Statistical Analysis of Amplitude-Quantized Sampled-Data Systems

B. WIDROW
ASSOCIATE MEMBER AIEE

QUANTIZATION, or round-off, occurs whenever physical quantities are represented numerically. The values of measurements may be designated by integers corresponding to their nearest numbers of units. Round-off errors have values between $\pm 1/2$ unit, and can be made small by choice of the basic unit. It is apparent, however, that the smaller the size of the unit, the larger will be the numbers required to represent the same physical quantities and the greater will be the difficulty and expense in storing and processing these numbers. Often, a balance has to be struck between accuracy and economy. In order to establish such a balance, it is necessary to have a means of evaluating quantitatively the distortion resulting from rough quantization. The analytical difficulty arises from the inherent nonlinearities of the quantization process.

Paper 60-1240, recommended by the AIEE Feedback Control Systems Committee and approved by the AIEE Technical Operations Department for presentation at the AIEE Fall General Meeting, Chicago, Ill., October 9-14, 1960. Manuscript submitted January 4, 1960; made available for printing September 14, 1960.

B. WIDROW is with Stanford University, Stanford, Calif.

Most of the material in this paper has appeared in the 1959 class notes of the Massachusetts Institute of Technology Course *6.54*, Pulsed Data Systems, and will appear in a book on sampled-data system theory by Linvill, Sittler, and Widrow. A treatise by the author on this subject is in preparation and will be published as a monograph.

Part I. Background

Definition of Quantizer

For purposes of analysis, it has been found convenient to define the quantizer as a nonlinear operator having the input–output relation shown in Fig. 1(A). Its output x' is a single-valued function of the input x, and has an "average gain" of unity. An input lying somewhere within a quantization "box" of width q will yield an output corresponding to the center of that box, i.e., the input is rounded off to the center of the box.

The quantizer symbol of Fig. 1(B) is useful in representing a rounding-off process as a dynamical system element whose inputs and outputs are signals in real time. As a mathematical operator, a quantizer may be defined as processing continuous signals to give a stepwise continuous output, or as processing sampled signals to give a sampled output.

The attention of most of this work will be focused upon the basic quantizer of Fig. 1. The analysis that develops will be applicable to a variety of different kinds of quantizers which can be represented in terms of this basic quantizer and other simple linear operators. For example, the quantizers shown in Fig. 2 are derived from the basic quantizer by changing input and output scales and by the addition of constants or d-c levels to input and output. Notice that these input–output characteristics would approach the dotted lines whose slopes are the average gains if the quantizer box sizes were made arbitrarily small.

Another kind of quantizer that can be represented in terms of the basic quantizer and some positive feedback is one having hysteresis at each step. The input–output characteristic is a staircase array of hysteresis loops, an example of which is shown in Fig. 3.

Two- and 3-level quantizers, which are more commonly called clippers, appear in "contactor" systems. They will be treated as ordinary quantizers whose inputs are confined to two and three levels respectively. Fig. 4 shows their input–output characteristics and their block-diagram symbols. Fig. 5 shows examples of how clippers with hysteresis can be represented as clippers with positive feedback.

Every physical quantizer is noisy to a certain extent, which means that the ability of a quantizer to resolve inputs which come very close to the box edges is limited. These box edges are actually smeared lines rather than infinitely sharp lines. If an input close to a box edge is randomly rounded up or down, the

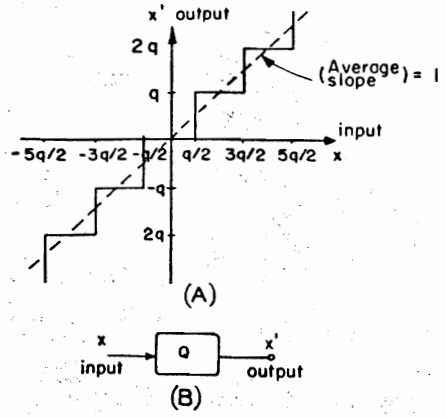


Fig. 1 (left). Basic quantizer

A—Input–output characteristic
B—Block-diagram symbol

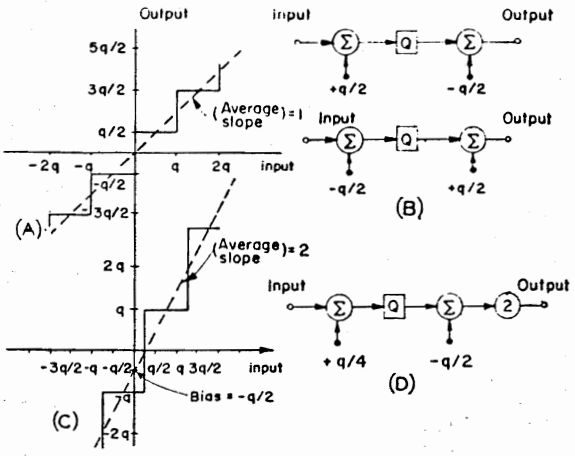


Fig. 2 (right). Effects of scale changes and addition of constants

A—Quantizer without dead zone
B—Equivalent representation of characteristic in A
C—Quantizer with scale and d-c level changes
D—Equivalent representation of characteristic in C

quantizer can be represented as an ideal (infinitely sharp) quantizer with a random noise added to its input.

Quantized systems result when quantizers are combined with dynamic elements. These systems may be open-looped or closed-looped, sampled, continuous, and linear or nonlinear (except for the quantizers). In a sense, quantized systems are more general quantizers. Several quantized feedback systems will be examined in part III.

Sampling and Quantization

A numerical description of a continuous function of an independent variable may be made by plotting the function on graph paper as in Fig. 6. The function $x(t)$ can be approximately represented over the range $0 \leq t \leq 10$ by a series of numerical values, its quantized samples: 1, 1, 2, 0, −1, −2, −2, −1, 0, 0, 1.

The plot of Fig. 6 suggests that quantization is like sampling in amplitude. Quantization is a sampling process that acts not upon the function itself, however, but upon its probability density distribution (DD).

Both sampling and quantizing are effected when signals are converted from analog to digital. Sampling and quantizing are mathematically commutable operations. It makes no difference whether a signal is first sampled and then the samples are quantized, or if the signal is quantized and the stepwise continuous signal is then sampled.

Both sampling and quantizing degrade the quality of a signal and irreversibly diminish knowledge of it. A reconstructed signal has a "dynamical" component of error which is proportional to signal amplitude and is due to the granularity in t, and a "static" component of error or a noise which is practically independent of signal amplitude and is due to quantization granularity.

Sampled-data systems behave very much like continuous systems in a macroscopic sense. They could be analyzed and designed as if they were conventional continuous systems by ignoring the effects of sampling. In order to take these effects into account, use must be made of the theory of sampled-data systems. Quantized systems, on the other hand, behave in a macroscopic sense very much like systems without quantization. They too could be analyzed and designed by ignoring the effects of quantization. These effects in turn could be reckoned with by applying the statistical theory of quantization.

Part II is a summary of the statistical theory of quantization which has already been presented in some detail in references 1 and 2. The material is given for review and, at the same time, represents a more advanced and simpler view of this theory.

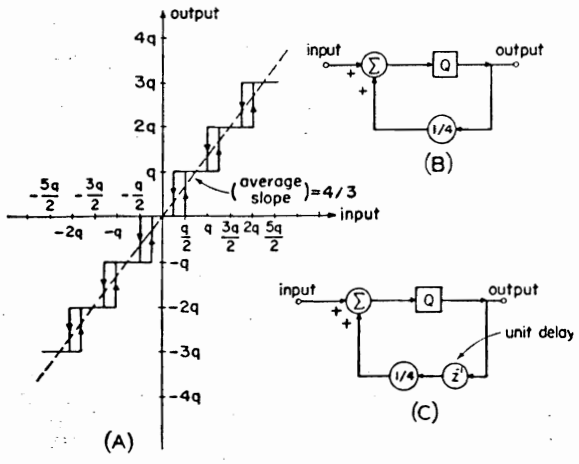


Fig. 3. Quantizer with hysteresis

A—Input–output characteristic
B—Equivalent representation for continuous signals
C—Equivalent representation for sampled signals

Part II. Statistical Theory of Amplitude Quantization

First-Order Probability DD of Quantizer Output

Let only the probability density distribution of the variable being quantized be considered. It will be seen that the DD of the quantized variable may be obtained by a linear sampling process on the DD of the unquantized variable. The analysis that follows will be developed for sampled signals because high-order random processes are more easily described in sampled form. How this analysis applies to continuous signals will be explained later.

If the samples of a continuous variable x are all independent of each other, a first-order probability density $w_x(x)$ completely describes the process. Its characteristic function (CF) is the Fourier transform, equation 1.

$$W_x(u) = \int_{-\infty}^{\infty} w_x(x) \epsilon^{-jxu} dx \qquad (1)$$

A quantizer input variable may take on a continuum of amplitudes, while the output assumes only discrete amplitudes. The probability density of the output $w_{x'}(x')$ consists of a series of impulses that are uniformly spaced along the amplitude axis, each one centered in a quantization box.

Fig. 7 shows how the output DD is derived from that of the input. Any input event (signal amplitude) occurring within a given quantization box is "reported" as being at the center of that box. Each impulse of the quantized DD must have an area equal to the area under

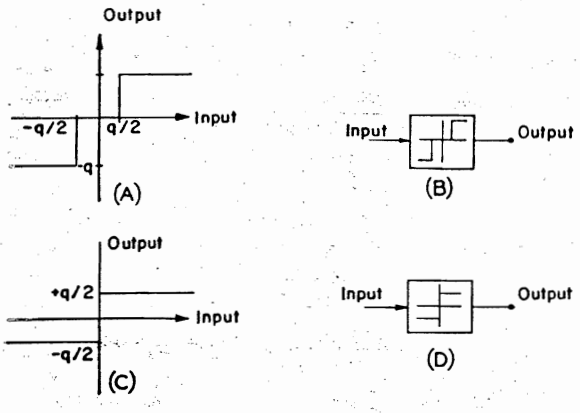


Fig. 4 (left). Saturating quantizers

A—Clipper with dead zone
B—Representation of A
C—Clipper with no dead zone
D—Representation of C

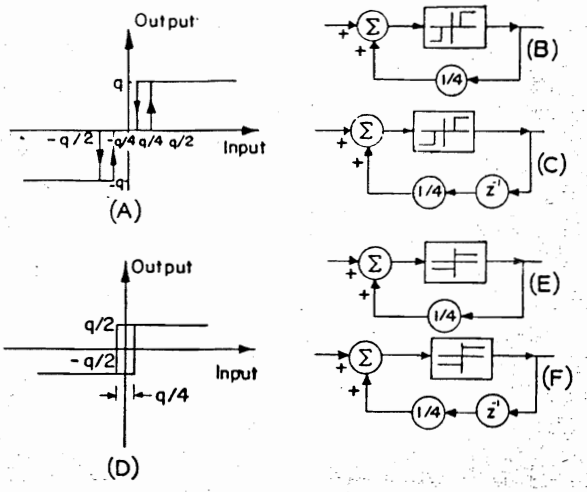


Fig. 5. Saturating quantizers with hysteresis

A—Three-level clipper with hysteresis
B—Continuous-data representation of A
C—Sampled-data representatation of A
D—Two-level clipper with hysteresis
E—Continuous-data representation of D
F—Sampled-data representatation of D

the probability density $w_x(x)$ within the bounds of the box. The probability of the quantized variable being at a certain level is equal to the total probability of the input being within the corresponding quantization box.

