Least squares quantization in PCM

Stuart Lloyd

Least Squares Quantization in PCM

STUART P. LLOYD

Abstract—**It has long been realized that in pulse-code modulation (PCM), with a given ensemble of signals to handle, the quantum values should be spaced more closely in the voltage regions where the signal amplitude is more likely to fall. It has been shown by Panter and Dite that, in the limit as the number of quanta becomes infinite, the asymptotic fractional density of quanta per unit voltage should vary as the one-third power of the probability density per unit voltage of signal amplitudes. In this paper the corresponding result for any finite number of quanta is derived; that is, necessary conditions are found that the quanta and associated quantization intervals of an optimum finite quantization scheme must satisfy. The optimization criterion used is that the average quantization noise power be a minimum. It is shown that the result obtained here goes over into the Panter and Dite result as the number of quanta become large. The optimum quantization schemes for 2^b quanta, $b = 1, 2, \cdots, 7$, are given numerically for Gaussian and for Laplacian distribution of signal amplitudes.**

I. Introduction

THE BASIC IDEAS in the pulse-code modulation (PCM) system [1], [2, ch. 19] are the Shannon–Nyquist sampling theorem and the notion of quantizing the sample values.

The sampling theorem asserts that a signal voltage $s(t)$, $-\infty < t < \infty$, containing only frequencies less than W cycles/s can be recovered from a sequence of its sample values according to

$$s(t) = \sum_{j=-\infty}^{\infty} s(t_j)K(t - t_j), \qquad -\infty < t < \infty, \quad (1)$$

where $s(t_j)$ is the value of s at the jth sampling instant

$$t_j = \frac{j}{2W}, \qquad -\infty < j < \infty,$$

and where

$$K(t) = \frac{\sin 2\pi Wt}{2\pi Wt}, \qquad -\infty < t < \infty, \quad (2)$$

is a "sin t/t" pulse of the appropriate width.

The pulse-amplitude modulation (PAM) system [2, ch. 16] is based on the sampling theorem alone. One sends over the system channel, instead of the signal values $s(t)$ for all times t, only a sequence

$$\cdots, s(t_{-1}), s(t_0), s(t_1), \cdots \quad (3)$$

of samples of the signal. The (idealized) receiver constructs the pulses $K(t - t_j)$ and adds them together with the received amplitudes $s(t_j)$, as in (1), to produce an exact reproduction of the original band-limited signal s.

PCM is a modification of this. Instead of sending the exact sample values (3), one partitions the voltage range of the signal into a finite number of subsets and transmits to the receiver only the information as to which subset a sample happens to fall in. Built into the receiver there is a source of fixed representative voltages—"quanta"—one for each of the subsets. When the receiver is informed that a certain sample fell in a certain subset, it uses its quantum for that subset as an approximation to the true sample value and constructs a band-limited signal based on these approximate sample values.

We define the *noise signal* as the difference between the receiver-output signal and the original signal and the *noise power* as the average square of the noise signal. The problem we consider is the following: given the number of quanta and certain statistical properties of the signal, determine the subsets and quanta that are best in minimizing the noise power.

II. Quantization

Let us formulate the quantization process more explicitly. A quantization scheme consists of a class of sets $\{Q_1, Q_2, \cdots, Q_\nu\}$ and a set of quanta $\{q_1, q_2, \cdots, q_\nu\}$. The $\{Q_\alpha\}$ are any ν disjoint subsets of the voltage axis which, taken together, cover the entire voltage axis. The $\{q_\alpha\}$ are any ν finite voltage values. The number ν of quanta is to be regarded throughout as a fixed finite preassigned number.

We associate with a partition $\{Q_\alpha\}$ a label function $\gamma(x)$, $-\infty < x < \infty$, defined for all (real) voltages x by

$$\begin{aligned} \gamma(x) &= 1 \quad \text{if} \quad x \quad \text{lies in} \quad Q_1, \\ \gamma(x) &= 2 \quad \text{if} \quad x \quad \text{lies in} \quad Q_2, \\ &\vdots \\ \gamma(x) &= \nu \quad \text{if} \quad x \quad \text{lies in} \quad Q_\nu. \end{aligned} \quad (4)$$

If $s(t_j)$ is the jth sample of the signal s, as in Section I, then we denote by a_j the label of the set that this sample falls in:

$$a_j = \gamma\big(s(t_j)\big), \qquad -\infty < j < \infty.$$

In PCM the signal sent over the channel is (in some code or another) the sequence of labels

$$\cdots, a_{-1}, a_0, a_1, \cdots, \quad (5)$$

each a_j being one of the integers $\{1, 2, \cdots, \nu\}$. The technology of this transmission does not concern us, except that

The author is with Bell Laboratories, Whippany Road, Whippany, NJ 07981.

we assume that such a sequence can be delivered to the receiver without error.

The receiver uses the fixed voltage q_α as an approximation to all sample voltages in Q_α, $\alpha = 1, 2, \cdots, \nu$. That is, the receiver, being given the value of a_j in the sequence (5), proceeds as if the jth sample of s had value q_{a_j} and produces the receiver-output signal

$$r(t) = \sum_{j=-\infty}^{\infty} q_{a_j} K(t - t_j), \qquad -\infty < t < \infty.$$

To put it another way, the system mutilates an actual sample voltage value x to the quantized value $y(x)$ given by

$$y(x) = q_{\gamma(x)}, \qquad -\infty < x < \infty, \tag{6}$$

and we may express the receiver output in terms of this as

$$r(t) = \sum_{j=-\infty}^{\infty} y\big(s(t_j)\big) K(t - t_j), \qquad -\infty < t < \infty. \tag{7}$$

Hence the noise signal, defined as

$$n(t) = r(t) - s(t), \qquad -\infty < t < \infty,$$

is given by

$$n(t) = \sum_{j=-\infty}^{\infty} z\big(s(t_j)\big) K(t - t_j), \qquad -\infty < t < \infty, \tag{8}$$

where

$$z(x) = y(x) - x, \qquad -\infty < x < \infty, \tag{9}$$

may be regarded as the quantization error added to a sample which has voltage value x.

