Lower Bound on the Mean-Squared Error in Oversampled Quantization of Periodic Signals Using Vector Quantization Analysis

Nguyen T. Thao and Martin Vetterli, *Fellow, IEEE*

Abstract— **Oversampled analog-to-digital conversion is a technique which permits high conversion resolution while using coarse quantization. Classically, by lowpass filtering the quantized oversampled signal, it is possible to reduce the quantization error power in proportion to the oversampling ratio R. In other words, the reconstruction mean-squared error (MSE) is in $\mathcal{O}(R^{-1})$. It was recently found that this error reduction is not optimal. Under certain conditions, it was shown on periodic bandlimited signals that an upper bound on the MSE of optimal reconstruction is in $\mathcal{O}(R^{-2})$ instead of $\mathcal{O}(R^{-1})$. In the present paper, we prove on the same type of signals that the order $\mathcal{O}(R^{-2})$ is the theoretical limit of reconstruction as an MSE lower bound. The proof is based on a vector-quantization approach with an analysis of partition cell density.**

Index Terms— **Oversampling, quantization, A/D conversion, optimal reconstruction, MSE lower bound, hyperplane, partition.**

I. Introduction

In oversampled analog-to-digital conversion (ADC), higher resolution in the discrete-time samples is achieved not by using a finer quantizer but by quantizing redundant, oversampled values of the continuous-time signal. In the simple version of oversampled ADC, a bandlimited signal $\boldsymbol{x}$ is sampled at R times its Nyquist rate, and using an ideal digital lowpass filter with passband $[-\frac{\pi}{R}, \frac{\pi}{R}]$, the out-of-band quantization noise is filtered out, leading to an approximation $\hat{\boldsymbol{x}}$ of the true signal $\boldsymbol{x}$ where the mean-squared error (MSE) behaves at best in $\mathcal{O}(R^{-1})$ [1], [2] (Fig. 1). The digital lowpass filter followed by the downsampling is usually considered as a part of the A/D converter. However, one can look at the ADC chain in a different way by saying that the real encoded signal is obtained after sampling and quantization, and that the lowpass filter is already a part of the reconstruction of the input signal. This leads to the block diagram of Fig. 2(a) which emphasizes the separation between the encoding part of the ADC process and the reconstruction part, or, decoder. Because of the lowpass filter, the decoder is based on a linear reconstruction algorithm and will be called *linear decoder*. While such a reconstruction is simple and easily implementable, the question of an *optimal reconstruction* from the oversampled and quantized signal, in the minimum mean-squared error sense, arises quite naturally (Fig. 2(c)). In [3], [4], we showed that for periodic bandlimited signals, there exists a *nonlinear reconstruction algorithm* which leads to an MSE of $\mathcal{O}(R^{-2})$ instead of $\mathcal{O}(R^{-1})$ in the linear reconstruction case (Fig. 2(b)). The method relies on finding estimates which are *consistent* with the bandlimitation characteristics and the oversampled data from the signal. The derivation of the MSE in $\mathcal{O}(R^{-2})$ is based on the fact that a bandlimited signal with bounded amplitude has a bounded slope, and consequently, errors of reconstruction in amplitude are of the same order as errors due to the time discretization, up to a bounded multiplicative coefficient. While the method of consistent reconstruction is clearly more complex than the straightforward linear reconstruction, it establishes an upper bound on the reconstruction MSE of $\mathcal{O}(R^{-2})$.

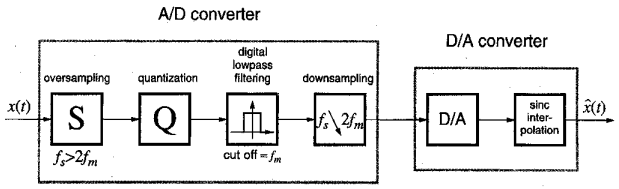

Fig. 1. Classical block diagram of oversampled ADC.

In the present paper, we are concerned about establishing a lower bound on the reconstruction error. Due to the difficulty of the general problem, we deal only with periodic and bandlimited input signals. For any given probability distribution of input signals, we show that the MSE in simple oversampled ADC is lower-bounded by $\mathcal{O}(R^{-2})$. Thus under the above assumption of input signals, an optimal decoder for oversampled ADC has a performance of $\mathcal{O}(R^{-2})$.

The outline of the paper is as follows. In Section II, we define precisely our space of periodic bandlimited signals. With this assumption, we show in Section III that the input signals belong to a finite-dimensional space. Thus they can be considered as vectors. As a consequence, the oversampled ADC chain can be analyzed as a vector quantizer with an encoding section and a decoding section. For a given probability distribution of input vectors, the performance of the optimal decoder all depends on the input space partition defined by

Manuscript received July 23, 1993; revised September 27, 1995. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Trondheim, Norway, June 27–July 1, 1994.

N. T. Thao is with the Department of Electrical and Electronic Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Koowloon, Hong Kong.

M. Vetterli is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720 USA, and the Département d'Electricité, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

Publisher Item Identifier S 0018-9448(96)01184-4.

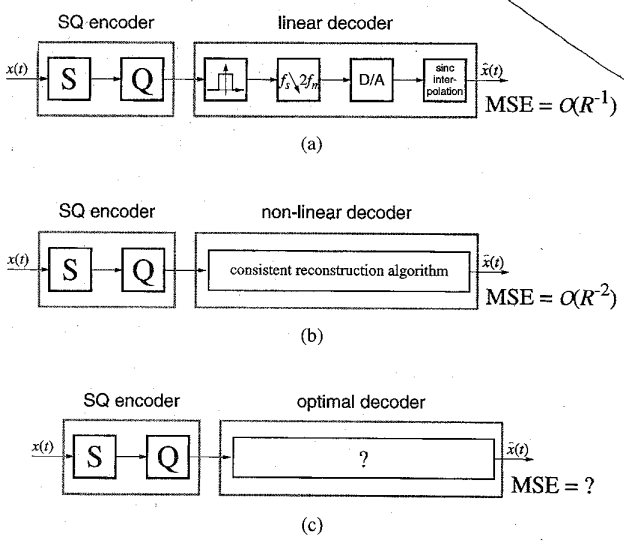

Fig. 2. Representation of the oversampled ADC process in terms of encoding-decoding. (a) Classical ADC chain. (b) ADC with consistent reconstruction. (c) ADC with optimal reconstruction.

the encoder. Assuming that the quantizer is uniform with an infinite input range, we show in Section IV that this partition has a particular structure called *hyperplane wave structure*. This partition structure is studied in the Appendix where an upper bound on the cell density is derived. A lower bound on the MSE can then be derived from this upper bound, thanks to Zador's formula [5]–[7]. We show finally that this MSE lower bound holds also when the quantizer has a finite number of levels.

II. Periodic Bandlimited Signals and Vector Representation

We assume that the continuous-time signals belong to $\mathcal{L}_2(0,T)$ (space of square-integrable functions on $[0,T]$) and we denote them by using bold face characters. The value of a signal $\boldsymbol{x}$ at time t will be denoted by $x(t)$. In $\mathcal{L}_2(0,T)$, we will consider the inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_T = \frac{1}{T} \int_0^T x(t) y(t)\, dt$$

and the norm

$$\|\boldsymbol{x}\|_T = \langle \boldsymbol{x}, \boldsymbol{x} \rangle_T^{1/2}.$$

Using real Fourier series, an orthonormal basis of $\mathcal{L}_2(0,T)$ is given by

$$u_1(t) = 1$$

$$\forall p \geq 1, \begin{cases} u_{2p}(t) = \sqrt{2}\cos\left(2\pi p \frac{t}{T}\right) \\ u_{2p+1}(t) = \sqrt{2}\sin\left(2\pi p \frac{t}{T}\right). \end{cases} \quad (1)$$

We recall that any signal $\boldsymbol{x} \in \mathcal{L}_2(0,T)$ can be written as

$$\boldsymbol{x} = \sum_{k=1}^{+\infty} X_k \boldsymbol{u}_k$$

where $X_k = \langle \boldsymbol{u}_k, \boldsymbol{x} \rangle_T$ for all $k \geq 1$. We call T-periodic bandlimited signals of bandwidth K, signals $\boldsymbol{x}$ of $\mathcal{L}_2(0,T)$ such that

$$\boldsymbol{x} = \sum_{k=1}^{K} X_k \boldsymbol{u}_k \quad \text{a.e.} \quad (2)$$

The set of such signals is a K-dimensional real vector space and is designated by $\mathcal{V}_K$. The term "T-periodic bandlimited" is justified as follows: if a signal $\boldsymbol{x}$ of $\mathcal{L}_2(0,T)$ is T-periodized, it has a discrete Fourier transform which is bandlimited.