The impulse distribution $w_{x'}(x')$ has a characteristic function $W_{x'}(u)$, which is periodic, being the Fourier transform of a series of impulses having uniform spacing q. The analysis techniques developed for the study of linear sampled-data systems will be used in the derivation of the characteristic function $W_{x'}(u)$ of the quantized variable.

The DD of a quantizer output $w_{x'}(x')$ consists of area samples of the input distribution density $w_x(x)$. The quantizer may be thought of as an area sampler acting upon the "signal," the probability density $w_x(x)$. Fig. 8 shows how $w_{x'}(x')$ may be constructed by sampling the differences $d(x+q/2)-d(x-q/2)$, where $d(x)$ is the input distribution, the integral of the input DD. Fig. 9 is a block-diagram model of this process, showing how $w_x(x)$ is first modified by a linear "filter" whose transfer function is:

$$\frac{\epsilon^{+juq/2}-\epsilon^{-juq/2}}{ju}=\frac{\sin(qu/2)}{(qu/2)} \qquad (2)$$

and then impulse modulated to give $w_{x'}(x')$. Using an asterisk notation to indicate sampling,

$$W_{x'}(u)=\left[W_x(u)q\frac{\sin(qu/2)}{(qu/2)}\right]^* \qquad (3)$$

The difference between area sampling and amplitude sampling is most clear in the frequency domain, where both give periodic transforms. The typical repeated section is the same as the transform of the envelope multiplied by $1/q$ in the case of amplitude sampling. For area sampling, the transform of the envelope is multiplied by $\sin(qu/2)/(qu/2)$, and then repeated at a frequency $\phi=2\pi/q$. Fig. 10 illustrates how $W_{x'}(u)$ is derived from $W_x(u)$.

When the quantization frequency or fineness $\phi=2\pi/q$ is twice as high as the highest frequency component contained in the shape of $w_x(x)$, the periodic sections of $W_{x'}(u)$ do not overlap, and it is possible to recover $w_x(x)$ from the quantized distribution $w_{x'}(x')$. (This can be done by inverse transforming the ratio of a typical section of $W_{x'}(u)$ divided by $\sin(qu/2)(qu/2)$, and is demonstrated in part III.) Henceforth, this will be called the quantizing theorem.

Consider a typical section of the quantized CF which is

$$W_x(u)\frac{\sin(\pi u/\phi)}{(\pi u/\phi)} \qquad (4)$$

This typical section is a product of two factors and could be thought of as a CF in its own right. It is well known that the CF of the sum of two statistically independent variables is the product of the individual CF's. The typical section is identical therefore to the CF of the sum of the quantizer input and a statistically independent noise whose CF is

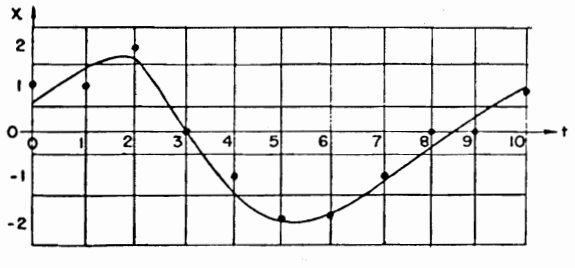


Fig. 6 (left). Sampling and quantization

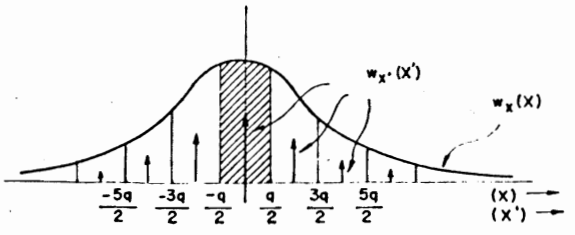
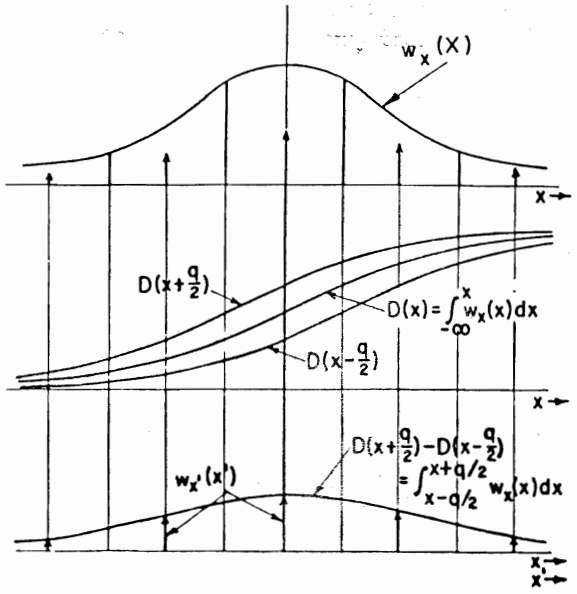


Fig. 7 (lower left). Area sampling, area of zero-th impulse equals shaded area

Fig. 8 (right). Construction of area samples



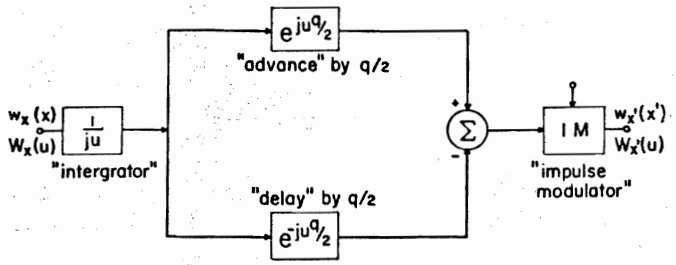


Fig. 9. Block diagram of area sampling

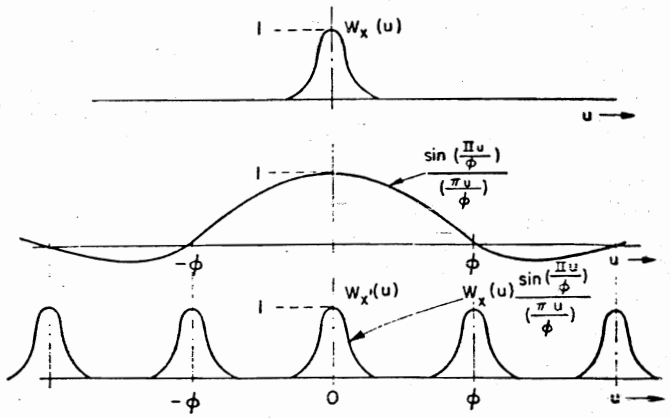


Fig. 10 (right). Formation of quantized CF

$$W_{fl}(u)=\frac{\sin(gu/2)}{(gu/2)}=\frac{\sin(\pi u/\phi)}{(\pi u/\phi)} \qquad (5)$$

This noise will ultimately be linked to quantization noise which has the flat-topped distribution density $w_{fl}(n)$, as shown in Fig. 11.

It is interesting to compare quantization with the addition of an independent random noise whose DD is $w_{fl}(n)$. This comparison is indicated in Fig. 12. The DD of the quantizer output is a series of impulses whose spacing is q. The DD of signal plus noise is continuous. The impulse DD consists of the impulse samples of the continuous DD under all conditions. The continuous DD can be uniquely derived from the impulse DD (as the sin x/x envelope) only when the quantizing theorem is satisfied.

It is well known that the moments of a random variable are given by the derivatives of its CF at the origin ($u=0$). For example, the kth moment of x or the average of x^k is given by:

$$\overline{x^k}=(j)^{-k}\frac{d^kW_x(u)}{du^k}\,u=0 \qquad (6)$$

The zero-th moment is $W_x(0)=1$; the area under $w_x(x)=1$.

If the quantizing theorem is satisfied, the periodic sections of the quantized CF do not overlap, and the derivatives of the quantized CF at the origin are the same as those of the typical section. Thus, the moments of the two output signals in Fig. 12 would be the same. Satisfaction of the quantizing theorem insures that the moments of a quantized signal are the same as those of the sum of the unquantized signal and a satistically independent noise uniformly distributed between plus and minus half a quantization box.

First-Order Probability Density of Quantization Noise

The investigation of the behavior of the quantizer would be complete at this point (knowledge of the output CF that results for a given input CF would be sufficient to determine the noise CF) if it were true that quantization noise were statistically independent of the quantizer input. The quantization noise signal is actually causally related to the input signal. Since the output of a quantizer is a single-valued function of the input, a given input yields a definite output and, consequently, a definite noise. The determination of the noise CF and DD is a new separate problem.

The causal tie between the input signal and the noise can be explored only when joint in–out distribution densities are derived.[1,2] In spite of the causal connection between signal and noise, the DD of the quantization noise is $w_{fl}(n)$, independent of the DD of the quantizer input, as long as the quantizing theorem is satisfied. Quantization noise is bounded and under all conditions must be distributed in some fashion between plus and minus half a box.

Quantization noise may be regarded as the difference between an input variable and the value of the center of the box which it falls into. The distribution of quantization noises resulting from inputs within the zero-th box may be constructed by plotting $w_x(x)$ between $-q/2<x<q/2$. The noise distribution resulting from inputs within the first box may be obtained by considering $w_x(x)$ for values $q/2<x<3q/2$ recentered to the origin. Events taking place in the various boxes are exclusive of each other. The probability of a given noise magnitude arising is, therefore, the sum of the probabilities of that noise arising from each box. Fig. 13 shows how the DD of quantization noise may be constructed graphically from a plot of $w_x(x)$. This technique leads to a simple analytical derivation of the quantization noise CF.

