Lower Bound on the Mean Squared Error in Multi-Loop $\Sigma\Delta$ Modulation with Periodic Bandlimited Signals

Nguyen T. Thao

Department of EEE
Hong Kong University of Science and Technology
Kowloon, Hong Kong

Martin Vetterli

Department of EECS
University of California, Berkeley
Berkeley, CA 94720

Abstract

Recent work has been devoted to optimal signal reconstruction in $\Sigma\Delta$ modulation. Classically, using linear filtering, the reconstruction mean squared error (MSE) asymptotically decreases with the oversampling ratio R in $R^{-(2n+1)}$, where n is the order of the modulator. Using some non-linear reconstruction scheme, an improvement on the MSE asymptotic behavior by 3 dB per octave of R has been recently observed on time-varying, periodic and bandlimited input signals. This implies an MSE in $\mathcal{O}(R^{-(2n+2)})$. In this paper, we show for the multi-loop configuration of $\Sigma\Delta$ modulation and the same kind of input signals that the asymptotic behavior of the reconstruction MSE cannot be less than $\mathcal{O}(R^{-(2n+2)})$, including optimal reconstruction.

1 Introduction

Signal reconstruction in $\Sigma\Delta$ modulation is usually performed by a linear filtering of the output bit stream [1]. The idea is to cancel the high frequency components of the shaped quantization noise. Thus, in the first order $\Sigma\Delta$ case, the mean squared error (MSE) of the remaining noise can be reduced by 9 dB per octave of oversampling [1]. In other words, the asymptotic behavior of the MSE with respect to the oversampling ratio R is of the order of $\mathcal{O}(R^{-3})$. In the n^{th} order case, the MSE decreases by $3(2n+1)$ dB per octave, using linear filtering [2, 3]. The corresponding asymptotic behavior is of the order of $\mathcal{O}(R^{-(2n+1)})$.

It was recently shown in [4] that linear filtering is not optimal in the case of constant inputs. However, although improvements in the MSE reduction were observed with an optimal reconstruction, the asymptotic behavior is still of the same order, that is $\mathcal{O}(R^{-(2n+1)})$, and was in fact shown to be a lower bound on the reconstruction MSE.

Yet, it was numerically observed in [5] that optimal reconstruction yields an improved asymptotic behavior as soon as the input signals are time varying. Indeed, a non-linear reconstruction scheme introduced in [5] applied on time-varying, periodic and bandlimited input signals experimentally yields an asymptotic behavior of the order of $\mathcal{O}(R^{-(2n+2)})$ which represents an improvement of the MSE reduction by 3 dB/octave. In this paper, we show that this achieved asymptotic behavior is in fact a lower bound to the reconstruction MSE, for time-varying, periodic and bandlimited inputs applied to an n^{th} order multi-loop $\Sigma\Delta$ modulator (see Figure 1).

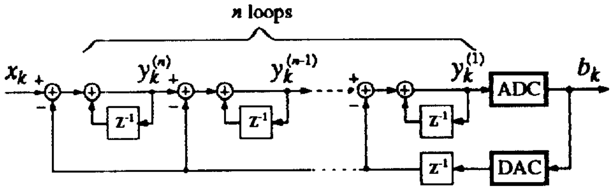

Figure 1: Block diagram of n-loop $\Sigma\Delta$ modulator

The proof is based on a vector quantization approach of $\Sigma\Delta$ modulation. Indeed, periodic and bandlimited signals necessarily belong to a finite dimensional space and can be considered as vectors. Thus, a $\Sigma\Delta$ modulator intrinsically defines a partition of the input vector space (Section 2). Assuming no quantization overloading, we show that this partition has the structure of a "hyperplane wave partition" (Sections 3 and 4). Thanks to this structure, an upper bound to the density of cells can be derived (Section 5). Using the work by Zador [6], a lower bound on the MSE is deduced. Its expression with respect to R gives the order $\mathcal{O}(R^{-(2n+2)})$ (Section 6). Finally, in the case of overloading, we explain qualitatively why this lower bound is still valid (Section 7).

1058-6393/95 $4.00 © 1995 IEEE

2 Vector quantization and partitioning approach

Optimal reconstruction was first introduced in $\Sigma\Delta$ modulation in the case of constant inputs in [4]. In this context, it was shown that the intrinsic behavior of a $\Sigma\Delta$ modulator is to define a one-to-one correspondence between intervals of constant amplitude values and the possible output bit streams. In other words, the encoding behavior of a $\Sigma\Delta$ modulator is characterized by the partition it defines in the space of input signals, which is one dimensional in the case of constant inputs. Optimal reconstruction simply consists in picking the center of the interval corresponding to the given output bit stream.

