OPTIMAL MSE SIGNAL RECONSTRUCTION IN OVERSAMPLED A/D CONVERSION USING CONVEXITY

Nguyen T. Thao and Martin Vetterli *

Department of Electrical Engineering
and Center for Telecommunications Research
Columbia University, New York, NY 10027-6699

ABSTRACT

Signal reconstruction in oversampled A/D conversion is classically performed by a lowpass filtering of the quantized signal. This leads to an MSE inversely proportional to R^{2n+1} where R is the oversampling rate and n is the order of the converter. We show that this reconstruction does not necessarily lead to a signal which gives the same digital sequence as that of the original input signal. If this were indeed the case, we show that we would obtain an MSE inversely proportional to R^{2n+2} instead. We propose a way to actually achieve such an estimate with the same bandwidth and same digital conversion. This enabled us to perform numerical experiments which confirm the R^{2n+2} behavior of the MSE.

1. Introduction

Oversampled A/D conversion (ADC) is a successful and alternative way to achieve high resolution data acquisition from bandlimited analog signals. This technique is based on the statistical properties of the error signal generated by quantizing the input signals. The assumption of white quantization noise has led to a good evaluation of the remaining noise power in the input bandwidth after oversampled ADC. It was found that when a converter is an n^{th} order noise shaper, such as n^{th} order $\Sigma\Delta$ modulators, the in-band noise power is proportional to $1/R^{2n+1}$, where R is the oversampling rate [1, 2]. Gray actually justified the white noise assumption up to the second moment on single-loop, multi-stage $\Sigma\Delta$ modulators of order greater than 2, with dc and sinusoidal inputs [3].

The first question we ask in this paper is the following: is lowpass filtering the quantized signal the best signal reconstruction we can perform from the digital sequence? Is the remaining in-band noise irreversible? Studying these questions, we noticed in [5, 6] that in the case of simple oversampled ADC, reduced to pure quantization, such reconstruction does not necessarily lead to a signal which gives the same digital sequence as that of the original input signal. This gave us the first hint that such signal reconstruction is not theoretically optimal. Then, assuming periodicity of input signals and certain conditions on threshold crossings, we showed analytically that an estimate with same bandwidth and same digital conversion as that of the analog input signal necessarily yields a mean square error (MSE) inversely proportional to R^2 instead of R; that is, an improvement of 6dB per octave of oversampling rather than only 3dB/octave.

*Work supported in part by the National Science Foundation under grants ECD-88-11111.

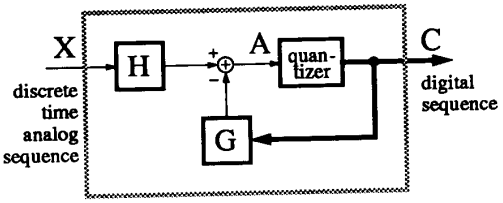

Figure 1: Block diagram of a single-quantizer A/D converter: H is an invertible affine operator and G a causal digital to analog operator.

In this paper, we try to answer the same questions in the case of n^{th} order oversampled A/D conversion. We first present a block diagram model that describes the mechanisms of most n^{th} order A/D converters currently known. Then, based on this model, we evaluate an upper bound on the MSE when taking as signal reconstruction an estimate which has the same bandwidth and gives the same digital sequence as that of the original input signal. Assuming that input signals are periodic, we immediately find an MSE upper bound of the order of $\mathcal{O}(R^{-2n})$. Then starting from the model of white and uniform quantization noise, we show mathematically that the MSE has an upper bound inversely proportional to R^{2n+2}, instead of R^{2n+1} in classical reconstruction. Due to the fact that the set of analog signals giving the same digital output sequence is convex, it is possible by projections to achieve such an estimate from the knowledge of the digital sequence. Our numerical experiments confirm the R^{2n+2} behavior of the MSE.

2. Model of n^{th} order A/D converter

We model an oversampled A/D converter by the block diagram of figure 1. For example, a 1^{st} order $\Sigma\Delta$ modulator can be reduced to figure 1 by taking H as a discrete integrator and G as the composition of a D/A converter, a discrete integrator and a sign change. We show in [6] how converters such as simple, dithered, predictive converters and multi-bit multi-loop multi-stage $\Sigma\Delta$ or interpolative modulators can be reduced to this structure. For multi-stage modulators, it is however necessary to collect the individual binary output of every in-built quantizer.

