A STUDY OF ROUGH AMPLITUDE QUANTIZATION BY MEANS OF NYQUIST SAMPLING THEORY

by

BERNARD WIDROW

S. B., Massachusetts Institute of Technology (1951)

S. M., Massachusetts Institute of Technology (1953)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
(June 1956)

Signature of Author ______________________
Department of Electrical Engineering

Certified by ______________________
Thesis Supervisor

Accepted by ______________________
Chairman, Departmental Committee on Graduate Students

May 7, 1956

Mr. Bernard Widrow
Room 10-397B
M.I.T.

Dear Mr. Widrow:

This letter is to give you permission to print additional copies of your thesis by the multilith process, and to submit to the Department of Electrical Engineering copies thus produced, in lieu of the typed copies normally required.

A copy of this letter is to be reproduced by the same process and is to be placed in each copy of the thesis immediately following its title page.

Sincerely yours,

S. H. Caldwell
for the
Department Graduate Committee

SHC/mm

A STUDY OF ROUGH AMPLITUDE QUANTIZATION BY MEANS OF NYQUIST SAMPLING THEORY

by
BERNARD WIDROW

Submitted to the Department of Electrical Engineering on May 21, 1956 in partial fulfillment of the requirements for the degree of Doctor of Science.

ABSTRACT

Quantization takes place whenever a physical quantity is represented numerically. This is common in P.C.M. communication systems and modern digital data processing systems where it is known as analog-to-digital conversion. In the numerical solution of equations it is called round-off. A quantizer is either a device or a mathematical operator that assigns a value to a variable equal to its closest integer. Quantization is a non-linear process that distorts in a definite manner. Since the distortion itself is usually undesirable, it will be described by and called an additive noise.

Let a signal be sampled and then quantized. The probability density distribution of the quantizer input is continuous while the probability density of the quantizer output signal is discrete, consisting of a series of impulses uniformly separated with spacing equal to the quantization box size q. The joint input-output probability density distribution, a most general distribution, is continuous along the input dimensions and sampled along the output dimensions. In characteristic function space, which is Fourier transformed from the multidimensional amplitude space of probability, the joint output-input c.f. is periodic with the radian fineness $\phi = 2\pi/q$ along the dimensions corresponding to the output variables and aperiodic along the equal number of input dimensions. The typical repeated section is identical to the joint input-output c.f. that results when the quantizer is replaced by a source of first order independent additive noise that is uniformly distributed over a box width. Actual quantization noise is not independent of the quantizer input but causally related to it (a given input gives a definite output and a definite noise). Outputs take on only discrete levels giving impulse probability densities, and characteristic functions are periodic along output dimensions rather than aperiodic as they would be if quantization noise were independent.

The typical sections do not overlap if a multidimensional Nyquist restriction upon the probability density of the high order input is met, i.e., if the c.f. of the input is zero (negligible) outside a hypercube of edge ϕ centered at the origin in c.f. space. This is a Nyquist restriction on the "bandwidth" of the probability density distribution of the quantizer input signal, and is to be distinguished from the Nyquist sampling restriction upon the bandwidth of that signal itself. The former pertains to quantization which is like sampling in amplitude while the latter pertains to the usual sampling in time. Under conditions of no overlap in c.f. space, all moments are the same as independent noise since they are determined by the derivatives of the joint c.f. at the origin in c.f. space. Also a most important result is realized when the Nyquist restriction is satisfied. The quantization noise itself is

uniformly distributed and first order (uncorrelated even though the input signal is of a high order process.

A qualitative measure of how rough the quantization may be while these results still hold may be obtained from the quantization of a Gaussian signal. The quantization box size may be as big as two standard deviations and there will be an error in the mean square of the quantization noise of 9%. If the correlation coefficient of the quantizer input is 0.9, that of the noise is about 0.3. When the quantization becomes this rough, the noise is no longer uncorrelated, although its correlation coefficient is considerably less than that of the input signal.

The first question to be settled in any quantizer system analysis problem is, does the Nyquist restriction hold at the input to all the quantizers? Sufficient but not necessary conditions are able to be derived for linear (linear except for the quantizers) sampled-data systems, but not for non-linear systems. However, the whole question practically reduces to whether or not each quantizer input takes on a continuum of levels spanning 3 or more quantization boxes. If so, the effects of the quantizers can be evaluated very precisely; if not, the same analysis will serve as a good "first try".

A linear quantizer system has an output consisting of the sum of two parts. The first part is the same as the output of the linear equivalent (an identical system except that the quantizers are replaced by unit gains) when excited by the given input. The second part is described statistically as the output of the linear equivalent having no input signal with the quantizers replaced by sources of uniformly-distributed independent first-order noises. The net noise output and the linear equivalent are both independent of the signal and completely characterize the system behavior for the large class of inputs giving statistics at the quantizers which obey the multidimensional c.f.restrictions.

Non-linear systems behave in like manner with respect to quantization noise as long as they are "small-signal linear". In cases where quantization is moderately fine, the Nyquist restriction is satisfied and the small quantization noises propagate to the system output through time-variable linear systems. The quantizers are replaced by independent noisy sources. This is so in numerical analysis, because roundoff will always be relatively fine here.

The performance of a continuous system may be determined by examination of samples of its output rather than the output itself. This is convenient for statistical analysis. The entire system is replaced by a sampled-data equivalent to within any desired precision (by adjustment of sampling rate) and the continuous problem is thereby converted to a problem that is already solvable.

Quantization noise is of a very simple nature in most practical problems, and this permits analysis and synthesis to proceed along usual lines with little change. Knowledge of the origin and the propagation of quantization noise in systems is useful because it gives the quality of a result in terms of the equipment used in generating it.

Thesis supervisor: William K. Linvill

Title: Associate Professor of Electrical Engineering

ACKNOWLEDGEMENT

The author is greatly indebted to Prof. W.K. Linvill for his assistance in the supervision of this thesis, and for his counselling and philosophizing during the past four years. His approach to engineering problems has been singularly most helpful.

The writer would like to thank his readers, Professors E. A. Guillemin and J. B. Wiesner. Their apprieciation of the problem and interest in it was very gratifying.

Innumerable conversations with the author's colleagues, Messrs. R.F. Jenney, M.A. Epstein, V.J. Mizel, M.J. Beran, and R. S. Tankin provided the opportunity for testing several of the ideas presented in this thesis. The proof shown in Figure 19 resulted from a conversation with R. F. Jenney.

The author would like to thank Miss Genevieve Tetreault for the typing of the final manuscript. Miss Mary Matas and the staff of the print and drafting rooms of Division VI of Lincoln Laboratory were very cooperative in getting it produced.

TABLE OF CONTENTS

TABLE OF CONTENTS

FIGURES

CHAPTER 1

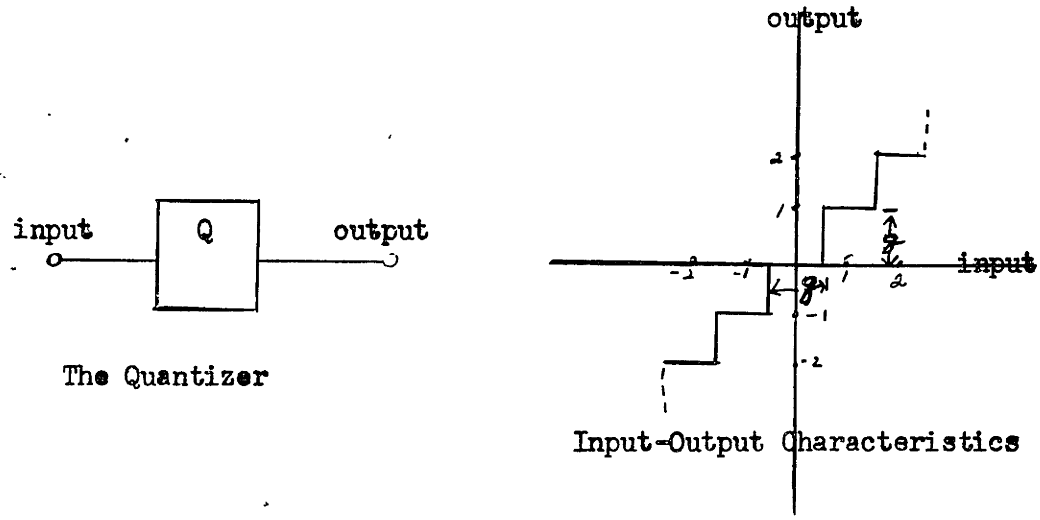
DESCRIPTION OF QUANTIZATION AND INTRODUCTION TO QUANTIZER SYSTEMS

Quantization or round-off occurs whenever physical quantities are represented numerically. The value of a measurement may be designated by an integer corresponding to the nearest number of units contained in the measured physical quantity. A round-off error thereby introduced must have value between plus and minus one half unit and can be made small by choice of the unit. It is apparent, however, that the smaller the size of the basic 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 made between accuracy and economy. This is particularly pronounced since the advent of the digital computer. In order to establish such a balance, it is necessary to have means of evaluating quantitatively the distortion resulting from rough quantization. The analytical difficulty arises from the inherent non-linearities of the quantization processes.

A. The Quantizer and Systems Containing Quantizers

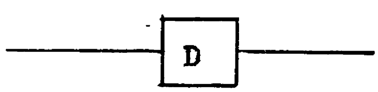
A rounding-off process may be represented symbolically as in Figure 1. For purposes of analysis, it has been found convenient to define the quantizer as a non-linear operator having the input-output relation shown in Figure 1. Its output is a single-valued function of the input, and it has an "average 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). More general quantizers could be obtained by preceeding and following the quantizer of Figure 1 with instantaneous linear amplifiers (multiplying factors) and adding dc levels to the quantizer input and output (tailoring averages).

Physical systems or processes involving quantization may now be represented as combinations of conventional system components with quantizers injected at points in the system where round-off naturally takes place. For the most part, the systems to be considered are



The Quantizer$_s$ Input-Output Characteristics

FIGURE 1



A General Linear
Sampled-Data Filter

gain $\equiv$ $D(Z) = \sum_{n=0}^{\infty} d_n Z^n$

where $Z \equiv e^{-sT}$

and T is period between samples

A General Linear Sampled-Data Filter

FIGURE 2

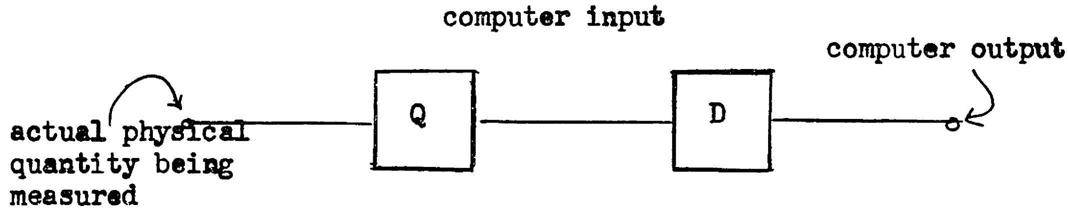
sampled-data types which are linear except for the quantizers. Other systems that will be considered belong to restricted classes of non-linear sampled-data quantizer systems and continuous quantizer systems.

B. Linear Sampled-Data Quantizer Systems

A general linear sampled-data filter is shown in Figure 2. Its present output sample is a linear combination of the present and past inputs. Often, a digital computer is programmed so that it may be represented as such.[1,5] A fine-grained digital computer programmed to be linear but having a coarse-grained input is shown in Figure 3. It may be noticed that the computer represented by D need not always be fine-grained but may be realized of discrete-state memory and arithmetic devices such as relays, diodes, and tubes because the filter is to be used to process inputs that take on only discrete states. Figure 3 could also represent a coarse-grained computer fed by a fine-grained input. The same kind of remarks apply to the system shown in Figure 4.

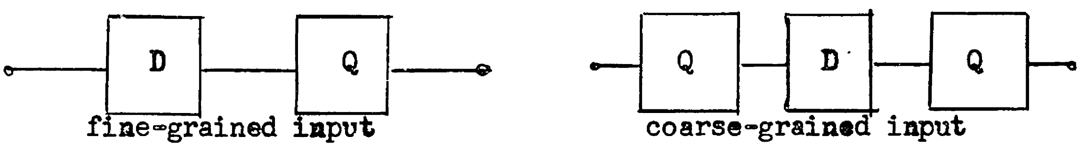
Systems of greater complexity are those in which quantization takes place within a feedback loop. Figure 5 shows systems with quantization in the feedforward and feedback sections. Other systems may be reduced to these forms. Figure 6 is an example of this. If a single quantizer is used in conjunction with many linear sampled-data filters and a feedback path through the quantizer exists, the system can be reduced to one of the forms of Figure 5 preceded and followed by linear sampled-data filters and quantizers not involved in feedback paths. The forms of Figure 5 are canonical and irreducible. They can, however, be expressed in terms of each other as shown in Figure 7. In order to be able to synthesize, analyze, and evaluate quantized sampled-data systems that are linearly programmed, one must deal with both open loop and closed loop quantization.

