Conf. on Information Sciences and Systems, The Johns Hopkins University, March 1991 (CISS'91)

Oversampled A/D Conversion using Alternate Projections

Nguyen T.Thao and Martin Vetterli
Department of Electrical Engineering
and Center for Telecommunications Research
Columbia University, New York, NY 10027-6699

Abstract

In this paper, it is shown that oversampled A/D conversion methods can be substantially improved by using iterative postprocessing involving **linear filtering** as well as **nonlinear operations**. The only condition is that the quantizer produces a so-called convex code, which is the case for the simple and dithered A/D conversion, as well as for n^{th} order single path $\Sigma\Delta$ modulation. The iterative algorithm, using alternate projection, searches for the intersection of a subspace (band limited signals) and a convex set (signals with same code). Convergence is guaranteed by convexity. We indicate practical algorithms for simple, dithered, 1^{st} and 2^{nd} order (single path) $\Sigma\Delta$ oversampled quantization. The improvements in SNR due to oversampling in these schemes which would normally be $3, 3, 9$ and $15dB$ per octave of oversampling respectively, are increased to $6, 6, 12$ and $18dB$. That is an increase of $3dB$ per octave.

1 Introduction

Oversampled A/D conversion (or ADC) is usually viewed as represented in figure 1. Consider a signal $x(t)$ band limited to the maximum frequency f_m. For the sake of our discussion, call it $x^{(0)}(t)$ since it will be our reference signal. In the scheme of figure 1, $x^{(0)}(t)$ is first sampled at a rate f_s larger then $2f_m$ to give a discrete-time and continuous amplitude signal $\{x_i^{(0)}\}$. For convenience, this sequence will be noted $\mathbf{x}^{(\mathbf{0})}$ (vector notation) and $x_i^{(0)}$ will be its value at time index i. This last signal is then approximated by a discrete-time and discrete amplitude signal $\{e_i^{(0)}\}$ (or $\mathbf{e}^{(\mathbf{0})}$). A reconstitution of the signal $x^{(0)}(t)$ is conceptually done by lowpass filtering the quantized signal $\mathbf{e}^{(\mathbf{0})}$ at the cutoff frequency f_m. In terms of information, $\mathbf{x}^{(\mathbf{0})}$ is completely equivalent to $x^{(0)}(t)$, according to the Whittaker-Shannon-Kotelnikov sampling theorem. However, $\mathbf{e}^{(\mathbf{0})}$ is a distorted transformation of $\mathbf{x}^{(\mathbf{0})}$ with an error called the quantization error. In the case of oversampling, this error signal, when viewed as a stochastic signal (like noise), has a power spectrum which spreads out over the whole frequency range f_s, so that only a portion of it remains added to the signal after the lowpass filtering. In the simple ADC, provided that the quantization error signal is uncorrelated with the input signal, the error signal power (or means square error, MSE) is reduced by a factor equal to the oversampling rate $R = f_s/2f_m$. If the quantizer is an n^{th} order single path $\Sigma\Delta$ modulator , the reduction is proportional to R^{2n+1} [1, 2, 3, 4]. But in general, let us point out that the error signal contained in $\mathbf{e}^{(\mathbf{0})}$ is reduced during the A/D conversion by manipulation of its power spectrum by mean of a linear time invariant (LTI) system (the lowpass filter). Now the question is: if we don't confine ourselves to the context of LTI processing, can we do any better than this and how ?

Before answering this question, it is first important to revise a little the understanding of the oversampled ADC, as shown in figure 3. We fundamentally separate the notion of coded signal $\mathbf{C}^{(\mathbf{0})}$ and quantized signal $\mathbf{e}^{(\mathbf{0})}$. $C_i^{(0)}$ is an abstract code (for example, a binary code) which identifies the quantization interval the input sample $x_i^{(0)}$ belongs to. $e_i^{(0)}$ is an **analog** signal sample chosen in the interval labeled $C_i^{(0)}$. If only the code sequence $\mathbf{C}^{(\mathbf{0})}$ is available, $e_i^{(0)}$ has to be chosen arbitrarily in the interval $C_i^{(0)}$ as an estimate of $x_i^{(0)}$ (traditionally at the center of the interval for different reasons). The quantization error appears then from this arbitrary estimation. The analog lowpass filter of figure 1 can be decomposed into two filters: a discrete lowpass filter of cutoff frequency f_m working at the oversampling frequency f_s, producing a discrete-time continuous signal $\mathbf{e}^{(\mathbf{1})}$, and a *sinc* interpolator producing the continuous time continuous amplitude $e^{(0)}(t)$. As signals $\mathbf{e}^{(\mathbf{1})}$ and $e^{(0)}(t)$ are equivalent, the actual error reduction lies in the discrete-time lowpass filter between $\mathbf{e}^{(\mathbf{0})}$ and $\mathbf{e}^{(\mathbf{1})}$. $\mathbf{e}^{(\mathbf{1})}$ is then a better estimate of $\mathbf{x}^{(\mathbf{0})}$ obtained by performing a filtering in the space of discrete signals.

The hint that $\mathbf{e}^{(\mathbf{1})}$ is not always the best estimate of $\mathbf{x}^{(\mathbf{0})}$ is that $\mathbf{e}^{(\mathbf{1})}$, when requantized, does not necessarily give the same code $\mathbf{C}^{(\mathbf{0})}$. Figure 6 shows an example made on a simple signal $x^{(0)}(t)$ oversampled by 4. One can see that samples 10 and 11 of $\mathbf{e}^{(\mathbf{1})}$ (lowpass version of $\mathbf{e}^{(\mathbf{0})}$) are not in the same quantization intervals as for $\mathbf{x}^{(\mathbf{0})}$. By projecting them to the border of the right interval (black arrow), we necessarily reduce the error between $\mathbf{e}^{(\mathbf{1})}$ and $\mathbf{x}^{(\mathbf{0})}$. We call this operation the **code projection**. We end up with an estimate $\mathbf{e}^{(\mathbf{2})}$ which would give the code sequence $\mathbf{C}^{(\mathbf{0})}$ if it were requantized, and which is necessarily better than $\mathbf{e}^{(\mathbf{1})}$. $\mathbf{e}^{(\mathbf{2})}$ is then the transformation of $\mathbf{e}^{(\mathbf{0})}$ by two consecutive operations:

(i) a lowpass filtering
(ii) a code projection.