Note that we assume that the receiver uses the nonrealizable pulses (2). If other pulses are used (e.g., step functions or other realizable pulses) there will be sampling noise, in general, even without quantization [3]. Our noise (8) is due strictly to quantization.

Finally we must emphasize that we assume that the $\{Q_\alpha\}$ and $\{q_\alpha\}$ are constant in time. In deltamodulation and its refinements the $\{Q_\alpha\}$ and $\{q_\alpha\}$ change from sampling instant to sampling instant, depending on the past behavior of the signal being handled. Such systems are very difficult to treat theoretically.

III. Noise Power

Instead of working with a particular band-limited signal, we assume that there is given a probabilistic family of such signals. That is, the s of the preceding sections and hence the various signals derived from it are to be regarded as stochastic processes [4]. We denote the underlying probability measure by $P\{\cdot\}$ and averages with respect to this measure (expectations) by $E\{\cdot\}$.

We use the following results of the probabilistic treatment. We assume that the s process is stationary, so that the cumulative probability distribution function of a sample,

$$F(x) = P\{s(t) \le x\}, \qquad -\infty < x < \infty,$$

is independent of t, $-\infty < t < \infty$, as indicated by the notation. Then the average power of the s process, assumed to be finite, is constant in time:

$$S = E\{s^2(t)\} = \int_{-\infty}^{\infty} x^2 \, dF(x), \qquad -\infty < t < \infty. \tag{10}$$

Moreover, the r and n processes have this same property; the average receiver-output power R is given by

$$R = E\{r^2(t)\} = \int_{-\infty}^{\infty} y^2(x) \, dF(x), \qquad -\infty < t < \infty, \tag{11}$$

where $y(x)$ is defined in (6), and the noise power N is

$$N = E\{n^2(t)\} = \int_{-\infty}^{\infty} z^2(x) \, dF(x), \qquad -\infty < t < \infty, \tag{12}$$

with $z(x)$ as in (9). (Detailed proofs of these statements, together with further assumptions used, are given in Appendix A.) The stochastic process problem is thus reduced to a problem in a single real variable: choose the $\{Q_\alpha\}$ and $\{q_\alpha\}$ so that the rightmost integral in (12) is as small as possible.

IV. The Best Quanta

We consider first the problem of minimizing N with respect to the quanta $\{q_\alpha\}$ when the $\{Q_\alpha\}$ are fixed preassigned sets.

The dF integral in (12) may be written more explicitly as

$$N = \sum_{\alpha=1}^{\nu} \int_{Q_\alpha} (q_\alpha - x)^2 \, dF(x). \tag{13}$$

(The sets $\{Q_\alpha\}$ must be measurable $[dF]$ if (11)–(13) are to have meaning, and we assume always that this is the case.) If we regard the given F as describing the distribution of unit probability "mass" on the voltage axis [5, p. 57], then (13) expresses N as the total "moment of inertia" of the sets $\{Q_\alpha\}$ around the respective points $\{q_\alpha\}$. It is a classical result that such a moment assumes its minimum value when each $\{q_\alpha\}$ is the center of mass of the corresponding $\{Q_\alpha\}$ (see, e.g., [5, p. 175]). That is,

$$q_\alpha = \frac{\int_{Q_\alpha} x \, dF(x)}{\int_{Q_\alpha} dF(x)}, \qquad \alpha = 1, 2, \cdots, \nu, \tag{14}$$

are the uniquely determined best quanta to use with a given partition $\{Q_\alpha\}$.

To avoid the continual mention of trivial cases we assume always that F is increasing at least by $\nu + 1$ points, so that the quantization noise does not vanish. Then none of the denominators in (14) will vanish, at least in an

optimum scheme. For if Q_α has vanishing mass it can be combined with some set Q_β of nonvanishing mass (discarding q_α) to give a scheme with $\nu - 1$ quanta and the same noise. Then one of the sets of this scheme can be divided into two sets and new quanta assigned to give a scheme with ν quanta and noise less than in the original scheme. (We omit the details.)

If the expression on the right in (14) is substituted for q_α in (13), there results

$$N = S - \sum_{\alpha=1}^{\nu} q_\alpha^2 \int_{Q_\alpha} dF(x),$$

where the $\{q_\alpha\}$ here are the optimum ones of (14). The sum on the right is the receiver-output power from (11). Hence when the $\{q_\alpha\}$ are centers of mass of the $\{Q_\alpha\}$, optimum or not, then $S = R + N$, which implies that the noise is orthogonal to the receiver output. One expects this in a least squares approximation, of course.

V. The Best Partition

Now we find the best sets $\{Q_\alpha\}$ to use with a fixed preassigned set of quanta $\{q_\alpha\}$. The considerations of this section are independent of those of the preceding section. In particular, the best $\{Q_\alpha\}$ for given $\{q_\alpha\}$ may not have the $\{q_\alpha\}$ as their centers of mass.

We assume that the given $\{q_\alpha\}$ are distinct since it will never happen in an optimum scheme that $q_\alpha = q_\beta$ for some $\alpha \neq \beta$. For if $q_\alpha = q_\beta$, then Q_α and Q_β are effectively one set $Q_\alpha \cup Q_\beta$ as far as the noise is concerned (13), and this set can be redivided into two sets and these two sets can be given distinct quantum values in such a way as to reduce the noise. (We omit the details.)

