Deterministic Analysis of Oversampled A/D Conversion and Decoding Improvement Based on Consistent Estimates

Nguyen T. Thao and Martin Vetterli, *Senior Member, IEEE*

Abstract—**This paper deals with the deterministic analysis of oversampled A/D conversion (ADC), the properties derivable from such an analysis, and the consequences on reconstruction using nonlinear decoding. Given a bandlimited input X producing a quantized version C, we consider the set of all input signals that are bandlimited and produce C. We call any element of this set a consistent estimate of X. Regardless of the type of encoder (simple, predictive, or noise-shaping), we show that this set is convex, and as a consequence, any nonconsistent estimate can be improved. We also show that the classical linear decoding estimates are not necessarily consistent. Numerical tests performed on simple ADC, single-loop, and multiloop $\Sigma\Delta$ modulation show that consistent estimates yield an MSE that decreases asymptotically with the oversampling ratio faster than the linear decoding MSE by approximately 3 dB/octave. This implies an asymptotic MSE of the order of $\mathcal{O}(R^{-(2n+2)})$ instead of $\mathcal{O}(R^{-(2n+1)})$ in linear decoding, where R is the oversampling ratio and n the order of the modulator. Methods of improvements of nonconsistent estimates based on the deterministic knowledge of the quantized signal are proposed for simple ADC, predictive ADC, single-loop, and multiloop $\Sigma\Delta$ modulation.**

I. Introduction

HIGH-resolution analog-to-digital conversion (ADC) of bandlimited signals is often achieved by oversampling, or sampling at higher than the Nyquist rate, rather than increasing the resolution of the amplitude quantization. In the simple ADC case, where the values of the oversampled discrete-time input signal are quantized individually, the signal reconstruction is based on the fact that only a fraction of the quantization error power lies in the input signal bandwidth. The global ADC resolution is then improved by low-pass filtering the quantized signal at the cut-off frequency equal to the maximum input frequency f_m. In the best case, where a certain number of conditions are verified [1], the quantization error signal can be satisfactorily modeled as white noise. In this situation, low-pass filtering reduces the mean square error (MSE) of the reconstructed signal by a factor equal to the oversampling ratio $R = (f_s/2f_m)$ (see Fig. 1(a)), where f_s is the sampling frequency. This method of reconstruction is the best that can be achieved with linear processing of the quantized signal. This will be called the linear decoding method.

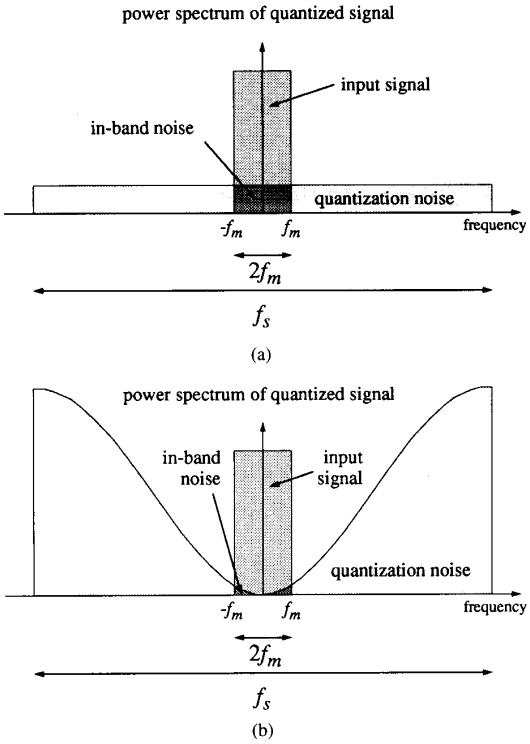

Fig. 1. Power spectrum of quantized signal with the assumption of white quantizing error: (a) Simple ADC; (b) single-loop $\Sigma\Delta$ modulation.

The question is whether the in-band error remaining in such a reconstruction is irreversible or can be further reduced. Note that increasing the oversampling by 2 decreases the error power by 2, whereas increasing the amplitude quantization by 2 leads to a reduction of the error power by 4 (since the error squared has a $(\Delta^2/12)$ behavior, where Δ is the quantization step size). This asymmetry is disappointing since a bandlimited signal with bounded amplitude has a limited slope. Thus, one expects that a variation along the amplitude dimension, or at least an upper bound, is linearly equivalent to a variation along

Manuscript received March 16, 1992; revised April 2, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. Barry Sullivan. This work was supported by the National Science Foundation under grant ECD-88-11111.

N. T. Thao is with the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong.

M. Vetterli was with the Department of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York, NY 10027-6699. He is now with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720

IEEE Log Number 9214661.

the time dimension. This lack of symmetry is a hint that linear decoding might be suboptimal not just by a fixed amount but rather by a factor dependent on R.

Very often, quantization is analyzed using statistical methods, even though it is essentially a deterministic operation. Using a deterministic analysis, we show in Section III that the information missing in the linear decoding method is the consistency constraint. We call a reconstruction of an analog input signal consistent when it reproduces the same quantized signal if it is to be requantized. Then, it is shown that a nonconsistent estimate can be necessarily improved due to convexity properties. We propose methods of improvements based on convex projections. Consistent estimates can be indeed approached by iterating these convex projections (alternating projection method). Numerical tests show that under certain conditions on the input quantization threshold crossings, the MSE of consistent estimates is proportional to R^{-2} instead of R^{-1} in linear decoding. Thus, we recover the symmetry between time and amplitude as expected. This implies an improvement of the MSE reduction by 3 dB per octave of oversampling. This result is proved in [2].

The same questions can be posed for more sophisticated types of encoding in oversampled ADC, such as predictive ADC and noise shaping ADC [3]. These encoders include an internal processing of the error made by the quantizer that reduces the overall spectral density of the error contained in the quantized output signal (predictive ADC) or the in-band portion of this error (noise-shaping ADC). Again, signal reconstruction is classically based on a linear filtering of the quantized signal. Fig. 1(b) shows the typical power spectral density of the quantized signal in single loop $\Sigma\Delta$ modulation. In the case of nth-order $\Sigma\Delta$ modulation and under the assumption of white quantization noise, it was shown in [3] that the in-band error included in the quantized signal has a power of the order of $\mathcal{O}(R^{-(2n+1)})$. Although the white quantization noise assumption does actually not hold [1], this approach leads to good prediction of the linear decoding performances. We show in Sections IV and V that the same deterministic approach that we used in simple ADC can be applied to predictive ADC and single- and multiloop $\Sigma\Delta$ modulation (which are examples of noise-shaping ADC). As a generalization of simple ADC, we show that the consistency constraint is not satisfied in the linear decoding schemes. However, the principles of improvement of nonconsistent estimates are exactly the same. Algorithms for convex projections are proposed for these more sophisticated encoding schemes. Numerical tests performed for the single- and multiloop cases (Section V) show that the MSE of consistent estimates has an asymptotic behavior of the order $R^{-(2n+2)}$ instead of $R^{-(2n+1)}$ in linear decoding. Under simplified assumptions, this result was shown in [4]. This again represents an asymptotic improvement of the MSE reduction by 3 dB per octave. A large portion of this improvement can be achieved with finite complexity algorithms.

Relation with Previous Work: Convex projection are standard in signal and image reconstruction [5]. In the field of image quantization, convex projections were used in [6] for the reconstruction of images given by multiple-level threshold crossings. In oversampled ADC and $\Sigma\Delta$ modulation, the notion of consistent estimates, the convexity of the set of such estimates, and the use of convex projections were introduced in [7], showing the potential for asymptotic improvements. Further studies of such schemes were done in [8]–[11]. The proofs of asymptotic improvement in simple ADC can be found in [7], [9], [2] and [11], and in [9], [4], and [11] for $\Sigma\Delta$ modulation with restrictive assumptions. Detailed proofs of convexity and algorithms for convex projections using the deterministic information of quantized signals are given in [9] and [11]. A study of the bandlimitation orthogonal projections is proposed in [10] and [12].

II. Mathematical Context and Notations

A. Continuous-Time and Discrete-Time Signals

Continuous-time signals, which are denoted by $X[t]$, are real signals and are assumed to be observed on a finite length time interval $]0, T_0]$. The error between two signals $X[t]$ and $X'[t]$ is measured by the MSE equal to $(1/T_0)\int_0^{T_0} |X'[t]-X[t]|^2 dt$. We assume that continuous-time signals are sampled N times in the interval $]0, T_0]$ at instants $(k/N)T_0, k = 1, \cdots, N$. The discrete-time signals are thus elements of $\boldsymbol{R}^N$ and denoted as $\boldsymbol{Y} = (Y(k))_{k=1,\cdots,N}$. The sampled version of a continuous-time signal $X[t]$ is the discrete-time signal $\boldsymbol{X}$ such that $X(k) = X[(k/N)T_0]$ for $k = 1, \cdots, N$.

B. Space $\boldsymbol{R}^N$ of Discrete-Time Signals

Notations relative to $\boldsymbol{R}^N$ are defined here. Subsets of $\boldsymbol{R}^N$ are designated by calligraphic letters (e.g., $\mathcal{A}, \mathcal{B}, \mathcal{C}$). If $\mathcal{C}$ is a subset of $\boldsymbol{R}^N, \mathcal{C} + \boldsymbol{X}$ denotes the translated version of $\mathcal{C}$ by $\boldsymbol{X}$, that is, $\mathcal{C} + \boldsymbol{X} = \{\boldsymbol{Y} + \boldsymbol{X}/\boldsymbol{Y} \in \mathcal{C}\}$. A mapping of $\boldsymbol{R}^N$ is a function H mapping any element of $\boldsymbol{R}^N$ into another unique element of $\boldsymbol{R}^N$ denoted $H[\boldsymbol{X}]$. The notation I denotes the identity mapping. If H is a one-to-one (or invertible) mapping, H^{-1} designates the inverse mapping. If H_1 and H_2 are two mappings of $\boldsymbol{R}^N, H_2H_1$ is the mapping such that $H_2H_1[\boldsymbol{X}] = H_2[\boldsymbol{H}_1[\boldsymbol{X}]]$. If $\mathcal{C}$ is a subset of $\boldsymbol{R}^N, H[\mathcal{C}]$ and $H^{-1}[\mathcal{C}]$ designate the forward and inverse images of $\mathcal{C}$ through H respectively (H need not be invertible). When $\mathcal{C}$ is reduced to a singleton $\mathcal{C} = \{\boldsymbol{C}\}$, by abuse of notation, $H^{-1}[\mathcal{C}]$ is denoted by $H^{-1}[\boldsymbol{C}]$. With this notation, when H is not invertible, $H^{-1}[C]$ is a subset of $\boldsymbol{R}^N$, that is, $H^{-1}[\boldsymbol{C}] = \{\boldsymbol{X} \in \boldsymbol{R}^N/H[\boldsymbol{X}] = \boldsymbol{C}\}$. For illustration purposes, we show here an example of use of these notations, which will be used later in this paper. If $\boldsymbol{C}$ is an element of $\boldsymbol{R}^N$ and H, G, Q are three mappings of $\boldsymbol{R}^N$ where H is invertible, then $H^{-1}[Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]]$ is a subset of $\boldsymbol{R}^N$, which is constructed as follows:

i) $Q^{-1}[\boldsymbol{C}]$ is a subset of $\boldsymbol{R}^N$, which is the inverse image of $\boldsymbol{C}$ through Q.

ii) $Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]$ is the subset $Q^{-1}[\boldsymbol{C}]$ translated by the fixed element $G[\boldsymbol{C}]$ of $\boldsymbol{R}^N$.

iii) $H^{-1}[Q^{-1}[\boldsymbol{C}]+G[\boldsymbol{C}]]$ is the forward image of the subset $Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]$ through the mapping H^{-1}.