What we have just done has a geometric interpretation. In the space $\mathcal{H}$ of all discrete signals $\mathbf{x}$ sampled at rate f_s, if we call V_0 the subspace of signals band limited by f_m, the lowpass filtering of $\mathbf{e}^{(\mathbf{0})}$ to get $\mathbf{e}^{(\mathbf{1})}$ is actually an orthogonal projection on V_0. It can be seen on figure 2 that this necessarily reduces the distance between the estimate and $\mathbf{x}^{(\mathbf{0})}$. But at the same time, we see that we confine our freedom of displacement to the direction perpendicular to V_0. The second transformation we introduced, the code projection, is indeed a projection in a different direction: it is the projection on the set Γ_0 of discrete signals which all give the same code $\mathbf{C}^{(\mathbf{0})}$ when quantized. If we repeat these alternate projections, we improve the estimate further and end up with an element $\mathbf{e}^{\infty}$ of $V_0 \cap \Gamma_0$, as shown in figure 2. Not only have we improved $\mathbf{e}^{(\mathbf{1})}$, but we have reached the theoretical limit of improvement, deterministically speaking. Indeed, as $\mathbf{x}^{(\mathbf{0})} \in V_0 \cap \Gamma_0$ is the only information we have, there is no deterministic reason to privilege a particular element of $V_0 \cap \Gamma_0$ rather than another one as an estimate of $\mathbf{x}^{(\mathbf{0})}$.

Starting from this idea, the goal of this paper is to improve a number of oversampled ADC schemes (namely classical and dithered A/D conversion, 1^{st} and 2^{nd} order single path $\Sigma\Delta$ modulation). As shown in part 2, the geometric interpretation of signals is based on the Hilbert space structure of $\mathcal{H}$ and the convexity property of V_0 and Γ_0. On a particular family of signals, we see in part 3 that, taking an estimate in $V_0 \cap \Gamma_0$ theoretically reduces the MSE by a coefficient proportional to R^2 instead of R. This means we obtain a gain of $3dB$ per octave of oversampling in the error reduction over the classical signal reconstruction. After designing alternate projections algorithms for the other more sophisticated ADC techniques (part 4 and 5), we show in part 6 the results of numerical tests: we still obtain a gain of $3dB$ per octave for the dithered ADC, the 1^{st} and 2^{nd} order $\Sigma\Delta$ which means that we improve their MSE reduction dependence from R, R^3, R^5 to R^2, R^4, R^6 respectively.

2 Alternate projections in simple ADC

The geometric interpretation of a signal is based on the fact that the traditional power measurement $\sum_{i\in\mathbf{Z}} |x_i|^2$ of a signal $\mathbf{x} \in \mathcal{H}$ is the norm associated with the inner product $\langle \mathbf{y}, \mathbf{z} \rangle = \sum_{i\in\mathbf{Z}} y_i z_i$, that is $||\mathbf{x}|| = (\sum_{i\in\mathbf{Z}} |x_i|^2)^{1/2}$. $\mathcal{H}$ then becomes a Hilbert space with regard to the inner product. Qualitatively speaking, $\mathcal{H}$ has the structure of a geometric space with the notion of orthogonality ($\mathbf{x} \perp \mathbf{y} \Leftrightarrow \langle \mathbf{x}, \mathbf{y} \rangle = 0$), and metric distance ($d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}||$). In this context, we basically use the following Hibert space property: if S_0 is a closed and convex subset and $\mathbf{y}^{(\mathbf{0})} \notin S_0$, there exists a unique $\mathbf{y}^{(\mathbf{1})} \in \mathcal{H}$ which minimizes $d(\mathbf{y}^{(\mathbf{0})}, \mathbf{y}^{(\mathbf{1})})$, and necessarily $\mathbf{y}^{(\mathbf{1})}$ is closer to any element of S_0 than $\mathbf{y}^{(\mathbf{0})}$. $\mathbf{y}^{(\mathbf{1})}$ is then called the projection of $\mathbf{y}^{(\mathbf{0})}$ on S_0. We recall that S_0 is convex if and only if for any pair of elements $x, y \in S_0$, the whole segment $[x, y]$ is included in S_0.

First of all, our two sets V_0 and Γ_0 are indeed convex. For V_0, this is trivial because it is a vector space. For Γ_0 it is enough to prove that:

$$\forall \mathbf{x}, \mathbf{y} \in \Gamma_0, \forall \theta \in [0,1], \quad \theta \mathbf{x} + (1-\theta)\mathbf{y} \in \Gamma_0$$

This is true because at every time index i, as both x_i, y_i belong to the quantization interval labeled $C_i^{(0)}$, then this is also true for $\theta x_i + (1-\theta) y_i$. We always consider the closure of V_0 and Γ_0.

The second fact is that the two signal transformations, the lowpass filter and the code projection, are indeed the projection operators on V_0 and Γ_0 respectively. For the lowpass filter, this can be seen by using the theorem of Pythagoras and the spectral expression of the inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) Y^*(\omega) e^{j\omega} d\omega$ from Parseval's theorem (X, Y are the Fourier transforms of $\mathbf{x}, \mathbf{y}$). For the code projection, figure 6 shows that the projected signal is indeed the closest signal to $\mathbf{e}^{(\mathbf{1})}$ of those which have the same code as $\mathbf{x}^{(\mathbf{0})}$.

