SVD AND SIGNAL PROCESSING, III
Algorithms, Architectures and Applications
M. Moonen and B. De Moor (Editors)

CONSISTENT SIGNAL RECONSTRUCTION AND CONVEX CODING

N.T. THAO
Department of Electrical and Electronic Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong

M. VETTERLI
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720
U.S.A.

ABSTRACT. The field of signal processing has known tremendous progress with the development of digital signal processing. The first foundation of digital signal processing is due to Shannon's sampling theorem which shows that any bandlimited analog signal can be reduced to a discrete-time signal. However, digital signals assume a second digitization operation in amplitude. While this operation, called quantization, is as deterministic as time sampling, it appears from the literature that no strong theory supports its analysis. By tradition, quantization is only approximately modeled as an additive source of uniformly distributed and independent white noise.

We propose a theoretical framework which genuinely treats quantization as a deterministic process, is based on Hilbert space analysis and overcomes some of the limitations of Fourier analysis. While, by tradition, a digital signal is considered as the representation of an approximate signal (the quantized signal), we show that it is in fact the representation of a deterministic convex set of analog signals in a Hilbert space. We call the elements of the set the analog estimates consistent with the digital signal. This view leads to a new framework of signal processing which is non-linear and based on convex projections in Hilbert spaces.

This approach has already proved effective in the field of high resolution A/D conversion (oversampling, Sigma-Delta modulation), by showing that the traditional approach only extracts partial information from the digital signal (3dB of SNR are "missed" for every octave of oversampling).

The more general motivation of this paper is to show that any discretization operation, including A/D conversion but also signal compression, amounts to encoding sets of signals, that is, associating digital signals with sets of analog signals. With this view and the framework presented in this paper, directions of research for the design of new types of high resolution A/D converters and new signal compression schemes can be proposed.

KEYWORDS. A/D conversion, digital representation, oversampling, quantization, Sigma-Delta modulation, consistent estimates, convex projections, set theoretic estimation, coding.

1 INTRODUCTION

Although numbers are usually thought of real continuous numbers in theory, signal processing is nowadays mostly performed digitally. Traditionally, digital signals are considered as the encoded version of an approximated analog signal. In many cases, the approximation error is considered negligible and digital numbers are thought of quasi-continuous. However, this assumption starts to be critical in more and more emerging fields such as high resolution data conversion (oversampled A/D conversion) and signal compression.

In this paper, we ask the basic question of the exact correspondence which exists between analog signals and there encoded digital signals. This starts by reviewing the existing foundations of analog-to-digital (A/D) conversion. It is known that A/D conversion consists of two discretization operations, that is, one in time and one in amplitude. While a strong theory (Shannon's sampling theorem) describes the operation of time discretization, we will see in Section 2 that the analysis of the amplitude discretization, or quantization, is only approximate and statistical. This approach turns out to be insufficient in fields such as oversampled A/D conversion. To find out what the exact analog information contained in a digital signal is, it is necessary to have a more precise description of the whole A/D conversion chain.

In Section 3 we define a theoretical framework which permits a more precise description of A/D conversion. To do this, we go back to the basic description of an analog signal as an element of a Hilbert space (or Euclidean space in finite dimension), and we describe any signal transformation geometrically in this space, instead of using the traditional Fourier analysis which is limited to time-invariant and linear transformations. In this framework, we show that the precise meaning of a digital signal is the representation of a deterministic convex set of analog signals. The elements of the set are called the analog estimates consistent with the digital signal. Because of the convexity property, we show that, given a digital signal, a consistent estimate must be picked as a necessary condition for optimal reconstruction.

With this new interpretation, digital signal processing implies a new framework of (nonlinear) signal processing based on convex projections in Hilbert spaces and presented in Section 4.

In fact, the case of A/D conversion which is thoroughly considered in this paper is only a particular case of digitization system. The more general motivation of this paper is to

show that the basic function of any digitization system, including high resolution data acquisition systems (Section 5) but also signal compression systems, is to associate digital representations with sets of analog signals, or, to encode sets of analog signals. Not only does this view give a genuine description of their functions, but it indicates new directions of research for the design of A/D converters and signal compression systems.

2 CLASSICAL PRESENTATION OF A/D CONVERSION

The term of "digital signal processing" often designates what should be actually be called "discrete-time signal processing" [1]. Thanks to Shannon's sampling theorem, it is known that any bandlimited analog signal can be reduced to a discrete-time signal, provided that the sampling rate is larger than or equal to the Nyquist rate, that is, twice the maximum frequency of the input signal. Mathematically speaking, there exists a invertible mapping between bandlimited continuous-time signals $x(t)$ and sequences $(x_k)_{k\in\mathbf{Z}}$ such that $x_k = x(kT_s)$, provided that $\frac{1}{T_s} = f_s \geq 2f_m$, where f_m is the maximum frequency of $x(t)$. Therefore, any processing of the continuous-time signal $x(t)$ can be performed in the discrete-time domain. This constitutes the foundation of discrete-time processing.

However, digital signal processing assumes that a second discretization in amplitude, or quantization, is performed on the samples, as indicated by Figure 1. The digital output

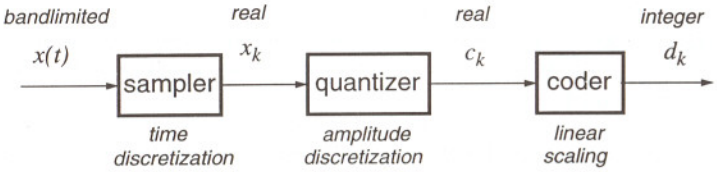


Figure 1: Analog-to-digital (A/D) conversion

sample d_k of an A/D converter is an integer representation of c_k which is a quantized version of the continuous-amplitude sample x_k. The transformation from x_k to c_k is known to introduce an error $e_k = c_k - x_k$, called the quantization error. While the time discretization process is supported by a solid theory, it appears from the literature that there only exists an approximate analysis of the quantization process. Either the quantization error is neglected and the quantization operation is considered as "transparent", or, when some close analysis is needed, it is commonly modeled as a uniformly distributed and independent white noise [2, 1]. This leads to the classical mean squared quantization error of $\frac{q^2}{12}$ where q is the quantization step size. However, this model, which is in fact only accurate under certain conditions [3, 4], does not take into account the deterministic nature of the quantization operation.