Finally, for $\boldsymbol{X} \in \boldsymbol{R}^N$, we define the norm $||\boldsymbol{X}|| = ((1/N)\Sigma_{k=1}^N |X(k)|^2)^{(1/2)}$.

C. Space $\mathcal{V}$ of Bandlimited Signals

The considered continuous-time signals are assumed to be perfectly bandlimited with cut-off frequency f_m. This implies that they have an infinite support or have nonzero values outside intervals of any finite length. However, as already stated they will be observed only on the finite interval $]0, T_0]$, as is always the case in practice. Even in the case of oversampling, Shannon's sampling theorem does not lead to uniqueness of reconstruction if the known samples are limited to the interval $]0, T_0]$. We therefore introduce an approximation of bandlimited signals that will allow us to recover Shannon's uniqueness. We assume that the energy outside $]0, T_0]$ of the considered signals is small enough and decays fast enough so that their restrictions to $]0, T_0]$ are "almost equal" to the periodized versions, which are obtained by subsequently adding their translated versions by multiples of T_0. By "almost," we mean that the error made by this approximation (or aliasing error) is at least small compared with other sources of errors existing in the ADC process, such as those due to quantization, for example. In the frequency domain, this means that the Fourier transform of a bandlimited signal is approximated by its discrete frequency version, where the period of discretization is equal to $f_0 = (1/T_0)$. Therefore, the approximation consists of saying that the restrictions of input signals to the interval $]0, T_0]$ are elements of the space $\mathcal{V}$ of all possible signals bandlimited by f_m and T_0 periodic.

We therefore propose to study the behavior of oversampled ADC when input signals are in general elements of $\mathcal{V}$. These signals have the form

$$X[t] = \sum_{i=-M}^{M} X_i e^{j2\pi i(t/T_0)}$$

where

$$X_{-i} = X_i^* \text{ for } i = 0, \cdots, M \tag{1}$$

and $M = (f_m/f_0)$ (we assume that f_0 is chosen to be an divisor of f_m). Therefore, $\mathcal{V}$ is a real vector space of dimension $2M+1$. As a finite dimension version of Shannon's sampling theorem, it can be shown that a signal $X[t]$ of the space $\mathcal{V}$ is uniquely represented by its sampled version $\boldsymbol{X}$ as soon as $N \geq 2M+1$ (N is analogous to the sampling frequency and $2M+1$ to the bandwidth). It can also be shown that when $N \geq 2M+1$, if $\boldsymbol{X}$ and $\boldsymbol{X}'$ are the sampled versions of two signals $X[t]$ and $X'[t]$ of $\mathcal{V}$, then $||\boldsymbol{X}' - \boldsymbol{X}||^2$ equals the MSE between $X[t]$ and $X'[t]$. Because of the uniqueness feature and this property of distance preservation, the bandlimited signals will be implicitly considered in their discrete-time version, and $\mathcal{V}$ will be considered as a subspace of $\boldsymbol{R}^N$. In the oversampled situation, that is $N > 2M+1$, $\mathcal{V}$ is a subspace of $\boldsymbol{R}^N$ in the strict sense. The oversampling ratio $R = N/(2M+1)$ gives the ratio between the dimensions of $\boldsymbol{R}^N$ and $\mathcal{V}$.

III. Simple Oversampled ADC

The basic mechanisms resulting from a deterministic analysis of oversampled ADC can be easily demonstrated in the case of simple ADC. Their detailed description is necessary in order to convey the basic concepts and prepare the framework for generalization to predictive and noise-shaping ADC.

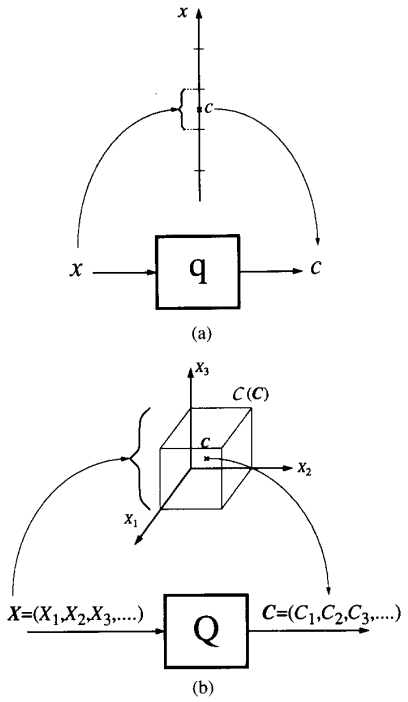

Fig. 2. Representation of quantization as many-to-one mapping: (a) Single sample quantization; (b) discrete-time signal quantization.

A. Deterministic Description of Quantization

In a deterministic approach, quantization is simply a *many-to-one* mapping of $\boldsymbol{R}^N$. In the particular case of $N = 1$, where signals are reduced to single sample values, quantization is a mapping of R (denoted by q) such that whole intervals of $\boldsymbol{R}$, which are called quantization intervals, are mapped into single discrete values, which are called quantization levels (see Fig. 2(a)). The quantization interval corresponding to a quantization level c is denoted by $q^{-1}[c]$. The second characteristic of the quantization mapping is that it is a consistent mapping. By this, we mean that for all quantization levels $c, q[c] = c$ or $c \in q^{-1}[c]$. In practice, c is typically chosen to be the center of the $q^{-1}[c]$.[1] This will be assumed in this paper. When quantization is uniform (this will not be necessarily assumed), the quantization intervals have a common length, which is denoted by Δ and called the quantization step size.[2]

In the general case where $N \geq 1$, quantization is a mapping Q of $\boldsymbol{R}^N$ such that

$$\forall \boldsymbol{X} \in \boldsymbol{R}^N,$$
$$\boldsymbol{C} = Q[\boldsymbol{X}] \Longleftrightarrow \forall k = 1, \cdots, N, C(k) = q[X(k)].$$

[1] Actually, a quantizer has a finite number of quantization levels, and the two extreme quantization intervals are necessarily infinite. If c is one of the two extreme quantization levels, it is typically chosen to be the center of $q^{-1}[c] \cap B$, where B is the specified bounded region of input samples, which is called the nonoverload region.

[2] For the extreme quantization intervals, it is implicitly $q^{-1}[c] \cap B$, which has length Δ.

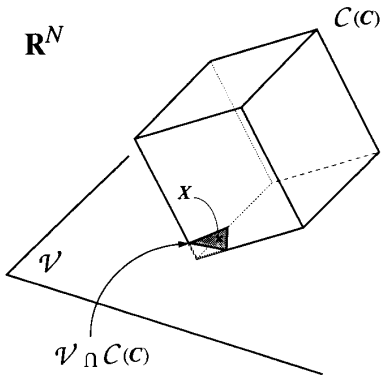

Fig. 3. Geometric representation of the information "$\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C}) \bigcap \mathcal{V}$" in simple ADC (uniform quantization).

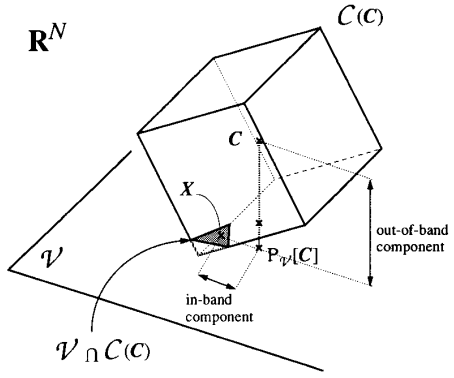

Fig. 4. Geometric representation of the nonconsistency of the linear decoding estimate $P_{\nu}[\boldsymbol{C}]$ in simple ADC.

We say that $\boldsymbol{C}$ is the quantized version of $\boldsymbol{X}$. Rigorously speaking, if an input signal $\boldsymbol{X} \in \boldsymbol{R}^N$ is known only by its quantized version $\boldsymbol{C}$, the full information available about $\boldsymbol{X}$ is "$\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C})$" where $\mathcal{C}(\boldsymbol{C})$ is the set of possible input signals with quantized version $\boldsymbol{C}$. By definition, we have $\mathcal{C}(\boldsymbol{C}) = Q^{-1}[\boldsymbol{C}]$, where

$$Q^{-1}[\boldsymbol{C}] = \{\boldsymbol{Y} \in \boldsymbol{R}^N / \forall k = 1, \cdots, N, Y(k) \in q^{-1}[C(k)]\}.$$

Therefore, when quantization is uniform, $\mathcal{C}(\boldsymbol{C})$ is a hypercube of $\boldsymbol{R}^N$ (see Fig. 2(b)) and C is its geometric center. In the more general case of nonuniform quantization, $\mathcal{C}(\boldsymbol{C})$ is a rectangular hyper-parallelepiped.[3]

Now, in oversampled ADC, we have the extra information about the input signal that it belongs to $\mathcal{V}$. We therefore have the following proposition.

Proposition 3.1: In oversampled ADC, when a bandlimited signal $\boldsymbol{X}$ is only known by the quantized signal $\boldsymbol{C}$ it produces through the encoder, the full information available about $\boldsymbol{X}$ is

$$\text{"}\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}\text{."}$$

This information is represented geometrically in Fig. 3.