Now, Youla and Webb prove in [5] that the alternate projections between two closed and convex sets always converge. Moreover, the limit is necessarily an element of this intersection, that is, $V_0 \cap \Gamma_0$ in our case.

Actually, this property remains true if, instead of considering the projection operators P_{V_0}, P_{Γ_0} on V_0, Γ_0, we used the following operators:

$$T_{V_0} = P_{V_0} + (1-\alpha)(P_{V_0} - 1) \ , \ T_{\Gamma_0} = P_{\Gamma_0} + (1-\beta)(P_{\Gamma_0} - 1)$$

where the relaxation coefficients α, β are chosen in the interval $]0, 2[$. These operators are intermediate between the identity operator (α or $\beta = 0$) and the symmetry operator with regard to the projection (α or $\beta = 2$). When $\alpha = \beta = 1$, they coincide with P_{V_0} and P_{Γ_0} respectively. Youla and Webb prove in [5] that the convergence to an element of $V_0 \cap \Gamma_0$ is still guaranteed with such operators. In practice, the speed of convergence can be dramatically enhanced by optimizing α and β in $]0,2[$. This is often done experimentally.

3 Improvement of the estimate $\mathbf{e}^{\infty}$ in simple ADC

To see if it is worth applying the alternate projections, we want to see how much better the estimate $\mathbf{e}^{\infty}$ is compared to $\mathbf{e}^{(\mathbf{1})}$. In order to have a first idea of the quality of such an estimate, we confine ourselves to periodic discrete signals of period T_0 and a sampling interval equal to T_s. For a given maximum discrete frequency $N < M/2$ (where $M = T_0/T_s$ is the number of samples per period), V_0 is the subspace of signals $\mathbf{x}$ where the k^{th} discrete Fourier component $X_k = \sum_{i=1}^{M} x_i e^{2\pi ik/M}$ is zero except for $-N \leq k \leq N$.

We want to evaluate the distance between any two elements of $V_0 \cap \Gamma_0$. Can we try to have a description of $V_0 \cap \Gamma_0$ in the time domain ? If we take the example of figure 6, from the knowledge of $C_i^{(0)}$, Γ_0 can be visually represented by the domain delimited by the heavy hashed lines in figure 7. This figure shows for example that the continuous version of a signal of Γ_0 should stay in the interval $[-\frac{q}{2}, \frac{q}{2}]$ between $4T_s$ and $6T_s$, in $[\frac{q}{2}, \frac{3q}{2}]$ between $7T_s$ and $9T_s$, and necessarily crosses the quantization threshold $\frac{q}{2}$ between $6T_s$ and $7T_s$. In other words, this representation gives an information on the quantization threshold crossings of the continuous time version of elements of Γ_0, with absolute precision on the amplitude, but uncertainty on the instant when it occurs. The time uncertainty is the sampling period T_s.

Suppose the original signal $\mathbf{x}^{(\mathbf{0})}$ is such that we observe more than $2N+1$ quantization threshold crossings from the coded sequence $C^{(0)}$. At the limit of an infinite sampling frequency, the instants $t_1^0, t_2^0, ..., t_{2N+1}^0$ are known with infinite precision. If $x^{(0)}(t)$ is the continuous time version of $\mathbf{x}^{(\mathbf{0})}$ and (X_k^0) the discrete Fourier transform of $\mathbf{x}^{(\mathbf{0})}$, then from:

$$x^{(0)}(t) = \frac{1}{2\pi} \sum_{k=-N}^{N} X_k^0 e^{j2\pi kt/T_0} \quad ,$$

we can write

$$\left[X_{-N}^0 \ ... \ X_N^0\right]^T = \left[\mathcal{M}(t_1^0, ..., t_{2N+1}^0)\right]^{-1} \cdot \left[L_1^0 \ ... \ L_{2N+1}^0\right]^T \tag{1}$$

where $\mathcal{M}(t_1, t_2, ..., t_{2N+1})$ is the matrix $\left[e^{j2\pi kt_i/T_0}\right]_{1\leq i\leq 2N+1, -N\leq k\leq N}$ and L_j^0 is the threshold that $x^{(0)}(t)$ crosses at time t_j^0. $\mathcal{M}$ is a Vandermonde matrix, and thus has a determinant different from zero if and only if all t_j^0 are distinct. If the sampling frequency is not infinite, any element of $V_0 \cap \Gamma_0$ will also be given by (1) but with an error bounded by T_s on $t_1^0, t_2^0, ..., t_{2N+1}^0$. This will induce an error on the knowledge of $\mathcal{M}$ which can be linearized for T_s small enough. This will the induce an error proportional to T_s on every coefficient X_k^0, $-N \leq k \leq N$. As, from Parseval's theorem, the power of a signal is also equal to the power of its Fourier transform, the mean square error of an element of $V_0 \cap \Gamma_0$ from $\mathbf{x}^{(\mathbf{0})}$ will be bounded by a function proportional to T_s^2, or inversely proportional to $R^2 = \left(\frac{T_0}{2NT_s}\right)^2$. A more detailed explanation is given in [6].

Conclusion : The quantization error power reduction is asymptotically proportional to R^2 instead of R, if $\mathbf{x}^{(\mathbf{0})}$ displays enough level crossings.

4 Generalization to other oversampled ADC systems

In part 2, we saw that the alternate projection principle is based on the fact that Γ_0 is convex (the convexity of V_0 is always true). We naturally want to extend this principle to other A/D conversion techniques where Γ_0 is still convex.