Consider the probability mass in a small interval around voltage value x. According to (13) any of this mass which is assigned to q_α (i.e., which lies in Q_α) will contribute to the noise at rate $(q_\alpha - x)^2$ per unit mass. To minimize the noise, then, any mass in the neighborhood of x should be assigned to a q_α for which $(q_\alpha - x)^2$ is the smallest of the numbers $(q_1 - x)^2$, $(q_2 - x)^2, \cdots, (q_\nu - x)^2$. In other words,

$$Q_\alpha \supset \left\{x : (q_\alpha - x)^2 < (q_\beta - x)^2 \text{ for all } \beta \neq \alpha\right\}, \quad \alpha = 1,, \cdots, \nu,$$

modulo sets of measure zero $[dF]$.[1] This simplifies to

$$Q_\alpha \supset \left\{x : (q_\beta - q_\beta)\left(x - \tfrac{1}{2}(q_\alpha + q_\beta)\right) < 0 \text{ for all } \beta \neq \alpha\right\}, \quad \alpha = 1, 2, \cdots, \nu. \quad (15)$$

It is straightforward that the best $\{Q_\alpha\}$ are determined by (15) as the intervals whose endpoints bisect the segments between successive $\{q_\alpha\}$, except that the assignment of the endpoints is not determined. To make matters definite we let the $\{Q_\alpha\}$ be left-open and right-closed, so that the best partition to use with the given quanta is

$$\begin{aligned} Q_1 &= \{x : -\infty < x \leq x_1\} \\ Q_2 &= \{x : x < x \leq x_2\} \\ &\vdots \\ Q_{\nu-1} &= \{x : x_{\nu-2} < x \leq x_{\nu-1}\} \\ Q_\nu &= \{x : x_{\nu-1} < x < \infty\}, \end{aligned} \quad (16)$$

where the endpoints $\{x_\alpha\}$ are given

$$\begin{aligned} x_1 &= \tfrac{1}{2}(q_1 + q_2) \\ x_2 &= \tfrac{1}{2}(q_2 + q_3) \\ &\vdots \\ x_{\nu-1} &= \tfrac{1}{2}(q_{\nu-1} + q_\nu). \end{aligned} \quad (17)$$

We have assumed, as we shall hereafter, that the indexing is such that $q_1 < q_2 < \cdots < q_\nu$.

VI. Quantization Procedures

From Sections IV and V we know that we may confine our attention to quantization schemes defined by $2\nu - 1$ numbers

$$q_1 < x_1 < q_2 < x_2 < \cdots < q_{\nu-1} < x_{\nu-1} < q_\nu, \quad (18)$$

where the $\{x_\alpha\}$ are the endpoints of the intervals $\{Q_\alpha\}$, as in (16), and the $\{q_\alpha\}$ are the corresponding quanta. We will regard such a set of numbers as the Cartesian coordinates of a point

$$\rho = (q_1, x_1, \cdots, q_\nu)$$

in $(2\nu - 1)$-dimensional Euclidean space $E_{2\nu-1}$. The noise as a function of ρ has the form

$$N(\rho) = \int_{-\infty}^{x_1} (q_1 - x)^2 \, dF(x) + \int_{x_1}^{x_2} (q_2 - x)^2 \, dF(x) + \cdots + \int_{x_{\nu-1}}^{\infty} (q_\nu - x)^2 \, dF(x). \quad (19)$$

In an optimum scheme the $\{q_\alpha\}$ will be centers of mass of the corresponding $\{Q_\alpha\}$, (14), and the $\{x_\alpha\}$ will lie midway between adjacent $\{q_\alpha\}$, (17). From the derivations these conditions are sufficient that $N(\rho)$ be a minimum with respect to variations in each coordinate separately and hence are necessary conditions at a minimum of $N(\rho)$. As it turns out, however, they are not sufficient conditions for a minimum of $N(\rho)$. Points at which (14) and (17) are satisfied, which we term *stationary points*, while never local maxima, may be saddle points of $N(\rho)$. Moreover, among the stationary points there may be several local minima, only one of which is the sought absolute minimum of $N(\rho)$. These complications are discussed further in Appendix B. The author has not been able to determine sufficient conditions for an absolute minimum.

The derivations suggest one trial-and-error method for finding stationary points. A trial point $\rho^{(1)}$ in $E_{2\nu-1}$ is

[1] If $C(x)$ is a condition on x, then $\{x: C(x)\}$ denotes the set of all x which satisfy $C(x)$.

chosen as follows. The endpoints

$$-\infty < x_1^{(1)} < x_2^{(1)} < \cdots < x_{\nu-1}^{(1)} < \infty$$

are chosen arbitrarily except that each of the resulting $\{Q_\alpha^{(1)}\}$ should have nonvanishing mass. Then the centers of mass of these sets are taken as the first trial quanta $\{q_\alpha^{(1)}\}$.

These values will not satisfy the midpoint conditions (17), in general, so that the second trial point $\rho^{(2)}$ is taken to be

$$q_\alpha^{(2)} = q_\alpha^{(1)}, \qquad \alpha = 1,2,\cdots,\nu$$
$$x_\alpha^{(2)} = \tfrac{1}{2}\left(q_\alpha^{(2)} + q_{\alpha+1}^{(2)}\right), \qquad \alpha = 1,2,\cdots,\nu-1,$$

with appropriate modifications if any of the resulting $\{Q_\alpha^{(2)}\}$ have vanishing mass. This step does not increase the noise, in view of the discussion in Section V; that is, $N(\rho^{(2)}) \leq N(\rho^{(1)})$.

The new $\{q_\alpha^{(2)}\}$, centers of mass (c.m.) of the old $\{Q_\alpha^{(1)}\}$, will not be centers of mass of the new $\{Q_\alpha^{(2)}\}$, in general; trial point $\rho^{(3)}$ is determined by

$$x_\alpha^{(3)} = x_\alpha^{(2)}, \qquad \alpha = 1,2,\cdots,\nu-1,$$
$$q_\alpha^{(3)} = \left(\text{c.m. of } Q_\alpha^{(3)}\right), \qquad \alpha = 1,2,\cdots,\nu.$$

For the resulting noise we have $N(\rho^{(3)}) \leq N(\rho^{(2)})$.

We continue in this way, imposing conditions (14) and (17) alternately. There results a sequence of trial points

$$\rho^{(1)}, \rho^{(2)}, \cdots \tag{20}$$

such that

$$N\left(\rho^{(1)}\right) \geq N\left(\rho^{(2)}\right) \geq \cdots.$$

The noise is nonnegative, so that $\lim_m N(\rho^{(m)})$ will exist, and we might hope that the sequence (20) had as a limit a local minimum of $N(\rho)$.