In this paper, we study the intrinsic behavior of a $\Sigma\Delta$ modulator for time-varying, periodic and bandlimited signals. Precisely, the signals we consider are defined and encoded on a time window $[0,T]$ and have the following finite Fourier expansion:

$$x(t)=X_1+\sum_{i=1}^{p} X_{2i}\sqrt{2}\cos\left(2\pi i\tfrac{t}{T}\right)+X_{2i+1}\sqrt{2}\sin\left(2\pi i\tfrac{t}{T}\right). \tag{1}$$

Writing $W=2p+1$, $x(t)$ belongs to the W dimensional space generated by the basis $(u_i(t))_{i\leq i\leq W}$ where

$$\begin{array}{rrcl} & u_1(t) & = & 1 \\ \forall i\geq 1, & \left\{ \begin{array}{r} u_{2i}(t) \\ u_{2i+1}(t) \end{array}\right. & \begin{array}{c} = \\ = \end{array} & \begin{array}{l} \sqrt{2}\cos(2\pi i\frac{t}{T}) \\ \sqrt{2}\sin(2\pi i\frac{t}{T}). \end{array} \end{array} \tag{2}$$

There is a one-to-one mapping between $x(t)$ and the vector $\vec{x}=(X_1,X_2,...,X_N)$ of $\mathbf{R}^W$. Therefore, $x(t)$ belongs to a W dimensional space.

As in the previous simple case, a $\Sigma\Delta$ modulator defines a partition of the input space where each cell comprises all input signals producing the same output bit stream. Figure 2 shows the partition defined by a single-loop $\Sigma\Delta$ modulator on the space generated by $(u_2(t),u_3(t))$. In the partitioning approach, it is important to see that the possible output bit streams have the function of "labeling" the cells. Then, as a known result in vector quantization [7], the optimal reconstruction of the input consists in picking the centroid of the corresponding cell. The MSE performance of the optimal reconstruction will give a lower bound on the reconstruction MSE in general. Thanks to the norm conservation $\int_0^T |x(t)|^2\,dt=\sum_{i=1}^{W}|X_i|^2$ which can be easily verified, the MSE can be entirely evaluated in the vector space $\mathbf{R}^W$ using its canonical norm.

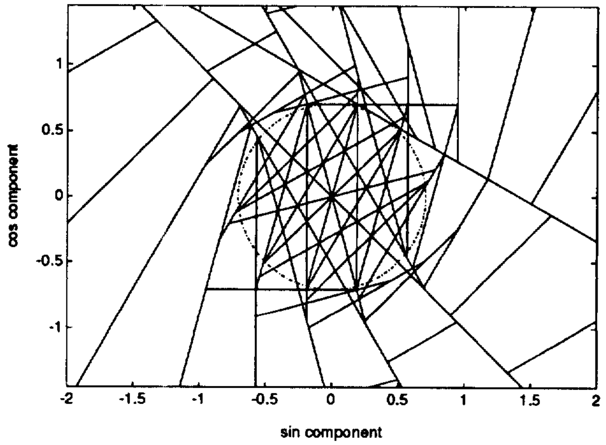

Figure 2: Partition defined by a single-bit single-loop $\Sigma\Delta$ modulator on the space of sinusoids of arbitrary phase and amplitude, sampled 12 times in $[0,T]$. The circle represents the no-overload region.

3 Analysis of the encoding process

Before deriving the partition defined by a multiloop $\Sigma\Delta$ modulator, we need to formalize the encoding process. It is composed of the sampling and followed by the $\Sigma\Delta$ encoding of the sampled values. We assume that $x(t)$ is sampled N times in $[0,T]$. Because an input is characterized by W values $X_1,X_2,...,X_W$ according to (1), $R=\frac{N}{W}$ represents the oversampling ratio. The N samples are denoted by $(x_1,x_2,...x_N)$. and are such that $x_k=x(\frac{k}{N}T)$. These samples can be directly expressed as a function of the vector $\vec{x}$ corresponding to $x(t)$. Indeed, let us define the following vector of $\mathbf{R}^W$:

$$\vec{f_k}=\left(u_1(\tfrac{k}{N}T),u_2(\tfrac{k}{N}T),...,u_W(\tfrac{k}{N}T)\right), \tag{3}$$

and denote the inner product of $\mathbf{R}^W$ by $\langle\vec{x},\vec{y}\rangle=\sum_{i=1}^{W}X_iY_i$, where X_i and Y_i are the i^{th} components of $\vec{x}$ and $\vec{y}$. Then, using (1), (2) and (3) we have

$$x_k=\sum_{i=1}^{W}X_i\;u_i(\tfrac{k}{N}T)=\left\langle\vec{x},\vec{f_k}\right\rangle. \tag{4}$$