Equation (2) defines a one-to-one linear mapping between $\boldsymbol{x} \in \mathcal{V}_K$ and the vector $\vec{x} = (X_1, \cdots, X_K) \in \mathbf{R}^K$. We will say that $\vec{x}$ is the vector representation of $\boldsymbol{x} \in \mathcal{V}_K$. We recall that $\mathbf{R}^K$ has an inner product

$$\langle \vec{x}, \vec{y} \rangle = \sum_{k=1}^{K} X_k Y_k$$

where $\vec{x} = (X_1, \cdots, X_K)$ and $\vec{y} = (Y_1, \cdots, Y_K)$, and a norm

$$\|\vec{x}\| = \left(\sum_{k=1}^{K} X_k^2 \right)^{1/2}.$$

Thanks to Parseval's equality, we have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_T = \langle \vec{x}, \vec{y} \rangle \quad \text{and} \quad \|\boldsymbol{x}\|_T = \|\vec{x}\|$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{R}^K$ are the vector representations of $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}_K$, respectively. As a consequence, the MSE between two signals $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}_K$ can be expressed in terms of their vector representation, since $\|\boldsymbol{y} - \boldsymbol{x}\|_T^2 = \|\vec{y} - \vec{x}\|^2$.

III. Vector Quantization Representation of the Oversampled ADC Chain

In this section, we show that the block diagrams of Fig. 2 can be studied in a vector quantization framework. We assume that signals are T-periodic bandlimited as described in the previous section, and thus belong to the space $\mathcal{V}_K$. We assume that they are regularly sampled N times in the time window $[0,T]$. The samples are then uniformly quantized into integers. Considering the case of Fig. 2(a), the linear decoder necessarily outputs a bandlimited signal, which is T-periodic, thus belonging to $\mathcal{V}_K$. In the case of Fig. 2(b) and (c), the decoder should naturally output signals of $\mathcal{V}_K$. For Fig. 2(b), this is because of the bandlimitation consistency with the input signal. For Fig. 2(c), optimality requires that the output signal belong to the same space as the input signal, as a necessary condition. Indeed, if this were not the case, the orthogonal projection of the output signal onto the space of the input signal would necessarily be a better estimate of the input, as a consequence of Pythagoras theorem. As summarized in Fig. 3, the block diagrams of Fig. 2 are composed of an encoder which maps a signal of $\mathcal{V}_K$ into an N-point sequence of integers, and a decoder which maps an N-point sequence of integers into some signal of $\mathcal{V}_K$. This scheme can be translated into vector-quantization terms, when replacing the analog input and the analog output by their vector representations, and considering an N-point sequence of integers $I = (i_1, \cdots, i_N)$ as an index

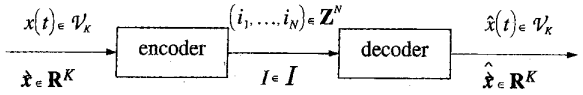

Fig. 3. Encoder–decoder decomposition of the oversampled ADC chain, with real signals (above the diagram), and with vector-quantization signal representation (below the diagram).

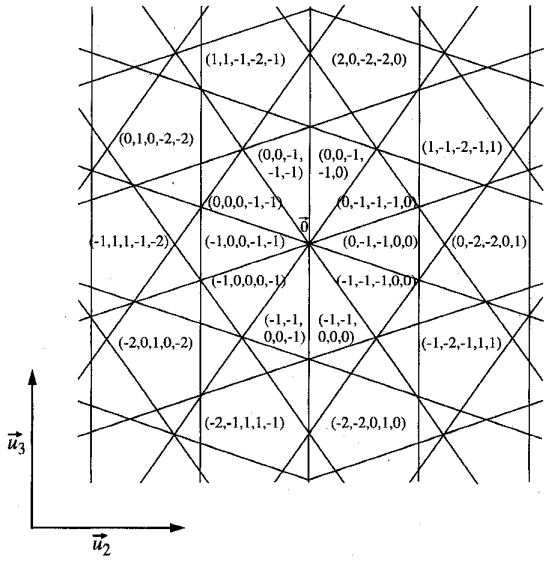

Fig. 4. Partition of the input signal space spanned by $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ defined by the sampling-quantization encoder when $N = 5$. Each cell comprises all input vectors giving the same N-tuple index through the sampling-quantization encoder. Examples of N-tuple indices are given for a selection of cells.

of the discrete set $\mathcal{I} = \mathbf{Z}^N$. This equivalent view is also shown in Fig. 3.

In this view, the encoder defines a partition of the space $\mathbf{R}^K$ into cells which comprise all input vectors giving a same index. In fact, the encoder is uniquely defined by this partition. Although each index has the form of an N-point sequence, it is not so much the explicit form of the index that matters, but its one-to-one correspondence with a cell of the partition. Fig. 4 shows the partition induced by the sampling-quantization encoder when the space of input signals is two-dimensional, spanned by $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ (space of sinusoids of period T with arbitrary phase and amplitude, see (1)), and the total number of samples is $N = 5$. For a selection of cells, the figure also indicates the corresponding indices in their original N-point sequence form. It is important to see that the dimension K of the encoded signal is intrinsic in the continuous-time analog signals of $\mathcal{V}_K$. This dimension is not the number of input samples N, contrary to the usual case in vector quantization. In particular, the components $X_1, \cdots, X_K$ of the input vectors $\vec{x}$ are not the samples of the analog signals. One should think of the sampling-quantization encoder as a "black box," where N only appears as an element of the internal process.

Because of the one-to-one correspondence between cells and indices, the role of the decoder amounts to mapping each cell into some vector of $\mathbf{R}^K$, called *code vector* according to the vector quantization terminology [8]. For a given probability

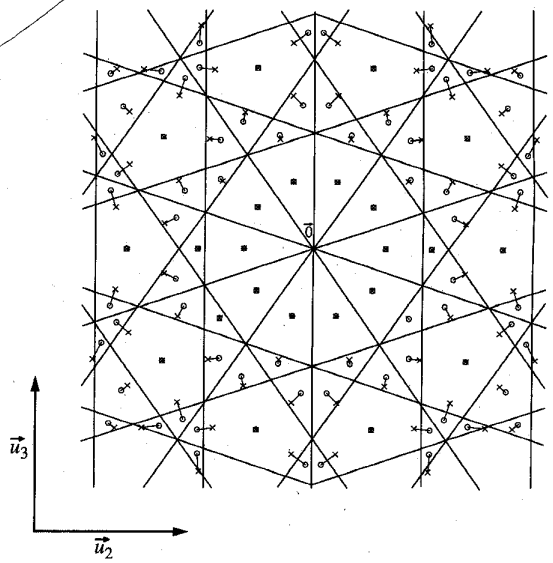

Fig. 5. Code vectors of the optimal decoder (circles) and the linear decoder (crosses) for the encoding configuration of Fig. 4.

distribution of input signals, any choice of decoder results in some expected MSE. As mentioned in [8], the decoder is optimal (that is, minimizes the expected MSE) when each code vector is the centroid of its associated cell. With the encoding conditions of Fig. 4, we show in Fig. 5 the code vectors of the optimal decoder in the case where the input vectors have a uniform probability distribution in a bounded region of $\mathbf{R}^K$. The figure also shows the code vectors of the linear decoder. One can see that these code vectors are not necessarily optimal, and not even necessarily consistent, since they sometimes lie outside their corresponding cells.

IV. Derivation of the Sampling-Quantization Partition

In this section, we show how the partition of the sampling-quantization encoder can be derived. We also present some of its properties. The vector quantization representation of the sampling-quantization encoder is shown in Fig. 6(a). The sampling operator originally maps continuous-time signals $\boldsymbol{x} \in \mathcal{V}_K$ into the N-point sequence $(x_1, \cdots, x_N)$ where for all $n = 1, \cdots, N$, $x_n = x(\frac{n}{N}T)$. Let $\vec{x} = (X_1, \cdots, X_K)$ be the vector representation of $\boldsymbol{x}$. Using (2), we have for each $n = 1, \cdots, N$

$$x_n = x(\tfrac{n}{N}T) = \sum_{k=1}^{K} X_k u_k(\tfrac{n}{N}T) = \left\langle \vec{f}_n, \vec{x} \right\rangle \tag{3}$$

where $\vec{f}_n$ is the vector of $\mathbf{R}^K$ defined by

$$\vec{f}_n = \left(u_1(\tfrac{n}{N}T), u_2(\tfrac{n}{N}T), \cdots, u_K(\tfrac{n}{N}T)\right), \quad \text{for } n = 1, \cdots, N. \tag{4}$$