If the sequence (20) has no limit points then some of the $\{x_\alpha^{(m)}\}$ must become infinite with m; this corresponds to quantizing into fewer than ν quanta. Since we have assumed that F increases at least by $\nu+1$ points there will be quantizing schemes with ν quanta for which the resulting noise is less than the optimum noise for $\nu-1$ quanta, obviously. If $\rho^{(1)}$ is such a scheme then (20) will have limit points, using the property that $N(\rho^{(m)})$ is a decreasing sequence.[2]

Suppose $\rho^{(\infty)}$ is such a limit point. If each of the coordinate values $\{x_\alpha^{(\infty)}\}$ of $\rho^{(\infty)}$ is a continuity point of F then it is easy to see that the coordinates of $\rho^{(\infty)}$ will satisfy both (14) and (17). In particular, if $N(\rho)$ has a unique stationary point ρ_0 (which is the minimum sought), then the sequence (20), unless it diverges, will converge to ρ_0.

Note, by the way, that at a local minimum of $N(\rho)$ the numbers $\{x_\alpha\}$ are necessarily continuity points of F. Suppose to the contrary that there is a nonvanishing amount of mass concentrated at one of the endpoints $\{x_\alpha\}$, and that the adjacent sets Q_α and $Q_{\alpha+1}$ are as in (16), so that the mass at x_α belongs to Q_α. The centers of mass q_α and $q_{\alpha+1}$ will lie equidistant from x_α (17), and from (19) the noise will not change if we reassign the mass at x_α to $Q_{\alpha+1}$, retaining the given $\{q_\alpha\}$ as quanta. But q_α and $q_{\alpha+1}$ are definitely not centers of mass of the corresponding modified sets, and the noise will strictly decrease as q_α and $q_{\alpha+1}$ are moved to the new centers of mass. Thus the given configuration is not a local minimum, contrary to assumption. From this result and (19) we see that $N(\rho)$ is continuous in a neighborhood of a local minimum. We have proved also that there is no essential loss of generality in assuming the form (16) for the $\{Q_\alpha\}$.

We refer to the above trial-and-error method as Method I. Another trial-and-error method is the following one, Method II. To simplify the discussion we assume for the moment that F is continuous and nowhere constant. We choose a trial value q_1 satisfying

$$q_1 < \int_{-\infty}^{\infty} x\, dF(x).$$

The condition that q_1 be the center of mass of Q_1 determines x_1 as the unique solution of

$$q_1 = \frac{\int_{-\infty}^{x_1} x\, dF(x)}{\int_{-\infty}^{x_1} dF(x)}.$$

The quantities q_1 and x_1 now being known, the first of conditions (17) determines q_2 as

$$q_2 = 2x_1 - q_1.$$

If this q_2 lies to the right of the center of mass of the interval (x_1, ∞) then the trial chain terminates, and we start over again with a different trial value q_1. Otherwise, x_1 and q_2 being known, the second of conditions (14):

$$q_2 = \frac{\int_{x_1}^{x_2} x\, dF(x)}{\int_{x_1}^{x_2} dF(x)}$$

serves to determine x_2 uniquely. Now the second of conditions (17) gives

$$q_3 = 2x_2 - q_2.$$

We continue in this way, obtaining successively $q_1, x_1, \cdots, q_{\nu-1}, x_{\nu-1}, q_\nu$; the last step is the determination of q_ν according to

$$q_\nu = 2x_{\nu-1} - q_{\nu-1}. \tag{21}$$

However in this procedure we have not used the last of conditions (14):

$$q_\nu = \frac{\int_{x_{\nu-1}}^{\infty} x\, dF(x)}{\int_{x_{\nu-1}}^{\infty} dF(x)}, \tag{22}$$

and the q_ν obtained from (21) will not satisfy (22) in general. The discrepancy between the right members of (21) and (22) will vary continuously with the starting value q_1, and the method consists of running through such chains

[2] It seems likely that this condition $N(\rho^{(1)}) \leq$ (optimum noise for $\nu-1$ quanta) is stronger than necessary for the nondivergence of (20).

using various starting values until the discrepancy is reduced to zero.

This method is applicable to more general F, with some obvious modifications. When F has intervals of constancy the $\{x_\alpha\}$ may not be uniquely determined by conditions (14), and a trial chain may involve several arbitrary parameters besides q_1. Discontinuities of F will cause no real trouble, since we know that the $\{x_\alpha\}$ of an optimum scheme are continuity points of F; a trial chain that does not have this property is discarded. We note that Method II may be used to locate all stationary points of $N(\rho)$.

VII. Examples

In all of the examples we now consider, the distribution of sample values is absolutely continuous with a sample probability density $f = F'$, which is an even function. If $N(\rho)$ has a unique stationary point, which we assume to be the case in the examples treated, then the optimum $\{q_\alpha\}$ and $\{x_\alpha\}$ will clearly be symmetrically distributed around the origin. In applications we are usually interested in having an even number of quanta, $\nu = 2\mu$, so we renumber the positive endpoints and quanta according to

$$0 = x_0 < q_1 < x_1 < \cdots < q_{\mu-1} < x_{\mu-1} < q_\mu; \quad (23)$$

the endpoints and quanta for the negative half-axis are the negatives of these.

We normalize to unit signal power $S = 1$. The $\{q_\alpha\}$ and $\{x_\alpha\}$ for other values of S are to be obtained by multiplying the numbers in the tables by $\sqrt{S}$.

The simplest case is the uniform distribution:

$$f(x) = \frac{1}{2\sqrt{3}}, \qquad -\sqrt{3} \le x \le \sqrt{3}$$

$$= 0, \qquad \sqrt{3} < |x| < \infty.$$

Method II of the preceding section shows that $N(\rho)$ in this case has a unique stationary point, which is necessarily an absolute minimum. The optimum scheme is the usual one with ν equal intervals of width $1/(2\nu\sqrt{3})$ each; the quanta being the midpoints of these intervals. The minimum value of the noise is the familiar $N = 1/\nu^2$.