Therefore, the whole encoding process can be represented as in Figure 3 where $(b_1,b_2,...,b_N)$ is the output bit stream.

Concerning the $\Sigma\Delta$ encoding, we assume that the quantizer (symbolized in Figure 1 by 'ADC') is uniform with a step size q. The transfer functions of the quantizer and the feedback D/A converter (symbolized in Figure 1 by 'DAC') are represented in Figure

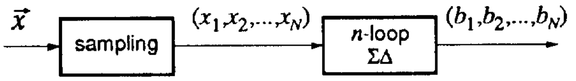

Figure 3: Representation of ΣΔ modulation with vector input

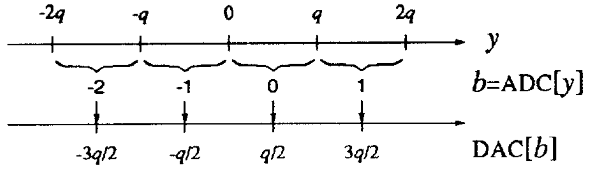

Figure 4: Transfer functions of built-in quantizer and D/A converter

4. Although the output of the quantizer is an integer, it is not so much the nominal value of this integer which matters, but its symbolic value as a codeword.

We assume for now that no quantization overloading occurs during the encoding process. In this situation, the transfer functions of the quantizer and the D/A converter can be mathematically expressed as

$$ADC[y] = \left\lfloor \frac{y}{q} \right\rfloor \quad \text{and} \quad DAC[b] = q\left(b + \tfrac{1}{2}\right), \qquad (5)$$

according to Figure 4, where $\lfloor w \rfloor$ denotes the largest integer which is lower or equal to w.

Based on these definitions, we show in the appendix that the block diagram of Figure 1 is entirely equivalent to that of Figure 5 where $\lfloor\cdot\rfloor$ denotes the operator which maps w into $\lfloor w \rfloor$. In this figure, for a given order n, $v_k^{(n)}$ is a deterministic sequence which does not depend on the input x_k and is recursively defined with n as follows: $\forall k \geq 1$, $v_k^{(0)} = 0$ and for $p = 1, ..., n$,

$$v_0^{(p)} = \frac{y_0^{(p)}}{q} - b_0 \text{ and } \forall k \geq 1,\ v_k^{(p)} = \tfrac{1}{2} + v_k^{(p-1)} - v_{k-1}^{(p-1)}. \qquad (6)$$

In this definition, $y_0^{(p)}$ is the initial value of the p^{th} accumulator of the ΣΔ modulator and $b_0 = ADC[y_0^{(1)}]$ (see Figure 1). Also, the n integrators in the equivalent diagram are initialized to zero at $k = 0$.

We propose now to give the exact content of the block diagram of Figure 3. Equation (4) gave us the direct expression of x_k in terms of the input vector $\vec{x}$. This permits the direct expression of z_k (see Figure 5) in terms of $\vec{x}$. Indeed, in the single-loop case ($n = 1$) for example, we have:

$$z_k = \sum_{j=1}^{k} \left(\frac{x_k}{q} - v_j^{(1)} \right) = \sum_{j=1}^{k} \left(\frac{1}{q} \left\langle \vec{f}_k, \vec{x} \right\rangle - v_j^{(1)} \right),$$

$$z_k = \left\langle \vec{d}_k, \vec{x} \right\rangle - C_k, \qquad (7)$$

where $\vec{d}_k = \frac{1}{q} \sum_{j=1}^{k} \vec{f}_j$ and $C_k = \sum_{j=1}^{k} v_j^{(1)}$. In the n^{th} order case, it is easy to show that (7) is still valid with

$$\vec{d}_k = \frac{1}{q} \sum_{j_n=1}^{k} \cdots \sum_{j_2=1}^{j_3} \sum_{j=1}^{j_2} \vec{f}_j \qquad (8)$$

$$\text{and} \quad C_k = \sum_{j_n=1}^{k} \cdots \sum_{j_2=1}^{j_3} \sum_{j=1}^{j_2} v_j^{(n)}.$$

Then, combining equation (7) and Figure 5, the whole encoding process of Figure 3 has the block diagram of Figure 6, where $\left\langle \vec{d}_i, \cdot \right\rangle - C_i$ denotes the operator which maps $\vec{x}$ into $\left\langle \vec{d}_i, \vec{x} \right\rangle - C_i$.

