# A DETERMINISTIC ANALYSIS OF OVERSAMPLED A/D CONVERSION AND ΣΔ MODULATION

*Nguyen T. Thao and Martin Vetterli* *

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

## ABSTRACT

As an alternative to the usual statistical analysis, we present a purely deterministic analysis of oversampled A/D conversion and ΣΔ modulation, which requires no assumption on the quantizer error. This leads to the notion of **consistent estimate** for the decoding, which is a necessary condition for optimality. Algorithms using convex projections are discussed, and the reduction of decoding error from $\mathcal{O}(R^{-(2n+1)})$ to $\mathcal{O}(R^{-(2n+2)})$ (where $R$ is the oversampling ratio and $n$ is the order of the encoder) is demonstrated experimentally for multi-loop and multi-stage ΣΔ modulators.

## 1. INTRODUCTION

Analog-to-digital conversion (ADC) is basically the discretization operation of an analog signal in time and amplitude. Following Shannon's well known sampling theorem, no information is lost in the time discretization operation, if the input signal is bandlimited to some maximum frequency $f_m$ and the sampling frequency $f_s$ is larger or equal to the Nyquist rate $2f_m$ (operation I in Figure 1). In this situation, the input signal can be uniquely recovered from its samples taken at the Nyquist rate (operation IV and V in Figure 1). However, some information is irreversibly lost when the samples are moreover encoded in amplitude (operation II). Recent techniques of ADC use amplitude quantization together with oversampling (choice of $f_s > 2f_m$), that is, quantization of a redundant set of samples [1, 2]. Because we have an oversampled set of quantized samples, a **decoder** (operation III) is used to take advantage of the redundancy, that is, reduce the quantization error. This decoding is the focus of our paper.

The classical way to analyze the effect of oversampling redundancy on amplitude encoding is to consider the quantizer as an additive source of error which is a white noise independent of the input [1, 2]. This permits a linearized analysis of the different encoding schemes (Figure 3) and leads to the conclusion that the encoded signal is the sum of the bandlimited input signal and an error signal which is not bandlimited and spreads out over the whole frequency range. This will be shown in Section 2. The redundancy due to oversampling is then exploited by canceling the out-of-band energy of the encoded signal, using a linear low-pass filter. This is the classical, linear decoding scheme. Although the white noise assumption is not theoretically justified, linear filtering leads to a good performance which is well predicted by the linear model analysis [1].

The basic question is whether the remaining in-band error is really irreversible? To analyze this question, we look at quantization from its basic definition as a deterministic operator (Section 3), that is, as defining a partition of the space of discrete-time signals. In oversampled ADC (Section 4) we show that the information contained in the encoded signal is a set of **consistent estimates** and that a decoded estimate of the input is not optimal as long as it is not consistent. We show that linear decoding estimates are not necessarily consistent. Numerical simulations (Section 5) show that consistent estimates asymptotically reduce the quantization error signal by 3 dB per octave of oversampling over linear decoding estimates. Some analytical evaluation done previously anticipated this result [3, 4, 5]. Finally, the deterministic analysis gives principles for finite complexity methods for non-consistent estimate improvement, which approach the performance of consistent ones.

## 2. BACKGROUND ON OVERSAMPLED ENCODERS

In oversampled ADC, there exists a large number of different encoding schemes [2]. We present here the basic structures which underline these schemes and will be sufficient for the presentation of our deterministic approach in the next sections.

The simplest version of encoding consists in the individual quantization of the input samples. We call it simple encoding. The transfer function of a quantizer is (in $z$-transform notations):

$$C(z) = X(z) + E(z), \tag{1}$$

where $X(z)$, $C(x)$ and $E(z)$ are respectively the input, the output and the quantizer error signal. With the white noise assumption, the linear decoding mean square error (MSE) is equal to $\frac{q^2}{12R}$, where $q$ is the quantization step size and $R = \frac{f_s}{2f_m}$ is the oversampling ratio.