It can be shown, based on this diagram, that the set $\Gamma(C_0)$ of all analog input signals giving a particular digital output sequence C_0 is convex. Indeed, the set of signals A giving C_0 through the quantizer is naturally convex. Then $\Gamma(C_0)$ is simply the image of this set through the transform $A \longmapsto H^{-1}[A + G[C_0]]$ which is affine, since $G[C_0]$ is a fixed signal. A more complete proof is shown in [6]. This convexity property gives us the first hint that, if an original signal X_0 gives a digital sequence C_0 ($X_0 \in \Gamma(C_0)$), then an estimate X of X_0 which does not yield C_0 as a digital

0-7803-0532-9/92 $3.00 © 1992 IEEE

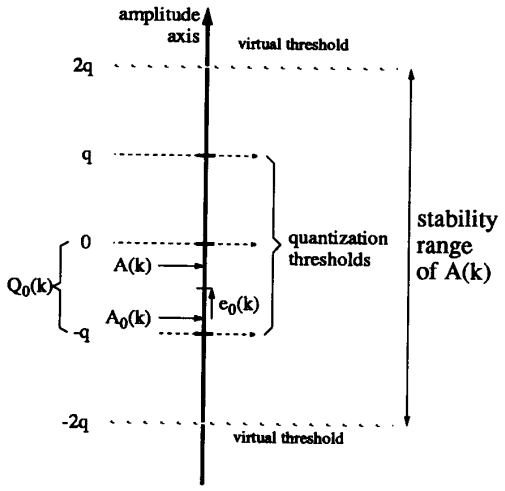

Figure 2: Amplitude subdivision of signal A for a 2 bit uniform quantization, with stability assumption.

output ($X \notin \Gamma(C_0)$) can always be improved. Indeed, because $\Gamma(C_0)$ is convex, X necessarily gets closer to X_0 when orthogonally projected on $\Gamma(C_0)$.

For most oversampled A/D converters, the operator H of the equivalent structure in figure 1 is a discrete integrator of a certain order n [6]. Qualitatively speaking, the signal A coded by the quantizer is an "amplified" version of the input signal. This explains intuitively why a high resolution description of the input signal can be achieved with a coarse resolution quantizer. However, the difficulty of such a coding system lies in the fact that the input range of the quantizer is limited. One of the achievements of $\Sigma\Delta$ modulators is to bring this integrated signal back to the amplitude range of the quantizer, thanks to a well designed feedback. This is typically the case of multi-stage $\Sigma\Delta$ modulators. Coming back to the structure of figure 1, the G output signal can be interpreted as a "smart" reference signal which insures that the integrated signal A will always fall into one of the finite length quantization intervals.

In this paper we assume that H is an n^{th} order discrete integrator, and that the quantizer is uniform with a step size equal to q. We include the particular case where the quantizer is reduced to a single threshold comparator (single-bit converters). We suppose that the coding process is stable in the sense that, thanks to the feedback operator G, the signal never goes beyond a distance equal to q from the extreme quantization thresholds. This is for example the case for single-bit 1^{st} order and single-loop multi-stage $\Sigma\Delta$ modulators when the feedback DAC output values are adjusted to $\pm\frac{q}{2}$ and input signals remain in the interval $[-\frac{q}{2}, \frac{q}{2}]$. Under this assumption, the amplitude range of possible values of A is divided into a finite number of intervals of length q. Moreover, when an input signal X gives a digital output sequence C_0, the digital value $C_0(k)$ of C_0 at time k determines which of these length q intervals the value $A(k)$ of A belongs to. Figure 2 shows an example of a 2 bit quantizer. As a consequence, if two input signals X_0 and X have the same digital output C_0, then the signals A_0 and A respectively seen by the quantizer are such that, at every instant k, $A_0(k)$ and $A(k)$ belong to the same quantization interval $Q_0(k)$, indicated by $C_0(k)$, and

$$|A(k) - A_0(k)| \leq q. \tag{1}$$

3. First MSE upper bound $\mathcal{O}(R^{-2n})$

Starting from the assumption that input signals are bandlimited and periodic, it appears from this model that the MSE between two signals having the same bandwidth and giving the same digital output has an upper bound inversely proportional to R^{2n}. Let us first define the notations and conditions of our analysis.