4 Derivation of the partition

The partition defined by the encoder of Figure 6 is basically due to the quantization operator $\lfloor\cdot\rfloor$. The portion of the block diagram, called "partitioning section", already defines a partition of the input space, where each cell is labeled by a possible sequence of discrete elements $(a_1, a_2, ..., a_N)$ output by the operator $\lfloor\cdot\rfloor$ (see Figure 6). The next portion of the block diagram, called "discrete section", only works as mapping from $(a_1, a_2, ..., a_N)$ to $(b_1, b_2, ..., b_N)$ which is also a sequence of discrete elements. One can easily verify that this mapping is a one-to-one correspondence. Therefore, it can be interpreted as a simple "relabeling" operator. Then, the partition respectively defined by the partitioning section and the whole encoder are necessarily the same, where the only difference lies in the label associated with each cell.

The partitioning section has a structure which has been recently studied in [8]. It was shown that this structure yields a type of partition, called "hyperplane wave partition", which has the following features:

(i) - the cells are formed from the division of the space by non-interrupted hyperplanes,

(ii) - these hyperplanes are perpendicular to $\vec{d}_1$, $\vec{d}_2$, ..., or $\vec{d}_N$.

(iii) - the hyperplanes perpendicular to a given vector $\vec{d}_k$ are equally spaced with a distance equal to $1/d_k$, where $d_k = \|\vec{d}_k\|$.

These properties can be observed on the partition of Figure 2 in the no-overload region.

5 Upper bound on the number of cells

Thanks to the properties of hyperplane wave partitions, an upper bound on the number of cells was

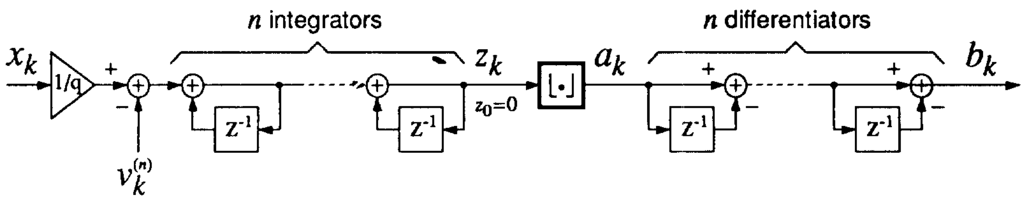

Figure 5: Equivalent block diagram for an n-loop $\Sigma\Delta$ modulator in the case of no overloading (see appendix)

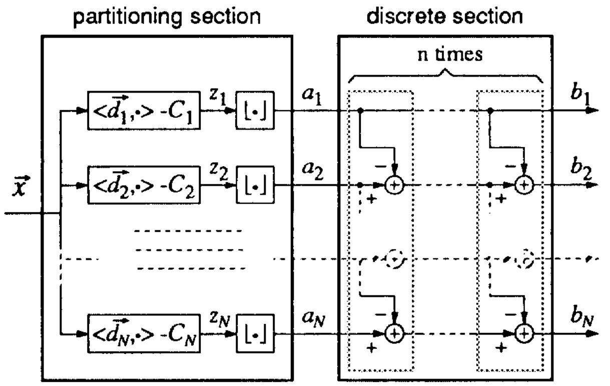

Figure 6: Vector quantization block diagram of an n-loop $\Sigma\Delta$ modulator

derived in [8]. This upper bound is given as follows:

Proposition 1 *Let M be the number of cells defined by the partitioning section of the encoder of Figure 6 within a region of $\mathbf{R}^W$ of diameter D and $d = \max(d_1, d_2, ..., d_N)$. Then*

$$M \leq \binom{N}{W} (Dd+2)^W. \qquad (9)$$

From (9), let us derive an upper bound for N asymptotically large. We first have $\binom{N}{W} = \frac{N(N-1)...(N-W+1)}{W!} \leq \frac{N^W}{W!}$. Then, we recall that $d_k = \|\vec{d_k}\|$ where $\vec{d_k}$ is given by (8). Using definitions (3) and (4), we have

$$d_k^2 = \|\vec{d_k}\|^2 = \frac{1}{q^2} \sum_{i=1}^{W} \left| \sum_{j_n=1}^{k} \cdots \sum_{j_2=1}^{j_3} \sum_{j=1}^{j_2} u_i(\tfrac{k}{N}T) \right|^2 .$$

Using the fact that $\left|u_i\left(\frac{j}{N}T\right)\right| \leq \sqrt{2}$ and $\forall k \leq N$, $\sum_{j_n=1}^{k} \cdots \sum_{j_2=1}^{j_3} \sum_{j=1}^{j_2} 1 \leq k^n \leq N^n$, we find that $d_k^2 \leq \frac{W}{q^2} \left|\sqrt{2}N^n\right|^2$. Therefore, $d \leq \frac{\sqrt{2W}}{q} N^n$. Then, from (9) we have for N asymptotically large:

$$M \leq \frac{N^W}{W!} \left(\frac{D}{q}\sqrt{2W}N^n + 2\right)^W \simeq \left(\frac{D}{q}\frac{\sqrt{2W}}{W!^{1/W}} N^{n+1}\right)^W . \qquad (10)$$

6 Derivation of the MSE lower bound

We assume that the input signals $x(t)$ have a standard deviation bounded by σ in the time window $[0, T]$. Using the norm conservation formula, this implies that the associated vector $\vec{x}$ is confined in a region of $\mathbf{R}^W$ of diameter $D = 2\sigma$. Let us call $\mathbf{p}$ the probability distribution of the vector $\vec{x}$ within this region.

Zador showed in [6] that the optimal reconstruction MSE can be lower bounded when the number M of cells defined by the given encoder is known. When M is large enough [8], he showed that

$$MSE_{opt} \geq C(W, \mathbf{p}) \cdot M^{-2/W}, \qquad (11)$$

where $C(W, \mathbf{p})$ is a coefficient which only depends on the dimension W of the input space and the probability distribution $\mathbf{p}$ of the input vectors. When the encoder is a multi-loop $\Sigma\Delta$ modulator, M is upper bounded as in (10). Then, we obtain from (11)

$$MSE_{opt} \geq C(W, \mathbf{p}) \left(\frac{q}{D}\right)^2 \frac{W!^{2/W}}{2W} \frac{1}{N^{2n+2}}.$$

Using $D = 2\sigma$ and $N = R \cdot W$, we derive that

$$MSE_{opt} \geq C(W, \mathbf{p}, n) \left(\frac{q}{\sigma}\right)^2 \frac{1}{R^{2n+2}}, \qquad (12)$$

where $C(W, \mathbf{p}, n) = \frac{W!^{2/W}}{8W^{2n+3}} C(W, \mathbf{p})$ which only depends on W, $\mathbf{p}$ and n.

The inequality (12) shows that the order $\mathcal{O}(R^{-(2n+2)})$ is asymptotically a lower bound to the reconstruction MSE.

7 Case of quantization overloading

In reality, the output of the quantizer is limited to a finite number of integers, from b_{min} to b_{max}. Overloading occurs when the input y of the quantizer is such that $\lfloor y/q \rfloor$ is outside the range $b_{min}, ..., b_{max}$. The quantizer behaves as if the function $\lfloor \cdot \rfloor$ was followed by a many-to-one integer mapping which assimilates every integer larger than b_{max} with b_{max} and every integer smaller than b_{min} with b_{min}. It can be derived that the equivalent block diagram of Figure 6 remains valid when including in the discrete section an extra many-to-one mapping of the discrete element sequences. The only possible effect on the overall partition is the merging of certain cells with each other. This can only decrease the effective number M of cells in a given region. Thus, the upper bound of (10) holds. So does the MSE lower bound.