Therefore, the sampling operator $\mathbf{S}$ of Fig. 6(a) which is directly applied on the vector representation $\vec{x}$ is expressed as

$$\mathbf{S}[\vec{x}] = \left(\left\langle \vec{f}_1, \vec{x} \right\rangle, \left\langle \vec{f}_2, \vec{x} \right\rangle, \cdots, \left\langle \vec{f}_N, \vec{x} \right\rangle\right) \in \mathbf{R}^N.$$

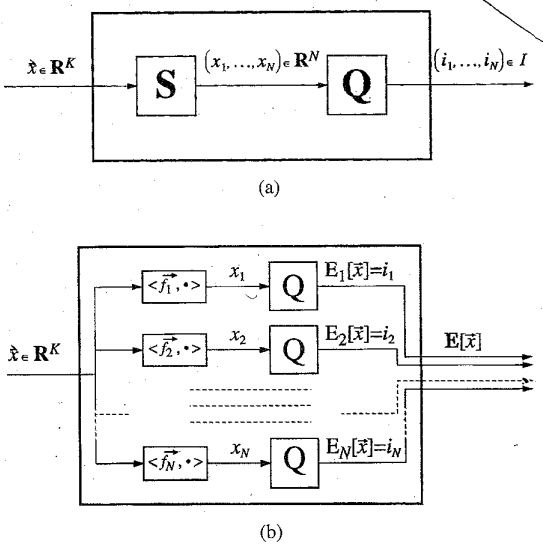

Fig. 6. Vector quantization representation of the sampling-quantization encoder. (a) Primitive form. (b) Parallel decomposition.

The quantization operator $\mathbf{Q}$ of Fig. 6 is simply defined by

$$\mathbf{Q}[(x_1, x_2, \cdots, x_N)] = (\mathrm{Q}[x_1], \mathrm{Q}[x_2], \cdots, \mathrm{Q}[x_N])$$

where Q is the scalar quantization operator. In most of the paper (except in the last paragraph), we assume that the quantizer is uniform on infinite input range, with step size q. More precisely, we assume that the quantizer splits the amplitude axis into intervals of the type $[iq, (i+1)q[$. Whenever an input sample falls into the interval $[iq, (i+1)q[$, the output of the quantizer is the integer i which serves as index of the interval. With these assumptions, Q has the following simple expression:[1]

$$\forall x \in \mathbf{R}, \mathrm{Q}[x] = \left\lfloor \frac{x}{q} \right\rfloor \tag{5}$$

where $\lfloor y \rfloor$ designates the greatest integer smaller than or equal to y.

Then, the whole encoder is defined by the mapping

$$\begin{aligned} \mathbf{E} : \mathbf{R}^K &\longrightarrow \mathbf{Z}^N \\ \vec{x} &\longmapsto \mathbf{E}[\vec{x}] = \left(\mathrm{Q}\left(\left\langle \vec{f}_1, \vec{x} \right\rangle\right), \cdots, \mathrm{Q}\left(\left\langle \vec{f}_N, \vec{x} \right\rangle\right)\right). \end{aligned}$$

The encoder $\mathbf{E}$ can be seen as the association of N subencoders $\mathrm{E}_1, \cdots, \mathrm{E}_N$ working "in parallel," such that, for each $n = 1, \cdots, N$

$$\begin{aligned} \mathrm{E}_n : \mathbf{R}^K &\longrightarrow \mathbf{Z} \\ \vec{x} &\longmapsto \mathrm{E}_n[\vec{x}] = \mathrm{Q}\left(\left\langle \vec{f}_n, \vec{x} \right\rangle\right). \end{aligned}$$

With this definition, we have

$$\forall \vec{x} \in \mathbf{R}^K, \mathbf{E}[\vec{x}] = (\mathrm{E}_1[\vec{x}], \cdots, \mathrm{E}_N[\vec{x}]). \tag{6}$$

The parallel decomposition of the sampling-quantization encoder is shown in Fig. 6(b). The partition $\mathcal{P}$ defined by the encoder $\mathbf{E}$ can be simply derived from the partitions $\mathcal{P}_1, \cdots, \mathcal{P}_N$ defined by the subencoders $\mathrm{E}_1, \cdots, \mathrm{E}_N$, respectively. Because of (6), $\mathcal{C}$ is a cell of $\mathcal{P}$ if and only if there exist cells $\mathcal{C}_1, \cdots, \mathcal{C}_N$ of $\mathcal{P}_1, \cdots, \mathcal{P}_N$, respectively, such that $\mathcal{C} = \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_N$. We will say by abuse of language that $\mathcal{P}$ is obtained by intersection of $\mathcal{P}_1, \cdots, \mathcal{P}_N$.

Therefore, we just need to concentrate on the study of $\mathcal{P}_n$ for each $n \in \{1, \cdots, N\}$. According to (5) we have

$$\mathrm{E}_n[\vec{x}] = \left\lfloor \frac{1}{q} \left\langle \vec{f}_n, \vec{x} \right\rangle \right\rfloor. \tag{7}$$

We propose an equivalent expression of E_n which will make the analysis of its partition easier. Let us define

$$\vec{d}_n = \frac{\vec{f}_n}{q}. \tag{8}$$

Then, combining (7) and (8), we find

$$\mathrm{E}_n[\vec{x}] = \left\lfloor \left\langle \vec{d}_n, \vec{x} \right\rangle \right\rfloor. \tag{9}$$

It is easy to see that the cells of the partition of E_n are of the type

$$\mathcal{C}_i = \left\{ \vec{x} \in \mathbf{R}^K \middle| i \le \left\langle \vec{d}_n, \vec{x} \right\rangle < i+1 \right\}, \quad i \in \mathbf{Z}$$

and are separated by the affine hyperplanes

$$\mathcal{H}_i = \left\{ \vec{x} \in \mathbf{R}^K \middle| \left\langle \vec{d}_n, \vec{x} \right\rangle = i \right\}, \quad i \in \mathbf{Z}.$$

An example of partition is shown in Fig. 7. The hyperplanes are perpendicular to $\vec{d}_n$ and are equally spaced with the period

$$q_n = \frac{1}{\|\vec{d}_n\|}.$$

We call the set of hyperplanes $\{\mathcal{H}_i | i \in \mathbf{Z}\}$ a *hyperplane wave* and say that $\vec{d}_n$ is the *density vector* of the wave. The corresponding partition is called a *hyperplane wave partition*, and is designated by $\mathcal{P}_K\left(\vec{d}_n\right)$, since it is uniquely defined by $\vec{d}_n$. Finally, the partition of the encoder $\mathbf{E}$ is obtained by intersection of the partitions

$$\mathcal{P}_K\left(\vec{d}_1\right), \cdots, \mathcal{P}_K\left(\vec{d}_N\right)$$

where each $\vec{d}_n$ is given by (8). Fig. 8 shows an example in the case where $N = 5$. By convention we designate the resulting partition by

$$\mathcal{P}_K\left(\vec{d}_1, \cdots, \vec{d}_N\right)$$

and call it the *hyperplane wave structured partition* of density vectors $\vec{d}_1, \cdots, \vec{d}_N$.

[1] In general, a quantizer may have an offset c such that $\mathrm{Q}[x] = \left\lfloor \frac{x-c}{q} \right\rfloor$. For the sake of simplicity, we assume that $c = 0$. The following derivations can be easily generalized for $c \neq 0$.

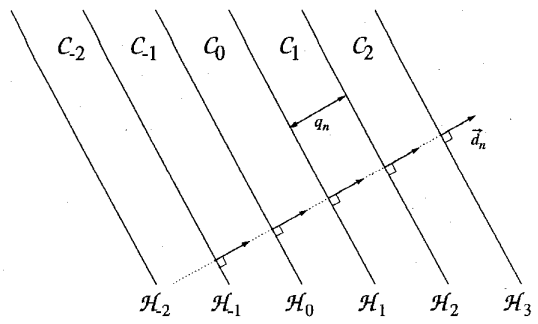

Fig. 7. Hyperplane wave partition.

V. Upper Bound on the Number of Cells

As mentioned in the Introduction, the method used to derive a lower bound on the optimal decoding MSE requires an upper bound on the number of cells of the partition. We propose such an upper bound based on the properties of hyperplane wave structured partitions.