B. Consistent Estimates

The goal of signal decoding in oversampled ADC is to reconstruct an estimate $\hat{\boldsymbol{X}}$ of an original input signal $\boldsymbol{X}$ from its quantized version $\boldsymbol{C}$. Since the knowledge of $\boldsymbol{C}$ implies the information "$\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$," it is tempting to pick as estimate of $\boldsymbol{X}$ an element $\hat{\boldsymbol{X}}$ of $\mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$. For reasons which will become clear later, we are particularly interested in the set of estimates $\overline{\mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}}$. We propose the following definition.

Definition 3.2: When $\boldsymbol{C}$ is the quantized signal produced by a bandlimited input signal $\boldsymbol{X}$, the elements of $\overline{\mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}}$ are called the consistent estimates of $\boldsymbol{X}$.

We show that when an estimate of $\boldsymbol{X}$ is not consistent, then by necessity it can be theoretically improved by a consistent estimate. This is based on the fact (which is easy to verify) that $\mathcal{C}(\boldsymbol{C})$ and $\mathcal{V}$ are convex sets and the two following lemmas.

[3]To be rigorous, $Q^{-1}[\boldsymbol{C}]$ may not be bounded. However, the picture of hypercube or hyperparallelepiped with C as geometric center holds when taking $\mathcal{C}(\boldsymbol{C}) = Q^{-1}[\boldsymbol{C}] \cap B^N$, which simply assumes that input signals $\boldsymbol{X}$ belong to the nonoverload region, or $X(k) \in B$ for $k = 1, \cdots, N$.

Lemma 3.3 [13]: Let $\boldsymbol{Y}$ be an element of $\boldsymbol{R}^N$ and $\mathcal{S}$ a closed set. There exists a unique element $\boldsymbol{Y}'$ of $\overline{\mathcal{S}}$ such that for all $Z \in \mathcal{S}, ||\boldsymbol{Y}' - \boldsymbol{Y}|| \leq ||\boldsymbol{Z} - \boldsymbol{Y}||$. The transformation from $\boldsymbol{Y}$ to $\boldsymbol{Y}'$ it then a mapping of $\boldsymbol{R}^N$ called the convex projection on $\mathcal{S}$ and is denoted by $P_{\mathcal{S}}$.

Lemma 3.4 [14]: If $\boldsymbol{Y}'$ is the convex projection of $\boldsymbol{Y}$ on a closed set $\overline{\mathcal{S}}$ and $\boldsymbol{Y} \notin \mathcal{S}$, then for all $\boldsymbol{Z} \in \mathcal{S}, ||\boldsymbol{Y}' - \boldsymbol{Z}|| < ||\boldsymbol{Y} - \boldsymbol{Z}||$.

It is easy to verify that $\mathcal{S} = \mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$ is convex. Applying Lemma 3.4 to $\mathcal{S} = \mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$ and using the fact that $\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$, we obtain the following property.

Property 3.5: Let $\boldsymbol{X}$ be a bandlimited signal producing the quantized signal $\boldsymbol{C}$. Let $\hat{\boldsymbol{X}}$ be a nonconsistent estimate of $\boldsymbol{X}$, that is, $\hat{\boldsymbol{X}} \notin \overline{\mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}}$. Then, the convex projection $\hat{\boldsymbol{X}}'$ of $\hat{\boldsymbol{X}}$ on $\mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$ is a consistent estimate of $\boldsymbol{X}$ and $||\hat{\boldsymbol{X}}' - \boldsymbol{X}|| < ||\hat{\boldsymbol{X}} - \boldsymbol{X}||$.

This implies that when $\hat{\boldsymbol{X}}$ is a nonconsistent estimate of $\boldsymbol{X}$, the knowledge of $\hat{\boldsymbol{X}}$ and $\boldsymbol{C}$ gives enough information to characterize a consistent estimate that is always better than $\hat{\boldsymbol{X}}$.

C. Nonconsistency of Linear Decoding

We show in this section that linear decoding in oversampled ADC does not necessarily provide consistent estimates. This can first be seen geometrically. Linear decoding consists of low-pass filtering the quantized signal $\boldsymbol{C}$ at cut-off frequency f_m. In the space $\boldsymbol{R}^N$, this amounts to performing an orthogonal projection of $\boldsymbol{C}$ on the subspace of bandlimited signals. In other words, the linear decoding scheme provides as estimate of $\boldsymbol{X}$ the projection on $\mathcal{V}$ of the center of the hypercube $\mathcal{C}(\boldsymbol{C})$. As indicated in Fig. 4, unless the cube $\mathcal{C}(\boldsymbol{C})$ lies at a particular "angle" with $\mathcal{V}$, there is no reason for this estimate to belong to $\mathcal{C}(\boldsymbol{C})$.

Fig. 5 shows a concrete example of this nonconsistency in the time domain. A bandlimited signal is generated numerically, sampled at the oversampling ratio $R = 4$, and quantized. The figure shows that the estimate $\hat{\boldsymbol{X}}$ obtained from low-pass filtering the quantized signal $\boldsymbol{C}$ does not belong to $\mathcal{C}(\boldsymbol{C})$ since, for example, the samples $\hat{X}(11)$ and $\hat{X}(12)$ do not belong to quantization intervals $q^{-1}[C(11)]$ and $q^{-1}[C(12)]$, respectively.

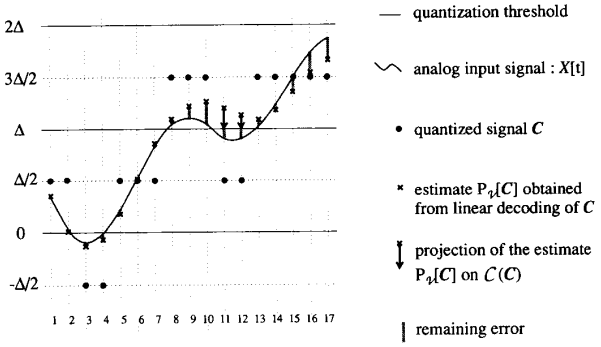

Fig. 5. Time representation of an example of nonconsistency of a linear decoding estimate $P_\nu[\boldsymbol{C}]$. The samples of $P\nu[\boldsymbol{C}]$ at time indices $k = 11, 12$ do not belong to the quantization intervals $q^{-1}[C(11)]$ and $q^{-1}[C(12)]$, respectively.

D. Methods of Improvement of Nonconsistent Estimates

Property 3.5 gives the mathematical justification for possible improvement of nonconsistent estimates but does not really provide a method to achieve it. In fact, a method for at least partial improvement can be obtained by applying Lemma 3.4 on either $\mathcal{S} = \mathcal{V}$ or $\mathcal{S} = \mathcal{C}(\boldsymbol{C})$. If the nonconsistent estimate $\hat{\boldsymbol{X}}$ does not belong to $\mathcal{V}$, then according to Lemma 3.4, it will be necessarily improved by a projection on $\mathcal{V}$, which is a low-pass filtering at cut-off frequency f_m. Similarly, if $\hat{\boldsymbol{X}} \notin \mathcal{C}(\boldsymbol{C})$, an improvement can be achieved by using a projection on $\mathcal{C}(\boldsymbol{C})$.[4] This projection is performed by the following algorithm, which is similar to the algorithm introduced in [6] for 2-D image reconstruction.

Algorithm 1

At every instant k

i) if $\hat{X}(k) \in q^{-1}[C(k)]$, take $\hat{X}'(k) = \hat{X}(k)$

ii) else, take $\hat{X}'(k)$ equal to the bound of the quantization interval $q^{-1}[C(k)]$ closest to $\hat{X}(k)$.

Qualitatively speaking, the algorithm consists in projecting each sample $\hat{X}(k)$ on the quantization interval $q^{-1}[C(k)]$ indicated by the quantized value $C(k)$ when $\hat{X}(k) \notin q^{-1}[C(k)]$. It is easy to verify that this leads to an estimate $\hat{X}'$, which is the projection of $\hat{X}$ on $\overline{\mathcal{C}(\boldsymbol{C})}$. In particular, this projection algorithm can be used for immediate improvement of the linear decoding estimate since this corresponds to the case $\hat{X}(k) \notin q^{-1}[C(k)]$. This improvement is illustrated in Fig. 5 by the dark arrows.

In fact, as long as an estimate does not belong to $\mathcal{C}(\boldsymbol{C})$ or $\mathcal{V}$, a projection on either $\mathcal{C}(\boldsymbol{C})$ or $\mathcal{V}$ can be reiterated, thus implying further reductions of the distance between the current estimate and $\boldsymbol{X}$. The obtained improvement will always be an increasing function of the number of iterations.

E. Conceptual Method for Consistent Reconstruction

It was shown in [14] that alternating projections infinitely between two intersecting convex sets necessarily converges to the closure of their intersection. This property became the basis of the algorithm of alternating projection classically used in signal processing and first introduced by Youla [15] for image reconstruction. In our case, this property ensures that the limit of alternating projections between $\mathcal{C}(\boldsymbol{C})$ and $\mathcal{V}$ constitutes a consistent estimate of $\boldsymbol{X}$. Numerically speaking, this implies that a consistent estimate can at least be approached using finite step alternating projections.

[4] To be precise, the improvement is strict when $\hat{\boldsymbol{X}} \notin \overline{\mathcal{C}(\boldsymbol{C})}$.

F. Analytical Comparison Between Linear and Consistent Decoding

Starting the alternating projection scheme from the estimate provided by linear decoding is a way to find a consistent estimate that automatically improves this decoding scheme. The question is now to have an analytical evaluation of this improvement. Our approach is to find an upper bound to any consistent estimate of $\boldsymbol{X}$ and compare it with the expected MSE of a classical estimate. In [2], we derived the following theorem:

Theorem 3.6: Let $\boldsymbol{X}$ be a bandlimited signal such that its continuous-time version $X[t]$ crosses the quantization thresholds at least $2M + 1$ times in the interval $]0, T_0]$. Then, there exists a constant $\alpha_x > 0$ that does not depend on the oversampling ratio R such that, for R high enough, if $\boldsymbol{C}$ is the quantized version of $X[t]$ at ratio R, then

$$\forall \hat{\boldsymbol{X}} \in \mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}, \quad ||\hat{\boldsymbol{X}} - \boldsymbol{X}||^2 \leq \frac{\alpha_x}{R^2}.$$

This is to be compared with the classical decoding MSE, which is proportional to $(1/R)$. This means that under the special condition on the quantization threshold crossings of Theorem 3.6, the consistent decoding MSE asymptotically decreases at least at the rate of 6 dB per octave of R instead of 3 dB per octave for the linear decoding MSE. Regardless of the value of α_x, this means that the speed of MSE reduction is improved by 3 dB per octave of R when R is high enough.