We are going to focus on a particular family of coding systems where Γ_0 can be equivalently described by the block diagram of figure 4. In this scheme, F and $\mathbf{d}^{(\mathbf{0})}$ are respectively a linear operator (not necessarily time invariant) and a discrete signal, and are assumed to be completely known to the user.

The classical ADC technique corresponds to the trivial case where $F = I$ (identity operator) and $\mathbf{d}^{(\mathbf{0})} = 0$. The dithered ADC also falls into this case where $F = I$ and $\mathbf{d}^{(\mathbf{0})}$ is the dither sequence. The only difference with the common use of a dither is that $\mathbf{d}^{(\mathbf{0})}$ is assumed to be completely known at every instant.

Less obvious is the fact that the n^{th} order single path $\Sigma\Delta$ modulation technique (figure 8) belongs to that family of coding systems. To see this, one can first be convinced that, if $\mathbf{C}^{(\mathbf{0})}$ is the code sequence given by $\mathbf{x}^{(\mathbf{0})}$, a signal $\mathbf{x}$ gives the code sequence $\mathbf{C}^{(\mathbf{0})}$ through the block diagram of figure 8 if and only if it also gives the code sequence $\mathbf{C}^{(\mathbf{0})}$ through the block diagram of figure 9. This last scheme is itself equivalent to figure 4 where F is the cascade of n accumulators (and is therefore linear), and $\mathbf{d}^{(\mathbf{0})}$ is the value of $-\mathbf{a}$ in figure 9 when $\mathbf{x}$ is forced to zero. $\mathbf{d}^{(\mathbf{0})}$ is known, as it can be computed from $\mathbf{C}^{(\mathbf{0})}$.

We are going to show that, for such coding systems, Γ_0 is always convex. If $\mathbf{x}^{(\mathbf{1})}$ and $\mathbf{x}^{(\mathbf{2})}$ give the same code sequence $\mathbf{C}^{(\mathbf{0})}$, then at every time index i, $a_i^{(1)}$ and $a_i^{(2)}$ will belong to quantization interval labeled $C_i^{(0)}$. As the signal at the quantizer input node corresponding to the input signal $\theta\mathbf{x}^{(\mathbf{1})} + (1-\theta)\mathbf{x}^{(\mathbf{2})}$ ($\theta \in [0,1]$) is equal to (using the fact that $1 = \theta + (1-\theta)$ and that F is linear):

$$\begin{array}{rl} & F(\theta\mathbf{x}^{(\mathbf{1})}+(1-\theta)\mathbf{x}^{(\mathbf{2})})-\mathbf{d}^{(\mathbf{0})} \\ = & \theta F(\mathbf{x}^{(\mathbf{1})})+(1-\theta)F(\mathbf{x}^{(\mathbf{2})})-\theta\mathbf{d}^{(\mathbf{0})}+(1-\theta)\mathbf{d}^{(\mathbf{0})} \\ = & \theta(F(\mathbf{x}^{(\mathbf{1})})-\mathbf{d}^{(\mathbf{0})})+(1-\theta)(F(\mathbf{x}^{(\mathbf{2})})-\mathbf{d}^{(\mathbf{0})}) \qquad , \\ = & \theta\mathbf{a}^{(\mathbf{1})}+(1-\theta)\mathbf{a}^{(\mathbf{2})} \end{array}$$

its value at time i also remains in the quantization interval $C_i^{(0)}$. This implies that $\theta\mathbf{x}^{(\mathbf{1})}+(1-\theta)\mathbf{x}^{(\mathbf{2})}$ also gives the code sequence $\mathbf{C}^{(\mathbf{0})}$. This proves that Γ_0 is convex.

The principle of alternate projections can then theoretically applied on such coding systems.

5 Algorithm design of the code projection for the dithered ADC, the 1^{st} and 2^{nd} order single path $\Sigma\Delta$ modulation

One should not forget that the mathematical existence of a solution is far from being the ultimate goal. If we have abandoned the context of LTI processing, we still have the constraint that any signal transformation should be progressive in time. By this we mean that the future is always unknown and the control over the past is limited.

In the simple ADC, we have seen that the code projection is straightforward, as every sample is processed independently. In the dithered ADC, this is also the case. The slight difference is that, from the view point of the input signal, the quantization intervals move from one time to another. But they are simply shifted by the value $d_i^{(0)}$ at time i, which is anyway known by the user.

If we want to deal with the 1^{st} order $\Sigma\Delta$ modulation (figure 5), the code projection can no longer be expressed as an individual transformation of the input samples. We need to design a more sophisticated algorithm. We base our study on the equivalent diagram of figure 4, where F is the accumulation function and $\mathbf{d}^{(\mathbf{0})}$ is equal to minus the accumulation of $\mathbf{C}^{(\mathbf{0})}$.

The projection of a signal $\mathbf{x}$ on Γ_0 is the signal $\mathbf{x}+\Delta\mathbf{x}$ such that $\mathbf{x}+\Delta\mathbf{x}\in\Gamma_0$ and the energy of $\Delta\mathbf{x}$ is minimized. If $\mathbf{a}=F(\mathbf{x})-\mathbf{d}^{(\mathbf{0})}$ is the signal at the input node of the quantizer, then the variation $\Delta\mathbf{x}$ on $\mathbf{x}$ will induce a variation $\Delta\mathbf{a}$ on $\mathbf{a}$ such that $\Delta\mathbf{a}=F(\Delta\mathbf{x})$. Saying that $\mathbf{x}+\Delta\mathbf{x}\in\Gamma_0$ is equivalent to saying that at every instant i, $a_i+\Delta a_i$ belongs to the quantization interval labeled $C_i^{(0)}$, or, Δa_i belongs to this interval shifted by $-a_i$. Let us call this shifted interval Q_i. Every time we want to project a signal on Γ_0, we can compute this interval sequence Q_i. An example is shown in figure 11, where at every time index i, the arrows symbolize the boundaries of Q_i. In this example, we assume we start coding at time $i=0$ where a_0 is initialized to 0 (Δa_0 is then constrained to be equal to 0). We also deal with the general case of nonuniform quantization: every Q_i can have different length, or even have only one boundary.