This is particularly critical when dealing with oversampled A/D conversion. Oversampling is commonly used in modern data conversion systems to increase the resolution of conversion while using coarse quantization. While the independent white noise model validity conditions become less and less valid with oversampling [4], it is still used as basic model to recover a high resolution estimate of the source signal from the oversampled and

coarsely quantized signal. With this model, a frequency analysis of the quantized signal shows in the frequency domain that only a portion of the quantization error energy lies in the input baseband region. Thus, the total energy of quantization noise can be reduced by the oversampling factor $R = \frac{f_s}{2f_m}$, by using a linear lowpass filter at cut off frequency f_m (see Figure 2). In practice, the lowpass filtering is performed digitally on the encoded

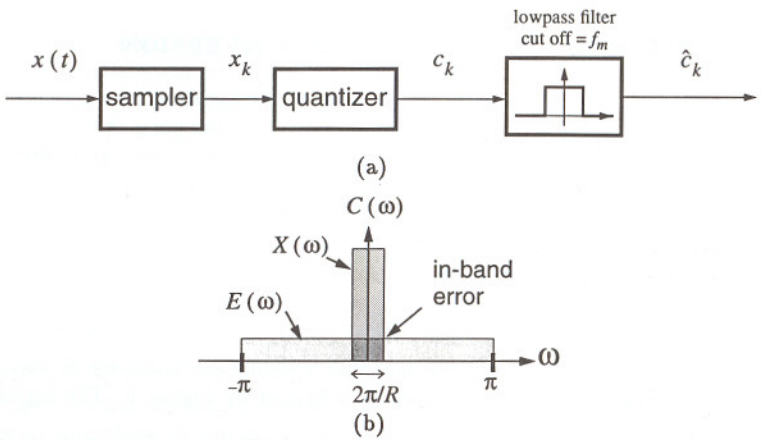


Figure 2: Oversampled A/D conversion. (a) Principle: the sampling is performed at the frequency $f_s > 2f_m$. (b) Power spectrum of the quantized signal $(c_k)_{k\in\mathbf{Z}}$ with the white quantization noise model.

version $(d_k)_{k\in\mathbf{Z}}$ of $(c_k)_{k\in\mathbf{Z}}$. On Figure 2(a), only the equivalent discrete-time operation is represented.

Although this noise reduction can be observed in practice under certain conditions, this does not tell us how much exactly we know about the analog source signal from the oversampled and quantized signal. We can already give a certain number of hints which tell us that a linear and statistical approach of the quantization process is not sufficient to give a full analysis of the signal content process.

First, it is not clear whether the in-band noise which cannot be canceled by the lowpass filter is definitely irreversible. Because quantization is a deterministic process, there does exist some correlation between the input signal and the quantization error signal, even after filtering. Second, with the linear filtering approach, it appears that the quantization mean squared error (MSE) has a non-homogeneous dependence with the time resolution and the amplitude resolution. Indeed, the MSE is divided by 4 when the amplitude resolution is multiplied by 2 (that is, q is divided by 2), whereas it is divided by 2 only when the time resolution is multiplied by 2 (that is, R is multiplied by 2). This is a little disappointing when thinking of A/D conversion as the two dimensional discretization of a continuous graph.

In fact, an example can already be given which shows by some straightforward mechanisms that the in-band noise is indeed not irreversible. Figure 3 shows a numerical example

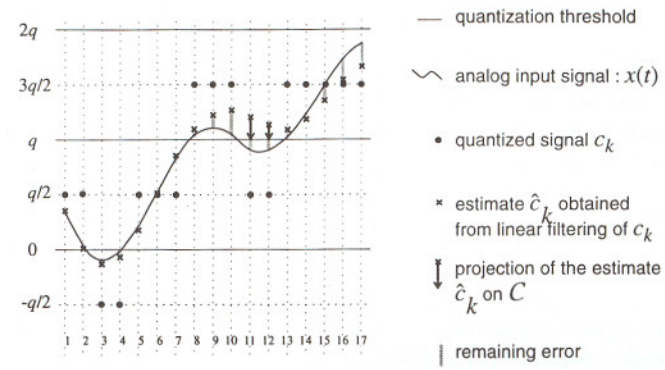


Figure 3: Example of oversampling and quantization of a bandlimited signal with reconstruction by linear filtering.

of a bandlimited signal $x(t)$, shown by a solid line, which is oversampled by 4 and quantized, giving a sequence of values $(c_k)_{k\in\mathbf{Z}}$ represented by black dots. The classical discrete-time reconstruction $(\hat{c}_k)_{k\in\mathbf{Z}}$ obtained by lowpass filtering $(c_k)_{k\in\mathbf{Z}}$ is shown by the sequence of crosses. Some error represented by grey shades can be observed between the signal reconstruction $(\hat{c}_k)_{k\in\mathbf{Z}}$ and the samples of the input signal. We know that this error forms a signal located in the frequency domain in the baseband region. However, some anomalies can be observed in the time domain. At instants 11 and 12, it can be seen that the values of $(\hat{c}_k)_{k\in\mathbf{Z}}$ are larger than q, while the given values of the quantized signal $(c_k)_{k\in\mathbf{Z}}$ tell us that the input signal's samples necessarily belong to the interval $[0, q]$. Not only is the sequence $(\hat{c}_k)_{k\in\mathbf{Z}}$ not consistent with the knowledge we actually have about the source input signal, but this knowledge also gives us a deterministic way to improve the reconstruction estimate $(\hat{c}_k)_{k\in\mathbf{Z}}$. Indeed, although we don't know where exactly the samples of the input signal are located within the interval $[0, q]$ at instants 11 and 12, we know that projecting the two respective samples of $(\hat{c}_k)_{k\in\mathbf{Z}}$ on the level q leads to a necessary reduction of the error (see Figure 3). This shows that the in-band error is not irreversible.

