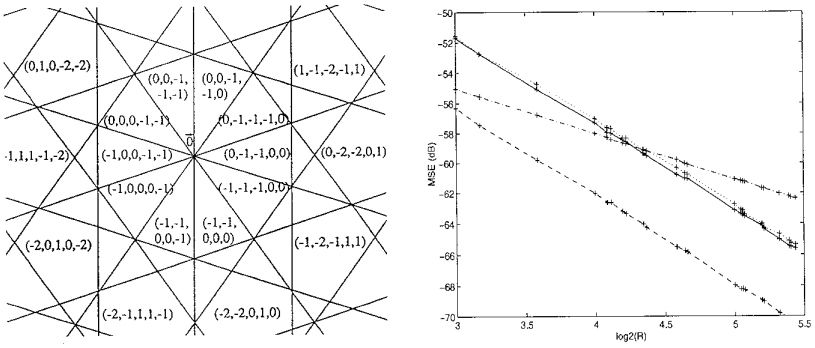
For quantization with good rate-distortion (R-D) performance, we do not expect to do well by quantizing a set of frame coefficients and retaining all of them. Furthermore, for relatively low rate coding, we expect that the best R-D performance would result from retaining a small number of coefficients. This motivates us to use a greedy algorithm to select a few inner products to retain.

Let $\mathcal{D} = \{\varphi_k\}_{k=1}^M \subset H$ be a frame . We impose the additional constraint that $||\varphi_k|| = 1\ \forall k$. We will call $\mathcal{D}$ our *dictionary* of vectors. Matching pursuit [4] 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. Hence we expect that after a few iterations we will have an efficient representation of f.

The algorithm begins by selecting k_0 such that $| < \varphi_{k_0}, f > |$ is maximized. Then f can be written as its projection onto φ_{k_0} and a residue $R_1 f$,

$$f =< \varphi_{k_0}, f > \varphi_{k_0} + R_1 f. \tag{10}$$

The algorithm is iterated by treating $R_1 f$ as the vector to be best approximated by



(a) (b)

Figure 2: (a) Partition of $\mathbb{R}^2$ induced by quantizing with $R = \frac{5}{2}$ tight frame. (b) Experimental results for reconstruction from quantized frame expansion. Dash-dot curve: "naive" reconstruction. Dotted curve: consistent reconstruction from expansion I. Solid curve: consistent reconstruction from expansion II. Dashed curve: improved consistent reconstruction.

a multiple of φ_{k_1}. Identifying $R_0 f = f$, we can write

$$f = \sum_{i=0}^{n-1} < \varphi_{k_i}, R_i f > \varphi_{k_i} + R_n f. \tag{11}$$

Hereafter we will denote $< \varphi_{k_i}, f >$ by α_i. Notice that since α_i is determined by projection, $\alpha_i \varphi_{k_i} \perp R_{i+1} f$. Thus we have the "energy conversation" equation

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

Since k_i is selected to maximize $|\alpha_i|$, the energy in the residue is strictly decreasing until f is exactly represented.

To use matching pursuit for quantization, at the ith stage we quantize α_i, yielding $\hat{\alpha}_i = Q[\alpha_i]$. The quantized version is used in determining the residual so that quantization errors do not propagate to subsequent iterations. Note that the coefficient quantization destroys the orthogonality of the projection and residual, so the analog of (12) does not hold.

At this point we have several design problems. We must choose a dictionary, design scalar quantizers, and decide how many quantized inner products to retain. In principle, one could optimize each of these for a given source distribution, distortion measure and rate measure.