The projection problem consists then in finding the sequence $\Delta\mathbf{a}$ that minimizes the energy of $\Delta\mathbf{x}=F^{-1}(\Delta\mathbf{a})$ with the constraint $\Delta a_i\in Q_i$ for every i (note that as F is the integration operator, F^{-1} is the first derivative operator). We show in [6] that this solution can be found by what we call the **"thread algorithm"**. Physically speaking, it amounts to attach a thread at node (0,0), and stretch it between the arrows in the direction of increasing time index by maintaining at tension. Taking Δa_i on the path of the resulting thread position gives the solution to the projection of $\mathbf{x}$ on Γ_0 (see figure 11).

For the 2^{nd} order $\Sigma\Delta$ (figure 10), the approach is similar, where this time F is the cascade of two accumulators. F^{-1} is consequently the second derivative operator. The algorithm is described in detail in [6].

6 Numerical results

We have concentrated our numerical tests (table 13) on ideal cases in order to validate the principle of alternate projections. to be able to perform ideal lowpass filtering, we have performed our simulations in the context of periodic signals. We have also dealt with uniform quantizers. For every test, we measure the remaining error contained in the estimate given by the algorithm. We compare it with the classical uniform quantization noise power $\frac{q^2}{12}$ and express the ratio in dB. If the MSE is proportional to R^n where n is a certain number, then every time R is doubled we will expect a gain of $10\log 2^n \simeq n\times 3dB$. For every test configuration, we fix $2N+1$ the number of non zero low frequency Fourier coefficients, and A the amplitude of the signal (peak to peak amplitude). we then apply the algorithm on a certain number of trial signals (between 780 and 1000) which are randomly generated and satisfy the conditions on the parameters N and A.

Simple ADC

In tests 1 and 2, we try to verify the theoretical factor R^2 in the MSE reduction we found in part 3, on simple sinusoidal signals having more than $2N+1=3$ quantization threshold crossings (they actually have exactly 4 crossings). We find a gain of $5.7dB$ as R is doubled. This is close to $6dB=2\times 3dB$ which corresponds to a R^2 behavior. In those tests, the iteration is stopped as soon as the step increment of gain is less than $5.10^{-3}dB$.

Adding the relaxation coefficients to alternate projections, we experimentally obtained the best results with $\alpha=\beta=2$. Test 3 gives a typical example of performance achieved in this case. With an oversampling rate of $R=64=2^6$, the classical signal reconstruction should give a quantization noise reduction of $6\times 3=18dB$: with our algorithm we improve it to more than $26dB$. Moreover, the iteration always explicitly ends up with an estimate in the interior of $V_0 \cap \Gamma_0$ in a finite number of steps (64 in average in this case).

In test 5 we deal with a nonideal filter. We use an FIR lowpass filter of length 2048 obtained by Hanning windowing the perfect filter. We limit the number of iterations to 32. We find that we only lose $0.1dB$ in average compared to the ideal case (test 4).

Dithered ADC

With the dithering technique, we get rid of the constraint on level crossings. In tests 6 and 7, we use a sinusoidal dither of amplitude q and frequency $2f_m$. Technologically speaking, this can be easily realized with high precision. The condition here is that every sample of the dither should be known by the user with precision. Test 6 is done with the same conditions as test 3 with this dither: it yields an even better error reduction ($29dB$) without any sign of weakness for signal of amplitudes going to zero (figure 12). This test is more thoroughly described in [6]. The R^2 dependency is confirmed in test 7, obtaining a gain of $11.5dB \simeq 4{\times}3dB$ as R is multiplied by $4 = 2^2$.

1^{st} order $\Sigma\Delta$

We have limited the quantizer to be a single threshold comparator, giving code $+1$ (resp. -1) when its input is positive (resp. negative). The in-built D/A outputs the analog value $+q$ (resp. $-q$) when it receives the code $+1$ (resp. -1). We fix $\alpha = \beta = 1$ to test the regular alternate projection. In the $\Sigma\Delta$ tests we also start with $\mathbf{e}^{(\mathbf{0})} = 0$ as the first estimate of the iteration. With the usual R^3 behavior, we would expect a gain of $3 \times 3 \times 3 = 27dB$ when R is multiplied by $8 = 2^3$. With test 8 and 9 we find $34.4dB$. This is close to $36dB = 4 \times 3 \times 3dB$ which corresponds to a R^4 behavior. Note the reasonable number of iterations.

2^{nd} order $\Sigma\Delta$

Tests 10 and 11 are similar to tests 8 and 9 with the 2^{nd} order $\Sigma\Delta$. The normal R^5 behavior would give a gain of $5 \times 3 \times 3 = 45dB$. We find $53.0dB$. This is close to $54dB = 6 \times 3 \times 3dB$ which corresponds to a R^6 behavior.

7 Conclusion

We have designed algorithms which systematically improve the signal reconstruction by $3dB$ per octave of oversampling in the following coding techniques: simple and dithered ADC, 1^{st} and 2^{nd} order single path $\Sigma\Delta$ modulation. They are based on the principle of alternate projections and the convexity of the coding system. We have theoretically justified this improvement on the simple ADC case for a certain family of signals.

In spite of their complexity, these algorithm have some implementation potential in the context of ADC precision improvement by off-line postprocessing. Note that the whole processing is done at the oversampling frequency f_s. The effect on the alternate projections of nonideal characteristics of implementable filters are under study.