Assume that the input signal $\vec{x}$ belongs to a region $\mathcal{B}$ of $\mathbf{R}^K$ such that

$$\forall \vec{x} \in \mathcal{B}, \|\vec{x}\| \leq B, \quad \text{where } B > 0. \tag{10}$$

When N is the total number of samples, the sampling-quantization encoder induces in the region a certain number of cells that we call $M_{SQ}(\mathcal{B}, N)$. According to the previous section, the partition defined by the sampling-quantization encoder is the hyperplane wave structured partition

$$\mathcal{P}_K\left(\vec{d}_1, \cdots, \vec{d}_N\right)$$

where $\vec{d}_n$ is given by (8). We designate by $N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N)$ the maximum number of cells that the partition

$$\mathcal{P}_K\left(\vec{d}_1, \cdots, \vec{d}_N\right)$$

can induce in a sphere of $\mathbf{R}^K$ of diameter D and of arbitrary origin. Since $\mathcal{B}$ is included in a sphere of diameter $2B$, we necessarily have

$$M_{SQ}(\mathcal{B}, N) \leq N_{K,2B}(\vec{d}_1, \cdots, \vec{d}_N). \tag{11}$$

Thanks to the particular properties of a hyperplane wave structured partition, an upper bound on the number of cells $N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N)$ is derived in the Appendix. The following theorem is shown:

Theorem 5.1: Let $\vec{d}_1, \cdots, \vec{d}_N$ be N vectors of $\mathbf{R}^K$ and $D > 0$. Let d be an upper bound on $\|\vec{d}_1\|, \cdots, \|\vec{d}_N\|$. Then

$$N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N) \leq \binom{N}{K}(\lceil Dd \rceil + 1)^K \tag{12}$$

where $\lceil y \rceil$ is the smallest integer greater than or equal to y.

In our hyperplane wave structured partition, the vectors $\vec{d}_n$ are given by (8). Using (1) and (4), it is easy to derive that

$$\|\vec{d}_n\|^2 = \frac{\|\vec{f}_n\|^2}{q^2} = \frac{\sum_{k=1}^{K} \left|u_k\left(\frac{n}{N}T\right)\right|^2}{q^2} \leq \frac{2K}{q^2}.$$

Therefore, $d = \frac{\sqrt{2K}}{q}$ is an upper bound on $\|\vec{d}_1\|, \cdots, \|\vec{d}_N\|$. Applying Theorem 5.1 with this upper bound and using the facts that $D = 2B$ and $\lceil y \rceil \leq y + 1$, we find

$$N_{K,2B}(\vec{d}_1, \cdots, \vec{d}_N) \leq \binom{N}{K}\left(2B\frac{\sqrt{2K}}{q} + 2\right)^K.$$

Applying (11) and using the fact that

$$\binom{N}{K} = \frac{N(N-1)\cdots(N-K+1)}{K!} \leq \frac{N^K}{K!}$$

we conclude that

$$M_{SQ}(\mathcal{B}, N) \leq \frac{1}{K!}\left[2N\left(\frac{B}{q}\sqrt{2K} + 1\right)\right]^K. \tag{13}$$

VI. Lower Bound on the Optimal Decoding MSE

In this section we assume that the input signals belong to the region $\mathcal{B}$ defined in (10) and have a certain probability distribution $\mathbf{p}$ in $\mathcal{B}$. The goal is to express a lower bound on the optimal decoding MSE in terms of the oversampling ratio R defined[2] by $R = \frac{N}{K}$. For convenience, we assume that N is chosen such that R is an integer ($R \in \mathbf{N}^*$). Let us designate by $\text{MSE}_{\text{opt}}(R)$ the optimal decoding MSE when the oversampling ratio is equal to R. For R given, we have $N = RK$ and the number of cells induced by the sampling-quantization encoder in $\mathcal{B}$ is equal to $M_{SQ}(\mathcal{B}, RK)$. According to (13) from the previous section, we have the following inequality:

$$M_{SQ}(\mathcal{B}, RK) \leq \frac{\left[2K\left(\frac{B}{q}\sqrt{2K} + 1\right)\right]^K}{K!} \times R^K. \tag{14}$$

Using Zador's formula [5], [6], we show how this upper bound on the number of cells can be used to derive a lower bound on $\text{MSE}_{\text{opt}}(R)$.

For a given probability density $\mathbf{p}$ of input signals in $\mathbf{R}^K$ and a given number of cells M, Zador analyzed the minimum MSE that can be achieved by a vector quantizer using M code vectors (or equivalently, defining a partition of M cells). Let us call this minimum $\text{MSE}_{\min}(K, \mathbf{p}, M)$. He showed that

$$\lim_{M \to +\infty} M^{2/K}\,\text{MSE}_{\min}(K, \mathbf{p}, M) = b_K \mathcal{N}(K, \mathbf{p}) \tag{15}$$

where b_K is a coefficient[3] which depends on K and

$$\mathcal{N}(K, \mathbf{p}) = \left(\int_{\mathbf{R}^K} (\mathbf{p}(\vec{x}))^{\frac{K}{K+2}} d\vec{x}\right)^{\frac{K+2}{K}}. \tag{16}$$

This limit says that when the number of cells M is large, $\text{MSE}_{\min}(K, \mathbf{p}, M)$ is of the order of $b_K \mathcal{N}(K, \mathbf{p}) M^{-2/K}$.

[2]This definition is consistent with the traditional definition of oversampling ratio. The periodic and bandlimited input signals have K nonzero consecutive discrete frequency components equally spaced by $1/T$. Their bandwidth in the traditional sense is therefore K/T, while the sampling frequency is N/T. The oversampling ratio in the traditional sense is then $\frac{N/T}{K/T} = \frac{N}{K}$

[3]Zador provides in [5] the following bounds on b_K:

$$\frac{1}{\pi}\frac{K}{K+2}\Gamma\left(\frac{K}{2}+1\right)^{2/K} \leq b_K \leq \frac{1}{\pi}\Gamma\left(\frac{K}{2}+1\right)^{2/K}\Gamma\left(1+\frac{2}{K}\right)$$

where Γ is the gamma function.

Since the sampling-quantization encoder defines $M_{SQ}(\mathcal{B}, RK)$ cells, it is clear that

$$\mathrm{MSE}_{\mathrm{opt}}(R) \geq \mathrm{MSE}_{\min}(K, \mathbf{p}, M_{SQ}(\mathcal{B}, RK)). \tag{17}$$

Intuitively speaking, when R is large, the number of cells $M_{SQ}(\mathcal{B}, RK)$ is large, and $\mathrm{MSE}_{\min}(K, \mathbf{p}, M_{SQ}(\mathcal{B}, RK))$ is of the order of $b_K \mathcal{N}(K, \mathbf{p}) M_{SQ}(\mathcal{B}, RK)^{-2/K}$. But using the upper bound on $M_{SQ}(\mathcal{B}, RK)$ in (14), this order can be lower-bounded by $c(K, \mathbf{p}, B, q) \cdot R^{-2}$, where $c(K, \mathbf{p}, B, q)$ is a coefficient which only depends on $K, \mathbf{p}, B, q$. This gives the intuition that $\mathcal{O}(R^{-2})$ is a lower bound on $\mathrm{MSE}_{\mathrm{opt}}(R)$. We formalize and prove this intuitive result in the following theorem.

Theorem 6.1: There exists a constant $c(K, \mathbf{p}, B, q)$ which may depend on $K, \mathbf{p}, B, q$ but not on R such that

$$\forall R \in \mathbf{N}^*, \mathrm{MSE}_{\mathrm{opt}}(R) \geq \frac{c(K, \mathbf{p}, B, q)}{R^2}.$$

Proof: Equation (15) implies that there exists a number $M_0 > 0$ which may depend on K and $\mathbf{p}$, such that

$$\forall M \geq M_0, M^{2/K} \mathrm{MSE}_{\min}(K, \mathbf{p}, M) \geq \frac{1}{2} b_K \mathcal{N}(K, \mathbf{p})$$

or

$$\forall M \geq M_0, \mathrm{MSE}_{\min}(K, \mathbf{p}, M) \geq \frac{1}{2} b_K \mathcal{N}(K, \mathbf{p}) M^{-2/K}. \tag{18}$$

It is clear that $\mathrm{MSE}_{\min}(K, \mathbf{p}, M)$ is a decreasing function of M, that is

$$M' \leq M \Longrightarrow \mathrm{MSE}_{\min}(K, \mathbf{p}, M') \geq \mathrm{MSE}_{\min}(K, \mathbf{p}, M). \tag{19}$$

In particular, for all $R \in \mathbf{N}^*$, this inequality can be applied to the following choice of M' and M:

$$M' = M_{SQ}(\mathcal{B}, RK) \tag{20}$$

and

$$M = \frac{\left[2K\left(\frac{B}{q}\sqrt{2K} + 1\right)\right]^K}{K!} \times R^K \tag{21}$$

due to the inequality (14). We would also like to apply (18) for M given by (21). This requires R to be larger than

$$R_0 = M_0 \times \frac{K!}{\left[2K\left(\frac{B}{q}\sqrt{2K} + 1\right)\right]^K} \tag{22}$$

which only depends on $K, \mathbf{p}, B, q$, since M_0 only depends on K and $\mathbf{p}$.