A question of the effect of the quantizer upon the stability of closed loop systems naturally arises. Although the average gain of the quantizer is unity, there are input values for which the incremental gain is infinite. Let a stable system be defined as one that gives a bounded output for a bounded input. This input could be applied to the actual system input, or at any other point in the system. Since the difference between the output and the input of the quantizer is never



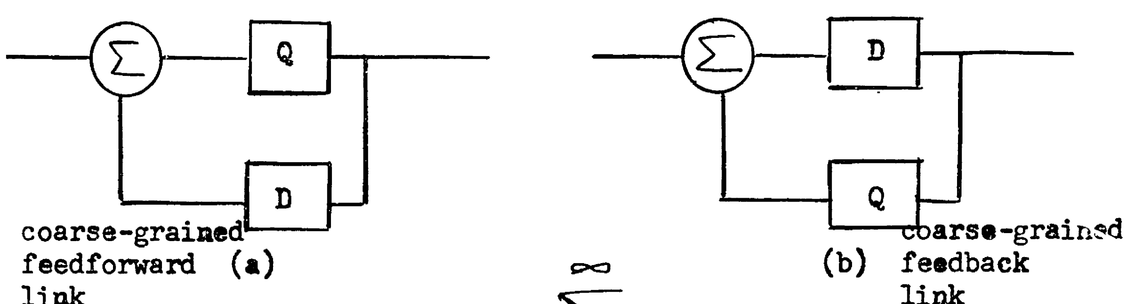
Linear Sampled-Data System with Coarse-Grained Input

FIGURE 3



Linear Sampled-Data System with Coarse-Grained Output

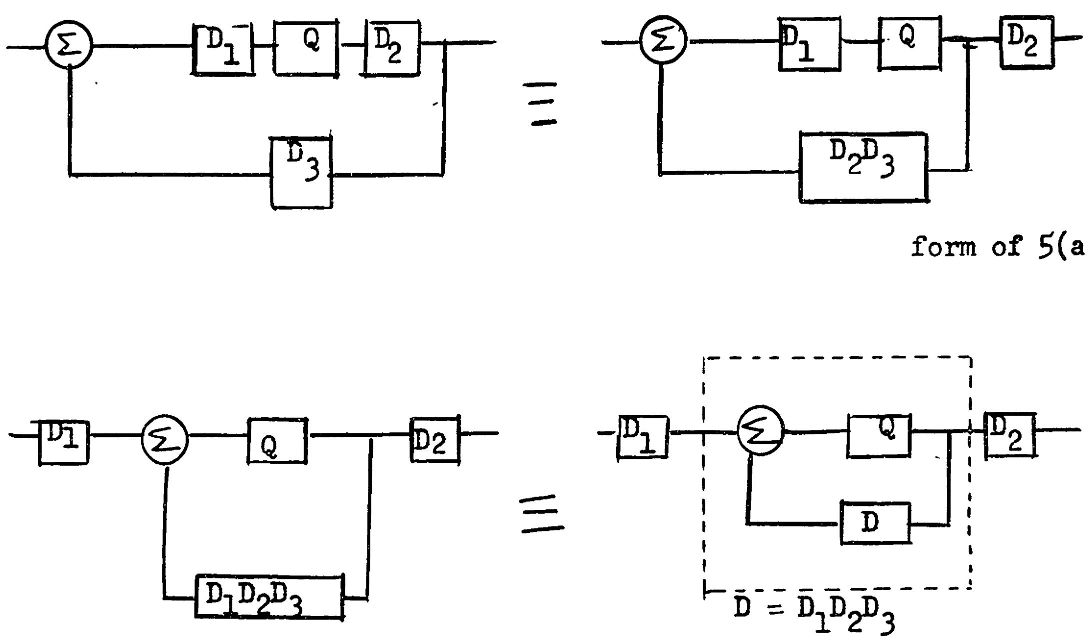
FIGURE 4



$$D = \sum_{0}^{\infty} d_n Z^n$$

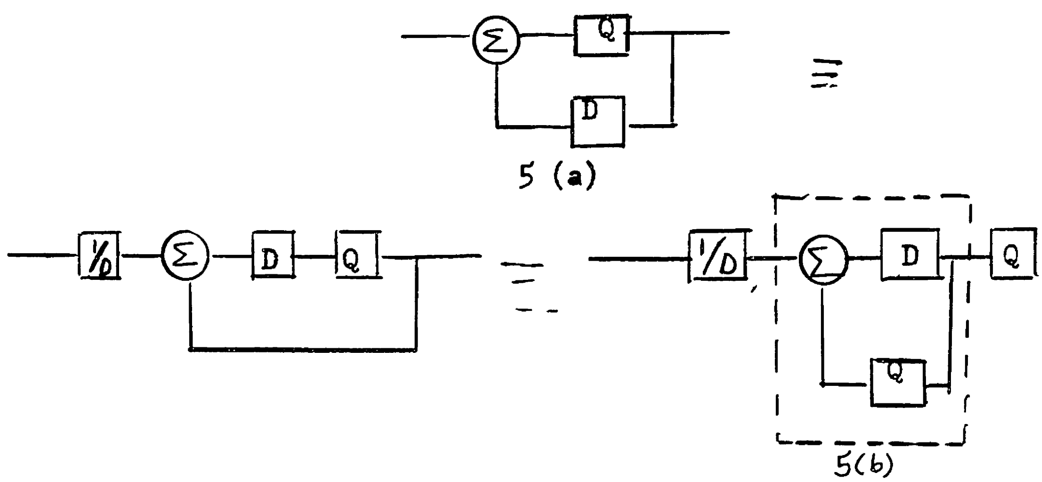
Quantizer Feedback Systems

FIGURE 5



Reduction of a System With Feedforward Quantization to the Basic Form of Figure 5(a)

FIGURE 6



Equivalence Between Systems With Feedforward Quantization and Feedback Quantization

FIGURE 7

greater than q/2, the quantization "noise" is bounded. It follows that if the linear equivalent (the system resulting from replacing the quantizer by a linear amplifier of gain 1) is stable, so is the quantizer system.

C. Non-Linear Sampled-Data Systems

Corresponding to every form of linear system shown, there are non-linear systems representing many kinds of situations including numerical solution of non-linear differential equations and non-linear control systems. The effect of quantization upon these systems is of interest and will be evaluated in certain useful respects.

D. Continuous Systems

Quantization occurs naturally in sampled-data systems and numerical processes. It appears also in continuous systems (consider contact servos, for example). The continuous system may be replaced by an equivalent sampled-data system if the "signals" at all points in the system are bandwidth limited. This condition is obviously not met where a continuous system is quantized; the quantizer output will be a series of nonuniformly spaced step-waves. Thus, sampled-data representation may be made only in cases where the quantizer output is first coupled to a linear continuous system of reasonably finite bandwidth.

The continuous problem is most often not very difficult from a sampled-data problem. Hence, analysis of sampled-data quantizer systems includes most of the possible types of quantizer systems. The analytical approach will be of a statistical nature, which turns out to be very clear-cut with sampled-data.

CHAPTER II

STATISTICAL DESCRIPTION OF THE QUANTIZATION PROCESS

How could one precisely evaluate the output signal of a quantizer in terms of its input? An output for a given input could be obtained by means of a point-by-point calculation in the "time-domain". Each input is a special case, however, and usually not much detail could be learned about new cases from a given time domain study. Superposition does not apply. Often, a quantizer input is random. It seems very unlikely that a point-by-point calculation scheme would be of any use here.

The type of a solution presented does not attempt to get precise descriptions of quantizer outputs in terms of their inputs. Rather than that, the solution proposed is a statistical one. This is exactly what is desired when the system input is given statistically, and will provide a sufficiently adequate picture for many cases of systems inputs that are causal.* The output of the quantizer differs from the input by some causal function of the input, which on the average, may be large or small depending chiefly upon the size q of the quantization box. An exact description of this difference is difficult to obtain. On the other hand, statistical methods will be presented that will allow relatively simple calculation of the probability density distribution of this difference, the quantization noise. The beauty of this method arises from the fact that a vast 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 then the sum of the input and noise of known statistics. In general, if another input is added to the first, the resulting output is the sum of two inputs plus a quantization noise which is a different waveform in time but has the same statistics as in the previous case. Thus, in a statistical sense, the quantizer is a linear device. Superposition applies.

It seems reasonable to expect that this statistical approach

---

* Causal means deterministic as opposed to stochastic. Yet a causal signal, a sine wave for example, has a probability density distribution

will simplify many analysis and synthesis problems because it gives average effects for large numbers of different inputs. The averages may be more useful than the effects themselves, because the effects, being so many and so varied, are often not accountable.

Instead of devoting attention to the variable (sampled "signal") being quantized, let only its probability density distribution be considered. It will be seen that the distribution of the quantized variable may be obtained by a linear sampling process upon the distribution of the unquantized variable.

A. First Order Statistics

If the samples of some continuous quantity are all independent of each other, a first order probability density W(X) is its characteristic function, the Fourier transform (2.1).

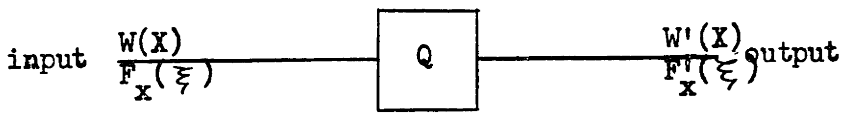
$$F(\xi) = \int_{-\infty}^{\infty} W(X)\, e^{jx\xi}\, dX \tag{2.1}$$

A quantizer input variable may take on a continuum of magnitudes (see Figure 8), while the output assumes only discrete states. The probability density of the output W'(X) consists of a series of impulses that are uniformly spaced along the amplitude axis, each one centered in a quantization box.

Figure 9 shows how the output distribution is derived from that of the input. Since any event occuring within a quantization box is always "reported" as being at the center of that box, each impulse has a magnitude equal to the area under the probability density W(X) within the bounds of the box. The impulse distribution W'(X) has a periodic characteristic function, being the Fourier transform of a series of impulses having uniform spacing q. The mathematical techniques developed by W. K. Linvill[1] for the study of linear sampled-data systems will be used in the derivation of the characteristic function of W'(X). The necessary aspects of this theory will be developed below.

1. Amplitude sampling treated as linear impulse modulation.

The process of periodically sampling some f(t) is the same as that of multiplying f(t) by a series of impulses of unit area which are spaced uniformly in t. Figure 10 shows as f(t) being sampled by an "impulse modulator". The impulse carrier of fundamental

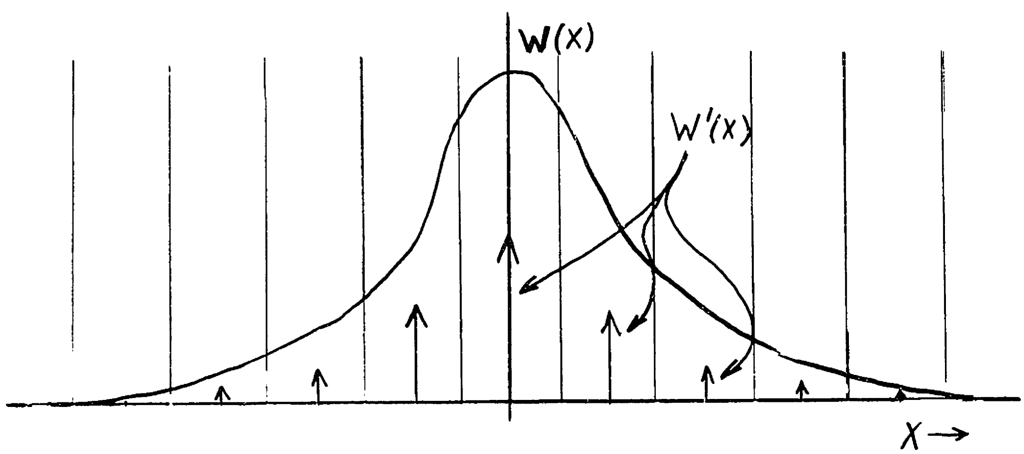


input : $F_x(\xi) = \int W(X)\, e^{jx} dX$

output : $F_x'(\xi) = \int W'(X)\, e^{jx} dX$

The Quantization Of A Random Sampled Process

FIGURE 8



Area Sampling Used In The Derivation Of W'(X)

FIGURE 9

frequency $\Omega = 2\pi/T$ is shown represented by its Fourier series (2.2).

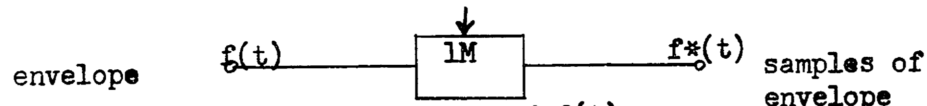
$$f^*(t) = \left[f(t)\right] \left[\text{impulse carrier}\right] = \left[f(t)\right] 1/T \sum_{-\infty}^{\infty} e^{jn\Omega t} \quad (2.2)$$

Each harmonic has the same amplitude 1/T, and is modulated by f(t). The envelope of f*(t) is f(t). The impulse carrier is thus the sum of an infinite number of sinusoidal carriers with uniform frequency spacing Ω which, when modulated by f(t), develop identical "sidebands" about each carrier. The pattern of these sidebands is the same as that of the Fourier transform of spectrum of f(t). The spectrum of f(t) and f*(t) are shown in Figure 11. F*(jω), the Fourier spectrum of the series of impulses f*(t), is the sum of a periodic array of sections, separated by the frequency Ω, where the typical repeated section is the same as F(jω), the spectrum of the envelope of the pulses. If it were possible to separate the zeroth section of F*(jω) from the rest, it would be possible to recover an envelope from its samples. This can be done with an "ideal low-pass filter" if the sections are distinct and do not overlap. The gain as a function of such a filter is shown in Figure 12. If F*(t) is applied to the input of the low pass filter of Figure 12, the output will be f(t). Since the impulse response of the ideal low pass filter is $\sin(\pi t/T)/(\pi t/T)$, it follows by linearity that the envelope of the impulses is a sum of these, properly weighed and spaced in time, as shown in Figure 13.

The low pass filter is an interpolater that yields f(t) as long as f(t) has no significant harmonic content at higher frequency than $\Omega/2$. This is the "Nyquist bandwidth restriction" of f(t). When the Nyquist restriction is not satisfied, the sampling frequency Ω is not sufficiently large and there is overlap among the periodic sections of F*(jω)(see Figures 11 and 12). The ideal low pass filter cannot separate out the spectrum of f(t) alone. These ideas will now be used in the de-ivation of the impulse distribution of a quantizer output.