Predictive encoders are more sophisticated encoders including a feedback loop, as shown in Figure 3(a), in order to minimize the amplitude of the input $A_k$ to the quantizer. For a quantizer of given complexity, this allows the use of a smaller step size $q$. The two built-in filters $H$ and $G$ are chosen so that the transfer function of the whole encoder in the linearized approach is of the type (1), where $E(z)$ is the error generated by the built-in quantizer. It can be shown [1] that this is verified with the following constraint:

$$H(z) = 1 + G(z). \tag{2}$$

The most popular example of predictive encoder is the Δ modulator where $H$ is an integrator, leading to

$$H(z) = \frac{1}{1 - z^{-1}} \quad \text{and} \quad G(z) = \frac{z^{-1}}{1 - z^{-1}}. \tag{3}$$

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



0-7803-0946-4/93 $3.00 

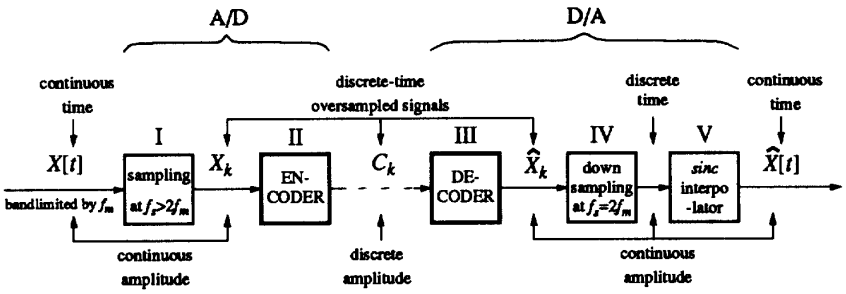


Figure 1: Block diagram of oversampled ADC

Currently, the most successful type of encoders are the noise-shaping encoders. They are derived from predictive encoders by putting the filter $H$ in front of the feedback loop, as shown in Figure 3(b), while the constraint (2) is maintained. When taking the choice of $H$ and $G$ of (3), we obtain the $1^{st}$ order $\Sigma\Delta$ modulator. In the general case of noise-shaping encoding, $H$ is typically an integrator and the input of the quantizer is no longer of small amplitude. However, it can be shown [1] that the constraint (2) implies the following relation:

$$C(z) = X(z) + H^{-1}(z)E(z). \tag{4}$$

Typically, $H^{-1}$ is a differentiator and filters out the low frequency components of the error signal $E_k$ (in $1^{st}$ order $\Sigma\Delta$, $H^{-1}(z) = 1 - z^{-1}$). In the classical approach, although the total variance of $E_k$ is large, its in-band portion is reduced by the "shaping" function $H^{-1}(z)$. With the white noise approach [1], the linear decoding MSE decreases with $R$ at the speed $\mathcal{O}(R^{-(2n+1)})$ in the case of $n^{th}$ order multi-loop $\Sigma\Delta$ modulation. Multi-stage $\Sigma\Delta$ modulators of order $n$, described in [2], also achieve this MSE performance.

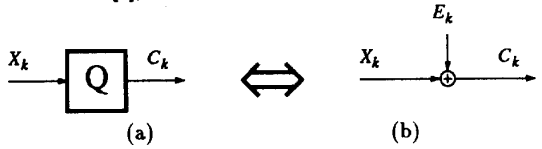


Figure 2: Simple encoding. (a) Quantization. (b) Additive error source model.

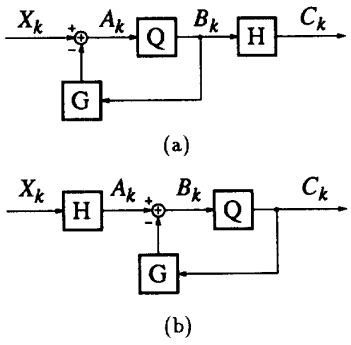


Figure 3: Encoding using two filters $H$ and $G$ satisfying the constraint (2). (a) Predictive encoding. (b) Noise-shaping encoding.

## 3. DETERMINISTIC ANALYSIS OF ENCODING

In practice, a finite number $N$ of samples of the input signal is processed through the encoder to give a quantized signal which is another sequence of $N$ points. The deterministic approach of encoding consists in analyzing the encoder as a many-to-one mapping from the space of $N$-point sequence $\mathbf{R}^N$ into a discrete subset of $\mathbf{R}^N$. When a sequence $X = (X_k)_{k=1,...,N}$ is only known by the encoded sequence $C = (C_k)_{k=1,...,N}$, the exact information we have about $X$ is that it belongs to the inverse image of $C$ through the encoding mapping. This inverse image is a whole subset of $\mathbf{R}^N$, denoted by $\mathcal{C}(C)$. We call its elements the consistent estimates of $X$. In other words, $\mathcal{C}(C)$ is the set of all signals of $\mathbf{R}^N$ having the same encoded signal $C$ as $X$.