These hints show that a new framework of analysis is necessary.

3 NEW ANALYSIS OF A/D CONVERSION

3.1 SIGNAL ANALYSIS FRAMEWORK

The goal is to define a framework where quantization can be analyzed in a deterministic way with the given definition of an error measure.

Bandlimited analog signals are usually formalized as elements of the space $\mathcal{L}^2(\mathbf{R})$ of square summable functions, where Fourier decomposition is applicable. Thanks to Shannon's sampling theorem, the analog signals $x(t)$ bandlimited by a maximum frequency f_m can be studied as elements of the space $\mathcal{L}^2(\mathbf{Z})$ of square summable sequences, thanks to the

invertible mapping

$$x(t) \longrightarrow (x_k)_{k\in\mathbf{Z}} \in \mathcal{L}^2(\mathbf{Z}) \text{ where } x_k = x(kT_s),$$

under the condition that $f_s = \frac{1}{T_s} \geq 2f_m$. Errors between the bandlimited analog signals are measured using the canonical norm of $\mathcal{L}^2(\mathbf{R})$ and can also be evaluated in the discrete-time space $\mathcal{L}^2(\mathbf{Z})$ using its own canonical norm, thanks to the relation:

$$\frac{1}{T_s}\int_{t\in\mathbf{R}} |x(t)|^2 dt = \sum_{k\in\mathbf{Z}} |x_k|^2.$$

Unfortunately, this framework cannot be used to study quantization because the quantized version $(c_k)_{k\in\mathbf{Z}}$ of an element $(x_k)_{k\in\mathbf{Z}}$ of $\mathcal{L}^2(\mathbf{Z})$ is not necessarily an element of $\mathcal{L}^2(\mathbf{Z})$ (or, is not necessarily square summable). For example, using the quantization configuration of Figure 3, although a sequence $(x_k)_{k\in\mathbf{Z}}$ may be decaying towards 0 when k goes to infinity, its quantized version $(c_k)_{k\in\mathbf{Z}}$ never goes below $\frac{q}{2}$ in absolute value. On the other hand, while the MSE type of error measure can be applied for the analysis of quantized signals, it cannot be applied to the elements of $\mathcal{L}^2(\mathbf{Z})$, since it would systematically lead to the value 0.

Therefore, we propose to confine ourselves to another space of bandlimited signals which can be entirely defined on a finite time window $[0, T_0]$. Precisely, we assume that the sinusoidal components of the Fourier series expansion of $x(t)$ on $[0, T_0]$ are zero as soon as the corresponding frequencies are larger than f_m. This is equivalent to saying that the T_0-periodized version of $x(t)$ defined on $[0, T_0]$ is bandlimited by the maximum frequency f_m. Under this assumption, we have a finite time version of Shannon's sampling theorem. It can be easily shown that, under the condition $\frac{N-1}{T_0} \geq 2f_m$ equivalent to the Nyquist condition, there is an invertible mapping between $x(t)$ and its discrete-time version $X = (x_k)_{1\leq k\leq N} \in \mathbf{R}^N$ where $x_k = x(k\frac{T_0}{N})$ for $k = 1, ..., N$ [5, 6]. In this context, we can evaluate the error between two bandlimited input signals using the mean squared sum:

$$MSE(x(t), x'(t)) = \frac{1}{T_0}\int_{t=0}^{T_0} |x'(t) - x(t)|^2 dt.$$

This error can be in fact evaluated in the discrete-time domain using the mean squared sum

$$MSE(X, X') = \|X' - X\|^2 = \frac{1}{N}\sum_{k=1}^{N} |x'_k - x_k|^2,$$

thanks to the relation, easy to show [5, 6]:

$$\frac{1}{T_0}\int_{t=0}^{T_0} |x'(t) - x(t)|^2 dt = \frac{1}{N}\sum_{k=1}^{N} |x'_k - x_k|^2.$$

Now, the quantized version of the discrete-time signal $X = (x_k)_{1\leq k\leq N}$ is an element $C = (c_k)_{1\leq k\leq N}$ of the same space $\mathbf{R}^N$, where $c_k = \mathrm{Q}[x_k]$ for $k = 1, ..., N$, and Q is the scalar quantizer function. Note that the MSE function can be applied to any element of $\mathbf{R}^N$ whether it is a continuous-amplitude signal or a quantized signal.

3.2 QUANTIZATION ANALYSIS

The quantization operation can be deterministically defined as a mapping $\mathbf{Q}$ from $\mathbf{R}^N$ to $\mathbf{R}^N$. But unlike the sampling operation, this is a many-to-one mapping. While, under the Nyquist condition, a discrete-time signal is characteristic of a unique analog bandlimited signal, a quantized signal C is characteristic of a whole set of possible continuous-amplitude and discrete-time signals. Mathematically, this set is simply the inverse image of C through the mapping $\mathbf{Q}$, usually denoted by $\mathbf{Q}^{-1}[C] \subset \mathbf{R}^N$. If a sequence X is only known by its quantized version C, the exact knowledge about X available from C is that X belongs to the set of signals $\mathbf{Q}^{-1}[C]$. We call the elements of the set $\mathcal{C} = \mathbf{Q}^{-1}[C]$ the estimates of $\mathbf{R}^N$ consistent with the quantized signal C.