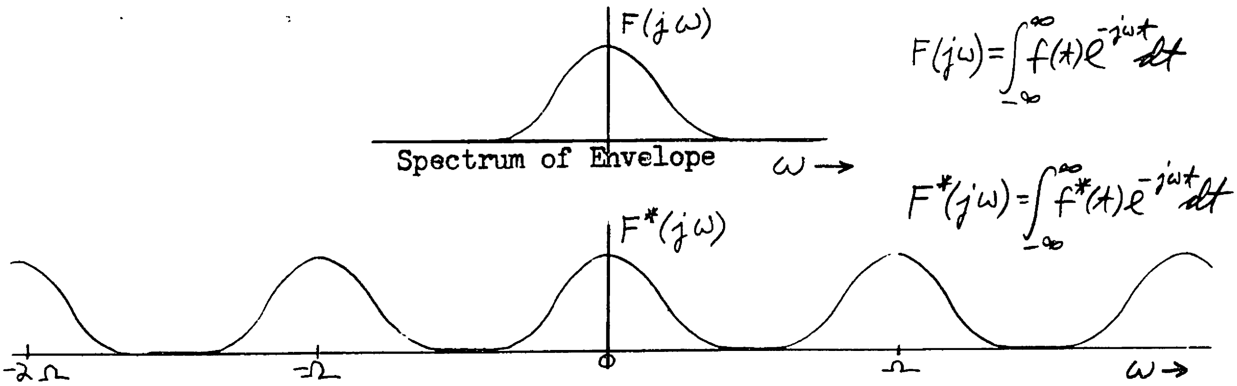
2. Derivation of the first order probability density of a quantized variable.

The distribution of a quantizer output W'(x) consists of "area samples" of the input distribution density W(x). The quantizer may be thought of as an area sampler acting upon the "signal", the probability density W(x). Figure 14 shows how W'(x) may be constructed by sampling the difference D(x + q/2) - D(x - q/2), where D(x) is the distribution,



Sampling of f(t)

FIGURE 10

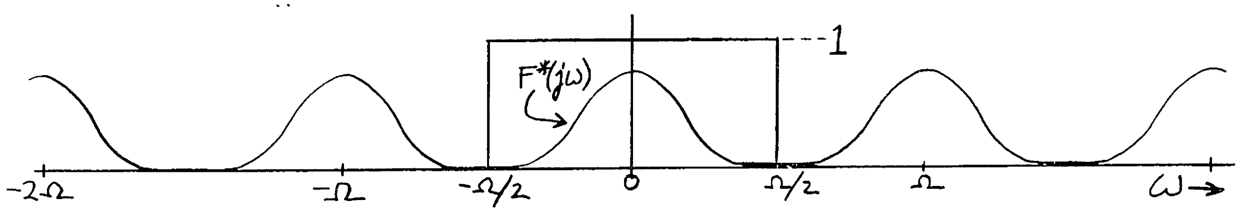


$$F(j\omega)=\int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt$$

$$F^*(j\omega)=\int_{-\infty}^{\infty} f^*(t)e^{-j\omega t}dt$$

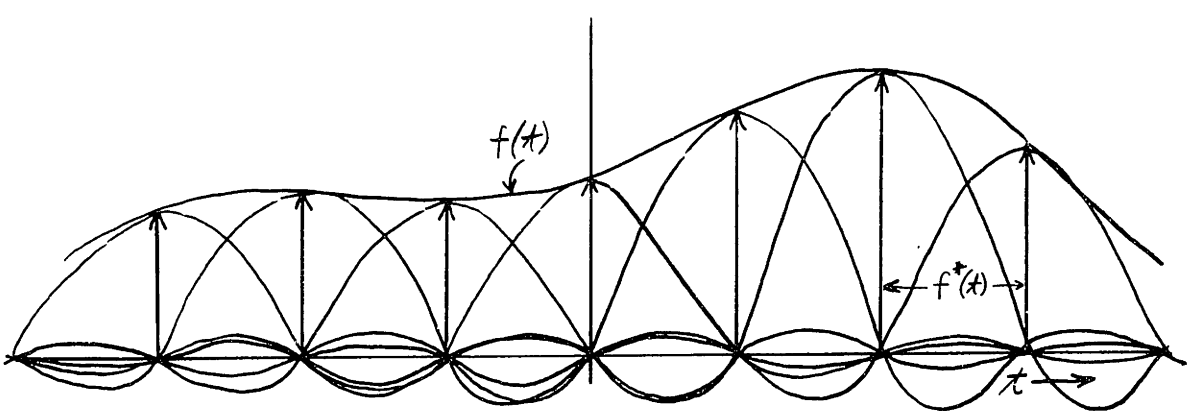
Spectrum of Samples of Envelopes

FIGURE 11



Recovery of Envelope From Samples In The Frequency Domain

FIGURE 12



Recovery of Envelope From Samples In The Time Domain

FIGURE 13

the integral of the distribution density. Figure 15 is a schematic diagram of this process, showing how W(X) is first modified by a linear filter of "gain" $\sin(q\xi/2)/(q\xi/2)$ and then sampled to give W'(X).

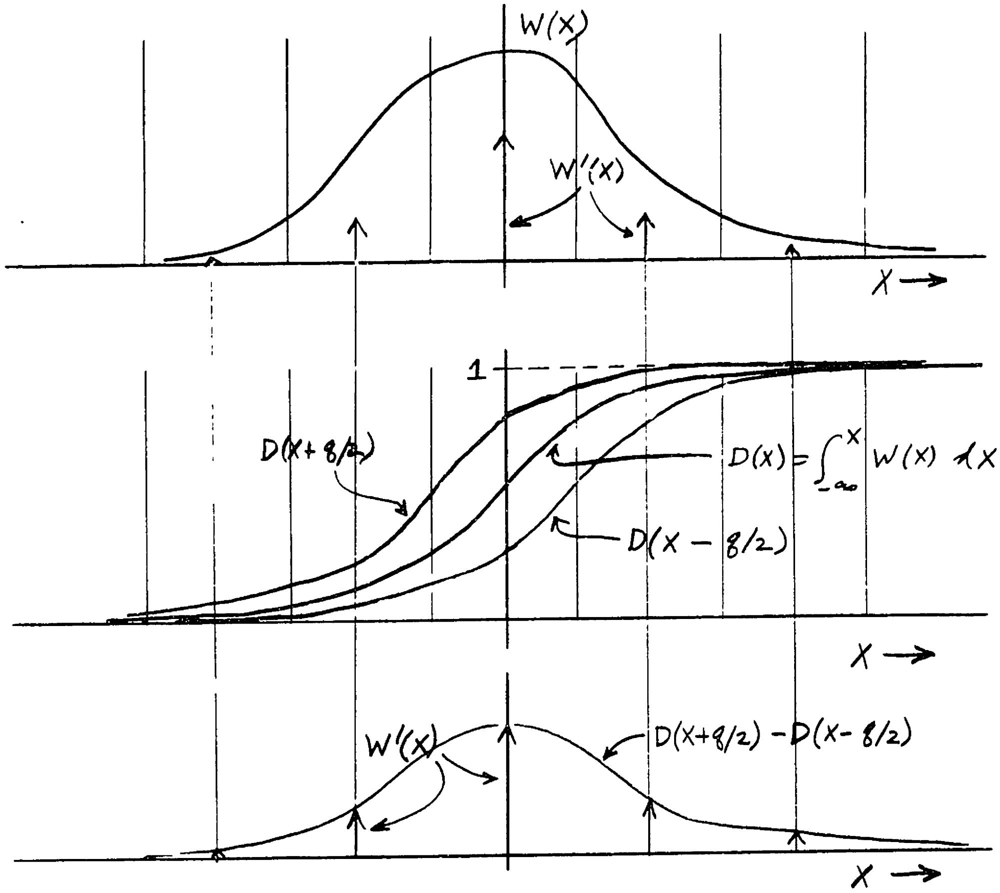
The difference between area sampling and the Linvill **amplitude** sampling is most clear in the "frequency" domain, where both give periodic transforms. The typical repeated section is the same as the transform of the envelope in the case of amplitude sampling. The transform of the envelope is first multiplied by $\sin(q\xi/q)/(q\xi/2)$.

When the quantization "fineness", the reciprocal of the box width, is twice as high as the "highest frequency" component contained in the shape of W(X), it is possible to recover W(X) from the quantized distribution W'(X) by inverse transforming the ratio of a typical section of $F_x'(\xi)$ and dividing by $\sin(q\xi/2)/(q\xi/2)$.

The characteristic function of the distribution density of the sum of two random independent variables is the product of the individual c.f.'s. Figure 16 shows the distribution Q(N) and its c.f. Q(N) will be shown to be the distribution of quantization noise. Its c.f. is $\sin(\pi\xi/\phi)/(\pi\xi/\phi) - \sin(q\xi/2)(q\xi/2)$. If purely random independent noise of distribution Q(N) were added to a signal of distribution W(X), their sum would have a c.f. $F_x(\xi)\sin(\pi\xi/\phi)/(\pi\xi/\phi)$ which is identical with the typical section of $F_x'(\xi)$. The derivatives of a c.f. at the origin determine moments. It follows that the moments of a quantized signal are the same as if the quantizer were a source of independent random additive noise of distribution Q(N) provided that Nyquist's restriction on W(X) is met. This correlates with work done by Sheppard as summarized in Appendix I.

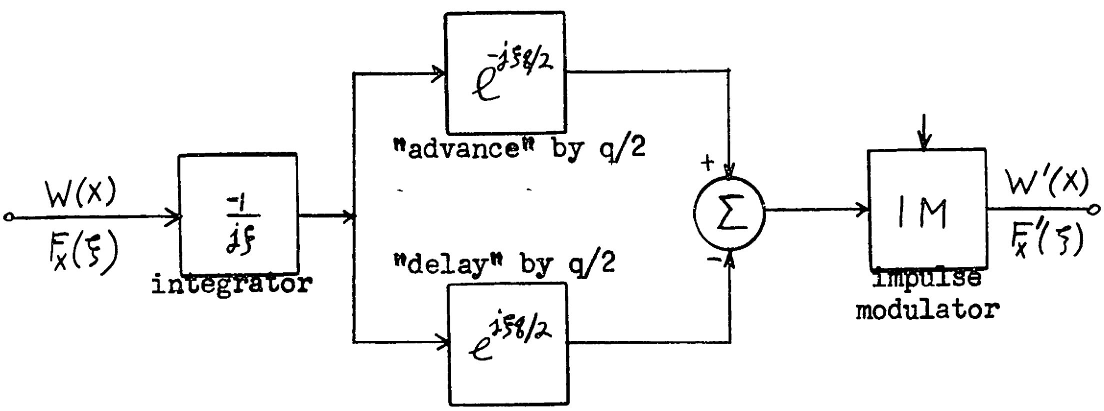
It is now known how W'(X) may be derived from W(X). The understanding of the quantizer would be complete if it were true that the quantization noise (the difference between input and output) were independent of the quantizer input. Not only is the quantization noise statistically related to the input, however, but it is also causally related. Since the output of a quantizer is a single valued function of the input, a given input yields a definite noise.

The distribution of the noise itself will be shown to be Q(N), independent of the distribution of the quantizer input (as long as the Nyquist restriction is satisfied). Their causal tie will show up later



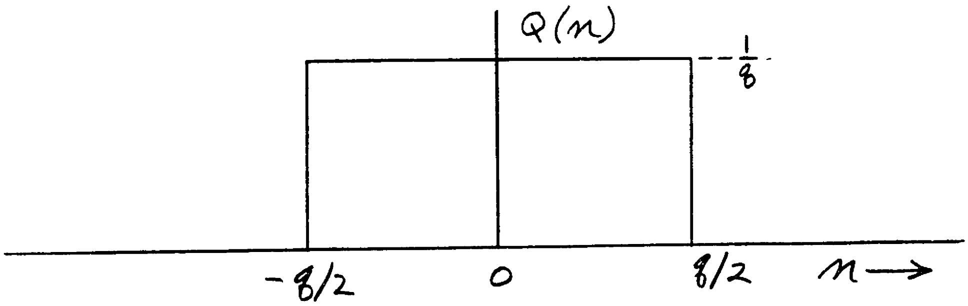
Construction Of Area Samples

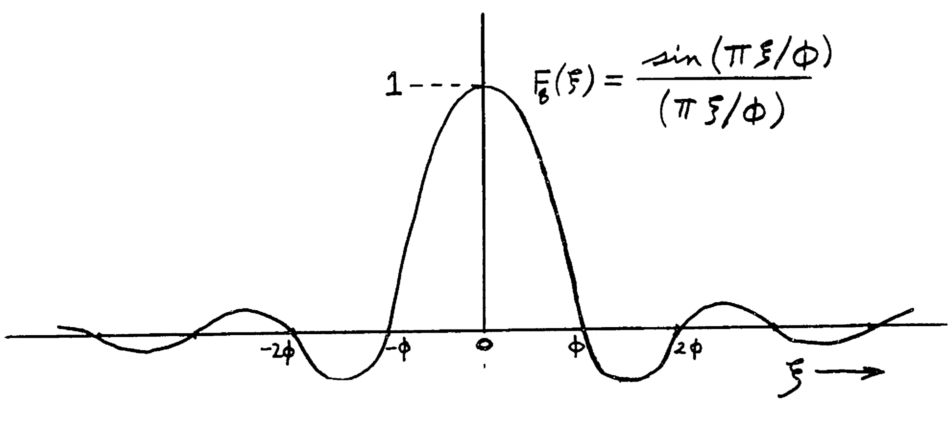
FIGURE 14



Block Diagram Of The Area Sampling Of Figure 14

FIGURE 15





The Distribution Of Quantization Noise
And Its Characteristic Function

FIGURE 16

when joint input-output distributions are derived.