The summing process illustrated in Fig. 13 could be achieved in the following way. The distribution density $w_x(x)$ could be added to itself, shifted a distance q to the right, shifted $2q$ to the right, etc., and shifted q to the left, $2q$ to the left, etc. This infinite sum could then be multiplied by a "window function" which has the value unity over the range $-q/2\leq x\leq q/2$ and zero elsewhere. The sum of the densities is represented by the expression of equation 7:

$$\sum_{m=-\infty}^{\infty} w_x(x+mq) \qquad (7)$$

The transform of this sum is equal to $\frac{1}{q}$ times the impulse samples of $W_x(u)$ and is given by:

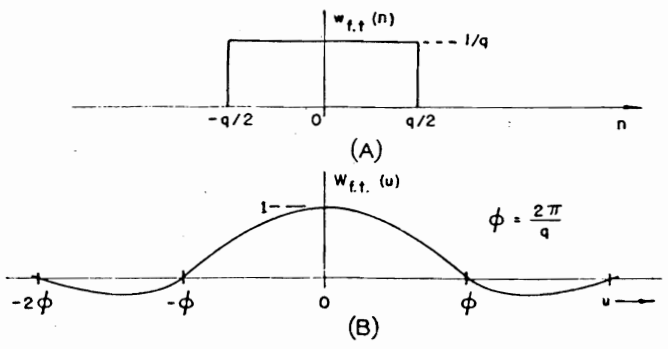
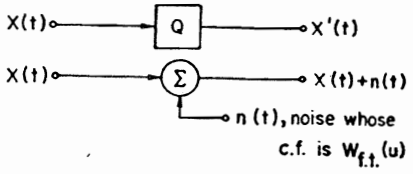


Fig. 11 (left). Flat-topped DD (A), and CF (B)



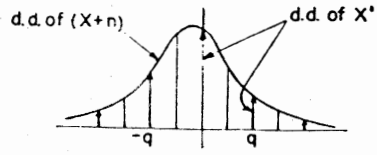


Fig. 12 (right). Comparison of quantization with noise addition

$$\sum_{-\infty}^{\infty} \phi W_x(k\phi)\delta(u-k\phi) \qquad (8)$$

Multiplication in the probability density domain by the window function corresponds in the CF domain to convolution with:

$$\frac{1}{\phi}\frac{\sin(\pi u/\phi)}{(\pi u/\phi)}$$

It follows, therefore, that the CF of the quantization noise is:

$$W_{(x-x')}(u) = \sum_{k=-\infty}^{\infty} W_x(k\phi)\frac{\sin \pi\left(\frac{u}{\phi}-k\right)}{\pi\left(\frac{u}{\phi}-k\right)} \qquad (9)$$

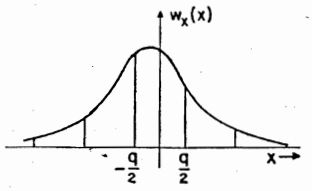
An examination of equation 9 makes apparent the fact that when the quantizing theorem is satisfied, the sum has only one nonzero term, which is:

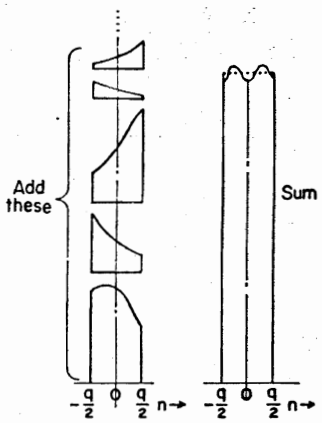
$$\frac{\sin(\pi u/\phi)}{(\pi u/\phi)}$$

This means that when the quantizing theorem is satisfied, the quantization noise has a sin u/u CF and a DD that is uniformly distributed between $\pm q/2$. A closer examination of equation 9 shows that the quantization noise will be precisely flat-topped even when the quantizing theorem is only one half satisfied, i.e., when the input CF $W_x(u)$ is zero for $|u|>\phi$, rather than for $|u|>\phi/2$. It is interesting to note that satisfaction of the quantizing theorem allows complete recovery of an original probability density given the quantized probability density. Half-satisfaction of the quantizing theorem allows only moments to be recovered. At the same time, half-satisfaction of the quantizing theorem insures a flat-topped quantization noise DD.

It should be noted that equation 9 gives the characteristic function of the quantization noise under all conditions.

Fig. 13. Construction of DD of quantization noise





Second-Order Probability DD of Quantizer Output

If a sampled input variable is a second-order process, the joint DD $w_{x_1,x_2}(x_1, x_2)$ is required to describe its statistics where x_1 and x_2 are respective first and second samples of an adjacent sample pair. This DD requires a 3-dimensional presentation (Fig. 14) in which quantization will be seen to give volume sampling rather than area sampling.

The quantization grid of Fig. 14 is square because the quantization box size is the same for both x_1 and x_2. The DD of the quantizer output $w_{x1',x2'}(x_1', x_2')$ consists of a set of impulses at the center of each square whose amplitudes (volumes) equal the volumes under $w_{x_1,x_2}(x_1, x_2)$ within the bounds of the squares. Associated with the input DD is its CF $W_{x_1,x_2}(u_1, u_2)$, given by equation 10 and shown in Fig. 15.

$$W_{x_1,x_2}(u_1, u) = \int_{-\infty}^{\infty}\int w_{x_1,x_2}(x_1, x_2)\, \epsilon^{-j(x_1u_1+x_2u_2)}dx_1dx_2 \qquad (10)$$

Inspection of Fig. 8 shows that area samples of the first-order DD $w_x(x)$ can be obtained by forming a new function of x and then taking its impulse samples. The new function of x, given by equation 11, equals the area under $w_x(x)$ within a range of width q centered at x.

(Area over range of width q)

$$= \int_{-\infty}^{x+q/2} w_x(x)dx - \int_{-\infty}^{x-q/2} w_x(x)dx \qquad (11)$$

$$= d(x+q/2) - d(x-q/2)$$

This scheme can be generalized to derive the volume samples $w_{x_1,x_2}(x_1, x_2)$. The first step is to form a new function of x_1 and x_2 which equals the volume under $w_{x_1,x_2}(x_1, x_2)$ within a $q \times q$ square in the x_1, x_2 plane whose edges are parallel to these axes and is centered at x_1, x_2.

(Volume over range of $q \times q$ square)

$$= \int_{-\infty}^{x_2+q/2}\left[\int_{-\infty}^{x_1+q/2} w_{x_1,x_2}(x_1, x_2)dx_1 - \int_{-\infty}^{x_1-q/2} w_{x_1,x_2}(x_1, x_2)dx_1\right]dx_2$$
$$- \int_{-\infty}^{x_2-q/2}\left[\int_{-\infty}^{x_1+q/2} w_{x_1,x_2}(x_1, x_2)dx_1 - \int_{-\infty}^{x_1-q/2} w_{x_1,x_2}(x_1, x_2)dx_1\right]dx_2 \qquad (12)$$

The next step is to take its impulse samples to get $w_{x_1',x_2'}(x_1', x_2')$. From this point on, it is best to examine the process in the 2-dimensional CF domain. The block diagram of Fig. 16, analogous to the 1-dimensional model of Fig. 9, is very helpful. Fig. 16(A) shows how the relation 12 can be implemented. Notice that orders of integration can be interchanged. The system of Fig. 16(B) results from algebraic simplifications of the cascade of Fig. 16(A).

The theory of 2-dimensional impulse modulation can be developed by generalization of the usual 1-dimensional theory. The results that bear upon the present discussion are: the CF of $w_{x_1',x_2'}(x_1', x_2')$ is periodic along the u_1 and u_2 axes with spacing $\phi = 2\pi/q$, being a sum of 2-dimensional "typical sections." Each typical section is identical to the Fourier trans-

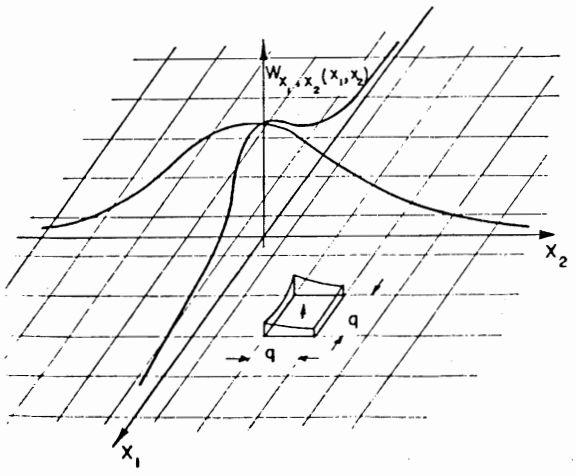


Fig. 14 (left). Volume sampling

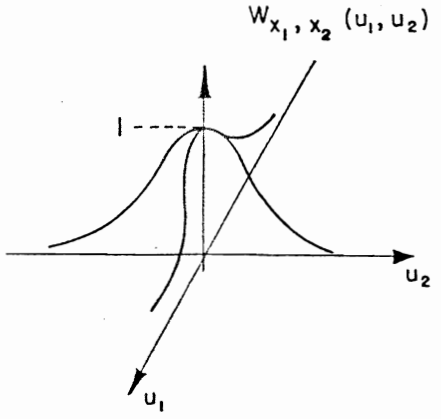


Fig. 15 (right). Two-dimensional input CF

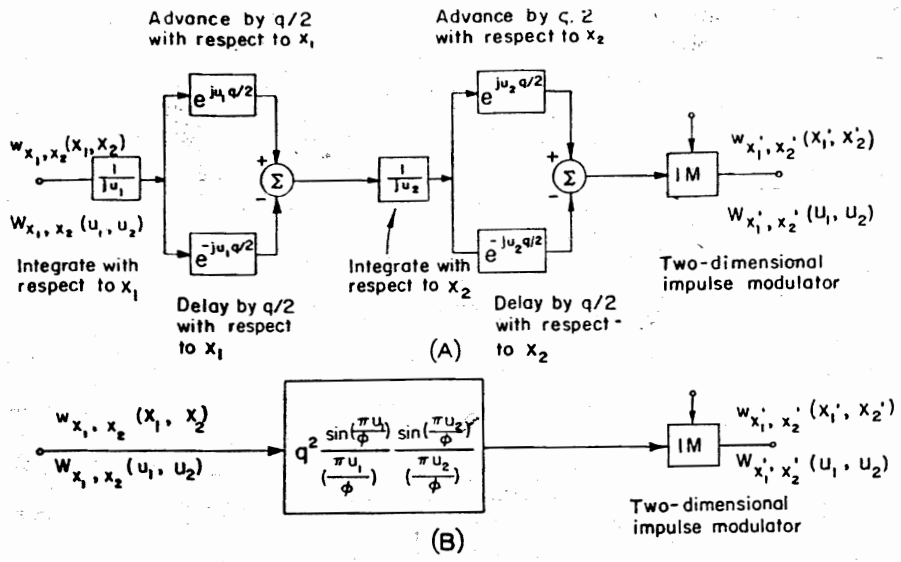


Fig. 16. Block diagram of volume sampling

A—Detailed process B—Simplified equivalent

form of the 2-dimensional signal presented to the impulse modulator, multiplied by $1/q^2$. It follows that the typical section of $W_{x_1',x_2'}(u_1, u_2)$ is the same as the CF of the sum of the quantizer input signal and a statistically independent noise having the CF.