G. Numerical Evaluation of the Performance of Consistent Reconstruction Using Alternating Projections

Numerical tests were performed to evaluate the performance of consistent estimates obtained by alternating projections with respect to the oversampling ratio R. For a given number $2M + 1$, bandlimited signals of the form (1) were randomly generated with the constraint that they cross the quantization thresholds at least $2M+1$ times. This is ensured by imposing a minimum to their peak-to-peak amplitude (PPA). For example, in uniform quantization with step size Δ, when $2M + 1 = 3$, the PPA is forced to be equal to 2Δ, ensuring that the input signals have at least three quantization threshold crossings. Then, for a fixed oversampling ratio R, the quantized version of each of these input signals was computed, as well as an approximately consistent estimate obtained by alternating projections. The linear decoding estimate was used as the first estimate in the projection iteration, and the alternating process was stopped as soon as the estimate MSE decrement per iteration was less than 0.001 dB. The resulting MSE is averaged over all randomly generated input signals. Although this MSE is not the MSE of rigorously consistent estimates, since the iteration of alternating projections is finite, one can

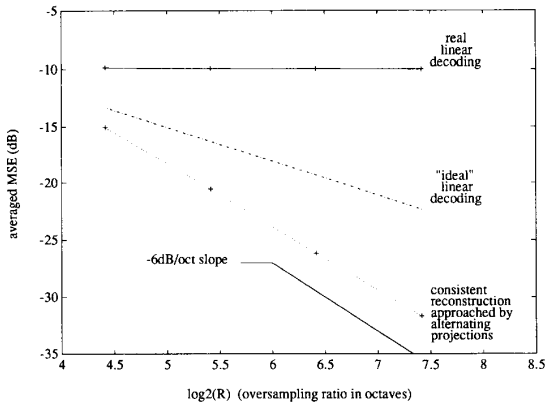

Fig. 6. Dependence of the MSE with the oversampling ratio R for real linear decoding, "ideal" linear decoding, and consistent decoding approached by alternating projections.

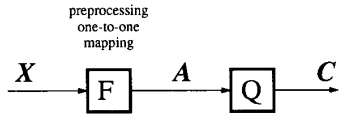

Fig. 7. Structure of generalized encoder.

be sure that the consistent estimates obtained from infinite iteration would necessarily yield an even smaller MSE.

The evolution of the averaged MSE versus the oversampling ratio is plotted in Fig. 6; in the case where $2M + 1 = 3$, quantization is uniform, PPA $= 2\Delta$, and R varies approximately between 20 and 170. The MSE is averaged over 1000 generated input signals. In the same figure, we plot the MSE predicted by the classical decoding model equal to $(\Delta^2/12)(1/R)$, whose slope is -3 dB/octave. Note that with PPA $= 2\Delta$, the number of the quantization level is limited to 3. With this very low quantization resolution, it is known that the classical decoding MSE no longer decreases by 3 dB/octave since the quantization error signal becomes correlated to the input signal. The solid curve of Fig. 6 obtained by computer simulation of the linear decoding shows that oversampling no longer reduces the real linear decoding MSE, which stagnates. For consistent estimates, there is no such stagnation, and the slope of -6 dB/octave is experimentally verified.

IV. Predictive Oversampled ADC

A. Encoding with Preprocessing

Predictive encoders belong to the more general family of encoders that include some analog preprocessing of the discrete-time input signals before quantization. We will show that this preprocessing can be described as a one-to-one mapping of $\boldsymbol{R}^N$, as shown in Fig. 7. Such an encoder can be seen as a many-to-one mapping, where for a given quantized signal $\boldsymbol{C}$, the set input signals producing $\boldsymbol{C}$ is $\mathcal{C}(\boldsymbol{C}) = F^{-1}[Q^{-1}[\boldsymbol{C}]]$. Then, the same approach of signal reconstruction as in simple ADC can be considered with this new expression of $\mathcal{C}(\boldsymbol{C})$.

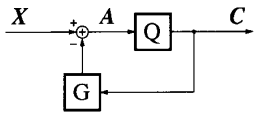

Fig. 8. General block diagram of a predictive encoder.

The general block diagram of a predictive encoder is shown in Fig. 8 [3], [16]. It includes the particular case of Δ modulation, when G is a discrete-time integrator. In general, we will only assume that G is a strictly causal mapping of $\boldsymbol{R}^N$, that is, when $\boldsymbol{D} = G[\boldsymbol{C}], D(k)$ only depends on $C(1), \cdots, C(k-1)$ for $k = 1, \cdots, N$. In particular, $D(1)$ is a constant G_1 independent of $\boldsymbol{C}$.

We show in Section IV-B that a predictive encoder indeed belongs to the family of encoders described by Fig. 7, where the one-to-one mapping F can be expressed in terms of G and Q. Then, the resulting expression of $\mathcal{C}(\boldsymbol{C})$ will show that $\mathcal{C}(\boldsymbol{C})$ is still a rectangular hyper-parallelepiped of $\boldsymbol{R}^N$ as in simple ADC. The notion of consistent estimation can be applied, and linear decoding nonconsistency and methods of improvement will be studied as extensions of the simple ADC case.

B. Deterministic Description of Predictive Encoding

The equivalence of a predictive encoder with the block diagram of Fig. 7 is based on the following lemma:

Lemma 4.1: $I + GQ$ is a one-to-one mapping of $\boldsymbol{R}^N$.

Proof: For a given $\boldsymbol{X} \in \boldsymbol{R}^N$, suppose there exists $\boldsymbol{A} \in \boldsymbol{R}^N$ such that $\boldsymbol{X} = (I + GQ)[\boldsymbol{A}]$. Let $\boldsymbol{D} = G[Q[\boldsymbol{A}]]$. This implies that $\boldsymbol{A} = \boldsymbol{X} - \boldsymbol{D}$. From the assumption of strict causality of G, we know that $D(1)$ is a constant G_1 independent of the input of G. Therefore, $A(1)$ is uniquely defined. Now, suppose that for a certain $k = 1, \cdots, N-1$, we have proved that $A(1), \cdots, A(k)$ are uniquely defined from the knowledge of $\boldsymbol{X}$. Then, $D(k+1)$ is uniquely defined from $q[A(1)], \cdots, q[A(k)]$ because G is strictly causal. This uniquely defines $A(k+1) = X(k+1) - D(k+1)$. We have therefore proved by induction that when $\boldsymbol{A}$ exists, it is uniquely defined. This induction actually shows an explicit construction of A and therefore gives at the same time the existence proof. Therefore $I + GQ$ is an invertible (or one-to-one) mapping. □

As a consequence, we have the following two propositions.

Proposition 4.2: The predictive encoder of Fig. 8 has the general structure of Fig. 7 where the preprocessing one-to-one mapping is

$$F = (I + GQ)^{-1}. \tag{2}$$

Proof: The signal $\boldsymbol{A}$ defined in Fig. 8 verifies $\boldsymbol{A} = \boldsymbol{X} - G[Q[\boldsymbol{A}]]$, which implies that $(I + GQ)[\boldsymbol{A}] = \boldsymbol{X}$. We have from Lemma 4.1 that $I + GQ$ is a one-to-one mapping. Therefore, $\boldsymbol{A} = F[\boldsymbol{X}]$, where $F = (I + GQ)^{-1}$. □

Proposition 4.3: The set of signals producing a quantized signal $\boldsymbol{C}$ through the predictive encoder of Fig. 8 is

$$\mathcal{C}(\boldsymbol{C}) = Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]. \tag{3}$$

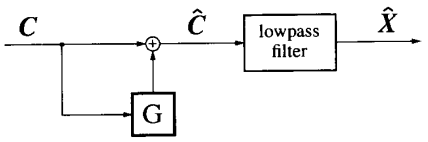

Fig. 9. Linear decoding in predictive ADC.

Proof: Applying Proposition 4.2, we have

$$\mathcal{C}(\boldsymbol{C}) = F^{-1}[Q^{-1}[\boldsymbol{C}]] = (I + GQ)[Q^{-1}[\boldsymbol{C}]]$$
$$= I[Q^{-1}[\boldsymbol{C}]] + GQ[Q^{-1}[\boldsymbol{C}]] = Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}].\square$$

This means that as in simple ADC, $\mathcal{C}(\boldsymbol{C})$ is a rectangular hyperparallelepiped of $\boldsymbol{R}^N$ (or a hypercube if quantization is uniform) since it is equal to the rectangular hyperparallelepiped $Q^{-1}[\boldsymbol{C}]$ translated by the vector $G[\boldsymbol{C}]$ of $\boldsymbol{R}^N$.

The main point is that Proposition 3.1 relative to the encoded information in the oversampling situation is still applicable in the case of predictive encoding, and the geometric representation of this information is still that of Fig. 3 since $\mathcal{C}(\boldsymbol{C})$ is a hypercube.

C. Consistent Estimates and Nonconsistency of Linear Decoding

We have the same notion of consistent estimates (Definition 3.2), and Property 3.5 is still valid. Let us show that the linear decoding scheme does not necessarily provide consistent estimates. The way to reconstruct an input signal $\boldsymbol{X}$ using the output $\boldsymbol{C}$ in linear decoding is to low-pass filter the signal $\hat{\boldsymbol{C}} = (I + G)[\boldsymbol{C}]$ as shown in Fig. 9 [3], [16]. Since $\boldsymbol{C}$ is the center of $Q^{-1}[\boldsymbol{C}]$, then $\hat{\boldsymbol{C}} = \boldsymbol{C} + G[\boldsymbol{C}]$ is the center of $Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]$ equal to $\mathcal{C}(\boldsymbol{C})$. Therefore, as in simple ADC, linear decoding consists in performing an orthogonal projection of the center of $\mathcal{C}(\boldsymbol{C})$ on $\mathcal{V}$. This is still represented by Fig. 4, where $\boldsymbol{C}$ has to be replaced by $\hat{\boldsymbol{C}}$. Therefore, linear decoding estimates are not necessarily consistent.