Figures 4(a) and (b) show the form of the set $\mathcal{C}$ of consistent estimates in the cases $N = 1$

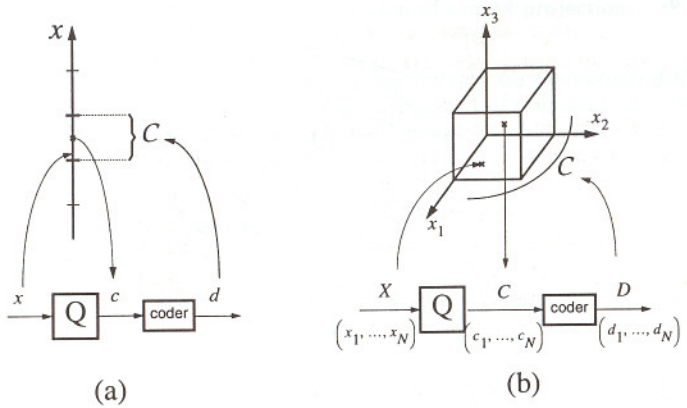


Figure 4: Quantization as a many-to-one mapping of $\mathbf{R}^N$. (a) Case $N = 1$. (b) Case $N > 1$.

and $N > 1$ respectively. In the case $N = 1$, $\mathcal{C}$ is equal to the whole quantization interval which contains the given quantized value c. Using the classical configuration of uniform quantization, the value c appears to be the particular consistent estimate located at the mid-point of the interval $\mathcal{C}$. In the traditional view point, the digital output d of a quantizer is a binary encoded version of the quantized value c. In our approach, we consider that d is a digital representation of the complete set $\mathcal{C}$. In the case $N > 1$, $\mathcal{C}$ is obviously the N dimensional cross-product of real intervals and therefore forms geometrically a hypercube of $\mathbf{R}^N$ parallel to the canonical axes. As a generalization of the case $N = 1$, the quantized signal C appears to be the particular consistent estimate located at the geometric center of $\mathcal{C}$. As in the case $N = 1$, we consider that the digital output D is the encoded version of the whole set $\mathcal{C}$, not of the signal C.

In the case of oversampled A/D conversion, the quantization operation is performed, not on any element of $\mathbf{R}^N$, but on the sampled version of bandlimited signals only. Indeed, it is easy to see from the assumption of Section 3.1 that the bandlimited signals have a finite Fourier series expansion containing not more than $2f_mT_0 + 1$ components. As a

consequence, they belong to a space of finite dimension equal to $W = \lceil 2f_m T_0 + 1 \rceil$, where $\lceil y \rceil$ designates the smallest integer greater than or equal to y. As a second consequence, their sampled version also belongs to a W dimensional space, since the sampling operation applied on bandlimited signals is a linear and invertible mapping. Because of the Nyquist rate condition $\frac{N-1}{T_0} \geq 2f_m$, note that we necessarily have $W \leq N$. Therefore, the sampled versions of the bandlimited signals belong to a W dimensional subspace $\mathcal{S}$ of $\mathbf{R}^N$. By abuse of language, we call $\mathcal{S}$ the space of bandlimited discrete-time signals. It can be shown that the dimensional ratio $\frac{N}{W}$ coincides approximately with the oversampling ratio R.

To recapitulate, in the oversampling context, the inputs to the quantizer are elements of the subspace $\mathcal{S} \subset \mathbf{R}^N$. Once $X \in \mathcal{S}$ is quantized into C, the complete knowledge which is available about X is that X belongs to the set $\mathcal{S} \cap \mathcal{C}$ where $\mathcal{C}$ =bf $\mathbf{Q}^{-1}[C]$. We will say that $\mathcal{S} \cap \mathcal{C}$ is the set of estimates consistent with C. This set is geometrically represented in Figure 5.

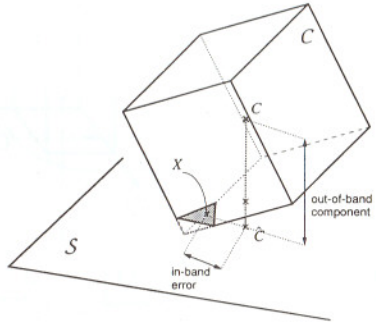


Figure 5: Geometric representation of oversampled A/D conversion.

3.3 NECESSITY FOR CONSISTENT RECONSTRUCTION

In the previous section, it was shown that when an input signal is quantized, the exact information which remains available to us is that it belongs to the set of consistent estimates. However, nothing tells us until now that we must pick a consistent estimate if we want to estimate the input signal from its quantized version. We show in this section that this is in fact the case in a certain sense.

It can be easily shown from the previous section that sets of consistent estimates are convex. We recall that $\mathcal{A}$ is a convex set if and only if for any couple of elements $X, Y \in \mathcal{A}$, the segment $[X, Y]$ is entirely included in $\mathcal{A}$. Because the considered norm $\|\cdot\|$ in $\mathbf{R}^N$ is a euclidean norm, the convexity property will appear to play an important role thanks to the following lemmas:

Lemma 3.1 *[7] Let X be an element of $\mathbf{R}^N$ and $\mathcal{A} \subset \mathbf{R}^N$ be a convex set. There exists a unique element X' of the closure $\overline{\mathcal{A}}$ of $\mathcal{A}$ such that for all $Y \in \mathcal{A}$, $\|X' - Y\| \leq \|X - Y\|$. The transformation from X to X' is then a mapping of $\mathbf{R}^N$ called the convex projection on $\mathcal{A}$.*

Lemma 3.2 *[8] If X' is the convex projection of X on a convex set $\mathcal{A} \subset \mathbf{R}^N$ and $X \notin \overline{\mathcal{A}}$, then for all $X_0 \in \mathcal{A}$, $\|X' - X_0\| < \|X - X_0\|$.*

These lemmas are illustrated by Figure 6. They lead to the following proposition.

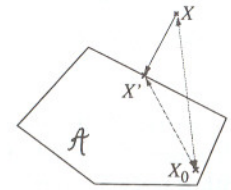


Figure 6: Geometric representation of convex projection.

Proposition 3.3 *Let $X_0 \in \mathbf{R}^N$ be the input of a quantizer, C_0 the output of the quantizer and X an element of $\mathbf{R}^N$ which does not belong to the set $\mathcal{A}$ of estimates consistent with C_0. Then, although we don't know where the input X_0 is located within the set $\mathcal{A}$, the distance between X and X_0 can be deterministically reduced[1] by a convex projection of X on $\mathcal{A}$.*

We recall that $\mathcal{A}$ is equal to $\mathbf{Q}^{-1}[C_0]$ without oversampling, and $\mathcal{S} \cap \mathbf{Q}^{-1}[C_0]$ with oversampling. In any case, the operation of convex projection on $\mathcal{A}$ is uniquely determined by the knowledge of C_0.