It is important to see that our algorithms do not require any modification to the A/D conversion process which can be done, for example, in real time with existing and traditional circuits. Actually, those algorithms can even take account of circuit defects such as quantizer nonuniformity provided they can be measured.

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 and O.J.Benjamin, "The structure of quantization noise from sigma-delta converters", IEEE Trans. Commun., vol. COM-29, pp.1316-1323, Sept.1981.

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

[4] R.M.Gray, "Spectral analysis of quantization noise in a single loop sigma-delta modulator with dc input", IEEE Trans. Commun., vol.37, pp.588-599, June 1989.

[5] D.C.Youla and H.Webb, "Image restoration by the method of convex projections: part 1 - theory", IEEE Trans. Medical Imaging, 1(2):81-94, Oct.1982.

[6] N.T.Thao and M.Vetterli, "Improvement of oversampled A/D conversion using alternate projections", CTR technical report, Columbia University, spring 1991.

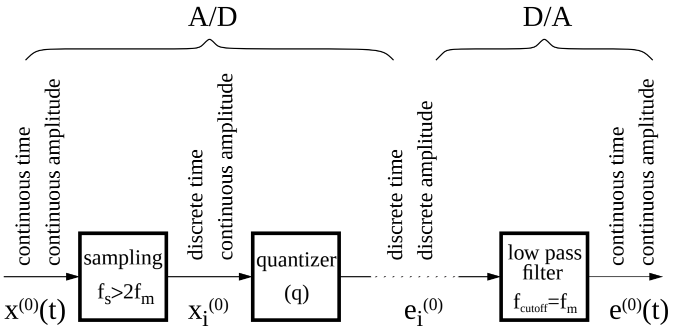


Figure 1: Oversampled A/D conversion principle

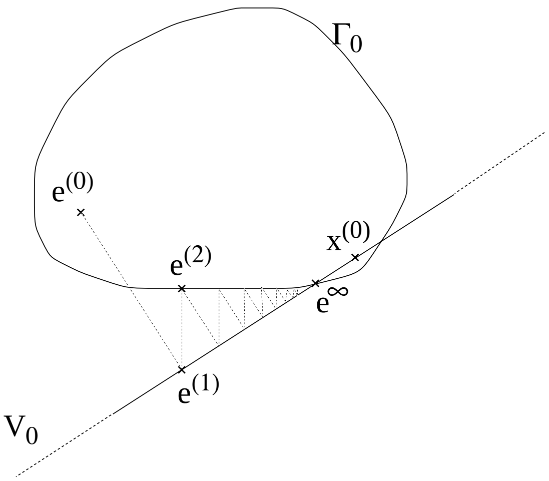


Figure 2: Geometric representation of alternate projections

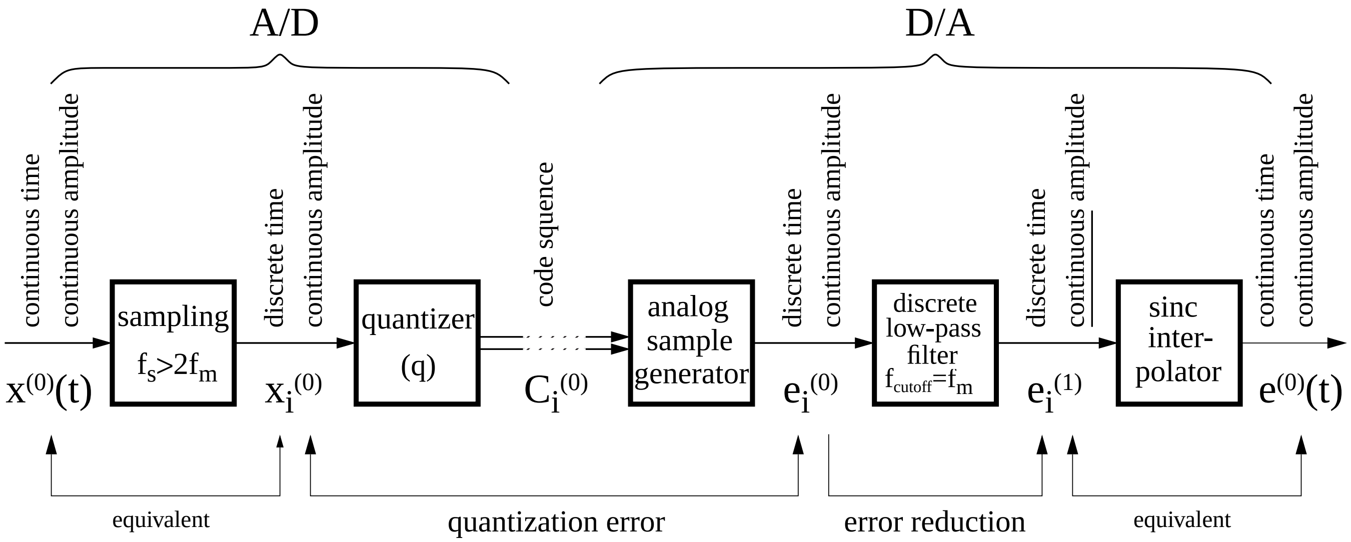
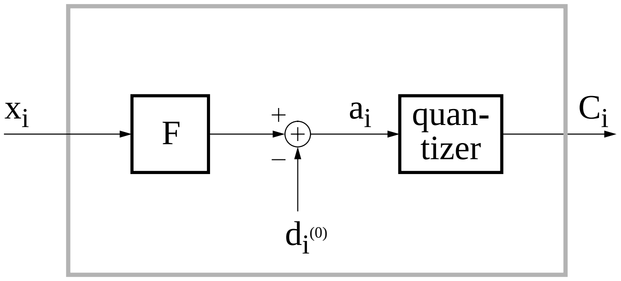