In the case of simple encoding, the many-to-one mapping is reduced to the quantization operation. In the case of single sample sequences $N = 1$, whole intervals of $\mathbf{R}$ (quantization intervals) are mapped into a discrete set of values of $\mathbf{R}$ (quantization levels) (Figure 4(a)). A quantization level $c$ is typically chosen to be the center of the corresponding quantization interval $\mathrm{q}^{-1}[c]$. When $x \in \mathrm{q}^{-1}[c]$, $c$ is a particular consistent estimate of $x$. In the general case of $N$-point sequences, a quantizer is a many-to-one mapping $\mathbf{Q}$ of $\mathbf{R}^N$. If $\mathbf{Q}[X] = C$, the set of consistent estimates of $X$ is:

$$\mathcal{C}(C) = \mathbf{Q}^{-1}[C], \tag{5}$$

where $\mathbf{Q}^{-1}[C]$ is the inverse image of $C$ through the mapping $\mathbf{Q}$, that is:

$$\mathbf{Q}^{-1}[C] = \left\{ Y \in \mathbf{R}^N \;/\; \forall k = 1,...,N, \;\; Y_k \in \mathrm{q}^{-1}[C_k] \right\}. \tag{6}$$

This set is typically a hyper cube of $\mathbf{R}^N$ whose geometric center is the quantized signal $C$.

This mapping analysis can be performed on predictive and noise-shaping encoders as well. The operators $H$ and $G$ are themselves mappings of $\mathbf{R}^N$ denoted by $\mathbf{H}$ and $\mathbf{G}$. We will use the fact that $\mathbf{H}$ is a linear and invertible mapping, and $\mathbf{G}$ is a strictly causal mapping[1]. From (2), we also have the mapping relationship:

$$\mathbf{H} = \mathbf{I} + \mathbf{G}, \tag{7}$$

where $\mathbf{I}$ is the identity mapping $\mathbf{R}^N$. Using these properties, we show in [6], in the case of predictive encoding, that:

$$\mathcal{C}(C) = \left(\mathbf{Q}^{-1}[B] - B\right) + C, \text{ where } B = \mathbf{H}^{-1}[C]. \tag{8}$$

[1] In practice, the feedback loop of a predictive or noise-shaping encoder necessarily has a delay. In our models, this delay is included in $\mathbf{G}$, not in $\mathbf{Q}$

In this relationship, $B$ is the fixed signal obtained from $C$ by the inverse mapping $\mathbf{H}^{-1}$ (see Figure 3(a)), and $(\mathbf{Q}^{-1}[B] - B) + C$ designates the subset $\mathbf{Q}^{-1}[B]$ successively translated by the fixed signals $-B$ and $C$. Since $\mathbf{Q}^{-1}[B]$ is a hyper-cube of $\mathbf{R}^N$ and $B$ is its center, $\mathcal{C}(C)$ is a hyper-cube of $\mathbf{R}^N$ with center $C$. For the case of noise-shaping encoding, we show in [6] that:

$$\mathcal{C}(C) = \mathbf{H}^{-1}\left[\mathbf{Q}^{-1}[C] - C\right] + C. \tag{9}$$

In this relation, $\mathbf{H}^{-1}\left[\mathbf{Q}^{-1}[C] - C\right]$ is the transformation through $\mathbf{H}^{-1}$ of $\mathbf{Q}^{-1}[C] - C$, which is a hyper-cube centered on $O$ (zero signal). Since $\mathbf{H}^{-1}$ is a linear mapping, $\mathbf{H}^{-1}\left[\mathbf{Q}^{-1}[C] - C\right]$ has the structure of a hyper-parallelepiped in $\mathbf{R}^N$, no longer cubic, but still centered on $O$. We conclude that in noise-shaping encoding, $\mathcal{C}(C)$ is a hyper-parallelepiped of $\mathbf{R}^N$ with center $C$.

Therefore, the deterministic approach leads to the following proposition:

**Proposition 1** *The encoded signal has the geometric property to be the center of the set of consistent estimates, regardless of the type of encoder.*

## 4. APPLICATION TO OVERSAMPLED ADC

In oversampled ADC, the discrete-time signals have the extra feature that they are the sampled versions of bandlimited signals. Therefore, they belong to a subset $\mathcal{V}$ of $\mathbf{R}^N$ which is a subspace. In this situation, the encoder is a mapping from $\mathcal{V}$ to $\mathbf{R}^N$. Also, when a signal $X \in \mathcal{V}$ is known by its encoded $C$, the exact information available about $X$ is: " $X \in \mathcal{C}(C) \cap \mathcal{V}$ ". We call the elements of $\mathcal{C}(C) \cap \mathcal{V}$ the **consistent estimates** of $X$. This set can be represented geometrically as shown in Figure 5.