$$W_{fl}(u_1, u_2)=\frac{q^2}{q^2}\frac{\sin(\pi u_1/\phi)}{(\pi u_1/\phi)}\frac{\sin(\pi u_2/\phi)}{(\pi u_2/\phi)} \quad (13)$$

Equation 14 is a formal expression for $W_{x_1',x_2'}(u_1, u_2)$.

$$W_{x_1',x_2'}(u_1, u_2) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} W_{x_1,x_2}(u_1+m\phi,\ u+n\phi)\times \frac{\sin\pi\left(\frac{u_1}{\phi}+m\right)\sin\pi\left(\frac{u_2}{\phi}+n\right)}{\pi\left(\frac{u_1}{\phi}+m\right)\pi\left(\frac{u_2}{\phi}+n\right)} \quad (14)$$

Fig. 17 illustrates the manner in which $W_{x_1',x_2'}(u_1, u_2)$ is derived from $W_{x_1,x_2}(u_1, u_2)$. Notice the analogy to Fig. 10. Equation 14 can be expressed in asterisk notation, where the asterisk represents 2-dimensional impulse modulation.

$$W_{x_1',x_2'}(u_1, u_2) = \left[q^2 W_{x_1,x_2}(u_1, u_2)\frac{\sin\frac{\pi u_1}{\phi}\sin\frac{\pi u_2}{\phi}}{\frac{\pi u_1}{\phi}\frac{\pi u_2}{\phi}}\right]^* \quad (15)$$

An examination of the CF of equation 16 shows it to be the CF of a degenerate second-order process, i.e., a first-order process.

$$\frac{\sin(\pi u_1/\phi)}{(\pi u_1/\phi)}\frac{\sin(\pi u_2/\phi)}{(\pi u_2/\phi)} \quad (16)$$

It consists of the product of independent factors. The random process that it describes could just as well be described by the following 1-dimensional CF.

$$\frac{\sin(\pi u/\phi)}{(\pi u/\phi)} \quad (17)$$

The CF of equation 16 is the Fourier transform of the 2-dimensional flat-topped DD shown in Fig. 18. It is apparent that the process described consists of noise samples that are statistically independent and distributed uniformly between $\pm q/2$.

It is again appropriate to compare quantization with the addition of an independent noise. Let the noise in this case be the same as the previous first-order flat-topped distributed independent noise. The 2-dimensional density of the quantizer output will be a 2-dimensional array of impulses, while the 2-dimensional probability density of the addition of the input signal x and the noise n will be continuous. The impulse DD will be the 2-dimensional impulse samples of the continuous DD.

If the CF of the quantizer input is band-limited in two dimensions, and if the grain size q is fine enough, i.e., if $W_{x_1,x_2}(u_1, u_2)=0$ for $u_1>\phi/2$ and/or for $u_2>\phi/2$, the sections of $W_{x_1',x_2'}(u_1, u_2)$ do not overlap. If this 2-dimensional quantizing theorem is satisfied, the 2-dimensional input density is recoverable

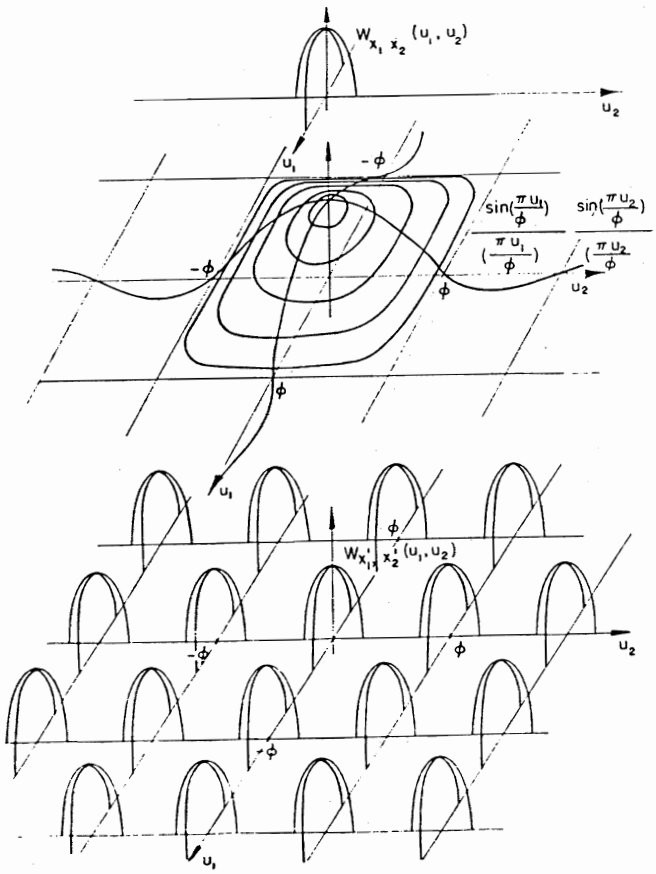


Fig. 17 (left). Formation of quantized 2-dimensional CF

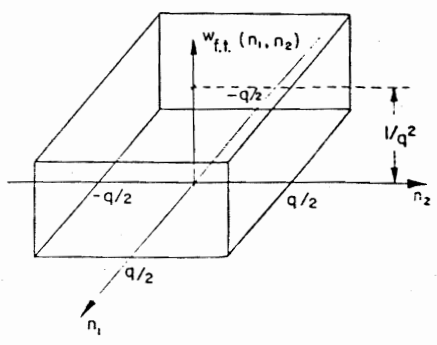


Fig. 18 (right). Two-dimensional flat-topped DD

from knowledge of the 2-dimensional output impulse density. The moments of x and all joint moments of x_1, x_2 may be derived from corresponding moments of x' and x_1', x_2', even when the 2-dimensional quantizing theorem is only half-satisfied. Under these conditions, all first-order and second-order (joint) moments of quantizer output signal are the same as if the quantizer were a source of additive, independent, uncorrelated flat-topped distributed noise. It should be recalled that although quantization noise is, in many respects, like independent noise, it is actually causally connected to the input signal. Its DD will next be shown to be independent of the signal's DD if the quantizing theorem is only half-satisfied.

Second-Order Probability Density of Quantization Noise

The derivation of the DD of quantization noise when the input signal is second-order proceeds by direct analogy to the 1-dimensional case. Following is a description of a 3-dimensional graphical procedure for getting the noise DD from the input DD. The joint input DD shown in Fig. 14 is sliced along the square quantization grid. The slices are then stacked and summed to give the 2-dimensional noise DD. Joint input events give joint noise events, whose probabilities are computed by summing their probabilities of occurrence within each box. This is analogous to the first-order procedure shown in Fig. 13, and again leads to a simple analytical derivation of the 2-dimensional characteristic function of the quantization noise.

The result of this derivation, given by equation 18, is seen to be a generalization of equation 9.

$$W_{(x_1-x_1'),(x_2-x_2')}(u_1, u_2) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} W_{x_1,x_2}(k\phi, l\phi) \times \frac{\sin \pi\left(\frac{u_1}{\phi}-k\right) \sin \pi\left(\frac{u_2}{\phi}-l\right)}{\pi\left(\frac{u_1}{\phi}-k\right)\pi\left(\frac{u_2}{\phi}-l\right)} \tag{18}$$

The sum of equation 18 consists of only the single term 19 when the 2-dimensional quantizing theorem is only half-satisfied.

$$\frac{\sin\left(\pi\frac{u_1}{\phi}\right)}{\left(\pi\frac{u_1}{\phi}\right)} \frac{\sin\left(\pi\frac{u_2}{\phi}\right)}{\left(\pi\frac{u_2}{\phi}\right)} \tag{19}$$

It follows that, under these conditions, although x_1 and x_2 are statistically connected, their corresponding quantization noise are statistically independent. These noises are uniformly distributed between plus and minus half a quantization box. Plausibility of a situation where statistically connected samples give independent quantization noises can be demonstrated by inspection of Fig. 6. Small changes in x cause large fractional changes in quantization noise. Noise samples are therefore far more erratic and show far less correlation than signal samples.

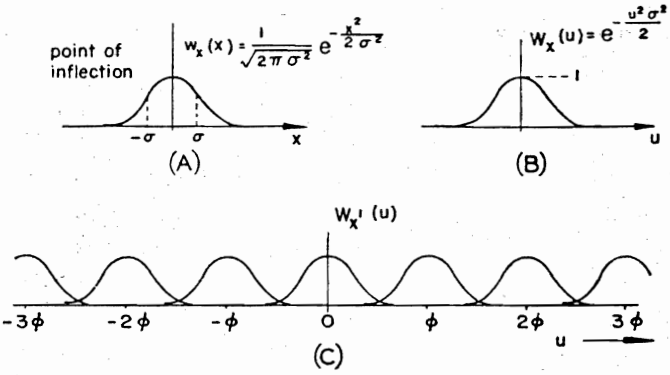


Fig. 19. Quantization of Gaussian process

A—Gaussian DD
B—Gaussian CF
C—CF after quantization

Satisfaction of Quantizing Theorem; Quantization of Gaussian Signals

Just as precisely band-limited signal waveforms occur infrequently in physical situations, precisely band-limited CF are rare also. From a practical standpoint, however, many signals are essentially band-limited, and at the same time, many CF's are almost band-limited. A few CF's and DD's will be examined.

A DD which is uniformly distributed over a certain range, like that of quantization noise, for example, has a CF of the form $(\sin u/u)$. This CF is not band-limited, but its amplitude decays with the factor $|u|$. A flat-topped CF, on the other hand, has a $(\sin x/x)$ DD. Although this CF is precisely band-limited, it could not be the CF of a physical process because it calls for negative probability density. Convolving two such CF's gives a triangular CF which is still precisely band-limited. Its DD is $(\sin x/x)^2$ which is never negative, but the moments are indeterminate.

A very interesting and important DD is the normal or Gaussian DD of equation 20. The standard deviation is σ, and $\bar{x}$ is the average or d-c level.

$$w_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \epsilon^{-(x-\bar{x})^2/2\sigma^2} \tag{20}$$

The corresponding CF is:

$$W_x(u) = \epsilon^{-\left(\frac{u^2\sigma^2}{2} - ju\bar{x}\right)} \tag{21}$$

Notice that the DD and CF amplitudes have the same shapes. The average $\bar{x}$ displaces $w_x(x)$ and causes a linear phase shift in its CF. Fig. 19 shows the Gaussian DD and CF for $x=0$. Notice that the Gaussian CF is not band-limited, but decays very sharply with the factor $\epsilon^{-u^2\sigma^2/2}$.