Suppose that $R \geq R_0$ and consider the numbers M' and M given in (20) and (21). Because of (17), (20), and (19), we have

$$\begin{aligned}\mathrm{MSE}_{\mathrm{opt}}(R) &\geq \mathrm{MSE}_{\min}(K, \mathbf{p}, M') \geq \mathrm{MSE}_{\min}(K, \mathbf{p}, M) \\ &\geq \frac{1}{2} b_K \mathcal{N}(K, \mathbf{p}) M^{-2/K}.\end{aligned}$$

Replacing M by its value given in (21), we derive that

$$\forall R \geq R_0, \quad \mathrm{MSE}_{\mathrm{opt}}(R) \geq \frac{1}{2} b_K \mathcal{N}(K, \mathbf{p}) \frac{K!^{2/K}}{\left[2K\left(\frac{B}{q}\sqrt{2K} + 1\right)\right]^2} \times \frac{1}{R^2}.$$

This can be rewritten as

$$\forall R \geq R_0, \mathrm{MSE}_{\mathrm{opt}}(R) \geq \frac{c_0}{R^2}$$

where c_0 depends only on $K, \mathbf{p}, B, q$. This inequality is not necessarily true for $R < R_0$. However, for each $R = 1, 2, \cdots, R_0$, we can always define the coefficient $c_R = R^2 \mathrm{MSE}_{\mathrm{opt}}(R)$ such that $\mathrm{MSE}_{\mathrm{opt}}(R) = \frac{c_R}{R^2}$. For each $R = 1, 2, \cdots, R_0$, c_R depends only on $K, \mathbf{p}, B, q$. Let us define $c(K, \mathbf{p}, B, q) = \min(c_0, c_1, c_2, \cdots, c_{R_0})$. It is clear that $c(K, \mathbf{p}, B, q)$ only depends on $K, \mathbf{p}, B, q$ (we recall from (22) that R_0 only depends on $K, \mathbf{p}, B, q$). Moreover, we have

$$\forall R \in \mathbf{N}^*, \mathrm{MSE}_{\mathrm{opt}}(R) \geq \frac{c(K, \mathbf{p}, B, q)}{R^2}. \qquad \square$$

Case of Finite Range Quantization

When the quantizer has a finite number of levels, the model of uniform quantization on an infinite input range is no longer valid. However, the upper bound on the number of cells of Theorem 5.1 remains valid. Indeed, the partition defined by a real quantizer is derived from the partition of the infinite range quantizer by merging all the cells located outside the range of the real quantizer. As a result, a certain number of cells of the partition defined by the encoder of Fig. 6(c) may be merged in the considered region of input signals. This can only result in decreasing the evaluated number of cells. Since Theorem 1 is valid, the lower bound on the MSE holds.

Appendix
Properties of Hyperplane Wave Structured Partitions

Although hyperplane wave structured partitions appear in this paper as a result of the sampling-quantization encoder in oversampled ADC, their definition is quite abstract, and their properties can be studied independently of this context. Also, it will be more convenient to study them in the more general context of *affine* spaces, where elements are not vectors but geometric points. As will be seen, the derivation of upper bounds on the number of cells will be recursive on the space dimension and will be based on partitions induced in affine hyperplanes. In Section A.1 we redefine the hyperplane wave partitions and the hyperplane wave structured partitions in affine spaces and show that vector spaces correspond to a particular case. We also demonstrate a property essential for the recursion. In Section A.2, we derive some upper bounds on cell densities. The affine space version of Theorem 5.1 will be shown in the end of this section (Theorem A.7).

A.1. General Definition and Properties in a Euclidean Affine Space

Let $\vec{\mathcal{W}}_K$ be a Euclidean vector space of dimension K associated with an inner product $\langle \cdot, \cdot \rangle$ and the norm $\| \cdot \|$. Let $\mathcal{W}_K$ be an affine space of direction $\vec{\mathcal{W}}_K$. This means that $\mathcal{W}_K$ is a set of points and that there exists a mapping

$$\begin{array}{ccc} \mathcal{W}_K \times \mathcal{W}_K & \longrightarrow & \vec{\mathcal{W}}_K \\ (A, B) & \longmapsto & \vec{v} = \vec{AB} \end{array} \tag{23}$$

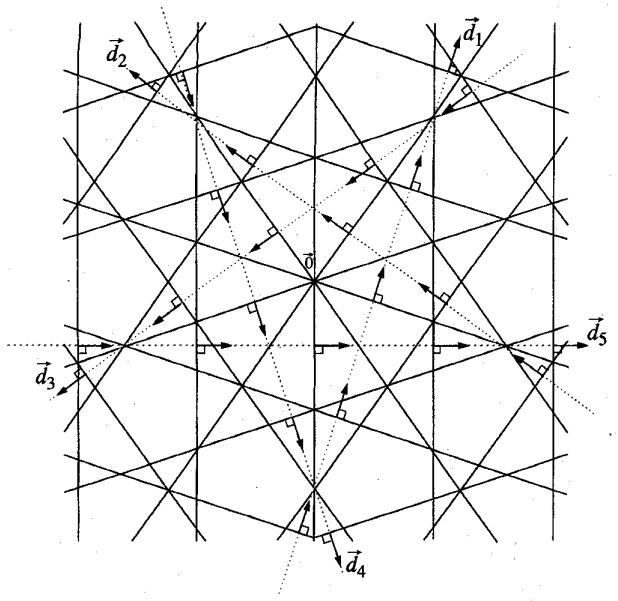

Fig. 8. Example of hyperplane wave structured partition with $N = 5$. One can see the five directions of hyperplane indicated by the five perpendicular vectors $\vec{d}_1, \cdots, \vec{d}_5$. For each direction, the hyperplanes are equidistant.

which satisfies the following properties:

$$\forall A, B \in \mathcal{W}_K, \vec{AB} = \vec{0} \Leftrightarrow A = B \tag{24}$$

$$\forall A, B, C \in \mathcal{W}_K, \vec{AB} + \vec{BC} = \vec{AC} \tag{25}$$

$$\forall A \in \mathcal{W}_K, \forall \vec{v} \in \vec{\mathcal{W}}_K, \exists! B \in \mathcal{W}_K, \vec{AB} = \vec{v}. \tag{26}$$

A.1.1) Hyperplane Wave Partition: We call the *hyperplane wave partition* of *density vector* $\vec{d} \in \vec{\mathcal{W}}_K$ and *origin* $C \in \mathcal{W}_K$, the partition obtained by inversion of the mapping

$$\begin{array}{rcl} \mathrm{E} : \mathcal{W}_K & \longrightarrow & \mathbf{Z} \\ X & \longmapsto & \left\lfloor \left\langle \vec{d}, \vec{CX} \right\rangle \right\rfloor \end{array} \tag{27}$$

where $\lfloor y \rfloor$ is the greatest integer smaller than or equal to y. We denote the partition by $\mathcal{P}_K(\vec{d}; C)$. The cells of $\mathcal{P}_K(\vec{d}; C)$ are separated by the affine hyperplanes

$$\mathcal{H}_i = \left\{ X \in \mathcal{W}_K \middle| \left\langle \vec{d}, \vec{CX} \right\rangle = i \right\}$$

where $i \in \mathbf{Z}$. These hyperplanes have as common direction the space orthogonal to $\vec{d}$. In the particular case where $\vec{d} = \vec{0}$, note that $\mathcal{P}_K(\vec{0}; C) = \{\mathcal{W}_K\}$ for any $C \in \mathcal{W}_K$. When $\vec{d} \neq \vec{0}$, the distance between two consecutive hyperplanes is constant and equal to $q = \frac{1}{\|\vec{d}\|}$.

Proposition A.1: Let $\vec{\mathcal{W}}_L$ be an L dimensional subspace of $\vec{\mathcal{W}}_K$, $\mathcal{W}_L$ be an L-dimensional affine subspace of $\mathcal{W}_K$ of direction $\vec{\mathcal{W}}_L$, and let $C \in \mathcal{W}_K$, $\vec{d} \in \vec{\mathcal{W}}_K$. The partition induced in $\mathcal{W}_L$ by $\mathcal{P}_K(\vec{d}; C)$ is a hyperplane wave partition $\mathcal{P}_L(\vec{d}'; C')$ of $\mathcal{W}_L$, where $\vec{d}' \in \vec{\mathcal{W}}_L$ is the orthogonal projection of $\vec{d} \in \vec{\mathcal{W}}_K$ on $\vec{\mathcal{W}}_L$ and C' is some point of $\mathcal{W}_L$.