This set has the particular property to be convex, since $\mathcal{C}(C)$ and $\mathcal{V}$ are both convex. As a consequence, we have the following property:

**Proposition 2** *If $C$ is the encoded signal of $X \in \mathcal{V}$, an estimate $\hat{X}$ of $X$ which is not consistent, is not optimal, since it can be theoretically improved by a convex projection on $\mathcal{C}(C) \cap \mathcal{V}$.*

This property is derived from the fact that projecting an element on a convex set (which does not contain the element) necessarily reduces its distance with any element of the set (Figure 6).

In linear decoding, the estimate $\hat{X}$ is obtained from the encoded signal $C$ by a cancellation of the out-of-band energy.

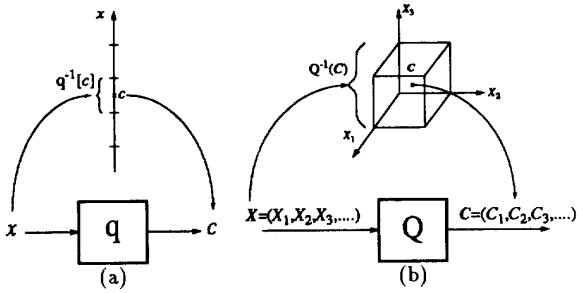


Figure 4: Representation of quantization as many-to-one mapping. (a) Single sample quantization. (b) Discrete-time signal quantization.

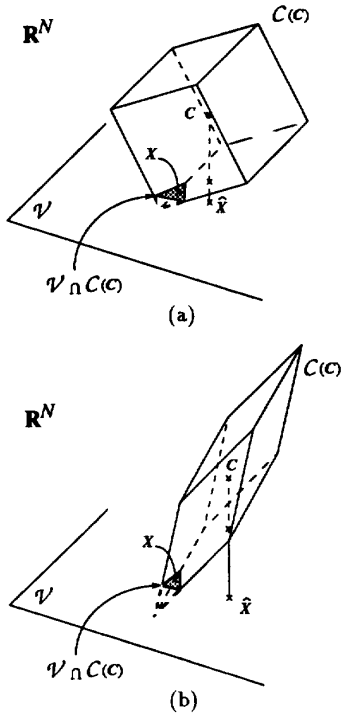


Figure 5: Geometric representation of encoding information in oversampled ADC and linear decoding scheme. (a) Simple and predictive encoding. (b) Noise-shaping encoding.

This is typically an orthogonal projection on the subspace $\mathcal{V}$. Therefore, the classical linear decoding has a geometric representation as shown in Figure 5. The figure shows that the projected estimate $\hat{X}$ does no longer belong to $\mathcal{C}(C)$. This indicates that the estimate obtained from the linear decoding scheme is not necessarily consistent and can be improved.

In [3, 5], some analytical evaluation of the improvement yielded by consistent decoding was given . For the case of simple encoding, it was shown in that under certain conditions on the quantization threshold crossings of the input signal, the mean square error (MSE) of consistent estimates decreases with $R$ in $\mathcal{O}(R^{-2})$, instead of $\mathcal{O}(R^{-1})$ for linear decoding. In the case of $n^{th}$ order $\Sigma\Delta$ modulation, starting on a certain model of quantization error signal, the order of $\mathcal{O}(R^{-(2n+2)})$ was derived, instead of $\mathcal{O}(R^{-(2n+1)})$ in linear decoding.

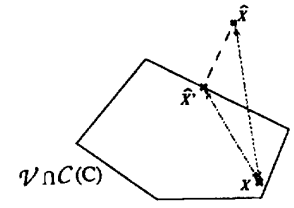


Figure 6: Geometric representation of Proposition 2

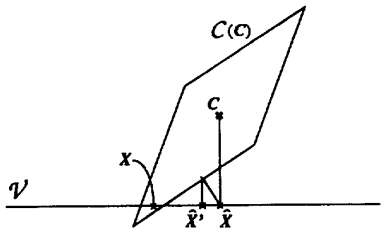


Figure 7: Geometric representation of the one-step improvement of a non-consistent estimate, on the example of the linear decoding estimate.

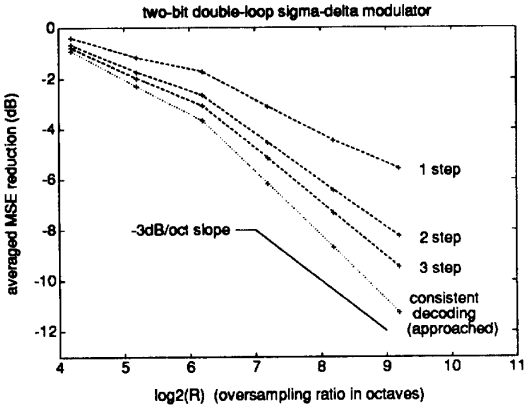


Figure 8: Numerical result of MSE reduction with respect to the linear decoding MSE obtained by finite step decoding improvement and consistent decoding approached by alternating projections, in two-bit double-loop $\Sigma\Delta$ modulation. The MSE is averaged over 600 periodic and lowpass signals containing 7 non-zero randomly generated discrete Fourier coefficients.