Another case of possible interest is the Gaussian:

$$f(x) = \frac{e^{-1/2x^2}}{\sqrt{2\pi}}, \qquad -\infty < x < \infty.$$

The optimum schemes for $\nu = 2^b$, $b = 1, 2, \cdots, 7$, are given in Tables I–VII,[3] respectively. The corresponding noise values appear in Table VIII together with the quantities $\nu^2 N$ and νx_1. The behavior of these latter with increasing ν hint at the existence of asymptotic properties; we examine this question in the next section.

[3]Since some of the tables were never completed, those tables although mentioned in text are not included in this paper.

TABLE I
GAUSSIAN, $\nu = 2$

α	q_α	x_α
1	0.7979	∞

TABLE II
GAUSSIAN, $\nu = 4$

α	q_α	x_α
1	0.4528	0.9816
2	1.5104	∞

TABLE III
GAUSSIAN, $\nu = 8$

α	q_α	x_α
1	0.2451	0.5006
2	0.7560	1.0500
3	1.3439	1.7480
4	2.1520	∞

TABLE IV
GAUSSIAN, $\nu = 16$

α	q_α	x_α
1	0.1284	0.2582
2	0.3880	0.5224
3	0.6568	0.7996
4	0.9423	1.0993
5	1.2562	1.4371
6	1.6181	1.8435
7	2.0690	2.4008
8	2.7326	∞

TABLE VIII
GAUSSIAN; OPTIMUM NOISE FOR VARIOUS VALUES OF ν

ν	N	$\nu^2 N$	νx_1
2	0.3634	1.452	
4	0.1175	1.880	3.93
8	3.455×10^{-2}	2.205	4.00
16	9.500×10^{-3}	2.430	4.13
32			
64			
128			
(∞)	(0)	(2.72)	(4.34)

For speech signals a distribution which has been found useful empirically is the Laplacian:[4]

$$f(x) = \frac{e^{-|x|\sqrt{2}}}{\sqrt{2}}, \qquad -\infty < x < \infty.$$

The optimum quantizing schemes for this distribution for $\nu = 2^b$, $b = 1, 2, \cdots, 7$, are given in Tables IX–XV, respectively. The corresponding N, $\nu^2 N$, and νx_1 values are given in Table XVI; again, we notice certain regularities.

VIII. Asymptotic Properties

Let us assume that the distribution F is absolutely continuous with density function $f = F'$, which is itself dif-

[4]The author is indebted to V. Vyssotsky of the Acoustics Research Group for this information (private communication).

ferentiable, and that for each ν there is a unique optimum quantization scheme. We revert to our original numbering (18).

Let the quantities $\{h_\alpha\}$ be defined by

$$h_\alpha = x_\alpha - q_\alpha = q_{\alpha+1} - x_\alpha, \qquad \alpha = 1,2,\cdots,\nu-1,$$

so that, for $\alpha = 2,3,\cdots,\nu-1$, Q_α consists of an interval of length h_α to the right of q_α together with an interval of length $h_{\alpha-1}$ to the left of q_α. We have already imposed the optimizing conditions (17) in the very definition of the $\{h_\alpha\}$. The center of mass conditions (14) (except for the first and last) may be written as

$$\int_{q_\alpha - h_{\alpha-1}}^{q_\alpha + h_\alpha} (x - q_\alpha) f(x)\, dx = 0, \qquad \alpha = 2,3,\cdots,\nu-1.$$

If we expand f here in Taylor's series around q_α, the integration gives

$$\tfrac{1}{2}\left(h_\alpha^2 - h_{\alpha-1}^2\right) f(q_\alpha) + \tfrac{1}{3}\left(h_\alpha^3 + h_{\alpha-1}^3\right) f'(q_\alpha) = o\left(h_\alpha^3\right) + o\left(h_{\alpha-1}^3\right), \qquad \alpha = 2,3,\cdots,\nu-1. \quad (24)$$

The numbers in Tables VIII and XVI suggest the existence of an asymptotic fractional density of quanta. Accordingly, we define the function $g_\nu(x)$, $-\infty < x < \infty$, by

$$\begin{aligned} g_\nu(x) &= 0, && -\infty < x \le q_1 \\ &= \frac{1}{2\nu h_\alpha}, && q_\alpha < x \le q_{\alpha+1}, \\ & && \alpha = 1,2,\cdots,\nu-1, \\ &= 0, && q_\nu < x < \infty. \end{aligned} \quad (25)$$

The definition is arranged so that (for given ν) the sets $Q_2, Q_3, \cdots, Q_{\nu-1}$ subtend equal areas of $1/\nu$ each under the graph of $g_\nu(x)$ versus x. We will proceed as if a limiting density,

$$g(x) = \lim_{\nu\to\infty} g_\nu(x), \qquad -\infty < x < \infty,$$

existed.

We wish to express g in terms of the given sample density function f. To do this we will use conditions (24), together with the following further assumptions. We assume that g has a derivative, and we assume that for given x and k the difference $\epsilon_\nu(x) = g_\nu(x) - g(x)$, $-\infty < x < \infty$, has the property[5]

$$\epsilon_\nu\left(x + \frac{k}{\nu}\right) - \epsilon_\nu(x) = o\left(\frac{1}{\nu}\right).$$

In (24), then, we may approximate $h_\alpha - h_{\alpha-1}$ by

$$\begin{aligned} h_\alpha - h_{\alpha-1} &= \frac{1}{2\nu}\left[\frac{1}{g_\nu(q_\alpha + h_\alpha)} - \frac{1}{g_\nu(q_\alpha - h_{\alpha-1})}\right] \\ &= -\frac{g'(q_\alpha)}{2\nu^2 g^3(q_\alpha)} + o\left(\frac{1}{\nu^2}\right), \qquad \alpha = 2,3,\cdots,\nu-1, \end{aligned}$$

and we find that the left-hand member of (24) is indeed $o(h^3) = o(1/\nu^3)$ provided that

$$\frac{g'(x)}{g(x)} = \frac{f'(x)}{3f(x)}, \qquad -\infty < x < \infty. \quad (26)$$

The normalized solution of (26) is

$$g(x) = \frac{f^{1/3}(x)}{\int_{-\infty}^{\infty} f^{1/3}(x')\, dx'}, \qquad -\infty < x < \infty, \quad (27)$$

provided that the integral in the denominator exists.