3. Derivation of the probability density of quantization noise.

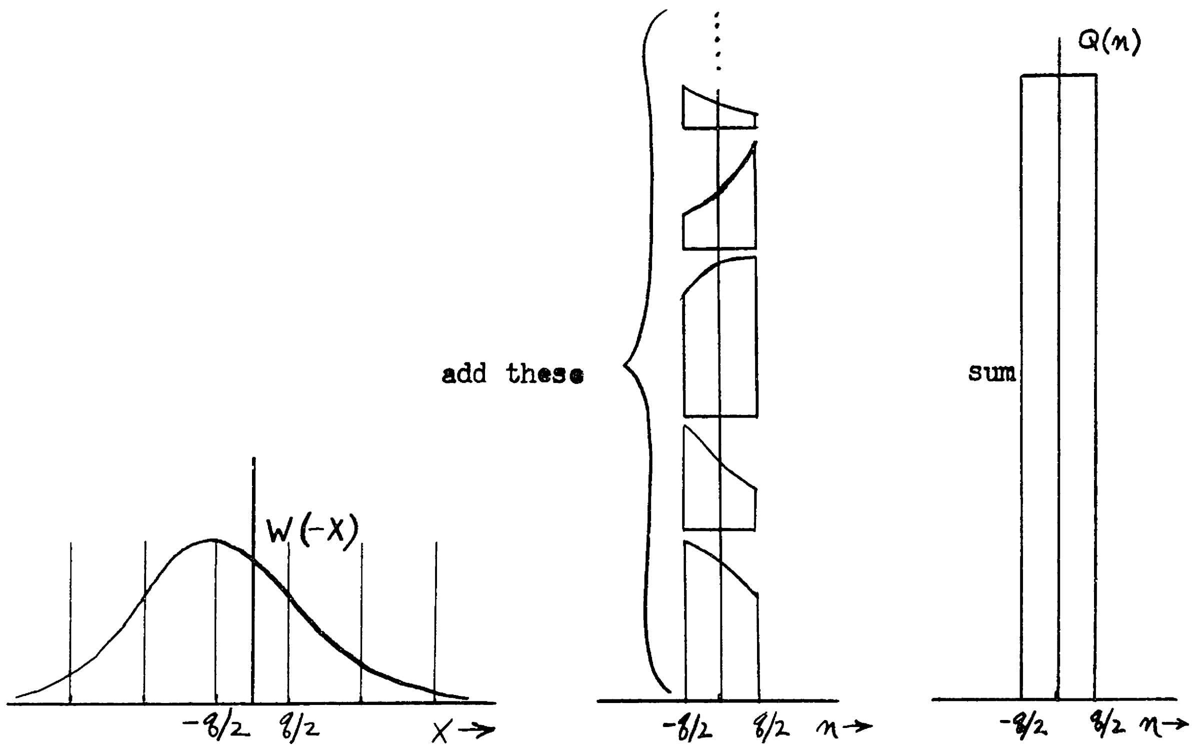
Quantization noise is always the difference between an input variable and the value of the box to which it has been assigned. The distribution of quantization noise resulting from events assigned to the zeroth box may be constructed by plotting W(-X) between $-q/2 < X < q/2$. The noise distribution resulting from events in the first box may be obtained by considering W(-X) for values $-3q/2 < X < -q/2$ 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 the sum of the probabilities of that noise from each box. Figure 17 shows how the distribution of quantization noise is constructed from W(-X).

If the quantization is sufficiently fine-grained, W(X) may be represented as the sum (2.3). In other words, W(X) may be represented by the "sin X/X" envelope of its samples if the Nyquist restriction is met. (See Figure 13.)

$$W(X) = \sum_{-\infty}^{\infty} W(X_n) \quad \frac{\sin\left[\frac{\pi(X - X_n)}{q}\right]}{\left[\frac{\pi(X - X_n)}{q}\right]} \qquad (2.3)$$

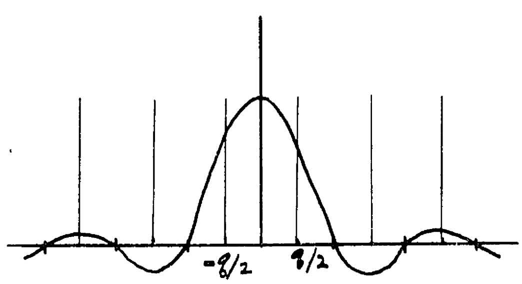
Since the development of the distribution of quantization noise as in Figure 17 is a linear process, the quantization noise distribution is the sum of the distributions of noise corresponding to constituents that are added to get W(X). All that needs be considered is the quantization of the basic form $\sin(\pi X/q)/(\pi X/q)$, as in Figure 18. The strips of Figure 18 are added in Figure 19 to give the quantization noise distribution which turns out to be flat-topped. Figure 19 shows how the constant unity may be composed of a sum of sin X/X's and how their sum over the range $-q/2 < X < q/2$ is the same as adding the strips of Figure 18. That the constant unity may be so represented is assured by the fact that its bandwidth being zero, is surely less than (1/2) ϕ.

An arbitrary distribution satisfying the Nyquist condition is the sum of a series of $\sin(\pi X/q)/(\pi X/q)$'s, where each gives a flat-topped distribution of quantization noise. The sum of flat-topped distributions is flat-topped. If the distribution density of a signal being quantized is W(X) and the quantization grain is fine enough to



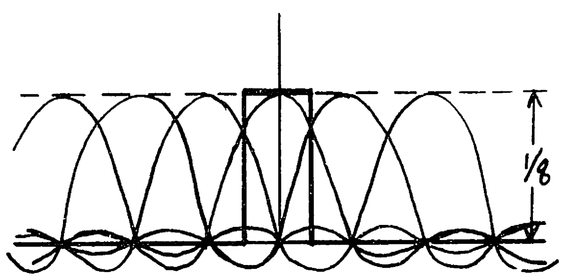
Construction Of Distribution Of Quantization Noise

FIGURE 17



"Sin X/X" Distribution

FIGURE 18



Quantization Noise Distribution For A Signal Having A "sin X/X"

FIGURE 19

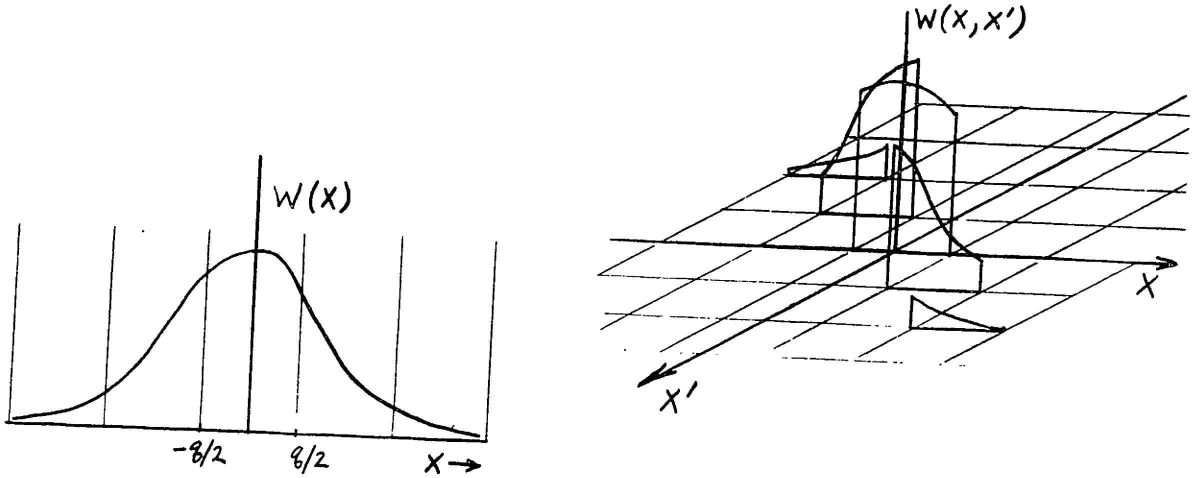
satisfy the Nyquist restriction, the distribution of the noise introduced by the quantizer will be flat-topped. This distribution is Q(N), shown in Figure 16.

4. Derivation of the joint probability density of the quantizer input and output.

A most general statistical description of a device having a random stationary output is the joint distribution between input and output. From this, the output distribution, input distribution, the difference (between input and output) distribution, the joint distribution between input and difference, and the joint distribution between output and difference may be determined. Any one of the joint distributions will determine all the rest, but at least one joint distribution need be known for a complete statistical understanding. Infinite numbers of joint distributions could give the same input, output, and difference distribution densities, so the latter are not sufficient for a complete understanding.

A study of Figure 20 shows how a joint in-out distribution W(X, X') is derived from a given input distribution W(X). The strips of W(X) are placed at the values of X' to which they correspond. Consider next the situation shown in Figure 21. For every value of X, all values of noise are possible between ± q/2 because the noise is independent of X. The joint distribution (Figure 22) between X and (X + n) shows this, whereas any plane parallel to the (X + n) and w axes cuts a flat-topped section from the surface of joint probability w(X, X + noise), the surface of joint probability is everywhere parallel to the X + n axis. The projection of the surface on the X - w plane has the same shape as W(X) and has an area of 1/q. The volume under the surface is unity.

A study of Figures 20 and 22 shows that the strips of Figure 20 are sections of the 3-dimensional surface of Figure 22 (if first multiplied in amplitude by q) cut by a series of planes parallel to the W and X axes with spacing q along the X + n axis. The strips of Figure 20 are thus the results of the amplitude modulation of a periodic carrier, a series of uniformly spaced impulse sheets, and an envelope which is the joint probability surface. It should be possible to deduce the joint characteristic function of the distribution W(X, X') from that of w(X, X + n). At the same time, the ways in which quantization is akin



Formation of Joint In-Out Distribution from Quantizer Input Distribution

FIGURE 20

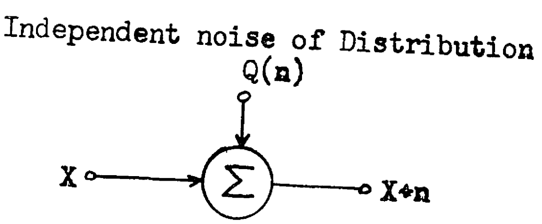
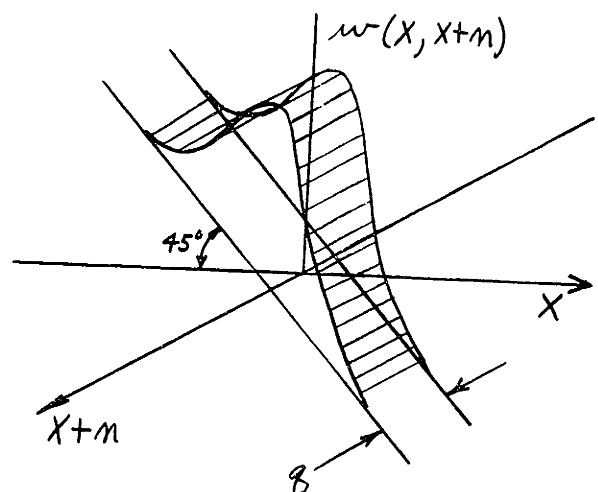


FIGURE 21



Joint In-out Distribution For A Quantizer And For A Source Of Additive Independent Quantization Noise

FIGURE 22

to the addition of random independent noise as in Figure 21 should be detected.

The methods of amplitude sampling may be readily generalized to handle sampling by impulse sheets. The first step is to get the two-dimensional Fourier Series of the impulse sheets (Figure 23). Each sheet extends to infinity in both directions, and has a unit volume per unit length. The Fourier series for Z(X, X') is (2.4).

$$Z(X, X') = \sum_{n=-\infty}^{\infty} 1/q \, e^{-jn\phi X'} \qquad (2\text{-}4)$$

Z appears to be one-dimensional, because there is no variation with X. If ξ_a is the variable that X transforms into, and ξ_b is the variable that X' transforms into, the two-dimensional spectrum of Z(X, X') is as indicated in Figure 24. If a carrier having this spectrum is modulated by an envelope w(X, X+n), the resulting spectrum (Figure 25) is periodic along ξ_b, and aperiodic along ξ_a. The shape of a typical section is the same as the spectrum of w(X, X+n), resulting from its convolution with the spectrum of Figure 24. Since the sections of w(X, X+n) are to be first multiplied by q, the factor 1/q is compensated for and the value of $F_{X,X'}(\xi_a, \xi_b)$ is 1 at the origin. All characteristic functions must have the value 1 at their origins in order that the total volume under their probability densitied be unity.

It is of interest to derive the typical section of $F_{X,X'}(\xi_a, \xi_b)$, the Fourier transform of w(X, X+n) which is the joint distribution between the input and output of Figure 21. A joint c.f. of two variables may be deduced from the characteristic functions resulting from sums of various proportions of the two variables. A block diagram of this technique, used many times in Appendix II, is given in Figure 26.