D. Methods of Improvement of Nonconsistent Estimates

As in simple ADC, the principle of projection, with the option of alternating projections, can be used to improve nonconsistent estimates. The projection on $\mathcal{C}(\boldsymbol{C})$ is slightly modified since the new expression of $\mathcal{C}(\boldsymbol{C})$ derived from Proposition 4.3 is

$$\mathcal{C}(\boldsymbol{C}) = \{\boldsymbol{Y} \in \boldsymbol{R}^N / \forall k = 1, \cdots, N,$$
$$Y(k) \in q^{-1}[C(k)] + D(k), \text{ where } \boldsymbol{D} = G[\boldsymbol{C}]\}.$$

In this expression, note that $q^{-1}[C(k)] + D(k)$ is simply the quantization interval $q^{-1}[C(k)]$ translated by the real value $D(k)$. This leads to the following algorithm.

Algorithm 2

Step 1) Calculate the signal $\boldsymbol{D} = G[\boldsymbol{C}]$.

Step 2) At every instant k

i) if $\hat{X}(k) \in q^{-1}[C(k)] + D(k)$, take $\hat{X}'(k) = \hat{X}(k)$

ii) else, take $\hat{X}'(k)$ equal to the bound of the interval $q^{-1}[C(k)] + D(k)$ closest to $\hat{X}(k)$.

We propose an equivalent form to Algorithm 2, which may look more complicated but whose principle will be used for future generalization to $\Sigma\Delta$ modulation. This consists of performing a change of variable by taking $\hat{\boldsymbol{X}}$ (the estimate to be projected) as the origin of the space $\boldsymbol{R}^N$. Then, looking for the signal $\hat{\boldsymbol{X}}'$, which is the projection of $\hat{\boldsymbol{X}}$ on $\mathcal{C}(\boldsymbol{C})$ amounts to looking for the signal $\boldsymbol{Y} = \hat{\boldsymbol{X}}' - \hat{\boldsymbol{X}}$, which is the projection of the zero signal on $\mathcal{C}(\boldsymbol{C}) - \hat{\boldsymbol{X}}$. Note that $\mathcal{C}(\boldsymbol{C}) - \hat{\boldsymbol{X}} = Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$, where $\boldsymbol{D} = G[\boldsymbol{C}] - \hat{\boldsymbol{X}}$. This leads to the following algorithm:

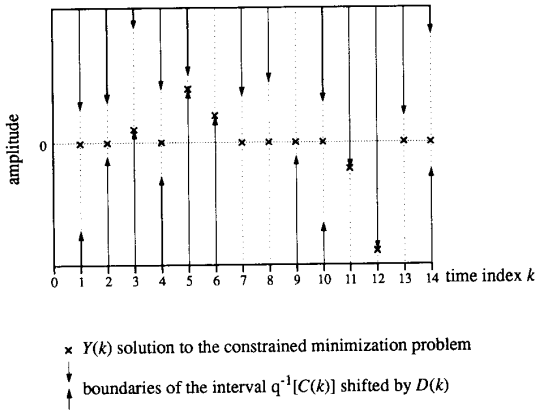

Fig. 10. Representation of the solution $\boldsymbol{Y}$ to Step 2 of Algorithm $2'$.

Algorithm 2′

Step 1) Calculate the signal $\boldsymbol{D} = G[\boldsymbol{C}] - \hat{\boldsymbol{X}}$.

Step 2) At every instant k

i) if $0 \in q^{-1}[C(k)] + D(k)$, take $Y(k) = 0$

ii) else, take $Y(k)$ equal to the bound of the interval $q^{-1}[C(k)] + D(k)$ closest to 0.

Step 3) Calculate the signal $\hat{\boldsymbol{X}}' = \boldsymbol{Y} + \hat{\boldsymbol{X}}$.

The computation of the signal $\boldsymbol{Y}$ in Step 2 has the simple graphic representation shown in Fig. 10.

E. Remarks on Predictive Encoding

The same notion of consistency and principles of decoding improvement as in simple ADC can be applied to predictive ADC, although this type of encoding includes a preprocessing transformation before quantization. The improvement of consistent decoding over linear decoding could also be studied in this case [11]. However, the main purpose of the section on predictive encoding is to prepare the study of the more important case of $\Sigma\Delta$ modulation on which we concentrate in the next section.

V. Single-Loop and Multiloop $\Sigma\Delta$ Modulation

A. Structure of Multiloop $\Sigma\Delta$ Modulation

The block diagram of an nth-order multiloop $\Sigma\Delta$ modulator, including the particular case of single-loop modulation (with $n = 1$), is shown in Fig. 11(a) [17]–[20], [3]. This is a particular case of noise-shaping encoding. As shown in Fig. 6 of [3], a noise-shaping encoder is, in general, equivalent to the composition of a linear operator H and a predictive encoder

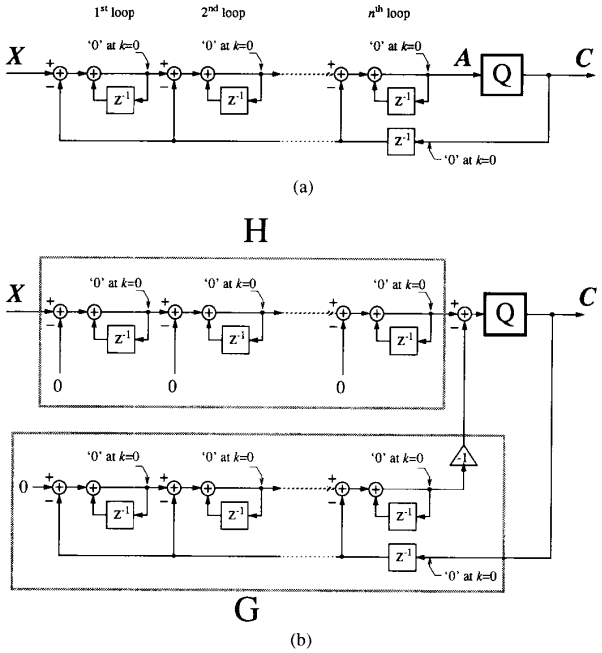

Fig. 11. Block diagrams of a multiloop $\Sigma\Delta$ modulator: (a) Original encoder (with zero initial conditions); (b) equivalent encoder.

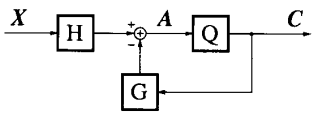

Fig. 12. General structure of a noise-shaping encoder [3].

of feedback operator G, with the particular relationship

$$H = I + G. \tag{4}$$

This is represented in the block diagram of Fig. 12. We show the details of this equivalence for multiloop $\Sigma\Delta$ modulation in Fig. 11(b). Note that every integrator node is initialized at value zero. The case where the initial condition is unknown will be considered in Section V-G. With the assumption of zero initial condition, the mapping H is an nth-order integrator, which is linear and invertible. The inverse H^{-1} is an nth-order discrete time differentiator. One can also check that G is strictly causal and that (4) is satisfied.

B. Deterministic Description of Multiloop $\Sigma\Delta$ Modulation Encoding

Using the properties of predictive encoders derived in Section IV, we immediately have the following propositions.

Proposition 5.1: The noise-shaping encoder of Fig. 12 has the structure of Fig. 7 where the preprocessing one-to-one mapping is

$$F = (I + GQ)^{-1}H. \tag{5}$$

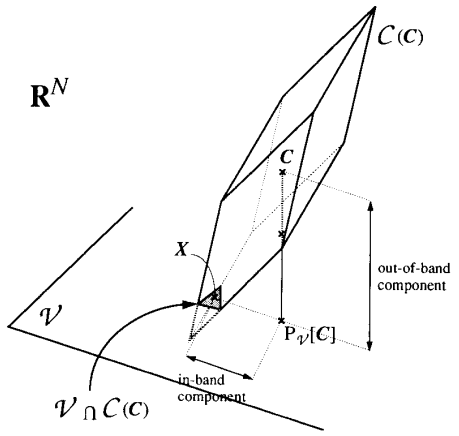

Fig. 13. Geometric representation of the information "$X \in \mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$" and nonconsistency of the linear decoding estimate $P_v[\boldsymbol{C}]$ in $\Sigma\Delta$ modulation.

Proposition 5.2: The set of signals producing a quantized signal $\boldsymbol{C}$ through the noise-shaping encoder of Fig. 12 is

$$\mathcal{C}(\boldsymbol{C}) = H^{-1}[Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]]. \tag{6}$$

Since $Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]$ is in general a rectangular hyperparallelepiped of $\boldsymbol{R}^N$, and H^{-1} a linear mapping, then $\mathcal{C}(\boldsymbol{C})$, although not necessarily rectangular, is still a hyperparallelepiped in $\boldsymbol{R}^N$. Proposition 3.1 relative to the encoded information in the oversampling situation is still applicable, but the representation of this information is modified as shown in Fig. 13. The deformation of $\mathcal{C}(\boldsymbol{C})$ is due to the mapping H^{-1}.

C. Consistent Estimates and Nonconsistency of Linear Decoding

We again use the notion of consistent estimates with its associated property (Property 3.5) and show once again that linear decoding is not necessarily consistent. We recall that in noise-shaping encoding, linear decoding consists of low-pass filtering the quantized signal $\boldsymbol{C}$. Once again, one would guess that $\boldsymbol{C}$ is the center of $\mathcal{C}(\boldsymbol{C})$. We have the following proposition.

Proposition 5.3: In noise-shaping encoding, $\boldsymbol{C}$ is the center of $\mathcal{C}(\boldsymbol{C})$.

Proof: An element $\boldsymbol{C}'$ is the center of the parallelepiped $H^{-1}[Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]]$ if and only if $H[\boldsymbol{C}']$ is the center of the parallelepiped $Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]$. This is because H is linear. However, the center of $Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]$ is simply $\boldsymbol{C} + G[\boldsymbol{C}]$. Then, using (4), we have $\boldsymbol{C}' = H^{-1}[\boldsymbol{C} + G[\boldsymbol{C}]] = H^{-1}(I + G)[\boldsymbol{C}] = H^{-1}H[\boldsymbol{C}] = \boldsymbol{C}$. □

Then as in simple ADC, linear decoding consists of performing the orthogonal projection of the center of $\mathcal{C}(\boldsymbol{C})$ on $\mathcal{V}$. Therefore, as can be seen in Fig. 13, linear decoding estimates are not necessarily consistent.