Figure 3: Revision of the oversampled A/D conversion principle



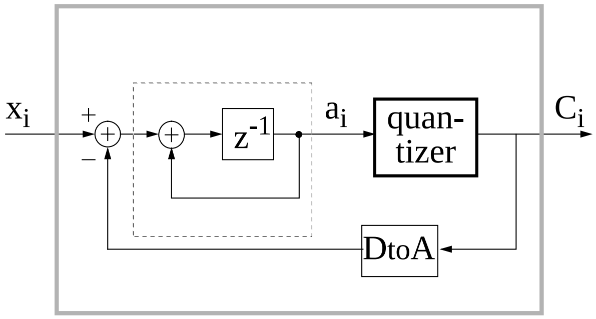


Figure 4: General coding system model

Figure 5: 1^{st} order $\Sigma\Delta$ modulation block diagram

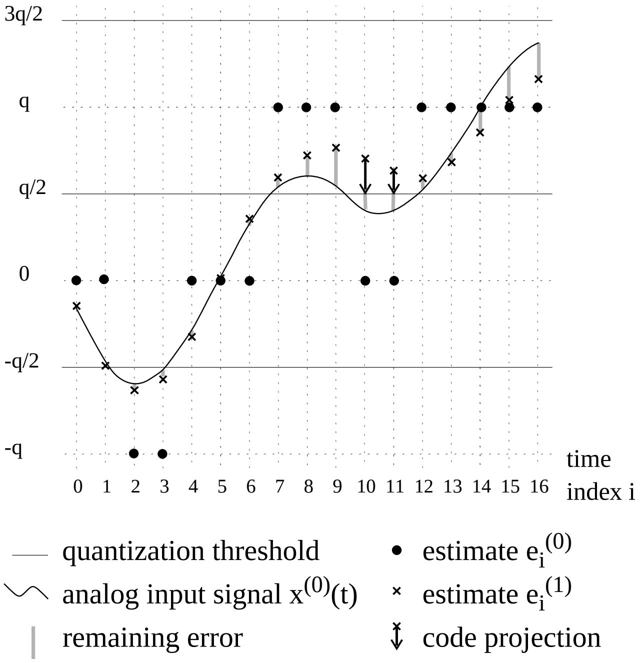


Figure 6: Code projection in simple ADC

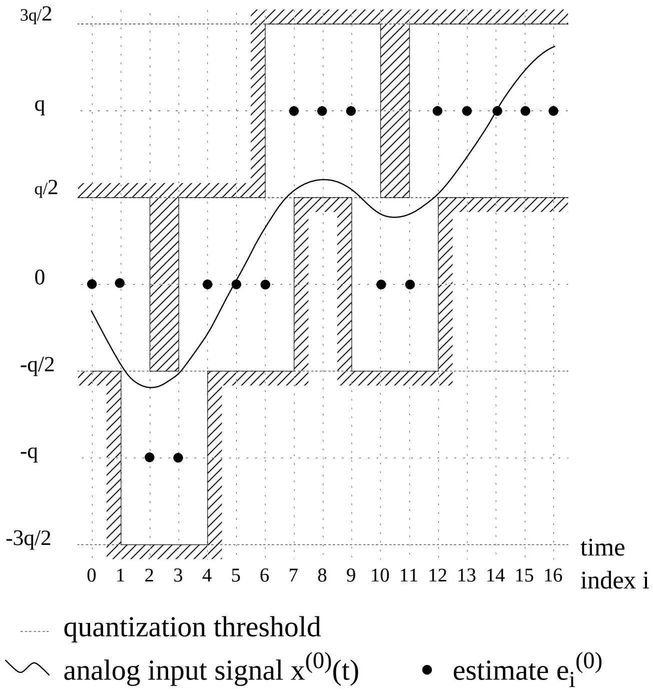


Figure 7: Time domain representation of Γ_0 in simple ADC (same signal as in figure 6)