If a signal having this DD and CF were quantized, the sections of the quantizer output CF would overlap, as shown in Fig. 19(C). A typical section is of the form:

$$\left(\epsilon^{-\frac{(u+n\phi)^2\sigma^2}{2}}\right)\left(\frac{\sin \pi\left(\frac{u}{\phi}+n\right)}{\pi\left(\frac{u}{\phi}+n\right)}\right) \tag{22}$$

Fig. 19 suggests a way in which the extent of the satisfaction of the quantizing theorem can be computed. The overlap on section *2*, essentially due to sections *1* and *3* although actually due to the infinite number of sections, contributes to the derivatives of the CF at its origin and thereby has effect upon the moments of the quantizer output. An analysis based on the idea that the quantization noise is uniformly distributed between plus and minus half a box is somewhat in error. The nature of this analysis error as reflected in the moments has been computed and depends only upon the ratio of standard deviation to grain size.

Since the random Gaussian variable x was chosen with zero mean, the CF of x' is an even function. The overlaps of sections *1* and *3* cancel for all odd derivatives and reinforce for all even derivatives. All odd input and output moments are zero and all errors in odd moments are zero.

An expression for the contribution to the second derivative at $u=0$ because of the overlap of sections *1* and *3* is:

$$2\sigma^2 \epsilon^{-2\pi^2\frac{\sigma^2}{\phi^2}}\left(2+\frac{\phi^2}{\sigma^2}\frac{1}{2\pi^2}\right) \tag{23}$$

This causes a decrease in second moment. An expression for the contribution to the fourth derivative at $u=0$ because of the overlap of sections *1* and *3* is:

Table I

Box Size	Errors in Mean Square: In Quantization Noise*	Errors in Mean Square: In Quantizer Output†	Errors in Mean Fourth: In Quantization Noise‡	Errors in Mean Fourth: In Quantizer Output§
$q=\sigma$	$2.6(10)^{-10}q^2$	$1.1(10)^{-8}\sigma^2$	$1.4(10)^{-11}q^4$	$0.8(10)^{-6}\sigma^4$
$q=2\sigma$	$7.2(10)^{-4}q^2$	$3.1(10)^{-2}\sigma^2$	$3.9(10)^{-5}q^4$	$0.53\sigma^4$
$q=3\sigma$	$1.1(10)^{-2}q^2$	$0.54\sigma^2$	$6(10)^{-2}q^4$	$2.4\sigma^4$

* Mean square of flat-topped quantization noise is $(1/12)q^2$.
† Mean square of input is σ^2; of output is $(\sigma^2+1/12)q^2$.
‡ Mean fourth of flat-topped quantization noise is $(1/80)q^4$.
§ Mean fourth of input is $3\sigma^4$; of output is $3\sigma^4+(1/80)q^4+(1/2)q^2\sigma^2$.

$$2\sigma^4\epsilon^{-2\pi^2\frac{\sigma^2}{q^2}}\left[\frac{16\pi^2\sigma^2}{q^2}-\frac{q^2}{\sigma^2}(0.7)-\frac{q^4}{\sigma^4}(0.025)\right] \tag{24}$$

This causes an increase in fourth moment. These expressions have been used to compute the errors in quantizer output moments as entered in Table I.

When the quantizing theorem is not perfectly half-satisfied, the DD of the quantization noise is not necessarily uniformly distributed (in special cases it might be; e.g., where the input DD is itself uniformly distributed over the range $\pm q/2$). If the signals and noises were mutually independent, errors in the output moments could be used directly to derive the errors in moments of the noise. These analysis errors are of great interest to the systems engineer, and can be derived from differentiation of the CF of the quantization noise as given by equation 9. For a band-limited $W_x(u)$, the only term of importance in the sum, equation 9 is the one for $k=0$. For a nonband-limited input CF, the overlap from sections *1* and *3* corresponds to the $k=1$ and -1 terms.

A general expression for the error in second derivative at $u=0$ of the noise CF is:

$$+\left[W_x\left(\frac{2\pi}{q}\right)+W_x\left(\frac{-2\pi}{q}\right)\right]\frac{q^2}{2\pi^2} \tag{25}$$

An expression for the error in fourth derivative at $u=0$ is:

$$+\left[W_x\left(\frac{2\pi}{q}\right)+W_x\left(\frac{-2\pi}{q}\right)\right]\left(\frac{3}{2\pi^4}-\frac{1}{8\pi^2}\right)q^4 \tag{26}$$

The errors in mean square and mean fourth entered in Table I pertain to a Gaussian input with zero mean.

When a quantizer input of arbitrary DD has a nonzero mean, a general expression for the error in mean is:

$$\left(\frac{q}{2\pi}\right)W_x\left(\frac{2\pi}{q}\right)+W_x\left(\frac{-2\pi}{q}\right)=\left(\frac{q}{\pi}\right)W_x\left(\frac{2\pi}{q}\right)\sin\left(\frac{2\pi\bar{x}}{q}\right) \tag{27}$$

$W_x(u)$ in expression 27 is the CF of the input with the mean removed. Errors in odd moments vary sinusoidally with input d-c level and are maximum for $x=q/4$.

Maximum errors in mean for a Gaussian input are $8.3(10)^{-10}\ q$ for $q=\sigma$, $2.3(10)^{-3}\ q$ for $q=2\sigma$, and $3.5(10)^{-2}\ q$ for $q=3\sigma$. Since the mean of uniformly distributed quantization noise is zero, these errors represent biases developed by quantization.

When $q=3\sigma$, the actual quantization noise DD is a highly distorted version of the flat-topped DD. The distortion is evidenced by the disparity of 48% in its mean fourth. The nature of such distortion is interesting, but generally not of great importance in system analysis. The mean and mean square, on the other hand, are of crucial importance. At worst, the mean is in error by only $3\frac{1}{2}\%$ of a quantum level, and the mean square is in error by only 13% of $(1/12)\ q^2$. These remarkable results suggest the possibility of simple and effective statistical analyses for crude systems which contain 2- and 3-level quantizers.

When a correlated Gaussian signal is quantized, a question of importance is: how does the correlation coefficient (the ratio of the correlation or joint first moment to the mean square) of the quantizer output and the correlation coefficient of the quantizer noise vary with that of the input and with the ratio of standard deviation to grain size? The 2-dimensional Gaussian DD and its CF are described by equations 28.

$$W_{x_1,x_2}(x_1,x_2)=\frac{1}{2\pi\sigma^2\sqrt{1-\rho}}\epsilon^{-\frac{(1-\rho^2)}{2\sigma^2}[x^2_1-2\rho x_1x_2+x_2^2]}$$

$$W_{x_1,x_2}(u_1,u_2)=\epsilon^{-\frac{\sigma^2}{2}[u^2x_1+2\rho u_{x_1}u_{x^2}+u^2x_2]}$$

$$\rho\equiv\frac{(\overline{x_1x_2})}{\sigma^2}\equiv\text{correlation coefficient} \tag{28}$$

The 2-dimensional CF of the signal after quantization is similar to the CF in Fig. 17. Of interest are the effects of overlap upon the partial and cross-partial derivatives at the origin, which cause errors in an analysis based on the assertion that quantization noise is uncorrelated. The important CF sections are shown at the top of Fig. 20. Equation 29 is a formula for the zero-th section.

$$\frac{\sin\left(\pi\frac{u_1}{\phi}\right)}{\left(\frac{\pi u_1}{\phi}\right)}\frac{\sin\left(\pi\frac{u_2}{\phi}\right)}{\left(\frac{\pi u_2}{\phi}\right)}\epsilon^{\frac{-\sigma^2(u_1^2+2\rho u_1u_2+u_2^2)}{2}} \tag{29}$$

The autocorrelation of the quantizer output is determined by the cross-partial derivative at the origin. It turns out that the contributions of sections *1* and *2* to this derivative are opposite at the origin and therefore cause no error in the output autocorrelation. Likewise, the contributions of sections *3* and *4* cancel. The contributions of *7* and *8* reinforce and are opposite, in a sense, to those of *5* and *6*. The latter four sections are mainly responsible for the error in the output autocorrelation. Sections *5* and *6* are particularly important and overlap heavily for high positive input correlations, when the sections become elongated as illustrated in Fig. 20. Sections *7* and *8* overlap heavily when input correlations are large and negative. Overlap is increased both by increase in grain size and increase in magnitude of input correlation.

The error in output autocorrelation due to overlap has been computed and is given very closely by the approximation:

(Error in output autocorrelation)

$$\approx\frac{q^2}{12}\epsilon^{-(1-\rho)4\pi^2\frac{\sigma^2}{q^2}} \tag{30}$$

When the 2-dimensional quantizing theorem is at least half-satisfied, the quantization noise itself is uncorrelated. The double sum, equation 18, has only a single term. When this condition is not met, the double sum has other terms

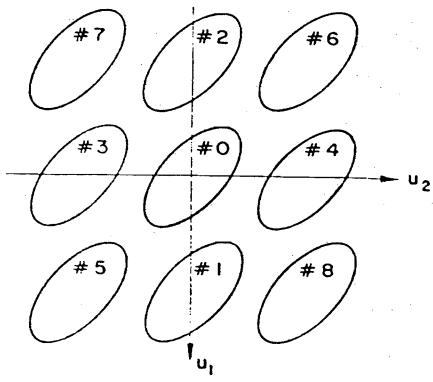


Fig. 20. Constant contours of quantized CF

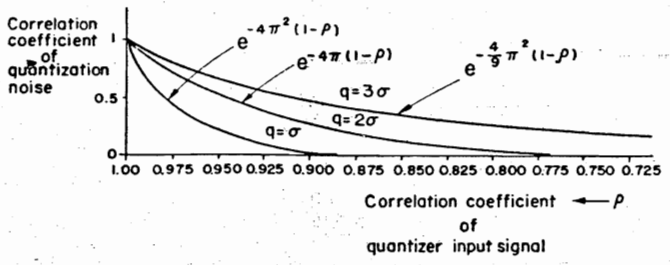


Fig. 21. Correlation of noise versus correlation of Gaussian input

which overlap with the zero, zero term, and cause errors in the autocorrelation of the quantization noise. These errors comprise the correlation of the quantization noise, since without these errors the noise is completely uncorrelated. The autocorrelation of the quantization noise can therefore be derived from differentiation of equation 18 at the origin. Such a procedure is not actually necessary, for it will next be shown that when the first-order quantizing theorem is satisfied to a good approximation, the error in the quantizer output correlation is equal to the error in the autocorrelation of the noise itself. The output autocorrelation is equal to the input autocorrelation plus the correlation of the noise.