We suppose that input signals belong to the subspace V_0 of bandlimited and periodic signals with a fixed period T_0. They can be written

$$X[t] = A + \sum_{j=1}^{N} B_j\sqrt{2}\cos(2\pi j\frac{t}{T_0}) + \sum_{j=1}^{N} C_j\sqrt{2}\sin(2\pi j\frac{t}{T_0})$$

where (A, B_j, C_j) are the real Fourier coefficients of X, and N the discrete maximum frequency. Therefore V_0 is $2N+1$ dimensional with the following norm

$$\|X\| = \left(A + \textstyle\sum_{j=1}^{N} |B_j|^2 + \sum_{j=1}^{N} |C_j|^2\right)^{1/2}.$$

We suppose that input signals are sampled M times within one period T_0. We will designate by $X(k) = X[\frac{k}{M}T_0]$ the k^{th} sample of X. We suppose that $M > 2N+1$. The oversampling rate is then $R = \frac{M}{2N+1}$ and is thus proportional to M. It can be shown from Parseval's equality that for any signal $X \in V_0$, $\forall M > 2N+1$, $\|X\|^2 = \frac{1}{M}\sum_{k=1}^{M} |X(k)|^2$. When two signals X_0 and X belong to V_0, the MSE between them is therefore

$$MSE(X_0, X) = \frac{1}{M}\sum_{k=1}^{M} |X(k) - X_0(k)|^2 = \|X - X_0\|^2.$$

The upper bound $\mathcal{O}(R^{-2n})$ comes as follows. Suppose that $X_0, X \in V_0$ give the same digital sequence C_0 and call A_0, A the corresponding signals entering the quantizer. Working on figure 1, it appears that

$$A - A_0 = H[X] - H[X_0] = H[X - X_0] \tag{2}$$

since the output of G is equal to $G[C_0]$ for both analog input signals. Therefore, the difference $A - A_0$ is the n^{th} order integration of $X - X_0$. But this difference is limited by equation (1). The following lemma says that, since $X - X_0$ is bandlimited, the maximum value of $|A - A_0|$ is of the order of M^n.

Lemma 1 *There exist two constants $0 < c_1 < c_2$ such that, for M large enough,*

$$\forall x \in V_0, \quad c_1\|x\|M^n \leq \max_{1\leq k\leq M} |a(k)| \leq c_2\|x\|M^n \tag{3}$$

where $a = H[x]$ is the n^{th} order discrete integration of the M point sequence $x(k) = x[\frac{k}{M}T_0]$.

Using equation (1) and applying the lower bound of equation (3) to $x = X - X_0$, we find that $\|X - X_0\| \leq \frac{q}{c_1 M^n} \propto \frac{1}{R^n}$ which implies that $MSE(X_0, X)$ has an upper bound of the order of R^{-2n}. For a proof of lemma 1, see [6].

4. Second MSE upper bound $\mathcal{O}(R^{-(2n+2)})$

In reality, the distance between $A_0(k)$ and $A(k)$ is more constrained than in (1). Suppose for a moment that we had the stronger constraint

$$|A(k) - A_0(k)| \leq q/M \tag{4}$$

instead of (1). Then going through the derivations of the previous paragraph we would find that $\|X - X_0\| \leq \frac{q}{c_1 M^{n+1}} \propto \frac{1}{R^{n+1}}$ which is equivalent to saying that $MSE(X_0, X)$ has an upper bound of the order of $R^{-(2n+2)}$.