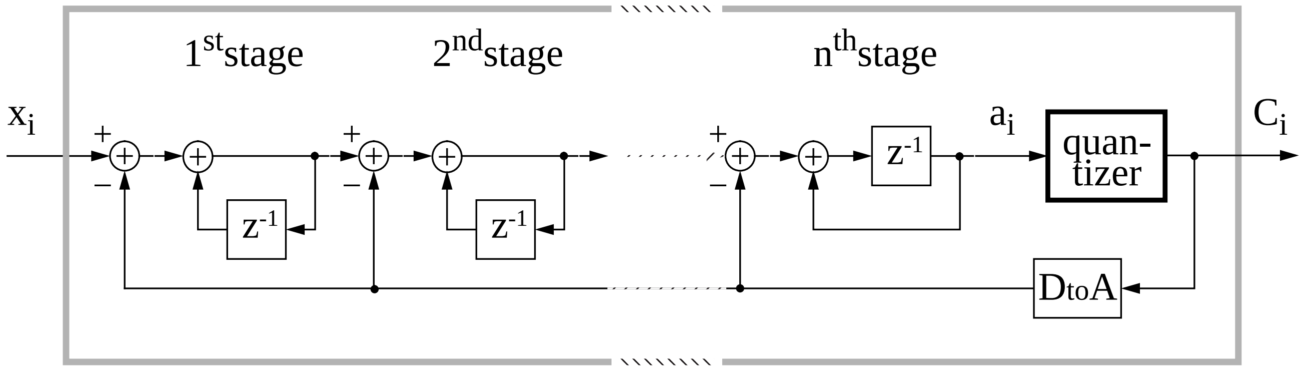


Figure 8: n^{th} order single path $\Sigma\Delta$ modulation block diagram

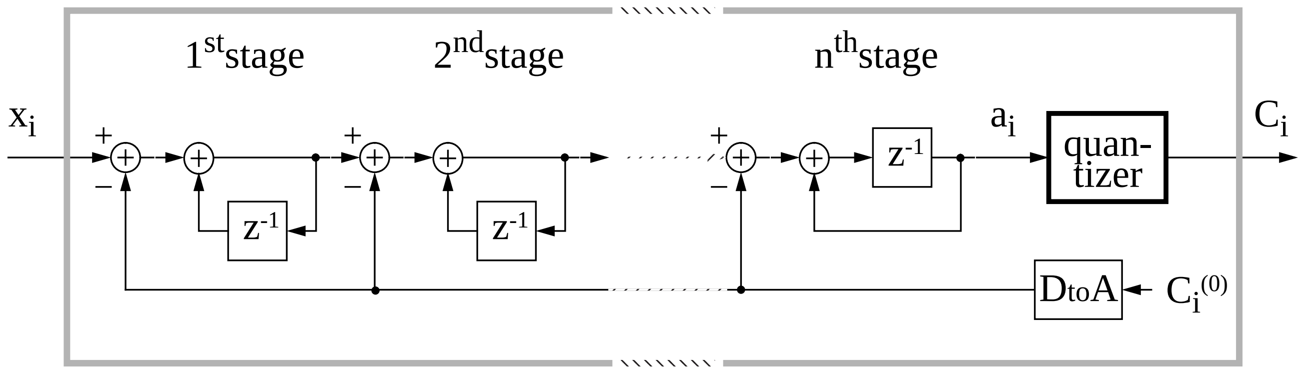


Figure 9: Virtual coding system for the n^{th} order $\Sigma\Delta$

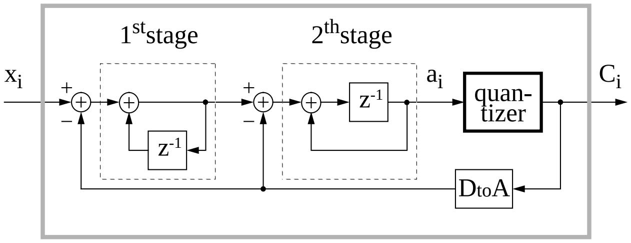


Figure 10: 2^{nd} order ΣΔ modulation block diagram

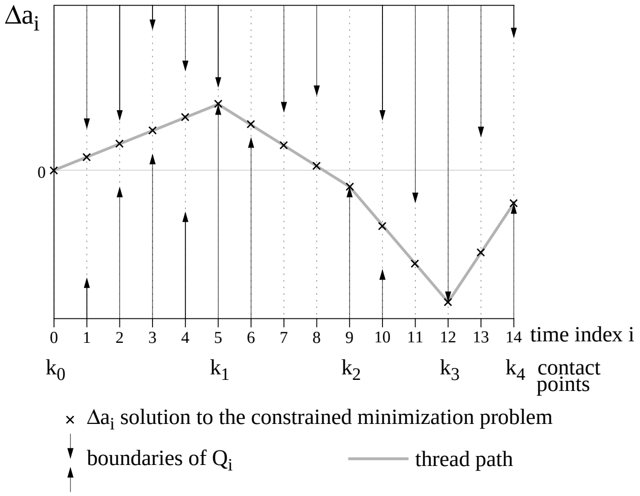


Figure 11: Constrained minimization of Δa in the 1^{st} order ΣΔ: the "thread algorithm"

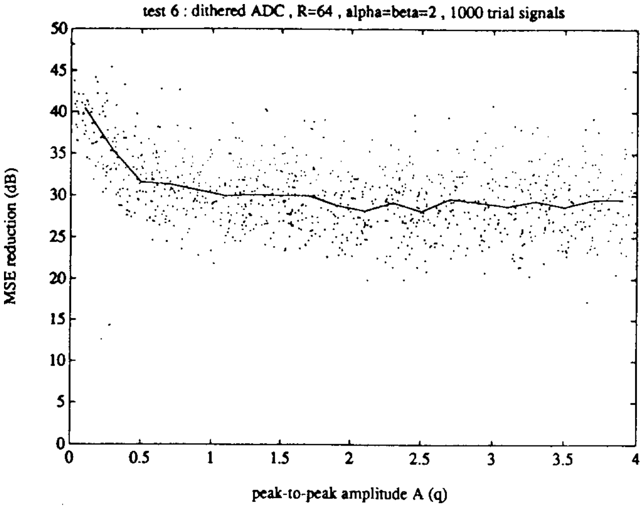


Figure 12: MSE reduction versus A obtained in test 6 (dithered ADC)

test number	A/D conversion technique	2N+1	A (q)	R	α/β	average number of iterations	error reduction average (dB)	error reduction standard dev. (dB)
1	simple	3	2	64	1/1	27	27.7 }Δ= 5.7	}Δ= 4.3
2				128		32	33.4	
3	simple	7	4	64	2/2	64	26.3	
4	simple	127	4	16	2/2	56	8.4 }Δ=-0.1	}Δ= 0.1
5						32	8.3	
6	dithered	7	0 to 4	64	2/2	173	30.4 }Δ=11.5	}Δ= 4.9
7				128		543	41.9	
8	1st order ΣΔ	3	1	64	1/1	21	28.5 }Δ=34.5	}Δ= 5.3
9				128		26	63.0	
10	2nd order ΣΔ	3	1	64	1/1	23	37.8 }Δ=53.0	}Δ= 6.0
11				128		26	90.8	

2N+1 : number of non-zero discrete low frequency components

A : peak-to peak amplitude (in multiples of the quantization step q)

R : oversampling rate

α/β : relaxation coefficients for the projections on V_0/Γ_0

Table 1: Numerical results