The correlation of a quantizer output signal is given by equation 31. The noises n_1 and n_2 are causally related to x_1 and x_2 respectively.

$$\overline{(x_1+n_1)(x_2+n_2)}=\overline{x_1x_2}+\overline{n_1x_2}+\overline{n_2x_1}+\overline{n_1n_2} \tag{31}$$

This assertion is proved if it can be shown that n_1x_2 and n_2x_1 are zero. When the correlation is very high, $\rho\rightarrow 1$ say, equation 31 becomes the same as equation 32, which gives the mean square of the quantizer output.

$$\overline{(x_1+n_1)(x_1+n_1)}=\overline{x_1^2}+\overline{2n_1x_1}+\overline{n_1^2} \tag{32}$$

Assuming that the first-order quantizing theorem is satisfied to a good approximation, the output mean square is essentially that of the input plus that of the noise. The cross correlation $\overline{2n_1x_1}$ is therefore nil. It follows that the average $\overline{2n_1x_2}$ is also nil, being even less than the former when ρ is less than 1. The same argument applies to $\overline{n_2x_1}$. This completes the proof.

The error in the output autocorrelation given by the relation 30 is the same as the correlation of quantization noise. A plot of this relation is shown in Fig. 21. The same curves can be used for negative as well as positive values of ρ. The correlation of quantization noise is positive if the input correlation is positive and negative if the input correlation is negative. The curves are normalized, and it should be noted that the absolute correlation of the noise equals the noise correlation coefficient multiplied by $q^2/12$. These results agree with correlations of Gaussian signals derived by Bennett[3] from a different point of view.

It is interesting to note that the multidimensional quantizing theorem can never be satisfied by a continuous signal. Its quantization noise cannot be uncorrelated for arbitrarily small time shifts. The noise autocorrelation function is more highly peaked than the input autocorrelation function, however, meaning that the noise is far more "random" than the signal. Its power density spectrum is broader.

It is also interesting to note that the autocorrelation function of the quantizer output equals the input autocorrelation function plus that of the noise when the first-order quantizing theorem is half-satisfied. In this respect, the causally connected noise adds to the signal as if it were statistically independent of it.

Part III. Systems Applications

Two categories of systems applications of the statistical theory of amplitude quantization will be discussed. These are open-loop and closed-loop applications.

The open-loop applications are concerned mainly with the making of statistical measurements. The basic problem is that of being able to recover actual process statistics from quantized measurements of the process. The recovery problem includes both DD's and moments. The closed-loop applications include the analysis of quantized sampled-data feedback systems, and the improvement of performance of quantized systems by injection of external "dither" signals which have the ability to insure satisfaction of the quantizing theorem, regardless of the nature of actual system input signals.

Sheppard's Corrections for Grouping

When the 1-dimensional quantizing theorem is half-satisfied, quantization noise acts like an independent noise which is uniformly distributed between $\pm q/2$, has mean square of $q^2/12$, mean fourth of $q^4/80$, and has zero odd moments. The moments that result when such a noise is added to the quantizer input signal x is expressed by the following equations:

$$\overline{(x+n)}=\bar{x}+\bar{n}=\bar{x}$$

$$\overline{(x+n)^2}=\overline{x^2}+2\bar{x}\,\bar{n}+\overline{n^2}=\overline{x^2}+\frac{1}{12}q^2$$

$$\overline{(x+n)^3}=\overline{x^3}+\overline{3x^2}\,\bar{n}+3\bar{x}\,\overline{n^2}+\overline{n^3}=\overline{x^3}+\frac{1}{4}q^2\bar{x}$$

$$\overline{(x+n)^4}=\overline{x^4}+\overline{4x^3}\,\bar{n}+\overline{6x^2n^2}+\overline{4\bar{x}\,n^3}+\overline{n^4}$$

$$=\overline{x^4}+\frac{1}{2}q^2\overline{x^2}+\frac{1}{80}q^4 \tag{33}$$

and so on.

These equations are based on the facts that the average of the sum is the sum of the averages, and that the average of the product is the product of the averages for statistically independent n. The moments of the unquantized DD given by equations 34 are obtained very directly from those of the quantized DD.

$$\bar{x}=\bar{x}'$$

$$\overline{x^2}=\overline{x'^2}+\left(-\frac{1}{12}q^2\right)$$

$$\overline{x^3}=\overline{x'^3}+\left(-\frac{1}{4}q^2\bar{x}'\right)$$

$$\overline{x^4}=\overline{x'^4}+\left(-\frac{1}{2}q^2\overline{x'^2}\frac{1}{240}q^4\right) \tag{34}$$

and so on.

The expressions in parentheses in equation 34 are known as Sheppard's corrections, first reported in 1898 by Dr. W. F. Sheppard.[4] His derivation made use of the Poisson sum formula. It should be noted that these corrections are quite accurate even for extremely rough quantization, when q is as big as three standard deviations for a smoothly distributed process.

Interpolation of First-Order DD from Histogram

When the quantizing theorem is satisfied to a good approximation, it is possible to recover a first-order DD either from its histogram or its quantized impulse DD. An inspection of Fig. 9 aids in recollection of the mechanism by which $w_{x'}(x')$ is formed from $w_x(x)$ and suggests a way of reversing the process. If the quantizing theorem is satisfied, the inverse of impulse modulation is sin x/x filtering. The inverse of the linear discrete filtering $(\epsilon^{juq/2}-\epsilon^{-juq/2})$ is linear discrete filtering via the following transfer function.

$$\left(\frac{1}{\epsilon^{juq/2}-\epsilon^{-juq/2}}\right)=\left(\frac{\epsilon^{-juq/2}}{1-\epsilon^{-juq}}\right) \qquad (35)$$

The inverse of integration is differentiation with the transfer function (ju).

Fig. 22(A), a block diagram the inverse of that of Fig. 9, shows how $w_x(x)$ can be recovered from $w_{x'}(x')$. The ordering of discrete filtering and sin x/x filtering has been reversed to give a form which can be realized graphically with greater facility. The remainder of Fig. 22 demonstrates the graphical implementation of this process as applied to a Gaussian signal which has been quantized to a granularity of $q = 2\sigma$. Since 99.7% of the area of the DD is contained between $\pm 3\sigma$, the histogram contains essentially three "bars," the impulse DD has three impulses.

In Fig. 22, the original DD is recovered fairly accurately from a very rough histogram containing only three bars. The same technique has been applied to a variety of experimental histograms with comparable success.

Recovery of Autocorrelation Function from Roughly Quantized Process Samples

Consider the autocorrelation function $\phi_{x'x'}(\tau)$ of a quantized sampled signal illustrated by Fig. 23. When the high order quantizing theorem is half-satisfied, this discrete correlation function is the same as that of the unquantized sampled signal, except for the $\tau = 0$ point. The mean square is increased by $q^2/12$ as a result of the addition of the flat-topped uncorrelated quantization noise. Since the autocorrelation of the quantization noise is zero except for zero shift, the only effect of quantization is upon the $\tau = 0$ point. This has been demonstrated many times by digital simulation. Data could be deliberately rounded off with larger and larger grain size, and the successive autocorrelation functions can be computed. Until the multidimensional quantizing theorem is violated, they differ only in the mean-square point. This effect is indicated in Fig. 23.

When the first-order quantizing theorem is satisfied to a good approximation, but the samples being quantized are so highly correlated that the quantization noises are correlated, the noise correlations add to the input correlations. The points on the autocorrelation curve showing the highest correlation are first affected as the grain size is increased, and this effect will be in accord with the curves of Fig. 21 if the quantizer input is Gaussian.

For a grain size of $q=3\sigma$, a Gaussian signal can be thought of as essentially quantized to two levels (the quantizer here is a no-dead-zone type). Making use of the curve for $q=3\sigma$ in Fig. 21, it can be shown with a small amount of calculation that correlation of quantization noise causes as much as a 15% change in the correlation of the quantizer output only if the correlation coefficient of the input is as high as 50%. This means that fairly accurate autocorrelation functions could be obtained from 2-level signals. In binary, each sample would be 1 bit of datum. A similar conclusion was drawn by Van Vleck,[5] who derived the autocorrelation function of clipped Gaussian noise with zero mean by an analytical method similar to the one used subsequently by Bennett.[3]

One-bit autocorrelation functions have been measured very successfully by Dr. James F. Kaiser, and his results are reported in reference 6.

A new area of application of 1-bit statistics will no doubt be in the field of space communications over millions of miles, where the real-time detection of threshold signals will be done by digital correlation techniques. Allowances in the complexity of computing and recording equipment are minimal here, and great advantage is to be obtained in being able to process, with very crude and simple digital apparatus, input signals which are completely buried in noise. Message information will be carried by the statistics of the transmitted signal amplitudes, rather than by the direct amplitudes. Detection will be based on moments of the received signals. As we have seen, moments may be recovered from roughly quantized measurements.

At my suggestion, the radar echo data from the planet Venus, originally used in its history-making detection,[7] has been deliberately rounded off to four levels and to two levels and rerun on the data reduction computer at Lincoln Laboratory, Massachusetts Institute of Technology. The radar returns were originally quantized to 64 levels. Preliminary tests have shown that detectability was not appreciably diminished in going to four levels

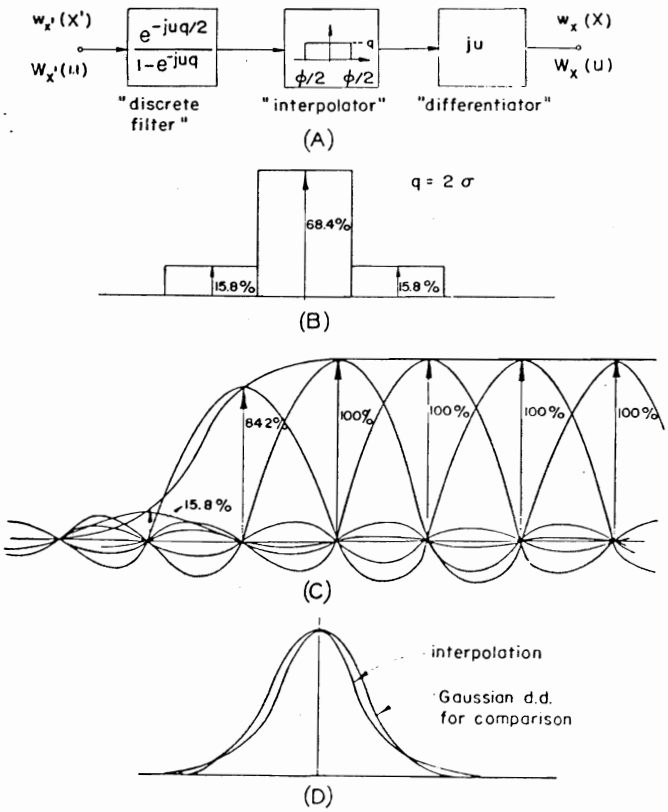
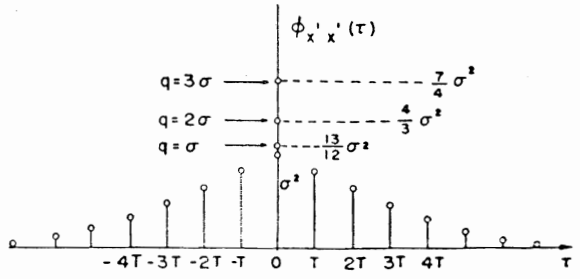


Fig. 22 (left). Interpolation of crude Gaussian histogram

A—Block diagram of interpolation
B—Histogram of Gaussian process
C—Summing and (sin X/X) interpolation
D—Derivative

Fig. 23 (below). Effect of quantization on autocorrelation function



and only began to show serious deterioration when signals were quantized to two levels. However, with the signals quantized to two levels, detection was still very positive; with some increase in sample size this might have been even more positive.