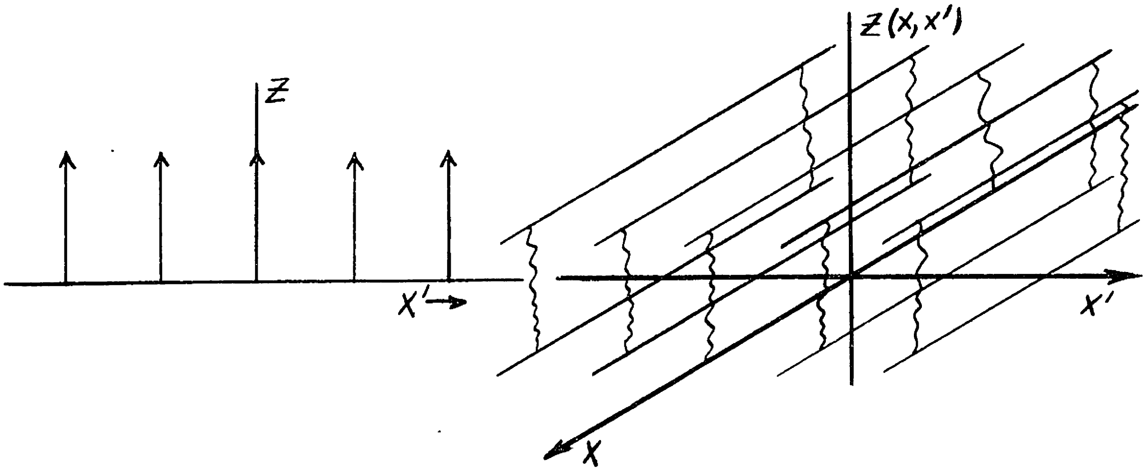
Formally,

$$F_{X,X+n}(\xi_a, \xi_b) = \iint_{-\infty}^{\infty} w(X, X+n) \, e^{j[X\xi_a + (X+n)\xi_b]} dX d(X+n) \qquad (2.5)$$

also,

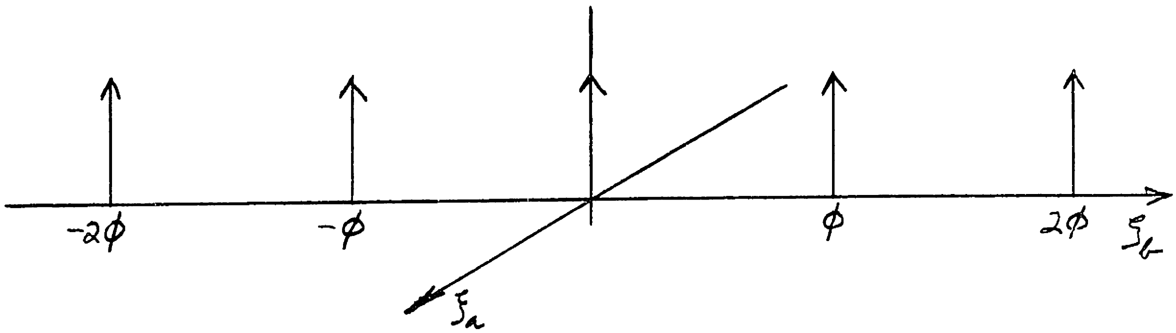
$$F_{\Sigma}(\xi) = \iint_{-\infty}^{\infty} w(X, X+n) \, e^{j[k_1\xi + k_2\xi(X+n)]} dX d(X+n) \qquad (2.6)$$

$F_{\Sigma}(\xi)$ can be readily evaluated and leads to $F_{X,X+n}(\xi_a, \xi_b)$ if the substitution is made, $k_1\xi \equiv \xi_a$, $k_2\xi \equiv \xi_b$. Any ξ_a, ξ_b can be obtained by choice of k_1, k_2, and ξ. The sum Σ equals $(k_1 + k_2)X + k_2 n$.



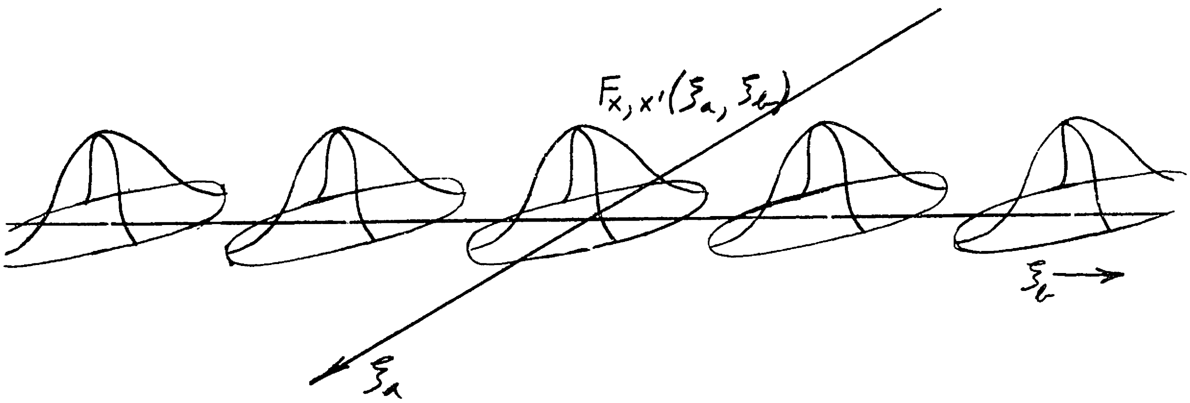
Periodic Impulse Sheet Carrier

FIGURE 23



Spectrum of Impulse Sheet Carrier

FIGURE 24



Joint In-Out Characteristic Function For A Quantizer

FIGURE 25

The c.f. of X is $F_x(\xi)$, and the c.f. of the independent noise is $F_q(\xi)$.

$$\therefore \quad F_{\Sigma}(\xi) = F_x\left[(k_1 + k_2)\xi\right] F_q(k_2\xi)$$

Whence:

$$F_{x,x+n}(\xi_a, \xi_b) = F_x(\xi_a + \xi_b) \qquad F_q(\xi_b)$$

$$= F_x(\xi_a + \xi_b) \frac{\sin(\pi\xi_b/\phi)}{(\pi\xi_b/\phi)} \tag{2.7}$$

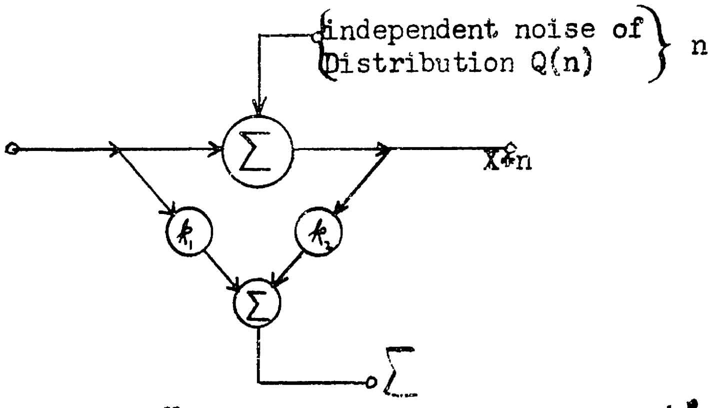
Let $F_x(\xi)$ be as sketched in Figure 27. It is finite over a width α along ξ_a. A top view of $F_{x,\,x+n}(\xi_a, \xi_b)$ that results is shown in Figure 28. Only the contours where the joint c.f. becomes insignificant are shown. When q is very small, n is small and $F_q(\xi) = 1$. $F_{x,x+n}(\xi_a, \xi_b) = F_x(\xi_a + \xi_b)$. The joint c.f. is finite only where $-\alpha/2 < (\xi_a + \xi_b) < \alpha/2$. This region is bounded by the two straight 45° lines in Figure 28. As the noise n is made larger, the finite regin can only become smaller because $F_q(\xi)$ becomes more significant. The curved contour is typical for such a case.

$F_{x,x+n}(\xi_a, \xi_b)$, the joint c.f. of the input and output of a quantizer is expressible in terms of the c.f. of the quantizer input, $F_x(\xi)$.

$$F_{x,x'}(\xi_a, \xi_b) = \sum_{n=-\infty}^{\infty} F_x(\xi_a + \xi_b + n\phi) \frac{\sin(\pi(\xi_b + n\phi)/\phi)}{(\pi(\xi_b + n\phi)/\phi)}$$

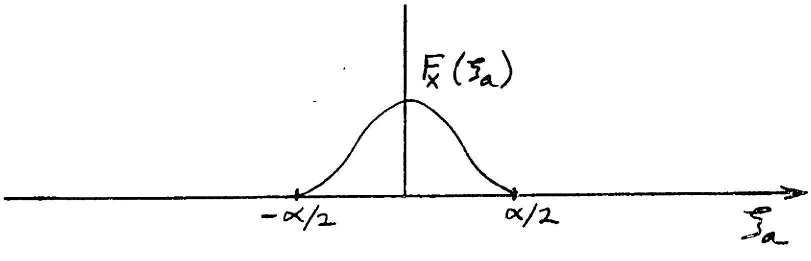
$$= \sum_{n=-\infty}^{\infty} F_x(\xi_a + \xi_b + n\phi) \frac{\sin(\pi\xi_b/\phi + n\pi)}{(\pi\xi_b/\phi + n\pi)}$$

This is a complete statistical description of the quantizer for first order (uncorrelated) statistics. A sketch of a joint c.f. where ϕ is somewhat greater than α (Figure 29) shows that this condition is sufficient to insure no overlap. The joint and self moments depend only upon the slopes (partial derivatives) of $F_{x,x'}(\xi_a, \xi_b)$ at the origin, which are unaffected by the periodicity of that function as long as there is no overlap. It may be concluded that with respect to all detachable moments, quantization is the same as addition of random independent noise of distribution Q(n), as long as the Nyquist restriction on W(X) is satisfied.



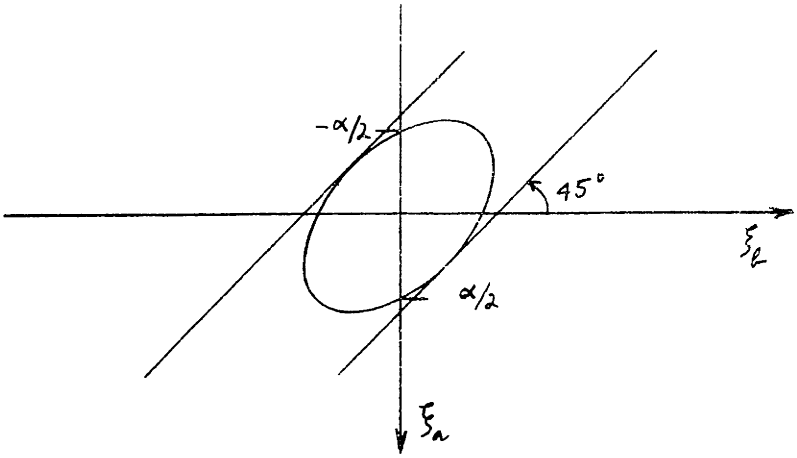
Flow Diagram Useful In Calculation of $F_{x,x+n}(\xi_a, \xi_b)$

FIGURE 26



C. F. Of First Order Quantizer Input Signal

FIGURE 27



Top View Of $F_{x,x+n}(\xi_a, \xi_b)$

FIGURE 28

The impulse distribution of the quantizer output, and the distribution of the quantizer noise Q(n) may be rederived readily from $F_{x,x'}(\xi_a, \xi_b)$. A plane perpendicular to the ξ_a, ξ_b plane through the ξ_b axis intersects the joint input-output c.f. $F_{x,x'}(\xi_a, \xi_b)$, giving a section which is $F_{x'}(\xi_b)$, the c.f. of the output (see Figure 30). As determined previously, the c.f. $F_{x'}(\xi_b)$ is periodic of frequency ϕ where each section is identical with the c.f. of the sum of the quantizer input and independent quantization noise. Likewise, a plane perpendicular to the ξ_a, ξ_b plane through the ξ_a axis gives an intersection which is $F_x(\xi_a)$, the c.f. of the quantizer input. Since this is a given property of the quantizer input and cannot be affected by the nature of the quantization, it appears from Figure 30 that some difficulties are to be encountered when taking the cut because of the periodicities of the joint input-output c.f. The issue is saved because of the factor $\sin(\pi\xi_b/\phi + n\pi)/(\pi\xi_b/\phi + n\pi)$ in each repeated section. This factor causes all sections except the one centered along the origin to be identically zero along the ξ_a axis (where $\xi_b = 0$) so that the periodicity of $F_{x,x'}(\xi_a, \xi_b)$ can have no effect upon $F_x(\xi_a)$. The c.f. of the difference between the quantizer output and input may be expressed as (2.8).

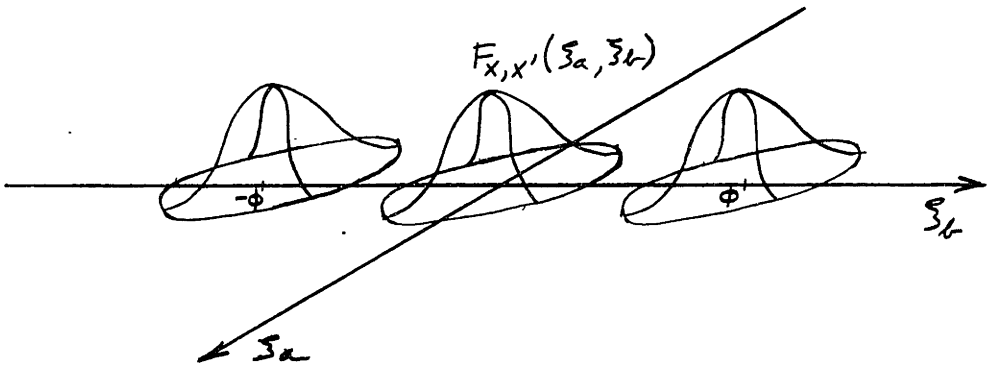
$$F_{(x'-x)}(\xi) = \iint_{-\infty}^{\infty} w(X,X') \, e^{j\xi(x'-x)} \, dXdX' \tag{2.8}$$

The joint input-output c.f. (2.9) gives (2.8) when the substitution (2.10) is made.

$$F_{x,x'}(\xi_a, \xi_b) = \iint_{-\infty}^{\infty} w(x,x') \, e^{j(x\xi_a + x'\xi_b)} \, dXdX' \tag{2.9}$$

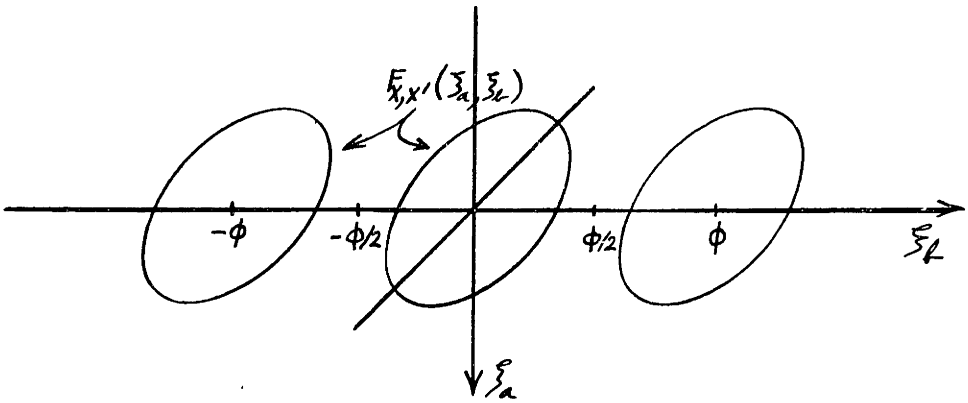
$$\begin{aligned} \xi_b &= \xi \\ \xi_a &= \xi \end{aligned} \tag{2.10}$$

$$\therefore F_{x,x'}(\xi_a, \xi_b) \Bigg|_{\substack{\xi_b = \xi \\ \xi_a = \xi}} = F_{(x'-x)}(\xi) \tag{2.11}$$



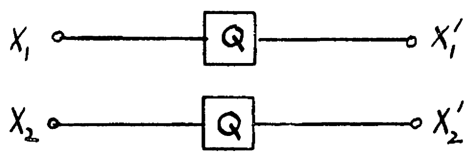
Joint Input-Output C.F. Of A Quantizer With A First Order Random Input

FIGURE 29



Determination Of Input, Output, and Difference Characteristic Functions From the Joint Input-Output Characteristic Function

FIGURE 30



Representation Of The Quantization Of A Second Order Signal As The Separate Quantization Of Two Jointly Related First Order Signals

FIGURE 31

Therefore, a section of $F_{x,x'}(\xi_a, \xi_b)$ through the F axis and a 45^o line in the $\xi_a - \xi_b$ plane as shown in Figure 30, when projected either upon the $F - \xi_a$ plane or upon the $F - \xi_b$ plane gives the c.f. of the distribution of quantization noise. That the cut is at 45^o insures that the periodicity of the joint c.f. can have no effect upon the distribution of quantizer noise. This distribution is therefore the same as the distribution of added noise in Figures 21 and 26, being Q(n), as shown previously.