EXAMPLE: Consider quantization of a source $\boldsymbol{x} = [\boldsymbol{X}_1, \boldsymbol{X}_2]^T$ with a uniform distribution on $[-1, 1]^2$. Suppose we want to use matching pursuit with a four element

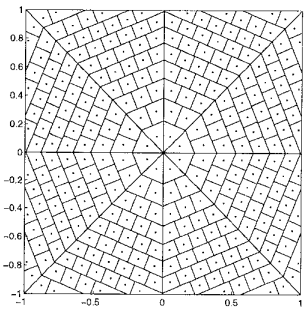


Figure 3: Partitioning of $[-1,1]^2$ by matching pursuit with four element dictionary

dictionary to design a codebook that minimizes MSE with a constrained codebook size. (Constraining codebook size is natural when a fixed rate coder will be applied to the quantizer output.) Guided by symmetry, we choose

$$\mathcal{D} = \left\{ [\cos\frac{k\pi}{8}, \ \sin\frac{k\pi}{8}]^T : k \in \{1,3,5,7\} \right\}. \tag{13}$$

It may seem more natural to use $k \in \{0,2,4,6\}$ in (13). The dictionary we have selected was determined to lead to a better design.

Having selected a dictionary, we can explicitly find the p.d.f. of α_0 to be

$$p_{\alpha_0}(y) = \begin{cases} 2(\sqrt{2}-1)|y| & |y| \le \frac{1}{2}\sqrt{2+\sqrt{2}} \\ -2(|y| - \sqrt{1+\sqrt{2}}) & \frac{1}{2}\sqrt{2+\sqrt{2}} < |y| \le \sqrt{1+\sqrt{2}} \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

We assume quantization is fine. Then the best codebook constrained quantizer for α_0 can be found analytically [2]. The distribution of α_1 given $\hat{\alpha_0}$ is approximately uniform on $[-|\hat{\alpha_0}|, |\hat{\alpha_0}|]$. Thus the optimal quantizer for α_1 is uniform.

We have yet to decide how to divide our bit rate between $\hat{\alpha_0}$ and $\hat{\alpha_1}$. Analysis shows that the strategy that is consistent with our optimality condition is to have the number of quantization levels for α_1 proportional to p_{α_0}. Using a codebook size of 304 and choosing the proportionality constant appropriately yields the codebook and partition shown in Figure 3.

This example was included to demonstrate several things. It illustrates 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. Lastly, Figure 3 shows that the partition generated by matching pursuit looks quite different than that generated by independent scalar quantization or quantization of tight frame coefficients.

In a practical situation, the source distribution is not known. Hence we endeavored to apply matching pursuit without making any 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 yet clear. The three methods described in Section 2 were used to generate dictionaries. Method I provides the most flexibility. However, large dictionaries of this type are not "evenly" distributed because they lie in the intersection of the unit sphere with the plane $x_1 = \frac{1}{\sqrt{N}}$. We present experimental results using each method.

Our 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. All inner product quantization was scalar, uniform and equal in each dimension. In addition to not relying on distributional assumptions, this is computationally easy and consistent with equally weighting the error in each direction. Distortion was measured by MSE and rate by summing the (scalar) entropies of the k_i's and α_i's retained. We denote the number of inner products retained by p and the quantization stepsize by Δ.

Figure 4 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 $p = 9$, respectively. The points along each dotted curve correspond various values of Δ. The dashed curve shows the performance of independently quantizing 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 choice p and Δ. The simulation results show that matching pursuit performs as well or better than independent scalar quantization for rates up to about 23 bits per vector (2.6 bits per source sample).

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

Simulations with $N = 25$ and R varied from 1 to 256 showed the best performance was achieved with R between 2 and 4. The matching pursuit quantizer outperformed the independent quantizer up to a rate of 75 bits per vector (3 bits per source sample).

Consider use of Method II. We select a dictionary of size eight in $\mathbb{R}^4$ from [3]. Figure 5(a) shows the R-D performance of four quantizers. The dashed curve results from using matching pursuit with separate entropy coding of each index and each coefficient. The solid 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 coefficients to retain changes. Independently quantizing in each dimension and scalar entropy coding gives the dotted curve. Replacing the scalar entropy coding by vector entropy coding gives the dash-dot curve.