As a conclusion, when reconstructing a discrete-time signal from its quantized version C_0, any non-consistent estimate is by necessity non-optimal and can be deterministically improved using the knowledge of C_0.

3.4 DETERMINISTIC ANALYSIS OF OVERSAMPLED A/D CONVERSION

We have already seen from Figure 3 an example where the reconstruction estimate $\hat{C} = (\hat{c}_k)_{1 \leq k \leq N}$ proposed by the classical and linear approach of oversampled A/D conversion is not necessarily consistent with the quantized signal information and can be improved. The previous section gave the formal reason why in general a non-consistent estimate can always be improved. Now the reason why the estimate $\hat{C}$ is not necessarily consistent can be seen geometrically in the euclidean space $\mathbf{R}^N$ as shown in Figure 5. The sequence $\hat{C} = (\hat{c}_k)_{1 \leq k \leq N}$ defined as the lowpass filtered version of $C = (c_k)_{1 \leq k \leq N}$ is more precisely the bandlimited discrete-time signal which coincides in the frequency domain with $C = (c_k)_{1 \leq k \leq N}$ in the baseband region. As a consequence $\hat{C}$ is the element of the space $\mathcal{A}$ of bandlimited discrete-time signals which is closest to C in the MSE sense, or equivalently, in the sense of the euclidean norm of $\mathbf{R}^N$. According to Lemma 3.1, $\hat{C}$ is in fact the convex projection of C on $\mathcal{S}$. As shown in Figure 5, while C is the geometric center of the hypercube $\mathcal{C}$, there is no reason for its convex projection $\hat{C}$ on $\mathcal{S}$ to remain necessarily in $\mathcal{S}$, and therefore, to be consistent.

[1] The reduction of distance is strict if X does not belong to the closure of $\mathcal{A}$, according to Lemma 3.2.

The second important question is now to know what the performance yielded by consistent estimates is in terms of MSE. The following result was recently shown in [5, 9]:

Theorem 3.4 *Let $x(t)$ be a bandlimited and T_0-periodic signal which has a time density of quantization threshold crossings larger than or equal to the Nyquist rate. At the oversampling ratio $R = \frac{N}{W}$, let $X \in \mathbf{R}^N$ be the sampled version of $x(t)$, C the quantized version of X and $X' \in \mathbf{R}^N$ any estimate consistent with C. Then, there exists a constant $\alpha > 0$ which only depends on $x(t)$ and the definition of the quantizer, such that*

$$MSE(X, X') \leq \frac{\alpha}{R^2}.$$

Qualitatively speaking, this theorem implies that under a certain condition on the input's quantization threshold crossings, signals chosen in the set of consistent estimates yield an MSE which asymptotically decreases with R in $\mathcal{O}(R^{-2})$, instead of $\mathcal{O}(R^{-1})$ as it is the case with the classical linear reconstruction. This represents a faster decrease of MSE over the classical method by 3dB per octave of R. With this new result, the symmetry of the MSE dependence with the amplitude and the time resolutions is recovered.

4 CONVEX PROJECTION BASED SIGNAL PROCESSING

We have seen that a digital signal is not the representation of a single estimate, but of a complete set of estimates called the consistent estimates. Although the convex set corresponding to a given digital signal is deterministically known, the problem of using this knowledge to retrieve a consistent estimate or at least partially improve a non-consistent estimate, is not trivial. Although the space of analysis $\mathbf{R}^N$ is of finite dimension, N may be "infinitely" large compared to the finite time window of operation of the working processor. For this reason, the existing algorithms derived from the field of *non-linear and linear programming* [10] to retrieve an estimate satisfying convex constraints, may be not feasible.

More feasible algorithms may be derived from the field of *set theoretic estimation* [11] in euclidean spaces or, for the case of infinite dimension, in Hilbert spaces. The basic idea is that a convex set $\mathcal{A}$ can be often decomposed as intersection of a certain number p of convex sets $\mathcal{C}_i$, $i = 1, ..., p$ with simple structure and on which the convex projections are implementable. For example, in oversampled A/D conversion, the convex set of consistent estimates is the intersection $\mathcal{S} \cap \mathcal{C}$, where $\mathcal{S}$ and $\mathcal{C}$ are two convex sets of relatively simple structure. While it is difficult to find directly an estimate in $\mathcal{S} \cap \mathcal{C}$, the convex projections on $\mathcal{S}$ and $\mathcal{C}$ respectively are easily defined. The hypercube $\mathcal{C}$ of $\mathbf{R}^N$ can be itself seen as the intersection of $2N$ convex sets which are half-spaces of $\mathbf{R}^N$. The projection on each of these convex sets is trivial and only implies local operations in time. Figure 7 shows in general how polygonal sets can be decomposed as intersection of half-spaces.

Assuming that the set of consistent estimates has the following decomposition $\mathcal{A} = \cap_{i=1}^{p} \mathcal{C}_i$ and that the convex projection on each set $\mathcal{C}_i$ is implementable, we already have a way to partially improve any non-consistent estimate. Indeed, if $X \notin \mathcal{A}$, there exists by necessity $i \in \{1, ..., p\}$ such that $X \notin \mathcal{C}_i$. The projection of X on $\mathcal{C}_i$ will reduce the distance of X with any element of $\mathcal{C}_i$, and therefore any element of $\mathcal{A}$, since $\mathcal{A} \subset \mathcal{C}_i$. This is exactly what was performed in the example of Figure 3. After noticing that $\hat{C}$ does not belong to $\mathbf{Q}^{-1}[C]$, it

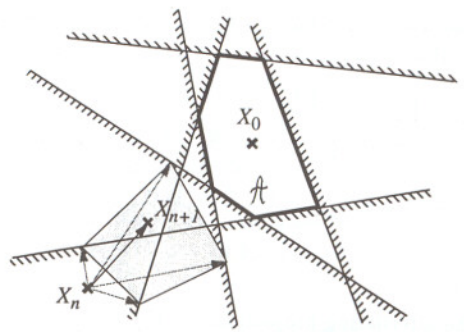


Figure 7: Decomposition of a polygon into half-spaces and representation of the parallel projection algorithm.

can be easily shown that the time domain operation indicated in Figure 3 is the projection of $\hat{C}$ on the convex set $\mathcal{C} = \mathbf{Q}^{-1}[C]$ and leads to a necessary improvement since $\mathcal{C}$ includes $\mathcal{S} \cap \mathcal{C}$. This process can be in general reiterated as long as the current estimate X does not belong to $\mathcal{A}$ (see Figure 8).