Proof: The partition induced by $\mathcal{P}_K(\vec{d}; C)$ in $\mathcal{W}_L$ is obtained by inverse image of the mapping

$$\mathrm{E} : X \mapsto \left\lfloor \left\langle \vec{d}, \vec{CX} \right\rangle \right\rfloor$$

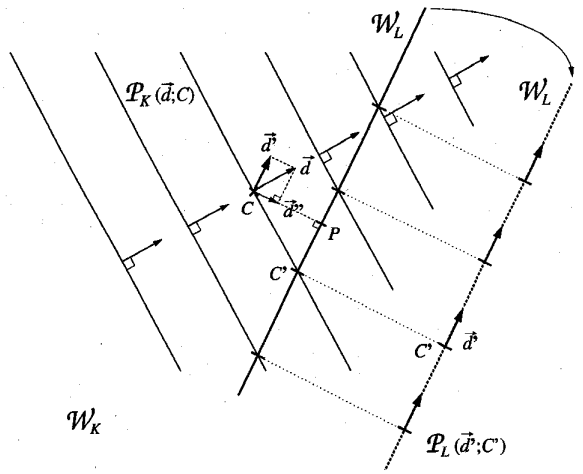

Fig. 9. Partition induced in an affine subspace $\mathcal{W}_L$ by a hyperplane wave partition $\mathcal{P}_K\left(\vec{d}; C\right)$.

restricted to $X \in \mathcal{W}_L$. Let P be the orthogonal projection of C on $\mathcal{W}_L$ and let us write $\vec{d} = \vec{d}' + \vec{d}''$, where $\vec{d}'' \in \vec{\mathcal{W}}_L^{\perp}$ (see Fig. 9). Then

$$\begin{aligned} \forall X \in \mathcal{W}_L, \left\langle \vec{d}, \vec{CX} \right\rangle &= \left\langle \vec{d}' + \vec{d}'', \vec{CP} + \vec{PX} \right\rangle \\ &= \left\langle \vec{d}', \vec{PX} \right\rangle + \left\langle \vec{d}'', \vec{CP} \right\rangle \end{aligned} \tag{28}$$

since $\vec{PX} \in \vec{\mathcal{W}}_L$ and $\vec{CP} \in \vec{\mathcal{W}}_L^{\perp}$.

Suppose $\vec{d}' = \vec{0}$. Then $\forall X \in \mathcal{W}_L$, $\mathrm{E}(X) = k$ where

$$k = \left\lfloor \left\langle \vec{d}'', \vec{CP} \right\rangle \right\rfloor$$

is a fixed integer. The inverse image of E restricted to $\mathcal{W}_L$ leads to the partition $\{\mathcal{W}_L\}$ which is equal to $\mathcal{P}_L(\vec{0}; C')$, where C' can be arbitrarily chosen in $\mathcal{W}_L$.

Suppose $\vec{d}' \neq \vec{0}$. Let C' be the unique point of $\mathcal{W}_L$ such that

$$\vec{PC'} = -\frac{\left\langle \vec{d}'', \vec{CP} \right\rangle}{\left\| \vec{d}' \right\|^2} \vec{d}'.$$

We have

$$\left\langle \vec{d}', \vec{PC'} \right\rangle = -\frac{\left\langle \vec{d}'', \vec{CP} \right\rangle}{\left\| \vec{d}' \right\|^2} \left\langle \vec{d}', \vec{d}' \right\rangle = -\left\langle \vec{d}'', \vec{CP} \right\rangle. \tag{29}$$

Combining (28) and (29), we find

$$\forall X \in \mathcal{W}_L, \left\langle \vec{d}, \vec{CX} \right\rangle = \left\langle \vec{d}', \vec{PX} \right\rangle - \left\langle \vec{d}', \vec{PC'} \right\rangle = \left\langle \vec{d}', \vec{C'X} \right\rangle.$$

Therefore

$$\forall X \in \mathcal{W}_L, \mathrm{E}(X) = \left\lfloor \left\langle \vec{d}', \vec{C'X} \right\rangle \right\rfloor.$$

The inverse image of E restricted to $\mathcal{W}_L$ generates then the partition $\mathcal{P}_L(\vec{d}'; C')$ □

A.1.2) Hyperplane Wave Structured Partition: Let $N \geq 1$ be an integer, $\vec{d}_1, \cdots, \vec{d}_N$ be N vectors of $\vec{\mathcal{W}}_K$ and $C_1, \cdots, C_N$ be N points of $\mathcal{W}_K$. We call the *hyperplane wave structured partition* of *density vectors* $\vec{d}_1, \cdots, \vec{d}_N$ and *origins* $C_1, \cdots, C_N$, the partition obtained by inverse image of the mapping

$$\begin{array}{rcl} \mathrm{E} : \mathcal{W}_K & \longrightarrow & \mathbf{Z}^N \\ X & \longmapsto & (\mathrm{E}_1(X), \cdots, \mathrm{E}_N(X)) \end{array} \tag{30}$$

where

$$\forall k = 1, \cdots, N, \forall X \in \mathcal{W}_K, \mathrm{E}_k(X) = \left\lfloor \left\langle \vec{d}_k, \overrightarrow{C_k X} \right\rangle \right\rfloor.$$

We denote this partition by $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N; C_1, \cdots, C_N)$. As in Section IV, $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N; C_1, \cdots, C_N)$ is obtained by intersection of the hyperplane wave partitions $\mathcal{P}_K(\vec{d}_1; C_1), \cdots, \mathcal{P}_K(\vec{d}_N; C_N)$. As a consequence of Proposition A.1 we have:

Proposition A.2: Let $\mathcal{W}_L$ be an L-dimensional affine subspace of $\mathcal{W}_K$ of direction $\vec{\mathcal{W}}_L$. The partition induced in $\mathcal{W}_L$ by $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N; C_1, \cdots, C_N)$ is a hyperplane wave structured partition $\mathcal{P}_L(\vec{d}'_1, \cdots, \vec{d}'_N; C'_1, \cdots, C'_N)$ where $\vec{d}'_1, \cdots, \vec{d}'_N$ are, respectively, the orthogonal projections of $\vec{d}_1, \cdots, \vec{d}_N$ on $\vec{\mathcal{W}}_L$ and $C'_1, \cdots, C'_N$ are N points of $\mathcal{W}_L$.

A.1.3) Particular Case of Affine Space $\mathcal{W}_K = \vec{\mathcal{W}}_K$: The definitions of hyperplane wave partition and hyperplane wave structured partition apply to the particular case where the affine space is $\vec{\mathcal{W}}_K$ itself. Indeed, $\vec{\mathcal{W}}_K$ is an affine space of direction $\vec{\mathcal{W}}_K$ when the mapping (23) is defined by

$$\begin{array}{rcl} \vec{\mathcal{W}}_K \times \vec{\mathcal{W}}_K & \longrightarrow & \vec{\mathcal{W}}_K \\ (\vec{a}, \vec{b}) & \longmapsto & \vec{v} = \vec{b} - \vec{a} \end{array} \tag{31}$$

One can check that this mapping satisfies (24)–(26). In this situation, a hyperplane wave partition is characterized by a density vector $\vec{d}$ of $\vec{\mathcal{W}}_K$ (as vector space) and an origin $\vec{c}$ of $\vec{\mathcal{W}}_K$ (as affine space), is obtained by inversion of the mapping

$$\begin{array}{rcl} \mathrm{E} : \vec{\mathcal{W}}_K & \longrightarrow & \mathbf{Z} \\ \vec{x} & \longmapsto & \left\lfloor \left\langle \vec{d}, \vec{x} - \vec{c} \right\rangle \right\rfloor \end{array} \tag{32}$$

and is denoted by $\mathcal{P}_K(\vec{d}; \vec{c})$. Similarly, a hyperplane wave structured partition is characterized by N density vectors $\vec{d}_1, \cdots, \vec{d}_N \in \vec{\mathcal{W}}_K$ and N origins $\vec{c}_1, \cdots, \vec{c}_N \in \vec{\mathcal{W}}_K$, and is denoted by $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N; \vec{c}_1, \cdots, \vec{c}_N)$. The main section deals with this present case where $\mathcal{W}_K = \vec{\mathcal{W}}_K = \mathbf{R}^K$ and uses the presentation of (32) in the case where $\vec{c} = \vec{0}$. The notation of the main section $\mathcal{P}_K(\vec{d})$ corresponds here to $\mathcal{P}_K(\vec{d}; \vec{0})$ and $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N)$ corresponds here to $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N; \vec{0}, \cdots, \vec{0})$.

A.2. Number of Cells Induced in a Sphere in Finite Dimension

We have already introduced in Section V the number $N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N)$ that we redefine here in the general context of affine spaces.