Analysis of Quantized Sampled-Data Feedback Systems

The stability of a quantized feedback system is unaffected by the presence of the quantizer.

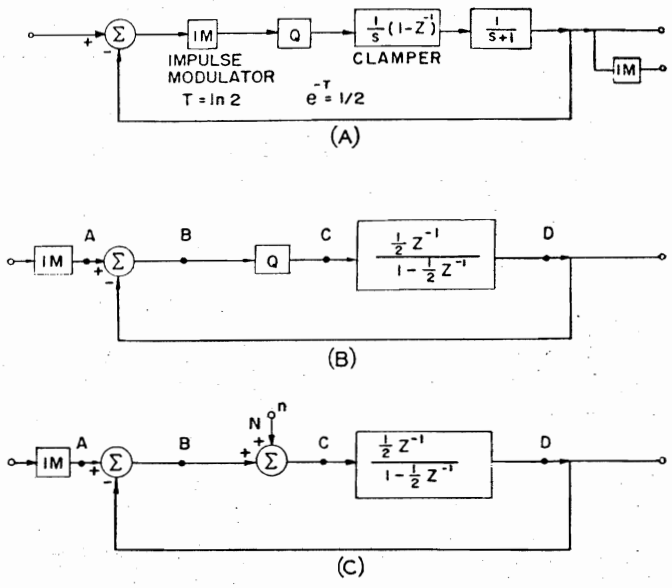
A test for stability may be made by replacing the quantizer with a gain of unity. If the resulting linear system is stable, then the quantized system is stable. The effect of quantization is to inject a bounded noise. A bounded noise in a stable system cannot cause an unbounded output. Stable-amplitude limit cycles are not precluded, however. These are quite possible and are most common when the quantizer is a no-dead-zone type.

The results of the quantization of Gaussian signals indicate that the first-order quantizing theorem would be half-satisfied to a good approximation by any signal having a fairly smooth DD whose dynamic range covers at least two or three quantization levels, and that the quantization noise will have correlation coefficients less than a few per cent even when the signal has correlation coefficients as great as 80%. These conditions are often met in quantizer control systems.

The output of a quantizer differs from the input by a quantity which varies between $+q/2$ and $-q/2$. An exact description of this difference is difficult to obtain. On the other hand, statistical methods will allow relatively simple calculation of the probability density distribution of this difference, the quantization noise. The beauty of the method arises from the fact that a large class of inputs to the quantizer, whose properties can be well defined, yield identical quantization noise statistics. These statistics characterize the quantization process. The quantizer output is the sum of the two inputs plus a quantization noise which is a different time waveform but has the same statistics as in the previous case. In this sense, superposition applies, and the quantizer acts like a linear device when the quantizing theorem is half-satisfied.

The nature of the quantization noise and the way in which this noise propagates and affects system outputs may be determined with great facility when the quantizing theorem is half-satisfied. It is necessary to be able to determine from the input and system characteristics the extent to which the quantizing theorem is satisfied by a quantizer embedded in a feedback loop.

Fig. 24. Simple quantized sampled-feedback system



The effect of the quantization noise itself upon the satisfaction of the quantizing theorem is twofold: to both help and hinder. It helps when it provides extra variance and extra dynamic range to the quantizer input signal, and it hinders when the wideband CF of the quantizer output is fed back to the quantizer input and broadens its CF. If the quantization noise were statistically independent of the quantizer input signal, then the noise would be guaranteed to aid in the satisfaction of the quantizing theorem. The characteristic functions all over the system would be narrower and the dynamic range of all signals would be increased.

The effects of the causal connection between signals and noise, as manifested in the impulsive nature of a quantizer output DD tends to be mollified in the course of the feedback process. Ordinary low-pass filtering within the feedback path can be shown to produce an effect on characteristic functions which is like low-pass filtering in the CF domain. Also the injection of the input signal has a low-pass filtering effect on CF's. Often these smoothing effects on the CF's are adequate so that the input signal to the quantizer has the same statistical characteristics as if the quantization noise were independent. On the other hand, there are cases where the causality between signals and noises has an appreciable effect on the CF of the quantizer input.

The approach to be taken here is the following one. The quantizing theorem will be assumed satisfied to a good approximation. An analysis of the noise appearing at the quantizer output based on this assumption will be made. Errors in this analysis will be calculated by testing for the satisfaction of the quantizing theorem at the quantizer input. The test will be made by ignoring the effects of the noise itself. Conditions will be derived where ignoring the quantization noise will be justified. Cases where this is not justified turn out to be rare and will not be treated here.

In Fig. 24(A) a sampled-data unity-feedback control system is shown. The error signal is both sampled and quantized and the sampling period $T=ln2$. Fig. 24(B) follows from (A) by the usual reduction procedures.[8] The output of Fig. 24(B) is the samples of the output of Fig. 24(A). The system of Fig. 24(C) follows from (B) when the quantizing theorem is half-satisfied.

Various transfer functions will be needed in the evaluation of the system output noise and in testing for the satisfaction of the quantizing theorem at the quantizer input. The discrete transmission from point A to point D in Fig. 24(C) is:

$$T_{AD}=\frac{1}{2}z^{-1} \tag{36}$$

The transmission from A to B is:

$$T_{AB}=1-\frac{1}{2}z^{-1} \tag{37}$$

The transmission from point N to point D is:

$$T_{ND}=T_{AD}=\frac{1}{2}z^{-1} \tag{38}$$

The transmission from N to B is:

$$T_{NB}=-\frac{1}{2}z^{-1} \tag{39}$$

Assume that the system is excited by a certain input and that the box size q is

sufficiently small so that the multidimensional quantizing theorem is satisfied. The DD of the noise component appearing at the output will be uniformly distributed between $\pm q/4$ and will be a first-order process. This follows from knowledge of T_{ND}. The noise component in the quantizer input signal which enters by the way of the feedback path is also uniformly distributed between $\pm q/4$ as may be seen from consideration of T_{NB}. As long as the variance of this noise component is small compared to the variance of the signal component of the quantizer input, the effect upon the satisfaction of the quantizing theorem of the quantization noise itself may be ignored. The variance of the quantization noise component at point B is $q^2/48$.

Let the system input be Gaussian where samples at point A are first-order and have the variance σ^2. The signal variance at point B equals σ^2 multiplied by the sum of the squares of the impulses of the transmission T_{AB}. This variance is therefore $(5/4)\sigma^2$. As long as the signal variance is, say, five times as great as that of the noise at point B, the effect of quantization noise upon the satisfaction of the quantizing theorem may be safely ignored. This occurs for $q<12\sigma$.

The signal component at point B is Gaussian, has a mean square of $5/4\ \sigma^2$, and the following power-density spectrum

$$\Phi_{BB}(z)=\sigma^2\left(\frac{5}{4}-\frac{1}{2}z^{-1}-\frac{1}{2}z\right) \qquad (40)$$

The first-order quantizing theorem will be satisfied to a good approximation when q is as big as three standard deviations of the signal at point B, that is; $q=(3/2)\ 5\sigma$. With this granularity, quantization noise has a mean square which is only 13% different from that of flat-topped distributed noise. Knowing the autocorrelation function of the signal and making use of equation 30 or the graph of Fig. 21, the correlation coefficient of the noise for a shift of one sample time is 8%. The power density spectrum of the noise at point N, is, therefore, to a good approximation:

$$\Phi_{nn}(z)=\frac{q^2}{12}\ (1-0.08z^{-1}-0.08z) \qquad (41)$$

The power spectrum of the samples of the noise at the system output is the same as this spectrum multiplied by one quarter.

This simple example illustrates the method which is usable in the analysis of quantized feedback control systems when inputs are Gaussian. Non-Gaussian inputs cause almost Gaussian signals within systems that provide adequate integration and smoothing in accord with the central limit theorem. Such cases may be treated by the foregoing mean square and autocorrelation techniques. A more exacting analysis would require knowledge of how CF's propagate in linear systems.[9] Testing for the satisfaction of the quantizing theorem would have to be done for the particular CF that develops at the quantizer input point.

Quantization noise has a standard deviation of $\sqrt{q^2/12}$ and this standard deviation "propagates" from where it arises to an output point via the square root of the sum of the squares of the impulse response of the transmission path. The noise is bounded between $\pm q/2$; this bound propagates via the sum of the magnitudes of the impulse response of the same transmission path.[10] The bound and the standard deviation are usually easy to calculate and together give an excellent picture of the quantization noise DD.

Linearization of Quantized Feedback Systems by Injection of External Dither

When the quantizing theorem is satisfied, the statistical performance of a quantized system is predictable and may be specified in general without regard to the nature of the system's input.

It is always possibe, at least theoretically, to inject an independent external signal to insure that the quantizing theorem will be satisfied to a good approximation. The external signal, or dither, could be made to assist the already present input signal in the satisfaction of the quantizing theorem or could be made to suffice alone.

Once an input component excites a system so that the quantizing theorem is satisfied, no other statistically independent input component could undo this condition because, when independent signals are added, CF's multiply. If the multidimensional CF of the quantizer input is already band-limited, then multiplication by another multidimensional CF yields a product which is still band-limited (perhaps even narrower).