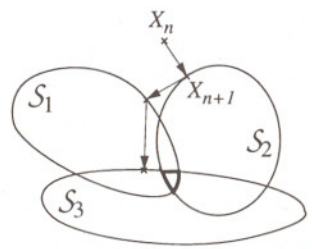


Figure 8: Geometric representation of the alternating projection algorithm.

It was in fact proved in [8] that by applying convex projections onto $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_p$ alternately and periodically, one converges to an element of the intersection[2]. We formalize this property as follows:

Theorem 4.1 *Let $\mathcal{C}_1, ..., \mathcal{C}_p$ be p convex sets in a Hilbert space $\mathcal{H}$, $\mathbf{P}_1, ..., \mathbf{P}_p$ be the convex projections on $\mathcal{C}_1, ..., \mathcal{C}_p$ respectively, and $(X_n)_{n \in \mathbf{N}}$ be a sequence in $\mathcal{H}$ such that*

$$X_{n+1} = \mathbf{P}_{n \bmod p+1}[X_n], \text{ for } n \in \mathbf{N}.$$

Then the sequence $(X_n)_{n \in \mathbf{N}}$ converges to an element of $\mathcal{A} = \cap_{i=1}^{p} \overline{\mathcal{C}_i}$ in the sense of the Hilbert norm of $\mathcal{H}$.

This is often called the algorithm of alternating projections or the POCS algorithm (Projection Onto Convex Sets). This algorithm became popular in signal processing with the work by Youla [12].

[2] Rigorously, one converges to an element of the intersection of $\overline{\mathcal{C}_1}, \overline{\mathcal{C}_2}, ..., \overline{\mathcal{C}_p}$, in the case where $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_p$ have not been specified as closed sets.

There exists a more general version of this algorithm including relaxation coefficients $(\alpha_n)_{n\in\mathbf{N}}$ and based on the following operation:

$$X_{n+1} = \alpha_n \cdot \mathbf{P}_{n \bmod p+1}[X_n] + (1-\alpha_n)\cdot X_n.$$

Note that the choice $\alpha_n = 1$ brings us back to the simple case of alternating projections. For the general case $\alpha_n \neq 1$, it is shown that this single operation reduces the distance[3] of the current estimate X_n with any element of $\mathcal{A}$. It is also shown that the infinite iteration converges to an element of $\mathcal{A}$ if there exists $\epsilon > 0$ such that $\forall n \in \mathbf{N}$, $\alpha_n \in [\epsilon, 2-\epsilon]$. In practice, the speed of convergence can be often accelerated by empirical adjustments of the coefficients α_n in $]1,2[$.

One drawback of the alternating projection algorithm is that it does not permit parallel processing. A new algorithm involving parallel projections was recently introduced by Combettes [13] and based on the following operation

$$X_{n+1} = \sum_{i\in I_n} w_{i,n}\cdot \mathbf{P}_i[X_n], \text{ where } w_{i,n} \geq 0 \text{ and } \sum_{i\in I_n} w_{i,n} = 1,$$

and where I_n is a subset of indices of $\{1,...,p\}$. Qualitatively speaking, at each step n, a certain number of sets among $\mathcal{C}_1,...,\mathcal{C}_p$ is selected (the set of the indices of the selected sets is called I_n) and the convex projections of X_n on these selected sets are applied. This forms a set of points $\{\mathbf{P}_i[X_n] \;/\; i \in I_n\}$ and X_{n+1} is chosen in the convex envelop of this set. These operations are illustrated in Figure 7. The distance of the estimate X_n with any element of $\mathcal{A}$ is shown to be reduced by this transformation [14] and the infinite iteration is proved to converge to an element of $\mathcal{A}$ under certain conditions on the sequence $(I_n)_{n\in\mathbf{N}}$ [13]. The admissible choices of $(I_n)_{n\in\mathbf{N}}$ include two particular cases:

(i) $I_n = \{1,...,p\}$: This is the case where all convex projections are performed in parallel at each step.

(ii) $I_n = \{n \bmod p+1\}$: This falls back to the case of alternating projections.

A version with relaxation coefficients is also introduced in [13] as:

$$X_{n+1} = \alpha_n \cdot \sum_{i\in I_n} w_{i,n}\mathbf{P}_i[X_n] + (1-\alpha_n)\cdot X_n.$$

The convergence to an element of $\mathcal{A}$ is shown to be guaranteed if $\exists\epsilon > 0$, $\forall n \in \mathbf{N}$, $\alpha_n \in [\epsilon, 2L_n - \epsilon]$ where

$$L_n = \frac{\sum_{i\in I_n} w_{i,n}\,\|\mathbf{P}_i[X_n] - X_n\|^2}{\left\|\sum_{i\in I_n} w_{i,n}\cdot \mathbf{P}_i[X_n] - X_n\right\|^2}.$$

5 APPLICATION TO HIGH RESOLUTION DATA CONVERSION

Although the deterministic approach was introduced on the simple version of oversampled A/D conversion in Section 3, it is also applicable to modern techniques of high resolution data conversion such as oversampled ΣΔ [15, 16]. The conversion scheme is similar to that of Figure 2, but the quantizer is replaced by a more sophisticated circuit called a ΣΔ

[3] The reduction of distance is strict when $X_n \notin \mathcal{S}_{n \bmod p+1}$ and $\alpha_n \in [0,2]$.

modulator, including an integrator, a quantizer and a feedback loop (see Figure 9). This