There is of course no reason for (4) to be true in general, but we want to give the intuition that (4) might happen sometimes. In practice, the relative position of $A_0(k)$ within the quantization interval $Q_0(k)$ which contains it, moves in a "random" way from one time index to another. If we suppose this relative position to be uniformly distributed within $Q_0(k)$, it can be shown [6] that the minimum distance of $A_0(k)$ to the upper (or lower) boundary of $Q_0(k)$, over the time range $k = 1, ..., M$, has an expectation of the order of q/M. On the other hand, $A - A_0$ is a slowly varying signal (being the n^{th} order integration of a bandlimited signal). Since $A(k)$ is constrained to remain between $A_0(k)$ and the boundaries of $Q_0(k)$, it will be "sometimes" forced to stay at a distance of $A_0(k)$ upper bounded by q/M.

The relative position of $A_0(k)$ within $Q_0(k)$ is in fact characterized by the quantization error $e_0(k)$ shown in figure 2 and defined as the difference between $A_0(k)$ and the center of interval $Q_0(k)$. We prove in this paragraph that, if we assume this error signal to be white and uniform, then the expectation of $MSE(X_0, X)$ will be actually bounded by $\mathcal{O}(R^{-(2n+2)})$. This will be based on the following lemma, proved in [6].

Lemma 2 *There exist two constants $0 < c_3 < c_4$ such that, for M large enough,*

$$\forall x \in V_0, \quad c_3 \|x\| M^n \leq \frac{1}{M} \sum_{k=1}^{M} |a(k)| \leq c_4 \|x\| M^n \tag{5}$$

where $a = H[x]$ is the n^{th} order discrete integration of the M point sequence $x(k) = x[\frac{k}{M} T_0]$.

Theorem 1 *Assume that for any $M > 2N + 1$, the quantization error values $e_0(1), ..., e_0(M)$ resulting from coding an input signal $X_0 \in V_0$ are independent random variables with a uniform probability density in $[-\frac{q}{2}, \frac{q}{2}]$. Then there exists a constant $\alpha_0 > 0$ such that for R large enough*

$$E[MSE(X_0, X)] \leq \frac{\alpha_0}{R^{2n+2}}$$

where $X_0 \in V_0$ is an input signal, $X \in V_0$ is a random estimate of X_0 giving the same digital sequence, and $E[MSE(X_0, X)]$ is the expectation of $MSE(X_0, X)$.

Proof : We first introduce some notations.
- $X /\!\!/ Y$ means $\exists \beta \geq 0, \quad X = \beta Y$
- $Prob(A)$ is the probability of event "A"
- $Prob(A \ / \ B)$ is the conditional probability of event "A" given the event "B".

Let U be a fixed element of V_0 such that $\|U\| = 1$. We designate by $E_U[MSE(X_0, X)]$ the conditional expectation of $MSE(X_0, X)$ given that $(X - X_0) /\!\!/ U$. We have

$$E_U[MSE(X_0, X)] = \int_{\lambda=0}^{+\infty} \lambda^2 f_U(\lambda) d\lambda \quad \text{where}$$

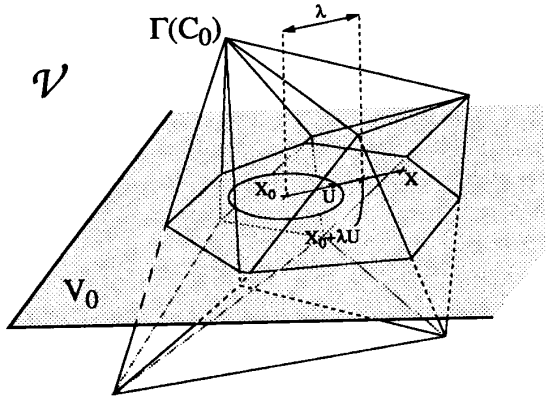

Figure 3: Space representation of signal X_0 and estimate X in $V_0 \cap \Gamma(C_0)$.