Some of the advantages gained by the use of external dither are that quantizer systems can be made to be "small-signal linear," average values of quantization noise can be made to be very close to zero, quantization noises can be made to have predictable variances, statistical bounds which might be much smaller than absolute bounds can be insured, and limit cycles and the sometimes associated mode switchings can be eliminated. Some of the arguments against the use of dither are that more equipment and/or more computing is necessary to generate and apply dither, and that it causes greater output noise in certain systems. The latter objection does not always apply. Usually the reverse is true, particularly in cases where the quantizer is followed by low-pass filtering.

Consider again the quantized sampled-data system shown in Fig. 24. It has been shown that the variance at the quantizer input (point B) because of quantization noise is $q^2/48$, and the variance at point B due to the first-order input of variance σ^2 is $5\sigma^2/4$. In order that the grain size q equal three standard deviations of the signal at point B, a minimal requirement for approximate satisfaction of the quantizing theorem, the standard deviation of a first-order Gaussian input (the dither in this case) must be:

$$\sigma=\frac{2\sqrt{5}}{15}q$$

The quantization noise output has a variance of $q^2/48$ and the dither output component has a variance of $q^2/45$, giving a net standard deviation of about $0.2q$. With no dither, the bound on the output noise is:

$$\frac{q}{2}\left(\frac{1}{2}\right)=0.25q$$

since the discrete impulse response from point B to the output is $z^{-1}/2$. Use of dither offers no real advantage here in reducing systems noise.

Suppose that the system of Fig. 24(A) is quiescent, with signals at all points zero. A constant input whose amplitude lies between $\pm q/2$ will go undetected. Small amplitude linearity could be achieved, on the other hand, by adding a continuous Gaussian signal at the system input whose samples are a first-order process, or by injecting first-order Gaussian samples directly at the quantizer input point. The same effect could be accomplished by injecting dither at any point in the system, but more output variance might result than if the dither were injected directly. Approximate satisfaction of the quantizing theorem as indicated above insures that the d-c bias developed across the quantizer will be less than 3.5% of q.

If the quantizer in the system of Fig. 24 is replaced by one having no dead zone, the system shown in Fig. 25(A) results. The conditions for the satisfaction of the quantizing theorem are the same and, when it is satisfied, there is very little difference between the systems. Marked

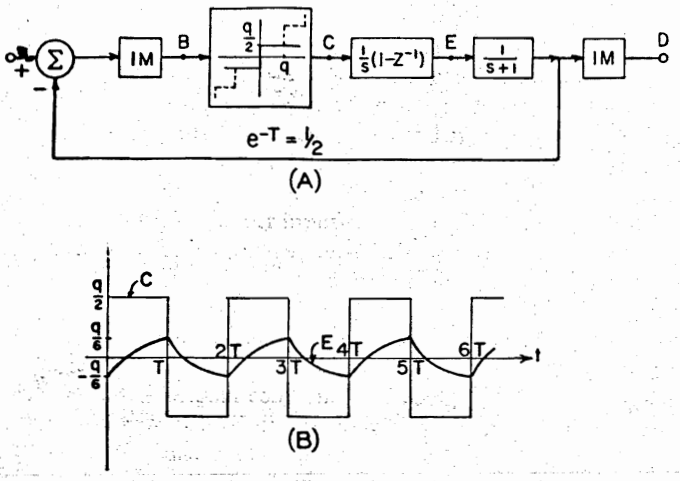


Fig. 25. Saturating no-dead-zone quantizer, zero-input limit cycle

differences appear when the quantizing theorem is not satisfied.

When the input in Fig. 25(A) is zero, a very interesting phenomenon takes place. A stable-amplitude oscillation or limit cycle of Fig. 25(B) develops.

It can be shown that a constant input whose amplitude varies between $\pm q/6$ would not change the nature of the limit cycle and therefore would remain undetected. Although the quantizer has no dead zone, the entire system has one. It is smaller by a factor of three than in the previous case.

If the system of Fig. 25 were dithered by a first-order Gaussian input whose standard deviation were greater than $(2\sqrt{5}/15)q$, the quantizing theorem would be approximately half-satisfied. The quantizer would generate uniformly distributed almost uncorrelated noise. Limit-cycle phenomena would not take place and the system would be linearized for small (or large) amplitude inputs.

Dithering may be effected in varying degrees by the actual system input signals. Outputs would consist of sums of irregular limit-cycle components and signal responses. The nearer the signals come to the half-satisfaction of the quantizing theorem, the more irregular and the closer to uncorrelated noise the limit-cycle components would be.

Use of external dither is especially indicated where a quantizer is followed by a low-pass filter. A broad-band dither signal having the required variance at the quantizer input point insures a broad-band $q^2/12$ quantization noise. Dither and quantization noise variances at a system output can be made arbitrarily small by making the dither bandwidth arbitrarily large. In a sampled system, the optimum dither signal at the quantizer input is first-order.

Dither signals need not necessarily be Gaussian or even random. Periodic signals, sine waves for example, are good possibilities. It can be shown that when peak to peak of a sinusoidal dither equals $2q$, the first-order quantizing theorem is reasonably well half-satisfied (25% error in mean square of the noise) and that the per-cent correlation of the noise is, at most, equal to that of the dither. In a continuous system, quantization noise can be made arbitrarily small by increasing dither frequency. In a sampled system, the best frequency is usually close to half the sampling rate. A considerable amount of analytical work on the quantization of sinusoids has been done by Furman.[11]

The system of Fig. 25 with a saturating quantizer (the solid-line characteristic) may be analyzed by the same methods, if the combination of dither and signal at the quantizer input has a DD which approximatey half-satisfies the quantizing theorem and has a dynamic range essentially covering only two quantizing levels. A Gaussian dither whose standard deviation equals $q/3$ and a signal bounded between $\pm q/2$ would cause a clipped output close to that which would result if the clipper were replaced by a gain of unity and an additive $q^2/12$ quantization noise. The same is true of a sine-wave dither of amplitude $(3/4)\ q$ and a signal bounded between $\pm q/3$. These ideas have been verified by simulation. Interesting comparisons have been made between the point of view of this paper and a describing function method of analysis for systems with saturating quantizers.[12]

Summary

1. A comparison of the addition of statistically independent first-order uniformly distributed noise (between $\pm q/2$) with quantization shows that a quantized DD consists of impulse samples of the signal-plus-noise DD. Quantization noise is causally connected to the quantizer input signal.

2. Satisfaction of the quantizing theorem (analogous to the sampling theorem) insures that a DD can be recovered from a quantized DD. Half-satisfaction of the quantizing theorem insures that moments can be recovered and that quantization noise is uniformly distributed between its bounds, $\pm q/2$.

3. When a Gaussian signal is quantized and the box size q is as big as three standard deviations, the quantizing theorem is half-satisfied to a good approximation. There is a maximum error of 3.5% of q in mean and 10% of $q^2/12$ in mean square. A signal with a correlation coefficient of 80% will cause a quantization noise with 30% correlation. A signal with 60% correlation will cause a noise with 9% correlation. Approximate satisfaction of the quantizing theorem will be achieved by almost any signal having a dynamic range covering as few as three quantum levels.

4. When the quantizing theorem is half-satisfied at a quantizer input point in a feedback system, the quantizer may be replaced by an additive noise source of mean zero and mean square $q^2/12$. A test for the satisfaction of the quantizing theorem can usually be made by ignoring the effects of the quantization noise itself.

5. An external dither signal can make a quantized feedback system be "statistically linear" for small and large signals, can eliminate low-frequency limit cycles, and can eliminate the effects of hysteresis and "dead zone." Random and periodic time functions make useful dither signals and have the effect of catalysts in improving system performance, yet do not appear appreciably in system outputs. Injection of an external dither can convert very crude control systems containing rough quantization, saturating quantization, or even hysteresis (as in contactor systems), to beautifully linear, almost noise-free control systems.[13]

References

1. A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory, B. Widrow. *Sc.D. Thesis*, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., June 1956.

2. A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory, B. Widrow. *Transactions*, Professional Group on Circuit Theory, Institute of Radio Engineers, New York, N. Y., vol. PGCT-3, no. 4, Dec. 1956.

3. Spectra of Quantized Signals, W. R. Bennett. *Bell System Technical Journal*, New York, N. Y., vol. 27, July 1948, pp. 446-72.

4. On the Calculation of the Most Probable Values of Frequency-Constants, for Data Arranged According to Equidistant Divisions

OF SCALE, W. F. Sheppard. *Proceedings*, London Mathematical Society, London, England, vol. 29, 1898, p. 353.

5. THE SPECTRUM OF CLIPPED NOISE, J. H. Van Vleck. *Report 51*, Radio Research Laboratory, Harvard University, Cambridge, Mass., July 21, 1943.

6. NEW TECHNIQUES AND EQUIPMENT FOR CORRELATION COMPUTATION, J. F. Kaiser, R. K. Angell. *Report 7668-TM-2*, Servomechanisms Laboratory, Massachusetts Institute of Technology, Dec. 1957.

7. RADAR ECHOES FROM VENUS, R. Price, P. E. Green, et al. *Science*, Washington, D. C., vol. 129, no. 3351, Mar. 20, 1959, pp. 751–53.

8. SAMPLED-DATA CONTROL SYSTEMS (book), J. R. Ragazzini, G. F. Franklin. McGraw-Hill Book Company, Inc., New York, N. Y., Sept. 1958.

9. PROPAGATION OF STATISTICS IN SYSTEMS, B. Widrow. *WESCON Convention Record*, Institute of Radio Engineers, pt. 2, 1957, pp. 114–21.

10. THE EFFECT OF QUANTIZATION IN SAMPLED-FEEDBACK SYSTEMS, J. E. Bertram. *AIEE Transactions*, pt. II (*Applications and Industry*), vol. 77, Sept. 1958, pp. 177–82.

11. IMPROVING PERFORMANCE OF QUANTIZED FEEDBACK SYSTEMS BY USE OF EXTERNAL DITHER, G. G. Furman. *S.M. Thesis*, Department of Electrical Engineering, Massachusetts Institute of Technology, Jan. 1959.

12. CAUSAL AND STATISTICAL ANALYSES OF DITHERED SYSTEMS CONTAINING THREE-LEVEL QUANTIZERS, R. C. Jaffe. *S.M. Thesis*, Department of Electrical Engineering, Massachusetts Institute of Technology, Aug. 24, 1959.

13. LINEARIZATION OF CONTACTOR CONTROL SYSTEMS BY EXTERNAL DITHER SIGNALS, T. Ishikawa. *Technical Report 2103-2*, Electronics Laboratories, Stanford University, Stanford, Calif. Oct. 1, 1960.

A reprint from APPLICATIONS AND INDUSTRY, published by American Institute of Electrical Engineers

Printed in the United States of America

January 1961 issue