## 5. METHODS OF IMPROVEMENT OF NON-CONSISTENT ESTIMATES

It is not necessary to look for a consistent estimate to obtain an immediate improvement of a non-consistent estimate. Indeed, if for example, $\hat{X}$ belongs to $\mathcal{V}$ but not to $\mathcal{C}(C)$, as it is the case in linear decoding, a single projection on $\mathcal{C}(C)$ will lead to an immediate improvement, since $\mathcal{C}(C)$ is itself a convex set and $X \in \mathcal{C}(C)$ (see Figure 7). This estimate can then be further improved by a second projection on $\mathcal{V}$ (lowpass filtering). We call this operation a one-step improvement of a non-consistent estimate. Algorithms performing the projection on $\mathcal{C}(C)$ were proposed in [3, 6] for different types of encoders. This projection has a very straightforward implementation in the time domain in the case of simple encoding [3]. It is performed in a similar way in predictive encoding, since the expression of $\mathcal{C}(C)$ in (8) is similar to (5) up to a signal translation. In the case of noise-shaping encoders, algorithms for multi-loop $\Sigma\Delta$ modulators were also proposed [6]. Figure 8 shows the numerical results obtained from one to three steps of improvements performed on the linear decoding estimates, for a 2 bit double-loop $\Sigma\Delta$ modulator.

In fact, from a result on alternating projections on convex sets [7], iterating the one-step improvement infinitely will automatically converge to a consistent estimate. We used this property to approach a consistent estimate numerically

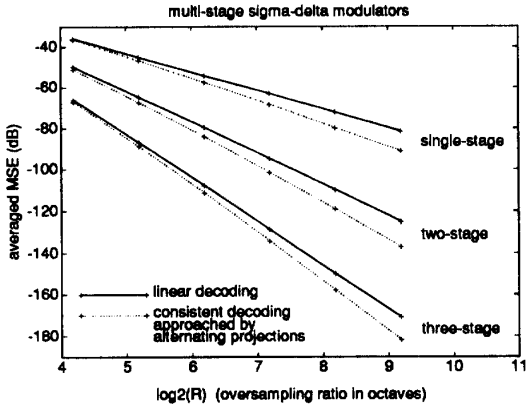


Figure 9: Comparison between the linear decoding MSE and the consistent decoding (approached by alternating projections) MSE, in single-bit single-loop, two-bit double-loop and three-bit triple-loop $\Sigma\Delta$ modulation. The experimental conditions are the same as in Figure 8.

by iterating the one-step improvement a large number of times. Infinite iteration schemes have also been investigated in [8], where the projection on $\mathcal{V}$ has been studied in particular. For double-loop $\Sigma\Delta$ modulation, the result is plotted in the same figure and shows an asymptotic improvement of at least 3 dB per octave of oversampling. Also, the comparison between the absolute MSE of linear decoding and consistent decoding (approached by alternating projections) is shown in Figure 9 for multi-stage $\Sigma\Delta$ modulators of order one to three. It confirms the asymptotic improvement of 3 dB/octave, regardless of the order of the modulator. This implies that the MSE is of the order of $\mathcal{O}(R^{-(2n+2)})$, versus $\mathcal{O}(R^{-(2n+1)})$ in linear decoding.

## References


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

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

[3] 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.

[4] N.T.Thao and M.Vetterli, "Optimal MSE signal reconstruction in oversampled A/D conversion using convexity," *Proc. IEEE Int. Conf. ASSP*, vol. IV, pp. 165–168, Mar. 1992.

[5] N.T.Thao and M.Vetterli, "Reduction of the MSE in $R$-times oversampled A/D conversion from $\mathcal{O}(1/R)$ to $\mathcal{O}(1/R^2)$," *IEEE Trans. on Signal Proc.* To appear.

[6] N.T.Thao and M.Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. on Signal Proc.* Submitted.

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

[8] S.Hein and A.Zakhor, "Reconstruction of oversampled band-limited signals from $\Sigma\Delta$ encoded binary sequences," *Proc. IEEE Int. Conf. ASSP*, vol. IV, pp. 161–164, Mar. 1992.