$f_U(\lambda)d\lambda = Prob\left(\|X - X_0\| \in [\lambda, \lambda + d\lambda] \ / \ (X - X_0) /\!\!/ U\right)$.
If we define the cumulative probability

$$P_U(\lambda) = Prob\left(\|X - X_0\| \geq \lambda \ / \ (X - X_0) /\!\!/ U\right) \tag{6}$$

$$\text{then} \quad f_U(\lambda) = -\tfrac{d}{d\lambda} P_U(\lambda). \tag{7}$$

Let us find an upper bound to $P_U(\lambda)$. Suppose that $P_U(\lambda) \neq 0$. Saying that $\|X - X_0\| \geq \lambda$ given that $(X - X_0) /\!\!/ U$ implies that $X_0 + \lambda U$ belongs to the segment $[X_0, X]$ (see figure 3). By convexity of $\Gamma(C_0)$ where C_0 is the digital sequence commonly output by X_0 and X, then $X_0 + \lambda U$ belongs to $\Gamma(C_0)$ and therefore gives the same digital sequence C_0. If A_0 and A are the signals seen by the quantizer when X_0 and $X_0 + \lambda U$ are respectively input, then equation (2) can be applied to X_0 and $X_0 + \lambda U$:

$$A - A_0 = H[(X_0 + \lambda U) - X_0] = \lambda H[U] = \lambda V$$

where $V = H[U]$ is the n^{th} order discrete integration of the M point sequence $U(k) = U[\frac{k}{M} T_0]$. Since $A(k)$ belongs to the same quantization interval as that of $A_0(k)$ at every instant k (see figure 2), then

$$|A(k) - (A_0(k) + e_0(k))| \leq q/2$$

which is equivalent to

$$|\lambda V(k) - e_0(k)| \leq q/2. \tag{8}$$

Since $|e_0(k)| \leq q/2$, this implies that

$$\forall k = 1, ..., M, \quad \lambda |V(k)| \leq q. \tag{9}$$

Then equation (8) is equivalent to $e_0(k) \in I_{\lambda V(k)}$, where $I_a = [-\frac{q}{2}, \frac{q}{2}] \cap [-\frac{q}{2} + a, \frac{q}{2} + a]$, $a \in \mathbf{R}$. Therefore, from (6)

$$P_U(\lambda) \leq Prob\left(\forall k = 1, ..., M, \quad e_0(k) \in I_{\lambda V(k)}\right).$$

When $|a| \leq q$, the length of I_a is $q - |a|$. Since $e_0(k), k = 1, .., M$ are independent random variables with a uniform distribution in $[-\frac{q}{2}, \frac{q}{2}]$, then

$$P_U(\lambda) \leq \prod_{k=1}^{M} Prob\left(e_0(k) \in I_{\lambda V(k)}\right) = \prod_{k=1}^{M} \frac{q - \lambda |V(k)|}{q}. \tag{10}$$

Using the inequality $1 - a \leq e^{-a}$ and applying the lower bound of equation (5) on $x = U$, then

$$P_U(\lambda) \leq \exp\left(-\frac{\lambda}{q}\sum_{k=1}^{M}|V(k)|\right) \leq e^{-c_3 M^{n+1}\lambda/q}. \quad (11)$$

Note that this inequality is trivially verified when $P_U(\lambda) = 0$. Therefore, $\lim_{\lambda\to+\infty} P_U(\lambda) = 0$, and, using (7), the conditional expectation can be calculated with an integration by parts

$$E_U\left[MSE(X_0, X)\right] = 2\int_0^{+\infty} \lambda P_U(\lambda) d\lambda.$$

Using (11) and performing another integration by parts, we find

$$E_U\left[MSE(X_0, X)\right] \leq \frac{2}{(c_3)^2}\frac{1}{M^{2n+2}}.$$

This upper bound does not depend on the choice of $U \in V_0$ with $\|U\| = 1$. Therefore, the unconditional expectation has the same bound:

$$E\left[MSE(X_0, X)\right] \leq \frac{2}{(c_3)^2}\frac{1}{M^{2n+2}} = \frac{\alpha_0}{R^{2n+2}}$$

where $\alpha_0 = 2(c_3)^{-2}(2N+1)^{-(2n+2)}$ □

The assumption of white and uniform quantization error signal has been used as a convenient way to obtain inequality (10). However, it may be possible to obtain this inequality with some multiplicative constant and a weaker probabilistic assumption about the quantization error signal (see [6]).