D. Methods of Improvement of Nonconsistent Estimates

We first describe a general scheme for the design of algorithms for the projection on $\mathcal{C}(\boldsymbol{C})$. Then, we will separately

consider the case of single-loop modulation and the case of higher order modulation. The formulation of the scheme starts with the same idea of change of variable proposed in Algorithm 2′ for predictive encoding, which will prove to be particularly convenient in the present case. The difference here is that we include the invertible mapping H in the change of variable. From Lemma 3.3, the projection of $\hat{\boldsymbol{X}}$ on $\mathcal{C}(\boldsymbol{C})$ is the signal $\hat{\boldsymbol{X}}'$, which minimizes $||\hat{\boldsymbol{X}}' - \hat{\boldsymbol{X}}||$ subject to the constraint $\hat{\boldsymbol{X}}' \in \mathcal{C}(\boldsymbol{C})$. By the change of variable $\boldsymbol{Y} = H[\hat{\boldsymbol{X}}' - \hat{\boldsymbol{X}}]$, this amounts to finding the signal $\boldsymbol{Y}$, which minimizes the functional $||H^{-1}[\boldsymbol{Y}]||$ subject to the constraint $\boldsymbol{Y} \in H[\mathcal{C}(\boldsymbol{C}) - \hat{\boldsymbol{X}}]$. Using the expression of $\mathcal{C}(\boldsymbol{C})$ from Proposition 5.2, we simply have $H[\mathcal{C}(\boldsymbol{C}) - \hat{\boldsymbol{X}}] = Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$, where $\boldsymbol{D} = G[\boldsymbol{C}] - H[\hat{\boldsymbol{X}}]$.

In addition, minimizing $||H^{-1}[\boldsymbol{Y}]||$ is minimizing $\phi(\boldsymbol{Y})$ where

$$\phi(\boldsymbol{Y}) = \frac{1}{2}||H^{-1}[\boldsymbol{Y}]||^2.$$

The projection algorithm is therefore as follows.

Algorithm 3

Step 1) Calculate the signal $\boldsymbol{D} = G[\boldsymbol{C}] - H[\hat{\boldsymbol{X}}]$.

Step 2) Find the minimum $\boldsymbol{Y}$ of ϕ subject to $\boldsymbol{Y} \in Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$.

Step 3) Calculate the signal $\hat{\boldsymbol{X}}' = H^{-1}[\boldsymbol{Y}] + \hat{\boldsymbol{X}}$.

Step 3 is essentially the computation of an nth-order discrete-time derivative. Step 2 is a problem of minimization of a quadratic functional under convex constraints. From the theory of convex analysis [21], there exists a characterization for the minimum in such a problem. Since the gradient of ϕ is involved in this characterization, we introduce the following notation.

Notation 5.4: ∇_ϕ denotes the mapping of $\boldsymbol{R}^N$ such that for any $\boldsymbol{Y} \in \boldsymbol{R}^N, \boldsymbol{Z} = \nabla_\phi[\boldsymbol{Y}]$ is the signal whose kth value $Z(k)$ is the partial derivative of $\phi(\boldsymbol{Y})$ with respect to $Y(k)$ or $Z(k) = (\partial\phi(\boldsymbol{Y})/\partial Y(k))$.

Then, the characterization of the minimum is given by the following lemma.

Lemma 5.5 [21]: A quadratic functional ϕ has a unique minimum in a convex set $\mathcal{S}$. It is the signal $\boldsymbol{Y}$ such that

$$\begin{aligned} &\boldsymbol{Y} \in \mathcal{S} \\ &\text{and} \quad \forall \boldsymbol{Y}' \in \mathcal{S}, \sum_{k=1}^{N} Z(k) \cdot (Y'(k) - Y(k)) \geq 0, \\ &\text{where} \quad \boldsymbol{Z} = \nabla_\phi[\boldsymbol{Y}]. \end{aligned} \tag{7}$$

1) Case of Single-Loop $\Sigma\Delta$ Modulation: In single-loop $\Sigma\Delta$ modulation, H is a first-order integrator. In this case, the expression of ∇_ϕ is given by the following property.

Property 5.6: For any $\boldsymbol{Y} \in \boldsymbol{R}^N$, if $\boldsymbol{Z} = \nabla_\phi[\boldsymbol{Y}]$, then

$$\begin{aligned} &\forall k = 1, \cdots, N, \; Z(k) \\ &\quad = -\{[Y(k+1) - Y(k)] - [Y(k) - Y(k-1)]\} \end{aligned} \tag{8}$$

with the convention $Y(0) = Y(N+1) = 0$.

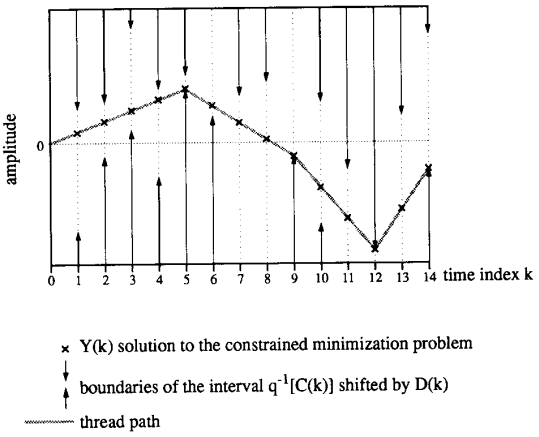

Fig. 14. Representation of the solution $\boldsymbol{Y}$ to the "Thread algorithm."

Proof: Since H^{-1} is the first-order discrete-time differentiator, then $\phi(\boldsymbol{Y}) = (1/2)||H^{-1}[\boldsymbol{Y}]||^2 = (1/2)\sum_{j=1}^{N} |Y(j) - Y(j-1)|^2$. For $k = 2, \cdots, N-1$

$$\begin{aligned} Z(k) &= \frac{\partial \phi(\boldsymbol{Y})}{\partial Y(k)} \\ &= [Y(k) - Y(k-1)] - [Y(k+1) - Y(k)]. \end{aligned}$$

One can check that $Z(1) = Y(1) - [Y(2) - Y(1)]$ and $Z(N) = [Y(N) - Y(N-1)] - Y(N)$. All these cases are summarized in (8). □

Then, when $\boldsymbol{Z} = \nabla_\phi[\boldsymbol{Y}], Z(k)$ is minus the change of slope of $\boldsymbol{Y}$ about time index k. Applying Lemma 5.5 on $\mathcal{S} = Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$ and using this property, we derive an algorithm for the computation of Step 2 of Algorithm 3. This algorithm was first introduced in [7]. We present it in a "physical" manner.

"Thread algorithm"

i) Represent graphically in the time domain the set of constraint $Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$ as sequence of the quantization intervals $q^{-1}[C(k)]$ translated by $D(k)$ (see Fig. 14).

ii) Attach a "thread" at node $(k = 0, Y(0) = 0)$, and pull it taut with a horizontal force between the constraints defined by these intervals (arrows in Fig. 14).

iii) For $k = 1, \cdots, N$, take $Y(k), k \geq 1$ on the path of the resulting thread position (see Fig. 14).

Let us show that the signal $\boldsymbol{Y}$ thus obtained is the minimum of ϕ subject to $Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$. It is sufficient to prove that $\boldsymbol{Y}$ satisfies the criterion (7). First, it is obvious that $\boldsymbol{Y} \in Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$. Then, whenever $\boldsymbol{Z} = \nabla_\phi[\boldsymbol{Y}]$ is nonzero or a change of slope occurs about time k, the thread touches a constraint (in Fig. 14, at $k = 5, 9, 12, 14$). However, the change of slope is always opposite to the direction from the constraint. For example, in Fig. 14, the change of slope is negative only when the arrow of contact goes up $(k = 5, 9)$ and vice-versa $(k = 12)$. When an arrow of contact goes up, for example at $k = 5$, then for any other admissible signal $\boldsymbol{Y}' \in Q^{-1}[\boldsymbol{C}] + \boldsymbol{D}$, we necessarily have $Y'(k) - Y(k) \geq 0$. Similarly, if an arrow of contact goes down at $k(k = 12)$, then $Y'(k) - Y(k) \leq 0$. Since $\boldsymbol{Z}$ gives the opposite of the slope's

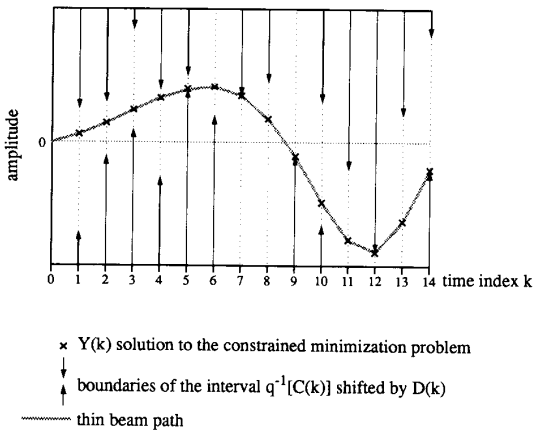

Fig. 15. Representation of the solution $\boldsymbol{Y}$ to Step 2 of Algorithm 3 for second-order $\Sigma\Delta$ modulation.

variation, we have just shown that whenever $Z(k) \neq 0$, then $Z(k)$ and $Y'(k) - Y(k)$ have the same sign. This implies (7).

Although it is presented in a physical manner, this algorithm can be easily translated in terms of computer operations.

2) Case of n-Loop $\Sigma\Delta$ Modulation with $n \geq 2$: The projection on $\mathcal{C}(\boldsymbol{C})$ in the general case of order $n \geq 2$ is quite involved. In this section, we will only sketch the main ideas used for the design of the algorithm. More detailed explanations can be found in [9] and [11].

With the single-loop configuration, the "thread algorithm" gave the signal $\boldsymbol{Y}$, which meets the constraints and, at the same time, minimizes the mean squared "slope." For an n -loop configuration in general, the functional ϕ to minimize is the nth-order discrete-time derivative. For example, when $n = 2$, we will have to minimize the "curvature" of the signal. This is similar to the problem of constrained thin flexible beam in the field of mechanics. Fig. 15 shows the solution minimizing the "curvature" under the same constraints as in Fig. 14.

Unfortunately, in the nth-order case, performing the direct projection on $\mathcal{C}(\boldsymbol{C})$ in one step as in the single-loop case is not possible. Iterative algorithms from the theory of nonlinear programming [22] could be used. However, the aim is to find a single transformation based on the knowledge of $\boldsymbol{C}$, which leads to a necessary improvement of the estimate. This would give us the freedom to use it once or to alternate it with the projection on $\mathcal{V}$.