The noise power becomes

$$\begin{aligned} N &= \frac{1}{12\nu^2}\int_{-\infty}^{\infty} \frac{f(x)}{g^2(x)}\, dx + o\left(\frac{1}{\nu^2}\right) \\ &= \frac{1}{12\nu^2}\left[\int_{-\infty}^{\infty} f^{1/3}(x)\, dx\right]^3 + o\left(\frac{1}{\nu^2}\right), \end{aligned} \quad (28)$$

neglecting the contributions from the end quanta.[6] For the Gaussian example then, the numbers $\nu^2 N$ of Table VIII should have a limit easily evaluated from (28) as $\nu^2 N \to \pi\sqrt{3}/2 (\approx 2.72)$, and in the Laplacian case, Table XVI, we find $\nu^2 N \to 9/2$.

The quantities denoted by νx_1 in Tables VIII and XVI should have the limiting value

$$\lim_{\substack{\nu\to\infty \\ q_\alpha \to 0}} \nu(h_{\alpha-1} + h_\alpha) = \frac{1}{g(0)},$$

comparing with (25). In the Gaussian example we find $1/g(0) = \sqrt{6\pi}\,(\approx 4.34)$, and for the Laplacian: $1/g(0) = 3\sqrt{2}\,(\approx 4.24)$.

For large values of ν the sets $\{Q_\alpha\}$ should subtend approximately equal areas of $1/\nu$ each under the graph of $g(x)$ versus x, so that the number b_ν defined by

$$\frac{1}{\nu} = \int_{b_\nu}^{\infty} g(x)\, dx$$

might be expected to be near the rightmost division point. Comparing Table XVII with Tables I–VII and IX–XVI we see that the approximation is surprisingly good, at least in the examples considered.

TABLE XVII
APPROXIMATE LAST ENDPOINT FROM ASYMPTOTIC FORMULA

ν	Gaussian b_ν	Laplacian b_ν
2	0	
4	1.168	
8	1.992	
16	2.657	
32		
64		
128		

[5] The notation $u(\nu) = o(v(\nu))$ means in our case $\lim_{\nu\to\infty} u(\nu)/v(\nu) = 0$.

[6] Other derivations of (27)–(28) are given in [6] and [7].

Acknowledgment

The numerical results presented in the tables are due to Miss M. C. Gray and her assistants in the Numerical Analysis and Digital Processes Group; the programming of Method I for the IBM-650 electronic computer was done by Miss C. A. Conn.

After substantial progress had been made on the work described here there appeared in [11] a review of a paper by J. Lukaszewicz and H. Steinhaus on optimum go/no-go gauge sets. The present author has not been able to obtain a copy of this paper, but it seems likely that these authors have treated a problem similar or identical to the one discussed in Sections IV–VI. M. P. Schützenberger in [12] examines the quantization problem in the case where $\nu = 2$ and where F increases at 3 or 4 points.

Appendix A

Suppose $s(t)$, $-\infty < t < \infty$, is a continuous parameter stochastic process, real, separable, measurable, stationary, and of finite power:

$$S = E\{s^2(t)\} = \int_{-\infty}^{\infty} x^2\, dF(x) < \infty, \qquad -\infty < t < \infty,$$

(where F is the first-order distribution of the process, Section III). Then s has a spectral representation

$$s(t) = \int_{-\infty}^{\infty} e^{2\pi i\lambda t} d\xi(\lambda), \qquad -\infty < t < \infty, \tag{29}$$

where the spectral process $\xi(\lambda)$, $-\infty < \lambda < \infty$, has orthogonal increments [4, p. 527]. To say that s is band-limited to the frequency band $-W \le \lambda \le W$ is to say that the ξ process has vanishing increments outside of this band with probability one, and (29) becomes

$$s(t) = \int_{-W-0}^{W+0} e^{2\pi i\lambda t} d\xi(\lambda), \qquad -\infty < t < \infty. \tag{30}$$

Since we are particularly concerned with the behavior of ξ at the band edges, we rewrite (30) as

$$s(t) = \int_{-W+0}^{W-0} e^{2\pi i\lambda t} d\xi(\lambda) + 2\delta_1 \cos 2\pi Wt - 2\delta_2 \sin 2\pi Wt, \qquad -\infty < t < \infty,$$

where the real random variables δ_1 and δ_2 describe the jumps of ξ at the band edges:

$$\xi(\pm W + 0) - \xi(\pm W - 0) = \delta_1 \pm i\delta_2.$$

For fixed t, the function $e^{2\pi i\lambda t}$, $-W \le \lambda \le W$, has Fourier coefficients

$$\begin{aligned} c_j &= \frac{1}{2W}\int_{-W}^{W} e^{-2\pi ij\lambda/(2W)} e^{2\pi i\lambda t} d\lambda \\ &= \frac{\sin 2\pi W(t - j/(2W))}{2\pi W(t - j/(2W))} \\ &= K(t - t_j), \qquad -\infty < j < \infty, \end{aligned}$$

in the notation of Section I. This function is of bounded variation, so that the partial sums

$$S_l(\lambda) = \sum_{j=-l}^{l} e^{2\pi ij\lambda/(2W)} K(t - t_j), \qquad -W \le \lambda \le W,$$

converge boundedly to

$$\begin{aligned} S(\lambda) &= \lim_{l\to\infty} S_l(\lambda) \\ &= e^{2\pi i\lambda t}, \qquad -W < \lambda < W, \\ &= \cos^{2\pi Wt}, \qquad \lambda = \pm W, \end{aligned}$$

from [8]. Hence, using the representation (30) for the samples, the sampling series