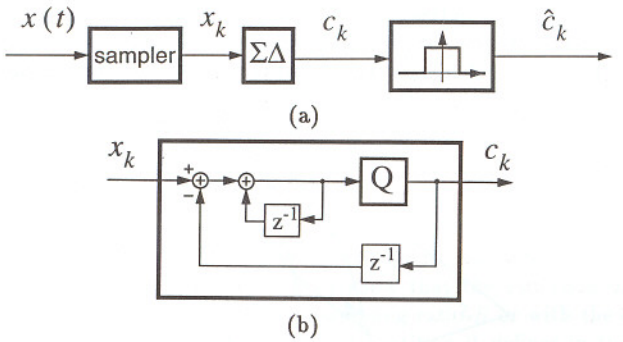


Figure 9: ΣΔ modulation. (a) Overall principle. (b) Detail of the ΣΔ modulator.

type of data conversion allows the use of very coarse quantization (down to one bit), and thus simple circuitry, while reproducing a high resolution estimate after lowpass filtering.

Although the conditions of validity of the white quantization noise model are not really applicable here [4], ΣΔ modulation is still classically analyzed using this model [16]. In this context, it is shown that a ΣΔ behaves like an additive source of independent noise whose spectrum is "shaped" as shown in Figure 10. Then, it is easy to show that the

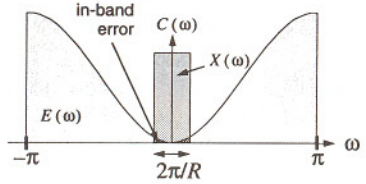


Figure 10: Power spectrum of the output of a ΣΔ modulator with the assumption of white quantization noise.

portion of noise energy contained in the baseband of the quantized signal decreases with the oversampling ratio R in R^{-3}, which represents a decrease of 9dB per octave of R. In spite of the limited validity of the assumed model, this result is observed in practice. More sophisticated architectures of ΣΔ modulation exist, which include a higher number of integrators [17]. In general, for an n^{th} order ΣΔ modulator, the noise energy remaining in the baseband of the quantized signal depends on R in $R^{-(2n+1)}$.

Now, the same kind of question as in Section 3 can be raised here. What do we know exactly about a bandlimited signal after it is oversampled and processed through a ΣΔ modulator ?

Like a single quantizer, a ΣΔ modulator can also be studied as a many-to-one mapping of $\mathbf{R}^N$. The set $\mathcal{C}$ of estimates consistent with the output of a ΣΔ modulator can be obtained

by inversion of this mapping. It is shown in [5, 6] that the set is no longer a hypercube, but a parallelepiped (the edges are no longer perpendicular). However, this is still a convex set, and it is shown that the quantized signal $C = (c_n)_{1 \le k \le N}$ is still located at its geometric center. As in Section 3, the set of consistent estimates is $\mathcal{S} \cap \mathcal{C}$. This is geometrically represented in Figure 11. Although the distance between X and $\hat{C}$, due to the in-band

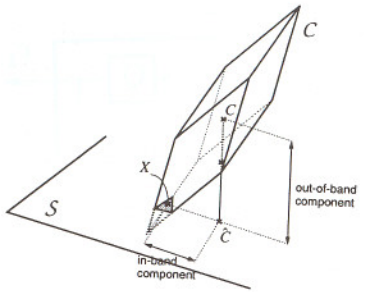


Figure 11: Geometric represetation of oversampled $\Sigma\Delta$ modulation.

error remaining in the quantized signal C, decreases with R faster than in the case of simple quantization, it appears that $\hat{C}$ is still not necessarily a consistent estimate. In fact, numerical experiments performed on bandlimited and T_0-periodic signals [5, 6] show that the MSE yielded by consistent estimates decreases in average with R in $\mathcal{O}(R^{-4})$ instead of $\mathcal{O}(R^{-3})$. In general, for an n^{th} order $\Sigma\Delta$ modulator, it was shown that the average MSE of consistent estimates behaves in $\mathcal{O}(R^{-(2n+2)})$ instead of $\mathcal{O}(R^{-(2n+1)})$, implying, as in the case of simple quantization, a faster decrease of MSE by 3dB per octave of R, regardless of the order n.

With a deterministic approach, these experiments show that the output of a $\Sigma\Delta$ modulator contains more information about the input signal than that recovered with the classical approach of A/D conversion.

6 CONCLUSION AND RELATED RESEARCH

The full meaning of a digital signal is obtained by a deterministic analysis of the digitization process as a many-to-one mapping. Thus, a digital signal is the representation of, not a single estimate, but a whole set of analog signals, called the set of consistent estimates. This set plays two roles:

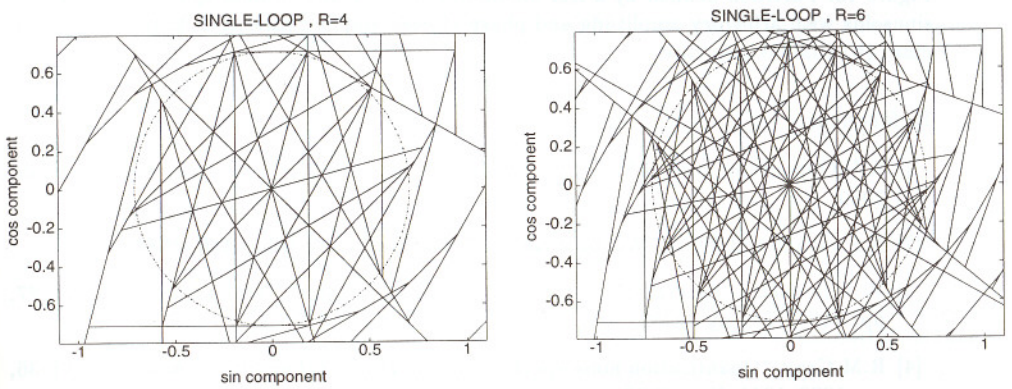
(i) it gives the exact knowledge of the possible locations of the original analog signal,

(ii) it is the set where a signal should be picked when estimating the original signal from its digital version.