The description of the quantizer response to first order statistics is complete and useful in itself. However, in order to understand the behavior of the quantizer in systems, particularly in feedback systems, it is necessary to consider how the quantizer reacts to correlated (high order) input samples. The methods already developed will be extended to handle multidimensional input distributions. It has been shown that, in many respects, quantization is the same as addition of a random independent noise. Conditions will be shown under which quantization of correlated samples will be very much like addition of random independent uncorrelated noise of distribution Q(n).

B. <u>Higher Order Statistical Quantizer Inputs</u>

If the random quantizer input variable X is second order, (the simplest Markov process), a joint distribution density $W(X_1,X_2)$ is required to completely describe its statistics. X_1 and X_2 are an adjacent sample pair. The distribution of the output is $W'(X_1,X_2)$. The joint distribution between output and input is $W(X_1, X_1', X_2, X_2')$ having c.f. $F_{x_1 x'_1 x_2 x_2'}(\xi_{1a} \xi_{2a} \xi_{1b} \xi_{2b})$. In order to sketch the joint distribution five dimensions are needed. Some other way to illustrate its significant features will be sought.

$W(X_1,X_2)$ may be resolved into a two-dimensional sin X/X series provided a two-dimensional Nyquist restriction is satisfied. The c.f. of $W(X_1X_2)$ is the sum of the separate components because of the linearity of the Fourier transform. The quantization process is a linear operation upon probability distributions and characteristic functions. As a matter of fact, any situation in which a stationary random signal is operated upon to produce another stationary randon signal, even though the operation upon the signals may be non-linear and have memory, has the characteristic that a linear operation is performed upon the input distribution to give the output distribution. The output and joint distribution of a quantizer are the sums of the corresponding distributions that could result from each component of the input distribution acting separately. It is necessary here to consider non-physical distributions that not only have areas and volumes different from unity, but also have regions of negative density.

The Fourier transform of $W(X_1, X_2)$ is (2.12).

$$F_{x_1 x_2}(\xi_1, \xi_2) = \iint_{-\infty}^{\infty} W(X_1, X_2)\, e^{j(x_1 \xi_1 + x_2 \xi_2)}\, dX_1 dX_2 \tag{2.12}$$

If this c.f. is negligible outside the range $-\phi/2 < \xi_1, \xi_2 < \phi/2$ where $\phi = 2\pi/q$, the Nyquist restriction is satisfied. $W(X_1, X_2)$ may be thought of as a sum (2.13) where each coefficient $A_{k\ell}$ is the value of

$$W(X_1, X_2) = \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} A_{k\ell} \frac{\sin \pi (X_1/q+k)}{\pi (X_1/q+k)} \quad \frac{\sin \pi (X_2/q+\ell)}{\pi (X_2/q+\ell)} \tag{2.13}$$

$W(X_1, X_2)$ at $X_1 = kq$ and $X_2 = \ell q$. All that needs be considered to be perfectly general is how the quantizer acts upon an input distribution such as the k, ℓ term of the above sum.

Start with a special case, the 0, 0 term, a two-dimensional sin X/X centered at the origin of Figure 29(a). Such a distribution is clearly that of first order statistics, already examined completely. The adjacent samples X_1 and X_2 are independent of each other.

Any first order process so described as a second order is a degenerate second order process. A degenerate second order distribution is formed as a product of the two first order distributions.

$$W(X_1, X_2) = W(X_1)\ W(X_2)$$

The double integral in the Fourier transform becomes the product of two double integrals so that the joint second order c.f. is the product of two integrals. Thus the c.f. is $F_{x_1,x_2}(\xi_1, \xi_2) = F_x(\xi_1)\ F_x(\xi_2)$. The joint second order distribution between the quantizer output and input is the product of the joint first order distributions.

$$W(X_1 X_1', X_2 X_2') = W(X_1 X_1')\ W(X_2 X_2')$$

The quadruple integral in the Fourier transform breaks up into the product of two double integrals so that the joint second order c.f. is the product of the two joint first order c.f.'s.

$$F_{x_1, x_1', x_2, x_2'}(\xi_{1_a}, \xi_{1_b}, \xi_{2_a}, \xi_{2_b}) = F_{x, x'}(\xi_{1_a} \xi_{1_b}) \, F_{x,x'}(\xi_{2_a} \xi_{2_b})$$

Periodicities of duration ϕ must exist for the joint degenerate second order c.f. along the ξ_{1_b} and ξ_{2_b} axes in five-dimensional space. The joint c.f. is aperiodic along the ξ_{1_a} and ξ_{2_a} axes. The typical repeated section in five-dimension space is the product of the typical sections of $F_{x,x'}(\xi_{1_a}, \xi_{1_b})$ and $F_{x,x'}(\xi_{2_a}, \xi_{2_b})$. Each of these is the joint c.f. between a variable X, and the sum of X plus quantization noise (see Page 29). The quantization noises are included with the factors $\sin(\pi/\phi \, \xi_{1_b})/(\pi/\phi \, \xi_{1_b})$ and $\sin(\pi/\phi \, \xi_{2_b})/(\pi/\phi \, \xi_{2_b})$ in the two joint first order c.f. typical sections. They are multiplied together in the formation of $F_{x_1, x_1', x_2, x_2'}(\xi_{1_a}, \xi_{1_b}, \xi_{2_a}, \xi_{2_b})$ and because of this, the quantization noise is first order.

This situation cannot be distinguished from that of quantizing (with identical quantizers) two first order variables X_1 and X_2 as shown in Figure 31, except that the possibility of having different first order densities for X_1 and X_2 is introduced. This could not arise in physical stationary processes where X_1 and X_2 are adjacent samples of the same random process but is a perfectly possible type of joint relation for a component of the joint distribution between the original input variables X_1 and X_2. If X_1 and X_2 are independent of each other, so will be the quantization noises.

The quantization of a process having a two-dimensional sin X/X distribution centered at the origin is now completely accounted for. The more general problem, that of the two-dimensional sin X/X distribution centered at $X_1 = kq$ and $X_2 = q$ presents no new analysis problems.

This distribution is $\sin \pi(X_1/q+k)/\pi(X_1/q+k)$ $\sin \pi(X_2/q+\ell)/\pi(X_2/q+\ell)$ which represents the statistics of two independent signals X_1 and X_2, because it is factorable. Again, the quantization noises are independent and the typical repeated section of the joint second order c.f. is the same as if the quantizers of Figure 31 were replaced by independent noise sources, each having distribution $Q(n)$.

Figure 31 may be modified to include 3, or more jointly related first order signals to represent higher order processes. By arguments similar to those of the second order process, a most general analysis may induced. If the probability density distribution of an n-th order quantizer input has a n-dimensional c.f. that is negligible outside the range $-\phi/2 < \xi_1, \xi_2 \ldots \xi_n < \phi/2$ the joint c.f. between the quantizer output and input, a function of 2n variables $\xi_{1a}, \xi_{1b}, \ldots \xi_{na}\ \xi_{nb}$ is periodic of radian fineness ϕ along the axes $\xi_{1b}, \xi_{1b}, \ldots \xi_{nb}$ and aperiodic along the the axes $\xi_{1a}, \xi_{2a} \ldots \xi_{na}$, having a typical repeated section which is the same as the joint c.f. between the quantizer input signal and that input plus independent first order noise of distribution Q(n). A sketch of this for a first order quantizer input is Figure 29(a). When the multidimensional Nyquist restriction is met, all self and joint moments are unaffected if the quantizer is replaced by a source of first order independent noise of distribution Q(n)

The description of the response of the quantizer to statistical inputs that satisfy the Nyquist restriction is complete. A question of how rough might the quantization be for specific distributions and have the restriction satisfied naturally arises. This will be answered for several cases of Gaussian statistics which are important in quantizer system analysis.

C. Practicality and Range Of Application of the Mathematical Results

(1) Quantization of first order Gaussian signals

Figure 32 shows the first order Gaussian distribution density centered at the origin, and its c.f. From the expression for the c.f., it is obvious that $F_x(\xi)$ will not go to zero outside of any finite band about its origin. However, it acquires negligible proportions very rapidly, going down with $e^{-\xi^2\sigma^2/2}$. σ is the root mean square of the Gaussian signal X. If we let the quantization nox size $q = \sigma$, then the error made by assuming that the c.f. obeys the Byquist restriction may be estimated from consideration of Figure 33, where the c.f. of quantized first order Gaussian statistics is shown. Each section, repeated with radian fineness $\phi = 2\pi/q = 2\pi/\sigma$, is of the form

$$e^{-\xi^2\sigma^2/2} \sin(\pi\xi/\phi)/(\pi\xi/\phi).$$

The errors in the moments of the quantized statistics when evaluated by assuming that the quantization noise is independent and of the distribution Q(n) are due to the contributions of the overlap to the derivatives of the typical section at the origin. Because X was chosen with zero average, the typical repeated c.f. section is even (symmetrical), causing the contributions to the odd derivatives to cancel, while the contributions to the even derivatives reinforce. The theoretical errors in all odd moments are zero. The errors in mean square and in mean fourth that result are tabulated, in Figure 34. The same calculation has been made for $q = 2\sigma$, and these results are also tabulated.

Errors in analysis are extremely small when $q = \sigma$. They remain moderately small when the quantization is as rough as $q = 2\sigma$, but increase rapidly as the roughness increases further. When $q = \sigma$, the error in the mean square is 10^{-6}% of the mean square of the input, and about 10^{-5}% of the mean square of the quantization noise. These percentages climb to 3% and 9% respectively, when q is increased to 2σ. Such errors are very tolerable, being suprizingly small for quantization that rough. The error in mean fourth is $3(10)^{-5}$% of the mean fourth of the quantizer input, $6(10)^{-2}$% of the mean fourth of Q(n) when $q = 2\sigma$.

The accuracy of this description of first order statistics as reflected in the accuracy of the moments of the quantizer output for Gaussian input is sufficiently great until the box size is as big as two standard deviations.

It was held that quantization noise is first order and uncorrelated although the quantizer input may be highly correlated, for fine quantization. Just how fine this has to be as a function of the correlation coefficient of a second order Gaussian input will give a general indication of the sensitivity of the statistical independence of quantization noise to quantization box size.

2. Quantization of second order Gaussian signals.

Figure 35 shows the two-dimensional second order Gaussian distribution, centered at the origin, and its two-dimensional c.f. Let this be the statistical distribution of the quantizer input. The two-dimensional c.f. of the resulting quantizer output, shown in

$$e^{-\xi^2\sigma^2/2} \sin(\pi\xi/\phi)/(\pi\xi/\phi).$$

The errors in the moments of the quantized statistics when evaluated by assuming that the quantization noise is independent and of the distribution Q(n) are due to the contributions of the overlap to the derivatives of the typical section at the origin. Because X was chosen with zero average, the typical repeated c.f. section is even (symmetrical), causing the contributions to the odd derivatives to cancel, while the contributions to the even derivatives reinforce. The theoretical errors in all odd moments are zero. The errors in mean square and in mean fourth that result are tabulated, in Figure 34. The same calculation has been made for $q = 2\sigma$, and these results are also tabulated.

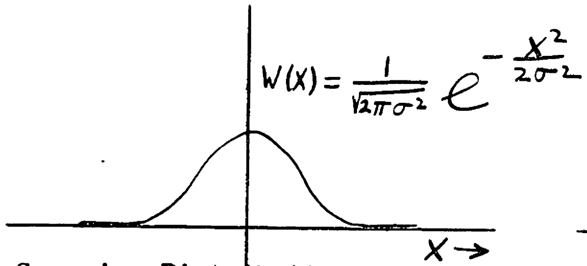
Errors in analysis are extremely small when $q = \sigma$. They remain moderately small when the quantization is as rough as $q = 2\sigma$, but increase rapidly as the roughness increases further. When $q = \sigma$, the error in the mean square is 10^{-6}% of the mean square of the input, and about 10^{-5}% of the mean square of the quantization noise. These percentages climb to 3% and 9% respectively, when q is increased to 2σ. Such errors are very tolerable, being suprizingly small for quantization that rough. The error in mean fourth is $3(10)^{-5}$% of the mean fourth of the quantizer input, $6(10)^{-2}$% of the mean fourth of Q(n) when $q = 2\sigma$.

The accuracy of this description of first order statistics as reflected in the accuracy of the moments of the quantizer output for Gaussian input is sufficiently great until the box size is as big as two standard deviations.