$$\hat{s}(t) = \sum_{j=-\infty}^{\infty} s(t_j) K(t - t_j) \tag{31}$$

converges (in stochastic mean square) to

$$\begin{aligned} \hat{s}(t) &= \text{l.i.m.}_{l\to\infty} \int_{-W-0}^{W+0} S_l(\lambda)\, d\xi(\lambda) \\ &= \int_{-W+0}^{W-0} e^{2\pi i\lambda t} d\xi(\lambda) + 2\delta_1 \cos 2\pi Wt \\ &= s(t) + 2\delta_2 \sin 2\pi Wt, \qquad -\infty < t < \infty, \end{aligned} \tag{32}$$

from [4, p. 429]. (The corresponding result for deterministic functions is given in [9].) Since the orthogonal increments property of ξ requires $E\{\delta_1\delta_2\} = 0$ together with

$$E = \{\delta_1^2\} = E\{\delta_2^2\} = \tfrac{1}{2}E\{|\xi(\pm W + 0) - \xi(\pm W - 0)|^2\},$$

we see from (32) that the sampling series (31) represents s with probability 1, if and only if, the ξ process has no power concentrated at the band edges,

$$E\{|\xi(\pm W + 0) - \xi(\pm W - 0)|^2\} = 0.$$

(Other proofs of this result appear in [3] and in [10].)

Let s be as above and suppose $\varphi(x)$, $-\infty < x < \infty$, is a Baire function. Then the random variables

$$\cdots, \varphi(s(t_{-1})), \varphi(s(t_0)), \varphi(s(t_1)), \cdots \tag{33}$$

constitute a stationary discrete-parameter stochastic process. If the number

$$\Phi = E\{\varphi^2(s(t_j))\} = \int_{-\infty}^{\infty} \varphi^2(x)\, dF(x), \qquad -\infty < j < \infty,$$

is finite then the process (33) admits a spectral representation

$$\varphi(s(t_j)) = \int_{-W}^{W} e^{2\pi ij\lambda/(2W)} d\eta(\lambda), \qquad -\infty < j < \infty,$$

where the η process has orthogonal increments ([4, p. 481], with a change of scale).

A certain continuous parameter stochastic process $\theta(t)$, $-\infty < t < \infty$, may be defined in terms of the η process by

$$\theta(t) = \int_{-W}^{W} e^{2\pi i\lambda t} d\eta(\lambda), \qquad -\infty < t < \infty.$$

This process is stationary in the wide sense and has the given process (33) as its samples, clearly

$$\theta(t_j) = \varphi(s(t_j)), \qquad -\infty < j < \infty.$$

Moreover,

$$E\{\theta^2(t)\} = \int_{-W}^{W} E\{|d\eta(\lambda)|^2\} = \int_{-\infty}^{\infty} \varphi^2(x) dF(x), \qquad -\infty < t < \infty.$$

The θ process is represented by the sampling series

$$\theta(t) = \sum_{j=-\infty}^{\infty} \varphi(s(t_j)) K(t - t_j), \quad -\infty < t < \infty, \tag{34}$$

if and only if, the spectral process η has no power concentrated at

the band edges. The arguments are identical to those given above for the s process itself.

The r and n processes of Sections II and III are of the form just described, since the functions $y(x)$, (6), and $z(x)$, (9), will differ from certain Baire functions only on sets of measure zero $[dF]$ when we assume, as we do, that the sets $\{Q_\alpha\}$ are measureable $[dF]$.

The well-known mean-erogodic property,

$$\text{l.i.m.}_{m''-m'\to\infty} \sum_{j=m'}^{m''} \frac{(-1)^j \varphi(s(t_j))}{m''-m'+1}$$
$$= \eta(W+0) - \eta(W-0) + \eta(-W+0) - \eta(-W-0)$$
$$= 2\,\mathrm{Re}[\eta(\pm W+0) - \eta(\pm W-0)],$$

([4, p. 491]) shows that the requirement that η have no power at the band edges is equivalent to the condition

$$\lim_{m''-m'\to\infty} E\left\{\left|\sum_{j=m'}^{m''} \frac{(-1)^j \varphi(s(t_j))}{m''-m'+1}\right|\right\}^2 = 0.$$

Finally we note that if the ξ process has a discrete component at frequencies $\pm\lambda_0$, then depending on the form of φ, the derived η process is likely to have discrete components at all of the harmonic frequencies $\pm m\lambda_0$ (modulo $2W$) of λ_0, $m = 1, 2, \cdots$. In particular, if λ_0 is rational then the η process may have a discrete component at the band edges, a possibility which must be excluded if (34) is to hold.

Appendix B

A simple example shows that the conditions (14) and (17) are not sufficient for an absolute minimum of N. Suppose F is absolutely continuous, with a density $f = F'$ as shown in Fig. 1, where $c_1(b_2 - b_1) + c_2(b_4 - b_3) = 1$. If $\nu > 1$ quanta are desired, let $\nu_1, \nu_2 > 0$ be any integers such that $\nu_1 + \nu_2 = \nu$, and divide the interval (b_1, b_2) into ν_1 equal intervals and (b_3, b_4) into ν_2 equal intervals; let the quanta be the midpoints of these ν intervals. If we suppose that $b_2 < \frac{1}{2}(b_1 + b_3)$ and $b_3 > \frac{1}{2}(b_2 + b_4)$ then the division point which separates the right-hand $\{Q_\alpha\}$ in (b_1, b_2) and the left-hand $\{Q_\alpha\}$ in (b_3, b_4) will lie in the interval (b_2, b_3), so that the conditions (14) and (17) will be satisfied. Thus we have $\nu - 1$ distinct local minima of N. (If c_1, respectively c_2, is small enough there may even be another minimum, corresponding to $\nu_1 = 0$, respectively $\nu_2 = 0$.) Which of these is the true minimum depends on the values of the parameters. (Explicitly, the noise has the value

$$N = \frac{c_1(b_2 - b_1)^3}{12\nu_1^2} + \frac{c_2(b_4 - b_3)^3}{12\nu_2^2},$$

and the ν_2/ν_1 for which this is a minimum is given by

$$\frac{\nu_2/(b_4 - b_3)}{\nu_1/(b_2 - b_1)} = \left(\frac{c_2}{c_1}\right)^{1/3},$$

agreeing with (27).)