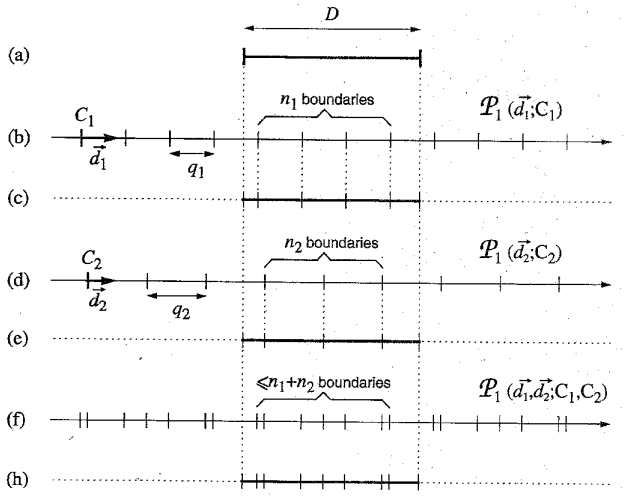

Fig. 10. Construction of the partition induced in a segment by a one-dimensional hyperplane wave structured partition.

Definition A.3: Let $\vec{d}_1, \cdots, \vec{d}_N$ be N vectors of $\vec{\mathcal{W}}_K$. The number $N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N)$ is the maximum number of cells that a partition $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N; C_1, \cdots, C_N)$ of arbitrary origins $C_1, \cdots, C_N$ can induce in a K-dimensional sphere of $\mathcal{W}_K$ of diameter D and arbitrary origin.

We have several remarks about this number. For a given choice of points $C_1, \cdots, C_N$ and sphere of diameter D, the number of cells induced by $\mathcal{P}_K(\vec{d}_1, \cdots, \vec{d}_N; C_1, \cdots, C_N)$ in the sphere is not necessarily equal to $N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N)$, but always upper-bounded by it. However, one can prove that there always exists a choice of points $C_1, \cdots, C_N$ and sphere of diameter D such that the induced number of cells is equal to $N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N)$.

From Proposition A.5 to Theorem A.7, we propose upper bounds on $N_{K,D}(\vec{d}_1, \cdots, \vec{d}_N)$ in various cases.

Proposition A.4: In the case $K = 1$ and $N \geq 1$

$$N_{1,D}(\vec{d}_1, \cdots, \vec{d}_N) \leq 1 + \sum_{k=1}^{N} \left\lceil D \|\vec{d}_k\| \right\rceil. \tag{33}$$

Proof: When $K = 1$, spheres of diameter D are simply segments of the real line of length D (see Fig. 10(a)). Each partition $\mathcal{P}_1(\vec{d}_k; C_k)$ such that $\vec{d}_k \neq \vec{0}$ divides the real line into intervals whose boundaries are equally spaced by $q_k = \frac{1}{\|\vec{d}_k\|}$. The number n_k of boundaries which can be found in a given segment of length D (Fig. 10(b)) is necessarily upper bounded as

$$n_k \leq \left\lceil \frac{D}{q_k} \right\rceil$$

which leads to

$$n_k \leq \left\lceil D \|\vec{d}_k\| \right\rceil. \tag{34}$$

When $\vec{d}_k = \vec{0}$, the whole real axis $\mathbf{R}$ is the unique cell of $\mathcal{P}_1(\vec{d}_k; C_k)$ and the number of boundaries induced in a segment is $n_k = 0$. Therefore, the inequality (34) is also true when $\vec{d}_k = \vec{0}$. Now, by intersection, the complete

partition $\mathcal{P}_1(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ divides the real line into intervals whose boundaries are the reunion of the boundaries of $\mathcal{P}_1(\vec{d}_k;C_k)$ from $k=1$ to N (see Fig. 10(f) for the case $N=2$). Therefore, the number of boundaries found in a segment of length D is less than

$$\sum_{k=1}^{N} n_k.$$

Then, the number of cells induced by $\mathcal{P}_1(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ in the segment is less than

$$1+\sum_{k=1}^{N} n_k$$

(Fig. 10(h)). This leads to (33) by using (34) □

Proposition A.5: In the case $K\geq 1$ and $N=K$

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_K)\leq \prod_{k=1}^{K}\left(\left\lceil D\|\vec{d}_k\|\right\rceil+1\right). \tag{35}$$

Proof: Let us consider the kth partition $\mathcal{P}_K(\vec{d}_k;C_k)$. The cells of this partition are separated by hyperplanes perpendicular to $\vec{d}_k$ and equally spaced by $q_k=\frac{1}{\|\vec{d}_k\|}$ (when $\vec{d}_k\neq\vec{0}$) (see Fig. 11). Let n_k be the number of hyperplanes which cut the sphere. The number n_k has the same upper bound

$$n_k\leq\left\lceil D\|\vec{d}_k\|\right\rceil, \tag{36}$$

as in (34) (see Fig. 11). Consequently, the partition $\mathcal{P}_K(\vec{d}_k;C_k)$ induces n_k+1 cells in the sphere. Consider now the full partition $\mathcal{P}_K\left(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N\right)$. Since this is the intersection of all the hyperplane wave partitions

$$\mathcal{P}_K\left(\vec{d}_k;C_k\right)\text{ for }k=1,\cdots,N$$

the number of cells of $\mathcal{P}_K(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ intersecting the sphere is upper-bounded by

$$\prod_{k=1}^{N}(n_k+1).$$

Then, using (36), we obtain

$$\begin{aligned}N_{K,D}(\vec{d}_1,\cdots,\vec{d}_K)&\leq\prod_{k=1}^{N}(n_k+1)\\&\leq\prod_{k=1}^{K}\left(\left\lceil D\|\vec{d}_k\|\right\rceil+1\right).\end{aligned}$$ □

The next proposition gives an inequality in terms of $N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N)$ which is recursive on K and N.

Proposition A.6: Let $K\geq 1$, $N\geq 1$, and $\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1}$ be $(N+1)$ density vectors of $\mathbf{R}^K$. Then

$$\begin{aligned}&N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1})\\&\leq N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N)+\left\lceil D\|\vec{d}_{N+1}\|\right\rceil\cdot N_{K-1,D}(\vec{d}'_1,\cdots,\vec{d}'_N)\end{aligned} \tag{37}$$

where $\vec{d}'_1,\cdots,\vec{d}'_N$ are the orthogonal projections of $\vec{d}_1,\cdots,\vec{d}_N$ on $<\vec{d}_{N+1}>^{\perp}$ (the space orthogonal to $\vec{d}_{N+1}$).

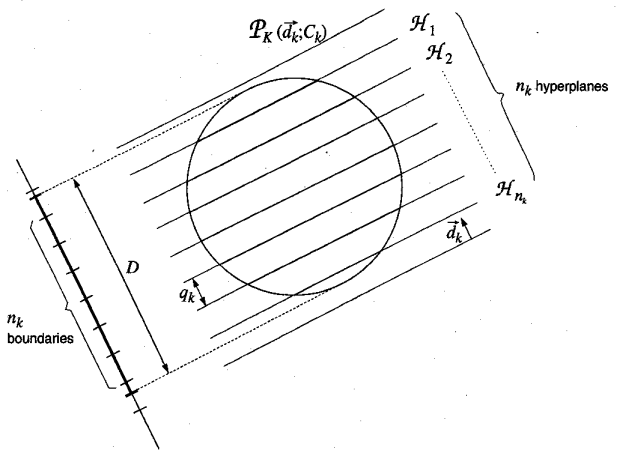

Fig. 11. Partition induced in a sphere by a hyperplane wave partition.

Proof: Let us consider $K\geq 1$ and $N\geq 1$. There exists a set of origins $(C_k)_{1\leq k\leq N+1}$ and a sphere $\mathcal{S}_{K,D}$ of dimension K and diameter D such that the number of cells induced by

$$\mathcal{P}_K\left(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1};C_1,\cdots,C_N,C_{N+1}\right)$$

in $\mathcal{S}_{K,D}$ is equal to the maximum number

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_{N+1}).$$

Let us fix this choice of origins and sphere. The partition

$$\mathcal{P}_K\left(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1};C_1,\cdots,C_N,C_{N+1}\right)$$

can be obtained by intersection of

$$\mathcal{P}_K\left(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N\right)$$

and

$$\mathcal{P}_K\left(\vec{d}_{N+1};C_{N+1}\right).$$

The first partition $\mathcal{P}_K(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ induces a certain number M_N of cells in $\mathcal{S}_{K,D}$ which necessarily satisfies

$$M_N\leq N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N) \tag{38}$$

by definition of $N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N)$ (Definition A.3). As already explained in the proof of Proposition A.5, the second partition $\mathcal{P}_K(\vec{d}_{N+1};C_{N+1})$ cuts the sphere with n_{N+1} hyperplanes perpendicular to $\vec{d}_{N+1}$ such that

$$n_{N+1}\leq\left\lceil D\|\vec{d}_{N+1}\|\right\rceil. \tag{39}$$

Let us call $\mathcal{H}_j$ these hyperplanes, from $j=1$ to n_{N+1}. They are all affine spaces of direction $<\vec{d}_{N+1}>^{\perp}$. To count the total number of cells $N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1})$, induced by

$$\mathcal{P}_K\left(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1};C_1,\cdots,C_N,C_{N+1}\right)$$

in $\mathcal{S}_{K,D}$, let us consider the following scheme: start from the partition $\mathcal{P}_K(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ which already induces M_N cells in $\mathcal{S}_{K,D}$, and count from $j=1$ to n_{N+1} the additional number of cells m_j induced in $\mathcal{S}_{K,D}$ obtained when