At rates up to about 6 bits per vector (1.5 bits per source sample), matching pursuit quantization outperforms independent scalar quantization with either entropy coding method. At these rates, only one quantized inner product is retained. The matching pursuit quantization does better than independent quantization with scalar entropy coding for rates up to about 12 bits per vector (3 bits per source sample).

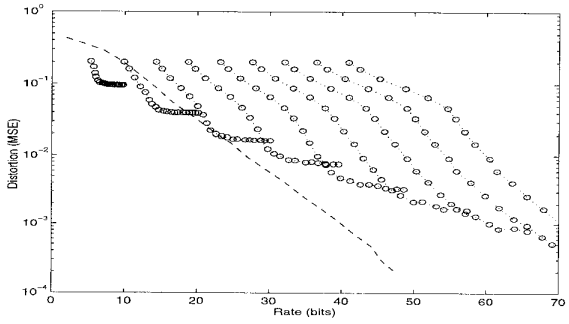


Figure 4: R-D performance of matching pursuit quantization with one to nine inner products retained. ($N = 9$, $R = 8$, dictionary generated using Method I.)

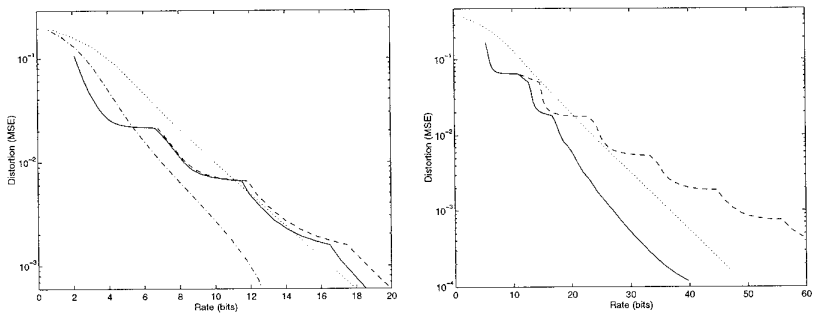
This simulation shows that vector entropy coding of indices gives improved performance at high rates. At first glance it may appear that at high rates independent quantization with vector entropy coding is far superior to other methods, but we must consider the complexity involved in the entropy coding. Consider operation at 8 bits/vector. The matching pursuit quantizer retains two coefficients, so the vector entropy code for the indices has $8^2 = 64$ symbols. The entropy codes for α_0 and α_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 5(b) shows simulations results using Method III with $N = 8$. The curve types are the same as in Figure 5(a). This simulation shows a great performance improvement in using vector entropy coding for the indices. The matching pursuit quantizer with vector entropy coded indices outperforms the independent scalar quantizer at all rates.

At this point a qualitative observation is in order. The advantage that we exploit over independent scalar quantization is that we represent the signal in the directions of maximal energy first and discard coefficients when they become disproportionately costly in an R-D sense. This is reminiscent of the KLT since the KLT transforms a signal to a representation where the coefficients are ordered to correspond to directions with decreasing energy. Thus we expect "on average" the dictionary elements chosen will correspond to the KLT. In simulations with $N = 2$ and R large, we find that histograms of k_0 and k_1 are sharply peaked at values corresponding to the first and second eigenvectors of the KLT, respectively.

References

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



(a) (b)

Figure 5: (a) Simulation results for $N = 4$ with 8 element dictionary obtained by Method II. (b) Simulation results for $N = 8$ with dictionary generated using Method III. Dotted curves: independent scalar quantization with scalar entropy coding. Dash-dot curve: independent scalar quantization with vector entropy coding. Dashed curves: matching pursuit quantization with scalar entropy coded indices. Solid curves: matching pursuit quantization with vector entropy coded indices.

[2] A. Gersho and R. M. Gray, "Vector Quantization and Signal Compression," Kluwer Academic Pub., Boston, 1992.

[3] 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/

[4] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. on SP*, Vol. 41, No. 12, pp. 3397–3415, Dec. 1993.

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

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

[7] 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. on SP*, Vol. 42, No. 1, pp. 200–203, Jan. 1994.

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