8 Appendix: Structure equivalence of multi-loop $\Sigma\Delta$ modulation

The equivalence between Figures 1 and 5 is based on definitions (5) and (6), and on the following proposition:

Proposition 2

$$\forall k \geq 1, \ \sum_{j_n=1}^{k}\cdots\sum_{j=1}^{j_2} b_j = \left\lfloor \sum_{j_n=1}^{k}\cdots\sum_{j=1}^{j_2}\left(\frac{x_j}{q} - v_j^{(n)}\right)\right\rfloor . \tag{13}$$

Sketch of the proof: We perform the proof by induction on n, using the recursive block diagram of an n-loop $\Sigma\Delta$ modulator from [3], as shown in Figure 7. Using the signal notations of this figure and definition (5), one can derive that

$$\frac{y_k^{(n)}}{q} = \sum_{j=1}^{k}\frac{x_j}{q} - \left(\frac{1}{2} - \frac{y_0^{(n)}}{q} + b_0\right) - \sum_{j=1}^{k-1} b_j, \tag{14}$$

In the case $n = 1$, the $(n-1)$-loop $\Sigma\Delta$ modulator is reduced to a quantizer. Therefore, $b_k = \lfloor y_k^{(1)}/q \rfloor$. In general, if $A, B \in \mathbf{R}$ such that B is an integer, then $\lfloor A + B \rfloor = \lfloor A \rfloor + B$. Because $\sum_{j=1}^{k-1} b_j$ is an integer, then (14) implies that

$$b_k = \left\lfloor \frac{y_k^{(1)}}{q} \right\rfloor = \left\lfloor \sum_{j=1}^{k}\frac{x_j}{q} - \left(\frac{1}{2} - \frac{y_0^{(1)}}{q} + b_0\right)\right\rfloor - \sum_{j=1}^{k-1} b_j. \tag{15}$$

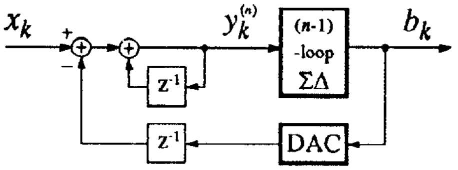

Figure 7: Recursive block diagram of an n-loop $\Sigma\Delta$ modulator [3]

Using (6), this gives $\sum_{j=1}^{k} b_j = \left\lfloor \sum_{j=1}^{k}\left(\frac{x_j}{q} - v_j^{(1)}\right)\right\rfloor$ which proves (13) for $n = 1$. To prove (13) for $n \geq 2$, the procedure is to assume that (13) is true for the $(n-1)$-loop $\Sigma\Delta$ modulator included in Figure 7. This gives an expression of (13) at the order $n - 1$, where the input x_j is replaced by $y_j^{(n)}$. In this expression, inject the expression of $y_j^{(n)}/q$ given in (14). This leads to equation (13) at the order n □

Let us define $z_k = \sum_{j_n=1}^{k}\cdots\sum_{j_2=1}^{j_3}\sum_{j=1}^{j_2}\left(\frac{x_j}{q} - v_j^{(n)}\right)$ and $a_k = \lfloor z_k \rfloor$. The computation of z_k and a_k can be indeed represented as shown in the block diagram of Figure 5. Because of (13), a_k is the n^{th} order discrete integration of the sequence b_k. Therefore b_k can be obtained by a n^{th} order discrete differentiation of a_k. Thus b_k is the output of the block diagram of Fig. 5.

References

[1] J. Candy and G.C.Temes, eds., *Oversampling delta-sigma data converters. Theory, design and simulation.* IEEE Press, 1992.

[2] W.Chou, P.-W.Wong, and R.M.Gray, "Multistage sigma-delta modulation," *IEEE Trans. Information Theory*, vol. 35, pp. 784–796, July 1989.

[3] N.He, F.Kuhlmann, and A.Buzo, "Multi-loop sigma-delta quantization," *IEEE Trans. Information Theory*, vol. IT-38, pp. 1015–1028, May 1992.

[4] S.Hein and A.Zakhor, "Lower bounds on the MSE of the single and double loop sigma delta modulators," *Proc. IEEE Int. Symp. Circ. and Systems*, pp. 1751–1755, May 1990.

[5] N.T.Thao and M.Vetterli, "Oversampled A/D conversion using alternate projections," *Conf. on Information Sciences and Systems, the Johns Hopkins University*, pp. 241–248, Mar. 1991.

[6] P.L.Zador, "Development and evaluation of procedures for quantizing multivariate distributions," *Ph.D. Dissertation, Stanford University*, 1963. Microfilm 64-9855.

[7] A.Gersho and R.M.Gray, *Vector quantization and signal compression.* Kluwer Academic Publishers, 1992.

[8] N.T.Thao and M.Vetterli, "Lower bound on the mean squared error in oversampled quantization of periodic signals," *IEEE Trans. Information Theory.* Submitted.