The algorithm we designed for the nth-order case is able to perform a projection on a convex set which is different from $\mathcal{C}(\boldsymbol{C})$ but satisfies the following two properties:

i) It includes $\mathcal{C}(\boldsymbol{C})$.

ii) It does not contain the estimate $\hat{\boldsymbol{X}}$.

Therefore, according to Lemma 3.4, projecting $\hat{\boldsymbol{X}}$ on such a convex set necessarily ensures the improvement of $\hat{\boldsymbol{X}}$ as estimate of $\boldsymbol{X}$ since $\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C})$ is an element of this convex set. Instead of projecting $\hat{\boldsymbol{X}}$ on $\mathcal{C}(\boldsymbol{C}) = H^{-1}[Q^{-1}[\boldsymbol{C}] + G[\boldsymbol{C}]]$, the idea is to project $\hat{\boldsymbol{X}}$ on the larger set

$$\mathcal{C}_K(\boldsymbol{C}) = H^{-1}[Q^{-1}[\mathcal{S}_K(\boldsymbol{C})] + G[\boldsymbol{C}]]$$

where $\mathcal{S}_K(\boldsymbol{C})$ is the set of signals that coincide with $\boldsymbol{C}$ on a certain subset of time indices $K \subset \{1, \cdots, N\}$. More precisely,

$$\mathcal{S}_K(\boldsymbol{C}) = \{\boldsymbol{C}' \in \boldsymbol{R}^N / \forall k \in K, C'(k) = C(k)\}.$$

It is easy to show that $\mathcal{C}_K(\boldsymbol{C})$ is a convex set that includes $\mathcal{C}(\boldsymbol{C})$. This is true for any choice of $K \subset \{1, \cdots, N\}$. For a given K, only Step 2 of Algorithm 3 has to be modified, where the minimization of $\phi(\boldsymbol{Y})$ is subject to the new constraint $\boldsymbol{Y} \in Q^{-1}[\mathcal{S}_K(\boldsymbol{C})] + \boldsymbol{D}$. We designed an algorithm which solves Step 2, where the subset K is progressively constructed in time and such that $\hat{\boldsymbol{X}} \notin \mathcal{C}_K(\boldsymbol{C})$. The improvement achieved by this algorithm has been numerically evaluated. The results are presented in Section V-F.

E. Analytical Comparison Between Linear Decoding and Consistent Decoding

Similarly to the simple ADC case, to measure the improvement achieved by applying alternating projections on the linear decoding estimate, we also propose to evaluate an upper bound to the MSE of any consistent estimate of an input signal $\boldsymbol{X}$ of $\mathcal{V}$. In the context of $\Sigma\Delta$ modulation, this analysis is quite difficult. Failing to describe the real behavior of consistent estimate analytically, a first approach is at least to derive the MSE behavior of a consistent estimates when making some simplifying assumptions.

When quantization is uniform with quantization step size Δ and when an input signal $\boldsymbol{X}$ is fed into the multiloop encoder of Fig. 11, the signal $\boldsymbol{C} - \boldsymbol{A} = (Q - I)F[\boldsymbol{X}]$, which is commonly called the quantization error signal, is an element of the subset $[-(\Delta/2), (\Delta/2)]^N$ of $\boldsymbol{R}^N$ (we assume that $\boldsymbol{X}$ is chosen so that the quantizer is not overloaded). As a theoretical test, we considered in [9], [4], and [11] what MSE would be obtained if, for a certain domain of input signals $\boldsymbol{X} \in \mathcal{D} \subset \mathcal{V}, (Q - I)F[\boldsymbol{X}]$ was assumed to have a uniform density in the subset $[-(\Delta/2), (\Delta/2)]^N$. It was found that the MSE of consistent estimates averaged over the input signals $\boldsymbol{X} \in \mathcal{D}$ is upper bounded by (β_0/N^{2n+2}), where N is assumed high enough, and β_0 is a constant independent of N. This implies the MSE upper bound (α_0/R^{2n+2}), where $\alpha_0 = \beta_0/(2M+1)^{2n+2}$. Although this result is based on an ideal assumption that is not necessarily verified in reality, it gives a good prediction of the results of numerical tests we performed.

F. Numerical Evaluation of the Performance of Consistent Reconstruction Using Alternating Projections

Experiments similar to the case of simple ADC were performed on single-bit single-loop, two-bit double-loop, and three-bit triple-loop $\Sigma\Delta$ modulation. The quantizers were uniform with step size Δ. Input signals were chosen to have $2M+1 = 7$ as bandwidth and were constrained to have $(\Delta/2)$ as maximum amplitude. The oversampling ratios were chosen between approximately 18 and 585. For each experiment, the decoding MSE was averaged over 600 randomly generated input signals. In Fig. 16(a), we show the comparison between the linear decoding and alternating projection schemes. The

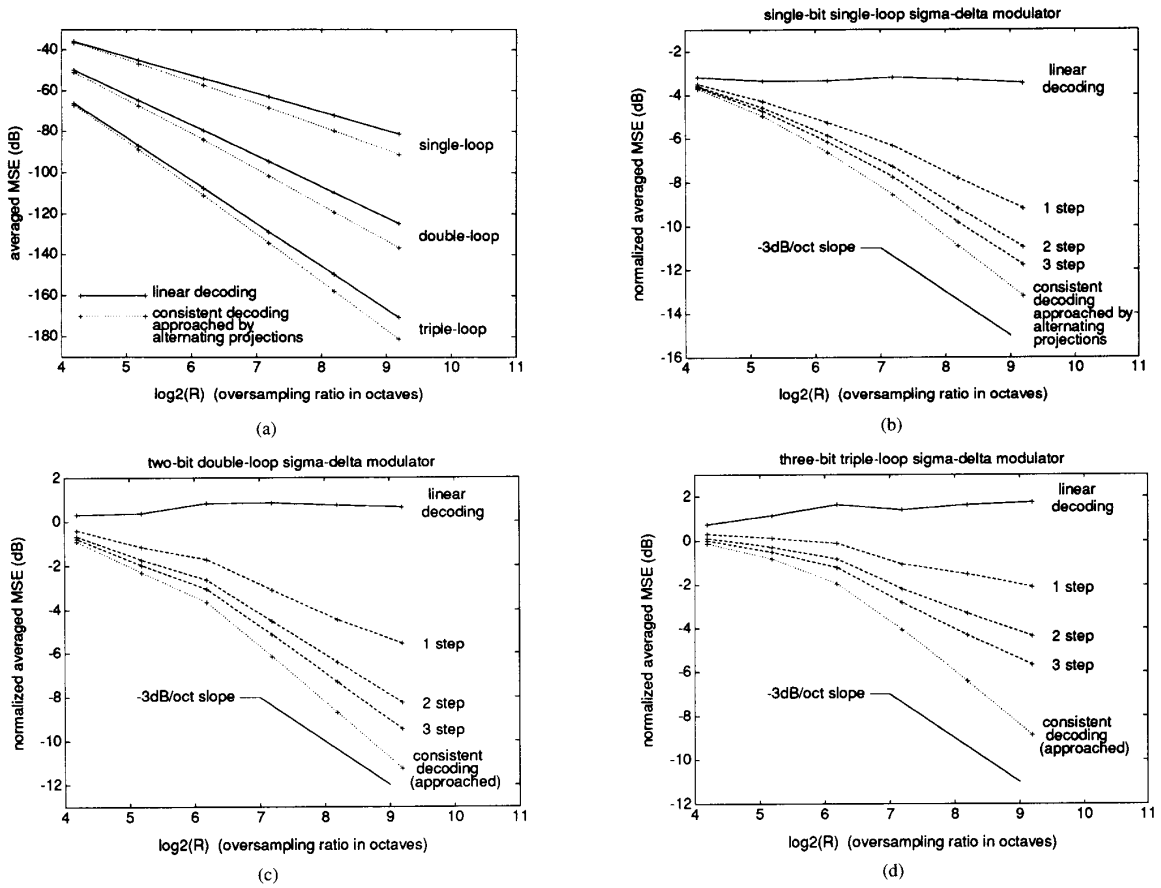

Fig. 16. Dependence of the decoding MSE with the oversampling ratio for single-bit single-loop, two-bit double-loop, and three-bit triple-loop $\Sigma\Delta$ modulation: (a) Comparison between linear decoding and consistent decoding approached by alternating projections; (b), (c), (d) decoding MSE normalized with the theoretical noise shaping equation (9) for linear decoding, consistent decoding approached by alternating projections, and finite step alternating projections: (b) Single-bit single-loop configuration; (c) 2-b double-loop configuration; (d) 3-b triple-loop configuration.

projection iteration was stopped as soon as the MSE decrement per step was less than 0.1 dB. Each step is a projection on $\mathcal{C}(\boldsymbol{C})$ followed by a projection on $\mathcal{V}$. This figure shows that the linear decoding scheme yields an MSE slope of $-(2n+1)\times 3$ dB/octave and that the alternating projection scheme increasingly improves this MSE reduction with the oversampling ratio, regardless of the encoder's order.

Fig. 16(b)–(d) shows this improvement separately for the three types of encoders. To emphasize the MSE slope tendency, we plotted the difference (normalized MSE) between the measured MSE and the theoretical linear decoding MSE given from [3] by

$$\text{MSE (linear decoding)} = \frac{\pi^{2n}}{2n+1}\cdot\frac{\sigma_\Delta^2}{R^{2n+1}},$$
$$\text{where } \sigma_\Delta^2 = \frac{\Delta^2}{12}. \tag{9}$$

One can see that when R is high enough, the MSE obtained from the alternating projection scheme decreases with R faster than the linear decoding MSE by approximately 3 dB/octave. This means that the global asymptotic MSE slope achieved by the alternating projection scheme is $-(2n+2)\times 3$ dB/octave. This confirms the improvement of the asymptotic behavior from $\mathcal{O}(R^{-(2n+1)})$ to $\mathcal{O}(R^{-(2n+2)})$ for consistent reconstruction, which is discussed in the previous section. Fig. 16(b)–(d) also show that a substantial fraction of the total improvement obtained by "infinite" alternating projection is achieved with very limited numbers of iteration (one to three steps).