5. Numerical experiments

Since the white and uniform quantization noise was only a model in our analysis, we did numerical experiments to verify this $\mathcal{O}(R^{-(2n+2)})$ upper bound in practice. We designed algorithms based on the fact that alternating projections between V_0 and $\Gamma(C_0)$ converges to an estimate in $V_0 \cap \Gamma(C_0)$. We worked with sinusoidal inputs drawn at random and calculated averaged MSE over 300 to 1000 experiments for each configuration of A/D conversion and oversampling rate. In figure 4, we compare our results with the theoretical calculation of SNR in classical reconstruction from lowpass filtering the quantized signal, which is proportional to R^{2n+1} [1]. We obtain a systematic improvement of signal reconstruction, regardless of the order n. The gain of SNR we obtained over classical reconstruction, shown in figure 5, yields an asymptotic slope of 3dB per octave of oversampling. This confirms that the MSE upper bound is inversely proportional to R^{2n+2}, instead of R^{2n+1}.

Hein and Zakhor have shown in [4] that for 1^{st} order $\Sigma\Delta$ modulation (n=1) with constant inputs (V_0 reduced to a one dimensional space) that the MSE is lower bounded by $\mathcal{O}(M^{-3})$. This corresponds to the classical behavior R^{2n+1}. However, we found that our predicted R^{2n+2} MSE behavior is recovered when V_0 is the subspace of sinusoids and input signals have an amplitude greater than $q/2000$.

6. Conclusion

Our reconstruction is optimal since it uses all available information (bandlimitedness, V_0, and digital sequence, $\Gamma(C_0)$) and picks an estimate belonging to the intersection $V_0 \cap \Gamma(C_0)$. We also give an upper bound on the performance of such a reconstruction, namely, MSE will go down at least as fast as $\mathcal{O}(R^{-(2n+2)})$.

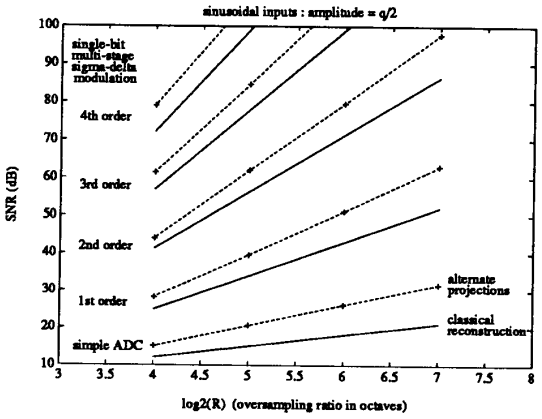

Figure 4: SNR of signal reconstruction versus oversampling rate, with classical and alternate projection methods.

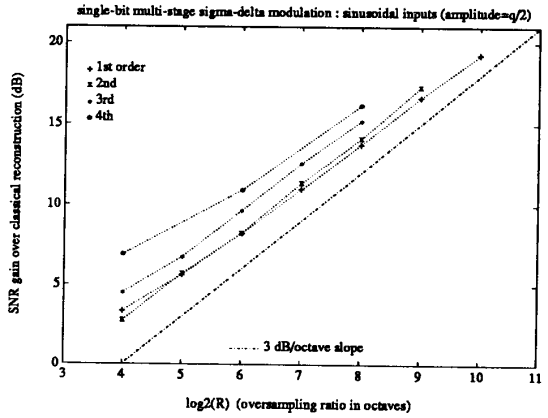

Figure 5: SNR gain of signal reconstruction with alternate projection over classical method, versus oversampling rate.

References

[1] S.K.Tewksbury and R.W.Hallock, "Oversampled, linear predictive and noise shaping coders of order $N > 1$", IEEE Trans. Circuits Syst., vol. CAS-25, pp.436-447, July 1978.

[2] J.C.Candy, "A use of double integration in sigma-delta modulation", IEEE Trans. Commun., vol. COM-33, pp.249-258, Mar.1985.

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

[4] S.Hein and A.Zakhor, "New properties of sigma delta modulators with dc input", to appear in IEEE Trans. Commun.

[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] N.T.Thao and M.Vetterli, "Convex coders and oversampled A/D conversion: theory and algorithms", CTR technical report, Columbia University, Fall 1991.