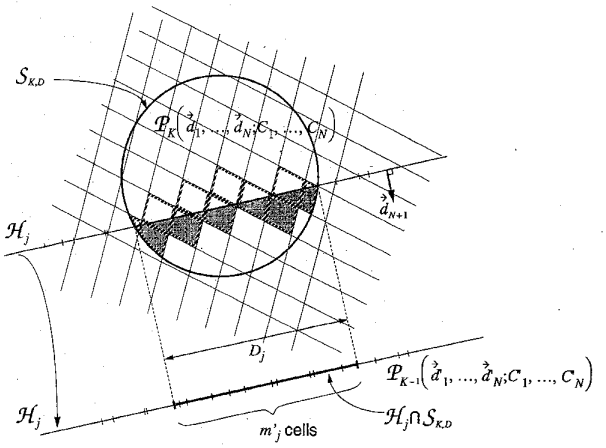

Fig. 12. Representation of the increased number of cells induced in a sphere due to the insertion of a hyperplane $\mathcal{H}_j$. The cells of the partition $\mathcal{P}_K(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ which are intersected by $\mathcal{H}_j$ are emphasized by a dashed contour. Their number is equal to m'_j.

inserting $\mathcal{H}_j$ respectively and successively. By this procedure we obtain

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1}) = M_N + \sum_{j=1}^{n_{N+1}} m_j. \tag{40}$$

Let us show that

$$m_j \le N_{K-1,D}(\vec{d}'_1,\cdots,\vec{d}'_N). \tag{41}$$

Fig. 12 shows what happens when inserting the hyperplane $\mathcal{H}_j$. A certain number of cells induced by $\mathcal{P}_K(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ in $\mathcal{S}_{K,D}$ are intersected by $\mathcal{H}_j$. They are represented by a shaded contour in Fig. 12. Let m'_j be their number. When inserting $\mathcal{H}_j$, these m'_j cells are cut into two parts (the bottom part of each split cell is emphasized by a shaded area). This implies that the total number of cells which already exist in the sphere has been increased by

$$m_j = m'_j. \tag{42}$$

Now, the intersection of these m'_j cells with $\mathcal{H}_j$ form m'_j cells of $\mathcal{H}_j \cap \mathcal{S}_{K,D}$ (see Fig. 12). These cells can be interpreted as follows. Because $\mathcal{H}_j$ is an affine subspace of $\mathcal{W}_K$ of dimension $K-1$, we know from Proposition A.2 that the partition induced by $\mathcal{P}_K(\vec{d}_1,\cdots,\vec{d}_N;C_1,\cdots,C_N)$ in $\mathcal{H}_j$ is a hyperplane wave structured partition $\mathcal{P}_{K-1}(\vec{d}'_1,\cdots,\vec{d}'_N;C'_1,\cdots,C'_N)$ where $\vec{d}'_1,\cdots,\vec{d}'_N$ are the orthogonal projections of $\vec{d}_1,\cdots,\vec{d}_N$ on $\vec{\mathcal{H}}_j =< \vec{d}_{N+1} >^{\perp}$ and $C'_1,\cdots,C'_N$ are N points of $\mathcal{H}_j$. The m'_j cells of $\mathcal{H}_j \cap \mathcal{S}_{K,D}$ are simply the cells induced by $\mathcal{P}_K(\vec{d}'_1,\cdots,\vec{d}'_N;C'_1,\cdots,C'_N)$ in $\mathcal{H}_j \cap \mathcal{S}_{K,D}$. The set $\mathcal{H}_j \cap \mathcal{S}_{K,D}$ is necessarily a $K-1$-dimensional sphere of a certain diameter $D_j \le D$. Therefore, $\mathcal{H}_j \cap \mathcal{S}_{K,D}$ is included in a $K-1$-dimensional sphere of diameter D. We conclude that, by necessity

$$m'_j \le N_{K-1,D}(\vec{d}'_1,\cdots,\vec{d}'_N). \tag{43}$$

Then (41) is a consequence of (42) and (43).

We can now use (40), (38), and (41) and find

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1}) \le N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N) + n_{N+1}\cdot N_{K-1,D}(\vec{d}'_1,\cdots,\vec{d}'_N)$$

which leads to (37) using (39) □

Theorem A.7: Let $K \ge 1$ and $N \ge K$. Let $\vec{d}_1,\cdots,\vec{d}_N$ be N vectors of $\vec{\mathcal{W}}_K$ and d be an upper bound on $\|\vec{d}_k\|$ for $k = 1,\cdots,N$. Then

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N) \le \binom{N}{K}(\lceil Dd\rceil + 1)^K. \tag{44}$$

Proof: We prove the proposition by double induction on K and N.

At $K = 1$, we have from Proposition A.4, $\forall N \ge 1$

$$N_{1,D}(\vec{d}_1,\cdots,\vec{d}_N) \le 1 + \sum_{k=1}^{N}\left\lceil D\|\vec{d}_k\|\right\rceil \le 1 + N\lceil Dd\rceil \le \binom{N}{1}(\lceil Dd\rceil + 1).$$

Therefore, (44) is proved for $K = 1$ and any $N \ge 1$.

Suppose that we have proved (44) at some $K-1 \ge 1$ and any $N \ge K-1$. Let us prove (44) at K by induction on $N \ge K$. The case $N = K$ is proved from Proposition A.5 as follows:

$$N_{K,D}\left(\vec{d}_1,\cdots,\vec{d}_K\right) \le \prod_{k=1}^{K}\left(\left\lceil D\|\vec{d}_k\|\right\rceil + 1\right) \le (\lceil Dd\rceil + 1)^K \le \binom{K}{K}(\lceil Dd\rceil + 1)^K.$$

Now, suppose we know that (44) is true at some $N \ge K$. Let $\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1}$ be $(N+1)$ density vectors with $\|\vec{d}_k\| \le d$, for all $k = 1,\cdots,N$. By recursive assumption on N we have

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N) \le \binom{N}{K}(\lceil Dd\rceil + 1)^K. \tag{45}$$

Let $\vec{d}'_k$ be the orthogonal projection of $\vec{d}_k$ on $< \vec{d}_{N+1} >^{\perp}$, for $k = 1,\cdots,N$. We necessarily have

$$\left\|\vec{d}'_k\right\| \le \left\|\vec{d}_k\right\| \le d.$$

Therefore, using the recursive assumption at $K-1$, we have

$$N_{K-1,D}(\vec{d}'_1,\cdots,\vec{d}'_N) \le \binom{N}{K-1}(\lceil Dd\rceil + 1)^{K-1}. \tag{46}$$

Then applying Proposition A.6 with (45) and (46), we find

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1}) \le \binom{N}{K}(\lceil Dd\rceil + 1)^K + \left\lceil D\|\vec{d}_{N+1}\|\right\rceil\binom{N}{K-1}(\lceil Dd\rceil + 1)^{K-1}.$$

However

$$\left\lceil D\|\vec{d}_{N+1}\|\right\rceil \le \lceil Dd\rceil \le \lceil Dd\rceil + 1.$$

Therefore

$$N_{K,D}(\vec{d}_1,\cdots,\vec{d}_N,\vec{d}_{N+1}) \le \left(\binom{N}{K}+\binom{N}{K-1}\right)(\lceil Dd\rceil+1)^K \le \binom{N+1}{K}(\lceil Dd\rceil+1)^K.$$

Therefore (44) is true at $N+1$. The double induction is completed □

Theorem 5.1 is the particular case of Theorem A.7 where the affine space $\mathcal{W}_K$ is equal to the vector space $\mathbf{R}^K$.

REFERENCES

[1] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, July 1948.

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

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

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

[5] P. L. Zador, "Development and evaluation of procedures for quantizing multivariate distributions," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1963.

[6] ——, "Asymptotic quantization error of continuous signals and their quantization dimension," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 139–149, Mar. 1982.

[7] J. H. Conway and N. J. A.Sloane, *Sphere Packings, Lattices and Groups.* New York: Springler-Verlag, 1988.

[8] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression.* Norwell, MA: Kluwer, 1992.