G. Unknown Initial Conditions

In this section, we show that the control of the initial conditions of a multiloop modulator is not essential for the deterministic approach of signal reconstruction. Suppose that the initial condition is no longer zero, but $A_0^{(0)},\cdots,A_0^{(n-1)},C_0$ as shown in Fig. 17. Then, it can be verified that feeding this encoder with $\boldsymbol{X}$ is equivalent to feeding the zero initial condition encoder of Fig. 11 with $\boldsymbol{X}+\boldsymbol{I}$, where $\boldsymbol{I}$ is the signal defined as follows: $\boldsymbol{I}$ is zero everywhere except at the n first instants $k=1,\cdots,n$ where $I(k)=\Sigma_{i=1}^{n-k+1}(-1)^{k-1}\binom{n-i}{k-1}(A_0^{(i-1)}-C_0)$. We call $\boldsymbol{I}$ the initial

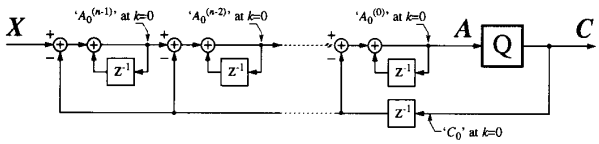

Fig. 17. Multiloop $\Sigma\Delta$ modulator with nonzero initial conditions.

condition signal. This means that, rigorously speaking, we no longer have $\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C})$, but $\boldsymbol{X} + \boldsymbol{I} \in \mathcal{C}(\boldsymbol{C})$ or $\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C}) - \boldsymbol{I}$. In practice, n is almost zero compared with the total number of samples N (typically $n \leq 4$). Therefore $\boldsymbol{I}$ is negligible and practically $\boldsymbol{X} \in \mathcal{C}(\boldsymbol{C})$.

At first, it might appear that unknown initial conditions make the use of the digital output sequence $\boldsymbol{C}$ impractical since they imply digital errors amplified by the feedback and the internal integrator. Indeed, the output sequence $\boldsymbol{C}$ may be digitally completely different from that obtained in the case of zero initial conditions. However, the important result is that the input $\boldsymbol{X}$ is close to $\mathcal{C}(\boldsymbol{C})$ in the MSE sense.

VI. Conclusion

The deterministic approach of ADC consists in saying that the exact information contained in the quantized version $\boldsymbol{C}$ of a discrete-time continuous amplitude signal $\boldsymbol{X}$ is that $\boldsymbol{X}$ belongs to the set $\mathcal{C}(\boldsymbol{C})$ of all possible signals producing $\boldsymbol{C}$ through the encoder. In the oversampled situation, we have the more restrictive information that $\boldsymbol{X}$ belongs to the set $\mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$, which is called the set of consistent estimates, where $\mathcal{V}$ is the space of bandlimited signals. When signals are sampled N times, we show that $\mathcal{C}(\boldsymbol{C})$ is a hyperparallelepiped of $\boldsymbol{R}^N$, regardless of the type of encoder (simple, predictive, single- or multiloop encoder). This implies the convexity of $\mathcal{C}(\boldsymbol{C}) \cap \mathcal{V}$ and the fact that any nonconsistent estimate can be necessarily improved. We point out that the classical linear decoding scheme consists of performing the orthogonal projection of the center of $\mathcal{C}(\boldsymbol{C})$ on $\mathcal{V}$, and, thus is not necessarily consistent. A basic method of improving a nonconsistent estimate $\hat{\boldsymbol{X}}$ consists of projecting it on any convex set that contains the original signal $\boldsymbol{X}$ but not $\hat{\boldsymbol{X}}$. We detailed, in particular, algorithms for the projection on $\mathcal{C}(\boldsymbol{C})$. One purpose of these algorithms is to show concretely how an estimate $\hat{\boldsymbol{X}}$ can be made closer to $\boldsymbol{X}$ using the knowledge of $\boldsymbol{C}$ although $\boldsymbol{X}$ is unknown. These algorithms are quite straightforward in simple and predictive ADC and relatively practical in single-loop $\Sigma\Delta$ modulation. For higher order encoding, where the direct projection on $\mathcal{C}(\boldsymbol{C})$ is difficult to perform, it is still possible to build from the knowledge of $\boldsymbol{C}$ convex sets containing $\mathcal{C}(\boldsymbol{C})$ on which the projection becomes feasible. These algorithms were used to show achievable improvements from finite iteration schemes. They were also used to approach consistent estimates from the principle of alternating projections. The numerical results obtained in the case of bandlimited and periodic signals show that consistent estimates yield an MSE whose reduction with the oversampling ratio is faster than in linear decoding by 3 dB per octave for simple ADC, single- and multiloop $\Sigma\Delta$ modulation (under certain conditions in simple ADC). This implies that the asymptotic order of the decoding MSE is $\mathcal{O}(R^{-(2n+2)})$ instead of $\mathcal{O}(R^{-(2n+1)})$ in linear decoding, where R is the oversampling ratio and n the order of the modulator.

With this deterministic approach, another contribution of this paper is to show a nonclassical description of predictive encoding and noise-shaping encoding (in the example of single- and multiloop $\Sigma\Delta$ modulation). Although these encoders were designed using statistical tools, they satisfy the deterministic interpretation of quantization, which are natural in simple ADC, including the fact that the quantized signal $\boldsymbol{C}$ is the geometric center of the set $\mathcal{C}(\boldsymbol{C})$.

References

[1] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1220–1244, Nov. 1990.

[2] 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. Signal Processing*, vol. 42, no. 1, pp. 200–203, Jan. 1994.

[3] 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.

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

[5] N. E. Hurt, *Phase Retrieval and Zero Crossings*. Boston: Kluwer, 1989.

[6] A. Zakhor, "Reconstruction of multidimensional signals from multiple level threshold crossings," Ph.D. Dissertation, Dept. of Elect. Eng., Mass. Inst. Technol., 1987.

[7] N. T. Thao and M. Vetterli, "Oversampled A/D conversion using alternate projections," in *Proc. Conf. Inform. Sci. Syst.*, (Johns Hopkins University), Mar. 1991, pp. 241–248.

[8] S. Hein and A. Zakhor, "Reconstruction of oversampled band-limited signals from $\Sigma\Delta$ encoded binary sequences," in *Proc. 25th Asilomar Conf. Signals Syst.*, (Pacific Grove, CA), Nov. 1991, pp. 866–872.

[9] N. T. Thao and M. Vetterli, "Convex coders and oversampled A/D conversion: Theory and algorithms," Tech. Rep. CTR, Columbia Univ., Dec. 1991.

[10] S. Hein and A. Zakhor, "Reconstruction of oversampled band-limited signals from $\Sigma\Delta$ encoded binary sequences," *IEEE Trans. Signal Processing*, to be published.

[11] T. T. Nguyen, "Deterministic analysis of oversampled A/D conversion and $\Sigma\Delta$ modulation, and decoding improvements using consistent estimates," Ph.D. dissertation, Dept. of Elect. Eng., Columbia Univ., Feb. 1993.

[12] S. Hein and A. Zakhor, "Theoretical and numerical aspects of an SVD-based method for band-limiting finite extent sequences," *IEEE Trans. Signal Processing*, to be published.

[13] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[14] L. M. Bregman, "The method of successive projection for finding a common point of convex sets," *Soviet Math. Doklady*, vol. 6, no. 3, pp. 688–692, May 1965.

[15] D. C. Youla and H. Webb, "Image restoration by the method of convex projections: Part 1–Theory," *IEEE Trans. Medical Imaging*, vol. 1, no. 2, pp. 81–94, Oct. 1982.

[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston: Kluwer, 1992.

[17] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proc. IEEE*, pp. 1524–1533, Nov. 1963.

[18] J. C. Candy, "A use of limit cycle oscillations to obtain robust analog-to-digital converters," *IEEE Trans. Commun.*, vol. COM-22, pp. 298–305, Mar. 1974.

[19] G. R. Ritchie, "Higher order interpolation analog to digital converters," Ph.D. dissertation, Univ. of Penna., 1977.

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

[21] I. Ekeland and R. Teman, *Convex Analysis and Variational Problems*. Amsterdam: North-Holland, 1976.

[22] D. G. Luenberger, *Linear and Nonlinear Programming*. New York: Wiley, 1984.

Nguyen T. Thao received the engineering degrees from Ecole Polytechnique, France, in 1984 and Ecole Nationale Supérieure des Télécommunication, France, in 1985, as well as the M.S. degree in electrical engineering from Princeton University in 1986 and the Ph.D. degree in electrical engineering from Columbia University in March 1993. He worked at the Center for Telecommunications Research (CTR) of Columbia University as a Research Assistant while he was a Ph.D. candidate and as a Postdoctural Fellow until September 1993. He then joined the Department of Electrical and Electronic Engineering as a lecturer.

After specializing in semiconductor physics and devices for his M.S. degree, he joined the GaAs development group of Thomoson-CSF, France, from 1986 to 1989. There, he participated to the development of an advanced GaAs MESFET analog circuit technology and the design of ultrafast GaAs A/D converters (1 GHz). At the CTR, he started focusing on the study of signal reconstruction in oversampled A/D conversion, including theoretical and practical aspects. His research interests also include signal analysis and wavelets, and signal compression.

Martin Vetterli (S'86–M'86–SM'90) was born in Switzerland in 1957. He received the Dipl. El.-Ing. degree from the Eidgenössische Technische Hochschule Zürich, Switzerland, in 1981, the master of science degree from Stanford University, Stanford, CA, in 1982, and the Doctoratés Science degree from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, in 1986.

In 1982, he was a Research Assistant at Stanford University, and from 1983 to 1986, he was a Researcher at the Ecole Polytechnique. He has worked for Siemens and AT&T Bell Laboratories. In 1986, he joined Columbia University New York, NY, where he was Associate Professor of Electrical Engineering, a member of the Center for Telecommunications Research, and Codirector of the Image and Advanced Television Laboratory. He is currently with the Department of Electrical Engineeing and Computer Science, University of California, Berkeley, CA. His research interests include multirate signal processing, wavelets, computational complexity, signal processing for telecommunications, and digital video processing.

Dr. Vetterli is a member of SIAM and ACM, a member of the MDSP Committee of the IEEE Signal Processing Society, and of the editorial boards of *Signal Processing, Image Communication, and Annals of Telecommunications*. He received the Best Paper Award of EURASIP in 1984 for his paper on multidimensional subband coding and the Research Prize of the Brown Bovery Corporation, Switzerland, in 1986 for his thesis and the IEEE Signal Processing Society's 1991 Senior Award (DSP Technical Area) for a 1989 transactions paper with D. LeGall on FIR filter banks.