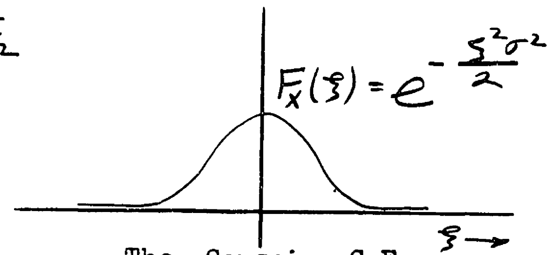
It was held that quantization noise is first order and uncorrelated although the quantizer input may be highly correlated, for fine quantization. Just how fine this has to be as a function of the correlation coefficient of a second order Gaussian input will give a general indication of the sensitivity of the statistical independence of quantization noise to quantization box size.

2. Quantization of second order Gaussian signals.

Figure 35 shows the two-dimensional second order Gaussian distribution, centered at the origin, and its two-dimensional c.f. Let this be the statistical distribution of the quantizer input. The two-dimensional c.f. of the resulting quantizer output, shown in



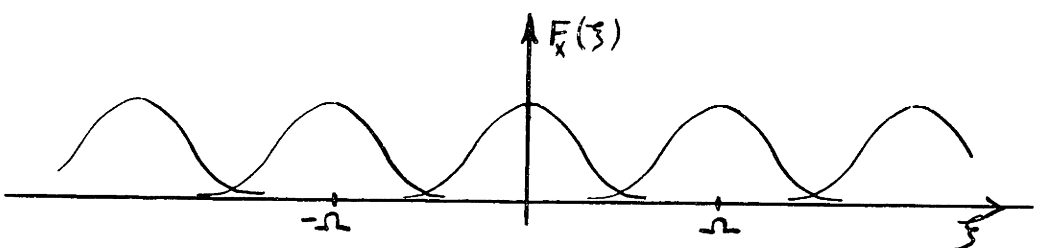
The Gaussian Distribution Density



The Gaussian C.F.

The Gaussian Distribution

FIGURE 32



C.F. Of First Order Gaussian Distribution After Quantization

FIGURE 33

Box size	error in mean square	mean sq. of input	mean square of quantization noise	error in mean fourth	mean fourth of input	mean fourth of quantization noise
$q = \sigma$	$(1.1)10^{-8}\sigma^2$	σ^2	$\sigma^2/12$	$(0.8)10^{-6}\sigma^4$	$3\sigma^4$	$(1.25)(10)^{-2}\sigma^4$
$q = 2\sigma$	$(3.1)10^{-2}\sigma^2$	σ^2	$\sigma^2/3$	$(0.5)\sigma^4$	$3\sigma^4$	$(0.2)\sigma^4$

note: errors in odd moments are zero.

Table of Accuracy of Analysis vs. Box Size

FIGURE 34

Figure 33. A top view is shown in Figure 37. Since the k, ℓ moment is

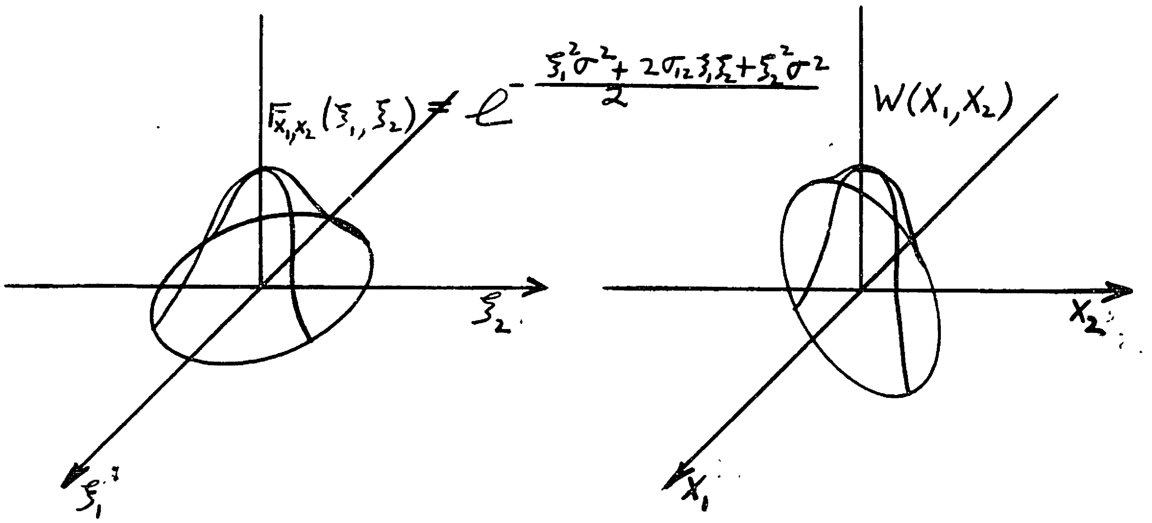
$$\overline{X_1^k X_2^\ell} = \frac{1}{j^{(k+\ell)}} \frac{\partial^{(k+\ell)}}{\partial \xi_1^k \partial \xi_2^\ell} F_{X_1, X_2}(\xi_1, \xi_2) \Big|_{\substack{\xi_1 = 0 \\ \xi_2 = 0}}$$

its errors are due to the contribution of the overlap (Figures 36 and 37) to this derivative. Of interest is the error in the correlation $\overline{(X_1 X_2)}$, the 1, 1 moment. To this, the contributions of sections #1 and #2 (see Figures 36 and 37) are opposite. Likewise, the contributions of #3 and #4 cancel. The contributions of #7 and #8 reinforce but are negligible compared to those of #5 and #6 which really are responsible for the error in the joint first moment of the quantizer output. This error is equal in magnitude to the correlation in the quantization noise. A plot of the normalized correlation of quantization noise (the ratio of the joint first moment to the mean square) as a function of the normalized correlation coefficient of the second order Gaussian distribution of the input (the ratio of the correlation coefficient $\overline{(X_1 X_2)} \equiv \sigma_{12}$ to the mean square σ^2) is shown in Figure 38. The general relation for the correlation of quantization noise is (2.14).

$$\text{(normalized correlation)} = e^{-4\pi \frac{\sigma^2}{q^2}(1 - \sigma_{12}/\sigma^2)} \tag{2.14}$$

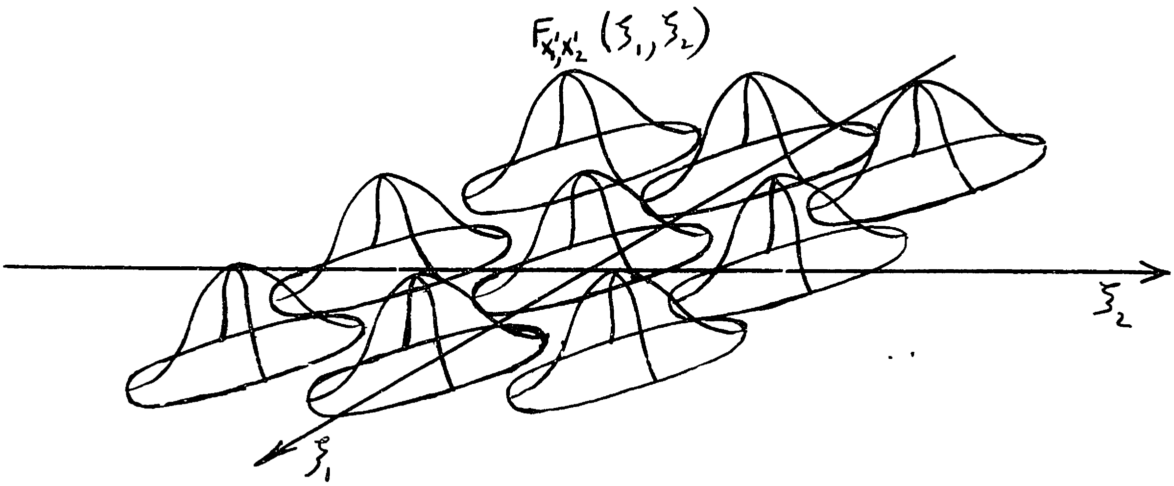
From Figure 38 it can be seen that quantization noise is practically uncorrelated until the box size is one standard deviation and the input correlation is 95% or until the box size is two standard deviations and the input correlation is 80%. A box size of two standard deviations corresponds to extremely rough quantization. The dynamic range of an input variable is practically three quantization levels. This is almost in the realm of switching circuits.

It can now be qualitatively stated that if the dynamic range of a variable being quantized extends over several boxes, the quantization noise will be uniformly distributed and will be a first order process. All moments will be the same as if the quantizer were a source of random independent noise. The causality between the input variable and the quantization noise is manifested in the form of



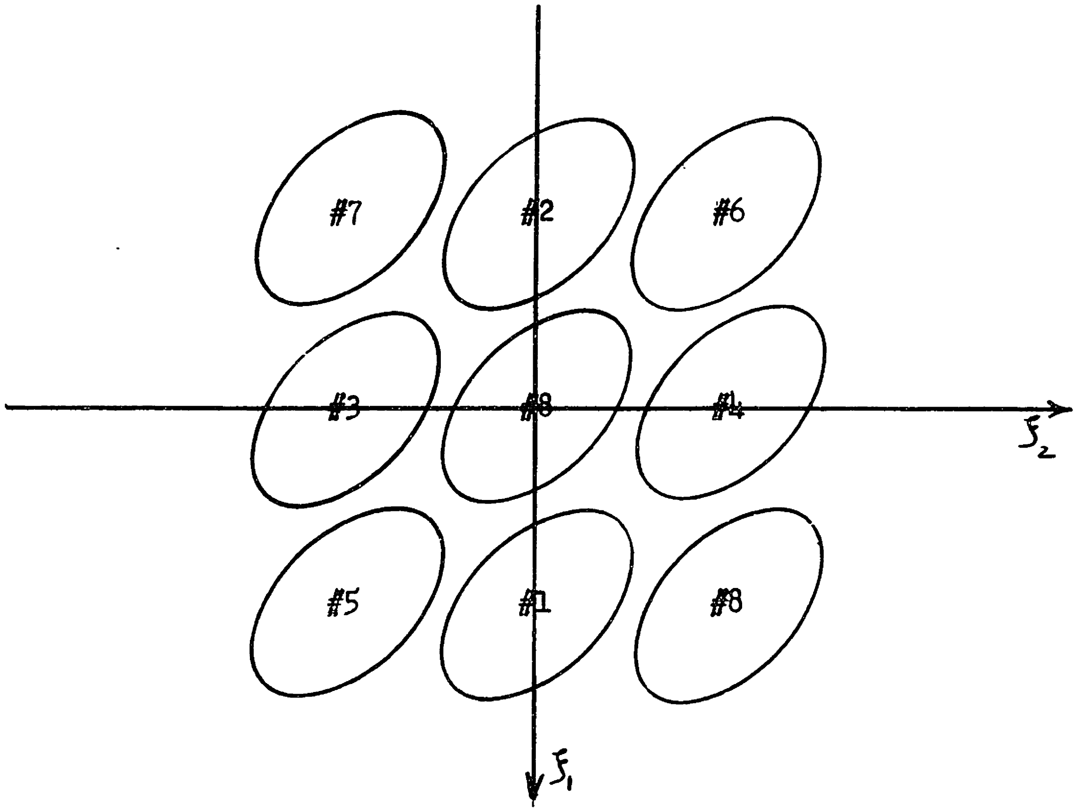
The Second Order Gaussian Distribution And Its C.F.

FIGURE 35



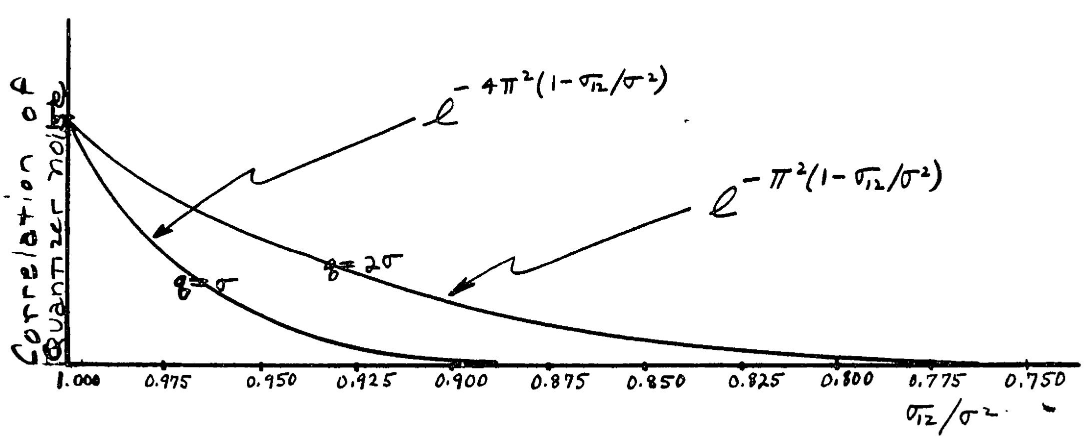
Two-Dimensional C.F. Of Quantizer Output For A Two-Dimensional Gaussian Input

FIGURE 36



Top View Of Figure 36 Showing Contour Of Constant Probability

FIGURE 37



Correlation Of Quantization Noise vs. Correlation Of Quantizer Input Signals

FIGURE 38

periodicities with radian fineness $\phi = 2\pi/q$ along the c.f. axes which are transformed from quantizer output variables $(X_1', X_2' \ldots. X_n')$ and one-sectioned along the axes transformed from the quantizer input variables $(X_1, X_2, \ldots. X_n)$. A single section corresponds to additive independent unrelated noise having the same statistical nature as the quantization noise itself. This picture of quantization noise greatly simplifies the analysis of quantizer systems.