The second item is due to the convexity of the set, as observed on classical quantization schemes. With this approach, not only is a more precise analysis of the A/D conversion process given, but, in the context of oversampling and $\Sigma\Delta$ modulation, it also leads to the conclusion that a digital output signal contains more analog information about the input

signal than that traditionally recovered by the classical analysis of A/D conversion. Namely, the MSE of consistent estimates decreases with the oversampling ratio R fastest than that of the classical linear reconstruction estimate by 3dB per octave. This new approach of digital signals implies a new framework of signal processing based on convex projections in Hilbert spaces, derived from the field of *set theoretic estimation.*

This past research leads to the new idea that the intrinsic function of an A/D converter is to split the space of analog input signals into convex sets, and assign a digital representation to each of them. For this reason, we say that an A/D converter is a *convex coder.* The intrinsic performance of a convex coder can be evaluated by its ability to split the input space into small sets with respect to the considered error measure. Recent research has been done to measure the intrinsic performance of an oversampled A/D converter or a $\Sigma\Delta$ modulator [18, 19, 20]. Figures 12 and 13 show that the evolution of the intrinsic performance of a $\Sigma\Delta$ modulator with the oversampling ratio R or with the order n of the modulator can be graphically observed by the set partition it defines in the input space. The intrinsic performance of the encoder can be measured by the average MSE of optimal

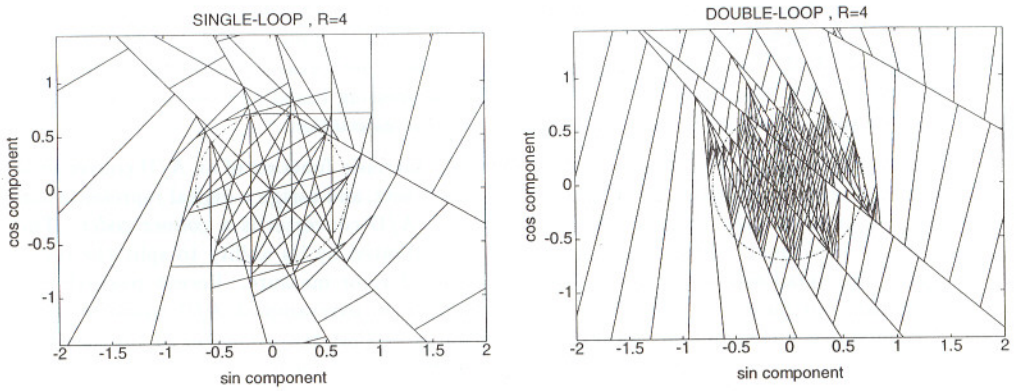


(a) (b)

Figure 12: Partition defined by a first order $\Sigma\Delta$ modulator in the 2 dimensional space of T_0- periodic sinusoids with arbitrary amplitude and phase: (a) Case of oversampling ratio $R = 4$. (b) Case $R = 6$.

reconstruction which consists of picking for each cell of the input space partition its centroid. It was shown in [19, 20] that optimal reconstruction yields the same MSE behavior in R as consistent reconstruction. This input space view can be a new direction for the design of high resolution data converters, traditionally designed using the noise shaping approach.

The convex coding approach can also be applied to signal compression [21]. Although this field implies a digital to digital transformation, the input signal is usually considered as quasi-continuous in amplitude. In this context, it is shown in [21] that classical signal compression schemes such as block DCT coding can be analyzed as convex coding schemes. An example of new signal compression scheme is proposed by a direct and active control of the encoded sets.



(a) (b)

Figure 13: Partition defined by a $\Sigma\Delta$ modulator in the 2 dimensional space of T_0-periodic sinusoids with arbitrary amplitude and phase at oversampling ratio $R = 4$. (a) Single-loop case. (b) Double-loop case.

References

[1] A.V.Oppenheim and R.W.Shafer, *Discrete-Time Signal Processing.* Prentice Hall, 1989.

[2] N.S.Jayant and P.Noll, *Digital Coding of Waveforms.* Prentice-Hall, 1984.

[3] W.R.Bennett, "Spectra of quantized signals," *Bell System Technical Journal*, vol. 27, pp. 446–472, July 1948.

[4] R.M.Gray, "Quantization noise spectra," *IEEE Trans. Information Theory*, vol. IT-36, pp. 1220–1244, Nov. 1990.

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

[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.*, vol. 42, pp. 519–531, Mar. 1994.

[7] D.G.Luenberger, *Optimization by vector space methods.* Wiley, 1969.

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

[9] 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.*, vol. 42, pp. 200–203, Jan. 1994.

[10] D.G.Luenberger, *Linear and nonlinear programming.* Wiley, 1984.

[11] P.L.Combettes, "The foundations of set theoretic estimation," *Proc. IEEE*, vol. 81, no. 2, pp. 1175–1186, Feb. 1993.

[12] 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.

[13] P.L.Combettes and H.Puh, "A fast parallel projection algorithm for set theoretic image recovery," *Proc. IEEE Int. Conf. ASSP*, vol. V, pp. 473–476, Apr. 1994.

[14] P.L.Combettes and H.Puh , *Personal communication*, May 1994.

[15] 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.

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

[17] 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.

[18] S.Hein, K.Ibraham, and A.Zakhor, "New properties of sigma-delta modulators with dc inputs," *IEEE Trans. Commun.*, vol. COM-40, pp. 1375–1387, Aug. 1992.

[19] N.T.Thao and M.Vetterli, "Lower bound on the mean squared error in multi-loop $\Sigma\Delta$ modulation with periodic bandlimited signals," *Proc. 27th Asilomar Conf. on Signals, Systems and Computers, Pacific Grove, CA*, Nov. 1993.

[20] N.T.Thao and M.Vetterli, "Lower bound on the mean squared error in oversampled quantization of periodic signals," *IEEE Trans. Information Theory.* Submitted in June 1993, revised in Sept. 1994.

[21] K.Asai, N.T.Thao, and M.Vetterli, "A study of convex coders with an application to image coding," *Proc. IEEE Int. Conf. ASSP*, vol. V, pp. 581–584, Apr. 1994.