The following interesting example is due to J. L. Kelly, Jr., of the Visual Research Group. Let the density f be as in Fig. 2, with $c_2 > c_1$, $c_1 + c_2 = 1$. The signal power is $S = 1/3$, independently of c_1 and c_2. Suppose $\nu = 2$. One configuration for which conditions (14) and (17) are satisfied is $q_1 = -\frac{1}{2}$, $x_1 = 0$, $q_2 = \frac{1}{2}$, clearly, and the resulting noise in $N = 1/12$. When $c_2 > 3c_1$; however, there is another solution—it is the one which in the limit $c_1 = 0$, $c_2 = 1$ goes into the scheme where $(0, 1)$ is divided

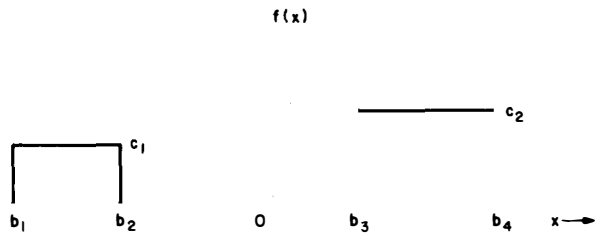

Fig. 1. The density $f(x)$ vanishes outside of the intervals (b_1, b_2), (b_3, b_4).

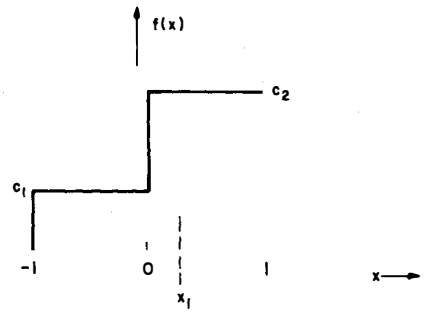

Fig. 2. The density $f(x)$ vanishes outside of the interval $(-1, 1)$.

into two equal parts and $(-1, 0)$ is ignored. The parameters for this configuration work out to be

$$q_1 = \frac{c_2 - 9c_1}{4c_2},$$
$$x_1 = \frac{c_2 - 3c_1}{2c_2},$$
$$q_2 = \frac{3(c_2 - c_1)}{4c_2},$$
$$N = \frac{1}{12} - \frac{(c_2 - 3c_1)^3}{16c_2^2}.$$

Hence if this configuration exists then it is better than the one first mentioned. We note, by the way, that method I of Section VI will converge to the $(-1/2, 0, 1/2)$ configuration, if the starting value $x_1^{(1)}$ is negative; if $c_2 > 3c_1$, however, this configuration is only a saddle point of N.

Author's Note 1981:

This is nearly a verbatim reproduction of a draft manuscript, which was circulated for comments at Bell Laboratories; the Mathematical Research Department log date is July 31, 1957. I wish to thank the editors for their invitation to publish this antique *samizdat* in the present issue.

The main reason the paper was not submitted for publication previously was that the numerical calculations were never completed. The Gaussian $\nu = 32$ case was done on the IBM 650 card programmable calculator; the Laplacian cases were done only for $\nu = 2$. Some time later the 650 was replaced by an IBM 701 electronic computer, but no quantizing program was written for it.

I was not satisfied with not having conditions for a unique minimum but would have published the paper

without this. Later, P. E. Fleischer of Bell Laboratories gave a neat sufficient condition in his paper [13].

In the examples of Appendix B, the direct current can be removed by changing the origin; the noise is not affected. The results of the paper are valid for the uncentered processes used.

I was aware when I wrote the paper that the methods for quantizing a real random variable extend to other loss functions. In the least squares case the process quantizing noise is just the noise per sample; the generalization is more complicated and was omitted.

References

[1] B. M. Oliver, J. R. Pierce, and C. E. Shannon, "The philosophy of PMC," *Proc. I.R.E.*, vol. 36, pp. 1324–1331, 1948.

[2] H. S. Black, *Modulation Theory*. Princeton, NJ: Van Nostrand, 1953.

[3] S. P. Lloyd and B. McMillan, "Linear least squares filtering and prediction of sampled signals," in *Proc. Symp. on Modern Network Synthesis*, vol. 5. Brooklyn, NY: Polytechnic Institute of Brooklyn, 1956, pp. 221–247.

[4] J. L. Doob, *Stochastic Processes*. New York: Wiley, 1953.

[5] H. Cramér, *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University, 1951.

[6] P. F. Panter and W. Dite, "Quantization distortion in pulse-count modulation with nonuniform spacing of levels," *Proc. I.R.E.*, vol. 39, pp. 44–48, 1951.

[7] B. Smith, "Instantaneous companding of quantized signals," *Bell Syst. Tech. J.*, vol. 36, pp. 653–709, 1957.

[8] A. Zygmund, *Trigonometrical Series*. New York: Dover, 1955, p. 47.

[9] H. P. Kramer, "A generalized sampling theorem," *Bull. Am. Math. Soc.* vol. 63, p. 117, 1957.

[10] E. Parzen, "A simple proof and some extensions of the sampling theorem," Department of Statistics, Stanford Univ., CA, Tech. Report 7, Dec. 1956.

[11] J. Kukaszewicz and H. Steinhaus, "On measuring by comparison," *Zastos. Mat.*, vol. 2, pp. 225–231, 1955; *Math. Reviews*, vol. 17, p. 757, 1956.

[12] M. P. Schützenberger, "Contribution aux applications statistiques de la théorie de l'information," *Pub. de l'Inst. de Statis.* de l'Université de Paris, vol. 3, Fasc. 1-2 pp. 56–69, 1954.

[13] P. E. Fleischer, "Sufficient conditions for achieving minimum distortion in quantizer," *IEEE Int. Convention Record*, part I, vol. 12, pp. 104–111, 1964.