CHAPTER III

QUANTIZER SYSTEM ANALYSIS

The techniques of quantizer system analysis will be developed specifically for sampled data systems. After this is done, how to use them, and when they can be used for the analysis of continuous systems will be indicated. Analysis will be the chief concern; synthesis proceeds along usual lines, ignoring quantization, and later introducing the same at various places for convenience or for economic reasons, allowing it to be as rough as can be tolerated. The systems to be considered, whether they are physical devices or purely mathematical situations, are described by conventional difference equations except for quantization. These difference equations will be "excited" by initial conditions and driving functions whose statistical properties will be given.

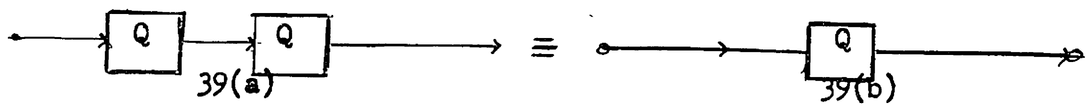
In analysis, it is always sought in what respect, if any, the quantizer behaves like a source of additional random first order noise of distribution density $Q(n)$. It is necessary to be able to determine whether or not the Nyquist restriction will be met by the probability density of the signal input to the quantizer. This will be accomplished analytically for "linear" quantizer systems, those that become linear sampled-data systems when the quantizer is replaced by a device having a gain of unity. This Nyquist test for "non-linear" quantizer systems cannot be made analytically. However, in many cases, such as numerical solutions of difference equations, quantization or round-off is fine enough so that variables surely make excursions over many quantization levels and thereby satisfy the restriction.

Systems containing more than one quantizer may often be treated as if each quantizer were an independent source of additive noise of distribution $Q(n)$ provided that the statistics of each quantizer input satisfies the Nyquist restriction. A case of a two-quantizer system, where the restriction is not met, is shown in Figure 39(a). The c.f. of the input to the second quantizer of Figure 39(a) is periodic and of infinite extent. This is difficult to analyze by the methods presented. However, other methods may often be used when a random signal having an impulse distribution is quantized. For example, an obvious equivalent of Figure 39(a) is Figure 39(b). If k is an integer, Figure 40(a) is identical to Figure 40(b). If k is 1/2 a marginal situation exists. The possibility arises for inputs to the second quantizer to lie exactly on the edges of the quantization boxes. If the rule is made that in case of doubt, the output is always upgraded, a plot of the output of the second quantizer versus the input of the first quantizer is shown in Figure 41. An equivalent one-quantizer system is shown in Figure 42. These procedures may not be useful when k is not rational, or when k is replaced by a more general linear sampled-data filter.

A. Analysis of "Linear Sampled-Data Quantizer Systems

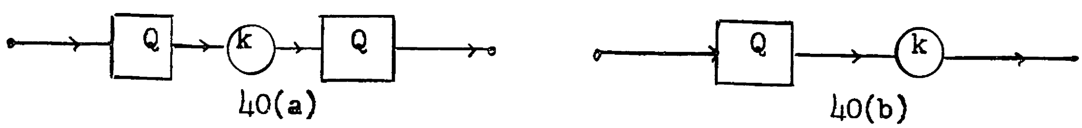
1. Qualitative aspects

If a sampled-data system which is linear except for a single quantizer contained somewhere within is excited so that the statistics of the quantizer input signal satisfy the Nyquist restriction, the difference between the quantizer output and input will be a first order process and will have the simple distribution Q(n). It is convenient to consider the linear "equivalent" to the quantizer system, driven by the same input excitation and identical in every respect except that the quantizer is replaced by a linear device having gain unity. The output of the quantizer system is the sum of two components, one identical with the output of the equivalent system, while the other is the result of driving the linear equivalent only by the quantization noise at the quantizer position. Since the rest of the quantizer system is linear, "signal" and "noise" may be treated separately; this is a wise procedure because the statistical nature of quantizer noise, the only aspect of it that is generally useful, is highly independent of the signal and it is clearly known à priori. Knowing the gain of the equivalent from the quantizer point to the output, it is possible to evaluate the statistical properties of the system output. Two very significant characteristics of a linear quantizer system are, then, the impulse response or gain of its linear equivalent, and the joint probability density distribution of the noise component in its output. An alternative to the latter would be knowledge of the linear gain from



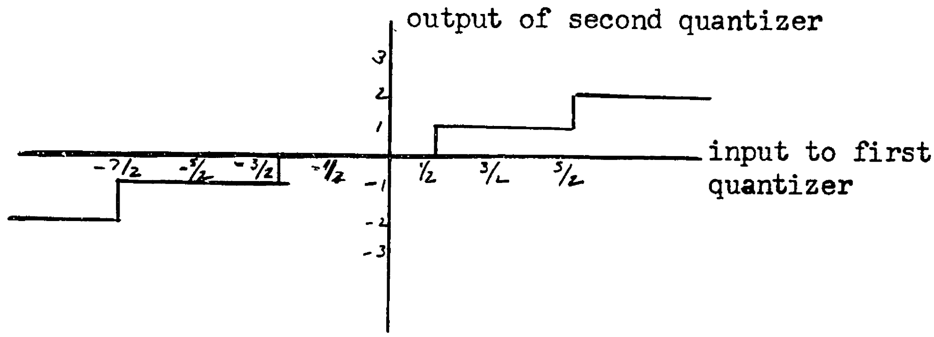
Two Cascaded Quantizers

FIGURE 39



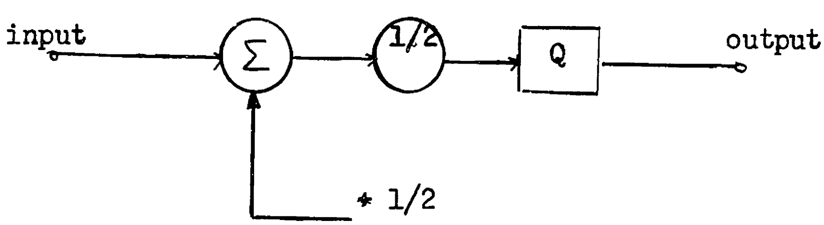
A Simple Quantizer System

FIGURE 40



Input-Output Relation For The System Of Figure 40(a) when k = 1/2

FIGURE 41



A One-Quantizer System Equivalent To That Of Figure 40(a) when k = 1/2

FIGURE 42

the quantizer to the output plus the quantization box size q.

Once a quantizer is driven with an input that satisfies the Nyquist restriction, addition of another independent signal cannot change this situation. When their respective c.f.'s are multiplied to give the c.f. of the sum, the result can be no wider than the narrower c.f. of the constituents. In general, it will be even narrower than this and the restriction will be met more easily.* In the amplitude domain, a quantizer having a sufficiently big dynamic range (extending over several quantization boxes) can only have it increased by the addition of another independent input. Since the output of a quantizer is the same as the input plus an additive noise of fixed distribution Q(n), the quantizer is "linearized" by any input component satisfying the Nyquist restriction. The same effects are realized with statistically related input components except where the addition of a component signal reduces the dynamic range already existing to one so small that the restriction is no longer met.

The system consequences of the "linearization" of a quantizer within a system are similar to those for the quantizer alone. Here, the entire system is "linearized". For two input components, the quantizer output consists of the sum of three parts. Two of them are the respective output components of the linear equivalent system when driven by the two inputs. The third is due to quantization noise. It has a different waveform in time after the addition of the second input component, but has the same statistical characteristics as before.

According to the Central Limit Theorem, the addition of a good number of independent random quantities of arbitrary distributions yields a random process that becomes closer and closer to Gaussian as the number of included variables is increased. The output of a sampled-data filter at a given sample time is a weighed sum of past inputs that are often of a first order process, so that statistical outputs of

---

* In conventional filter theory, this appears in the fact that the linear filtering of a signal can only reduce its bandwidth, or at best, allow it to be unchanged.

"long-memory" sampled-data systems are almost Gaussian. In particular, if the impulse response from a quantizer point to the output contains a half dozen samples or more, a given noise output, the sum of that many independent past noises, is nearly Gaussian. All that is needed to specify the first order distribution of the system output component due to quantization noise, then is its mean square. Since the original quantization noise samples are independent, mean squares add. The second moment of the distribution Q(n), the mean square of quantizer noise, is $1/12\ q^2$. The mean square system noise is then $1/12\ q^2$ times the sum of the squares of the impulse magnitudes of the response at the output to a unit impulse applied at the quantizer position.

This is as far as can be gone qualitatively. The above results are perhaps the most useful - surely the most simple. The problems remaining deal chiefly in analytical fine points:

(a) Propagation of statistics through linear sampled-data systems. This is treated in detail in Appendix II.

(b) Tests to determine whether the Nyquist restriction will be met at the input of the quantizer (quantizers). This can usually be done by inspection, keeping in mind the results of Chapter II, Section C. More precise ways of doing this will be developed below.

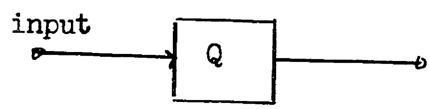
(c) Derivation of moments of system outputs, and of joint moments between system output and input will be done below.

2. Tests on linear quantizer systems to determine whether the Nyquist restriction is met at the quantizer input.

If the c.f. of the random input to the quantizer of Figure 43 is finite within a hypercube of edge A, and centered at the origin in multidimensional c.f. space, and zero (negligible) elsewhere, the Nyquist restriction will be satisfied if $\phi < A$, where $\phi = 2\pi/q$. If the input of Figure 43 is first applied to a device having the magnification k (as in Figure 44, the Nyquist restriction will be obeyed as long as $\phi < A/k$.

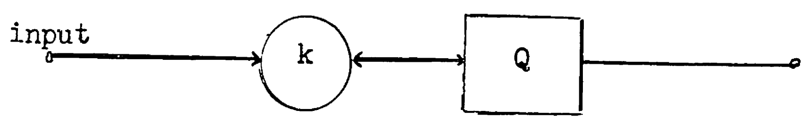
The next more complicated quantizer system is shown in Figure 45. The c.f. of the quantizer input is not simply a scaled version of the filter output.

From Appendix II, the two-dimensional c.f. of the output of the



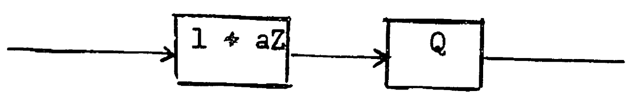
A Quantizer Having A Box Size q

FIGURE 43



A Quantizer Preceeded By a Multiplier

FIGURE 44



A Quantizer Preceeded By A Sampled-Data Filter Of One Memory State

FIGURE 45

linear filter is (3.1).

$$F_{0_1,0_2}(\xi_1,\xi_2) = \left[F_x(\xi_1)\right] \left[F_x(a\xi_1+\xi_2)\right] \left[F_x(a\xi_2)\right] \quad (3.1)$$

The input c.f. is $F_x(\xi)$. From this, it is possible, but not particularly easy, to explore the value of $F_{0_1,0_2}(\xi_1.\xi_2)$ and to see how big a square in the ξ_1, ξ_2 plane would be necessary to enclose its finite region. Let the edge of this square be β, and the test is surely positive as long as $\phi > \beta$.

Let the input signal in Figure 45 have a c.f. that is only finite within a hypercube of edge A. The addition of an independent first order signal having the c.f. $F_{\ell p}(\xi)$ as shown in Figure 46 does not change the statistics of the system input as long as $\alpha > A$. If the first order process indicated in Figure 46 is treated as a degenerate process of the same order as the original input, the c.f. of the sum, the product of the c.f.'s, is not changed because it is multiplied by unity everywhere in c.f. space where it is finite and significant. In the analogy between the c.f. domain and the frequency domain of conventional linear system analysis, $F_{\ell p}(\xi)$ resembles an ideal low-pass filter characteristic, whence the notation.

Addition of the independent signal having c.f. $\equiv F_{\ell p}(\xi)$ caused no change in the input distribution and therefore could have no effect upon the output distribution. The output component due to the signal of c.f. $\equiv F_{\ell p}(\xi)$ could have been added in directly at the filter output because the filter is linear, again without changing the statistics of the signal there. Examination of the expression for the two-dimensional output c.f. of the added independent output signal has values of either 0 or 1, with a connected 1 region symmetrically clustered about the origin. This 1 region must completely envelope the multidimensional c.f. of the output signal due to the original input in order that their multiplication yield the original c.f. Therefore, the spread of the c.f. of the input having c.f. $\equiv F_{\ell p}(\xi)$ in propagating through the filter sets a limit on the spread of the c.f. of the original signal and always allows a conservative estimate of this for the types of characteristic functions of interest (those that are band-limited). A linear filter is conservatively characterized then by the greatest extent of the multidimensional c.f. that results when the input is driven by a random signal of c.f. $\equiv F_{\ell p}(\xi)$. In this way, a ratio is obtainable for a given filter that will yield a conservative estimate of the maximum allowable quantization box size for the satisfaction of the Nyquist restriction given the maximum allowable box size for the input signal if it were to be quantized. To be more specific and evaluate the errors in analysis, as was done for the quantizer alone in Chapter II, section C, would require separate calculation for every input distribution, first order and higher order.

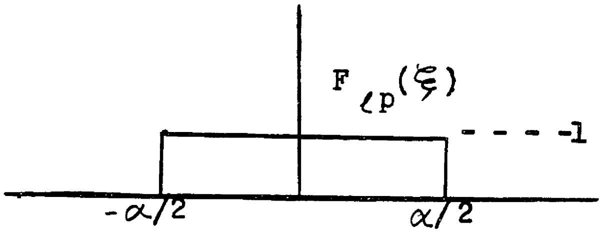
As an example, consider the sampled data filter of gain $(1+Z/2)$ to have an input signal of c.f. $\equiv F_{\ell p}(\xi)$ where $F_{\ell p}(\xi) = 1 \quad -\alpha < \xi < \alpha$; $= 0$ elsewhere, as in Fig. 45

The c.f. of the resulting random output is (3.2).

$$F_{0_1, 0_2}(\xi_1, \xi_2) = \left[F_i(\xi_1)\right]\left[F_i(1/2\,\xi_1 + \xi_2)\right]\left[F_i(1/2\,\xi_2)\right] \qquad (3.2)$$

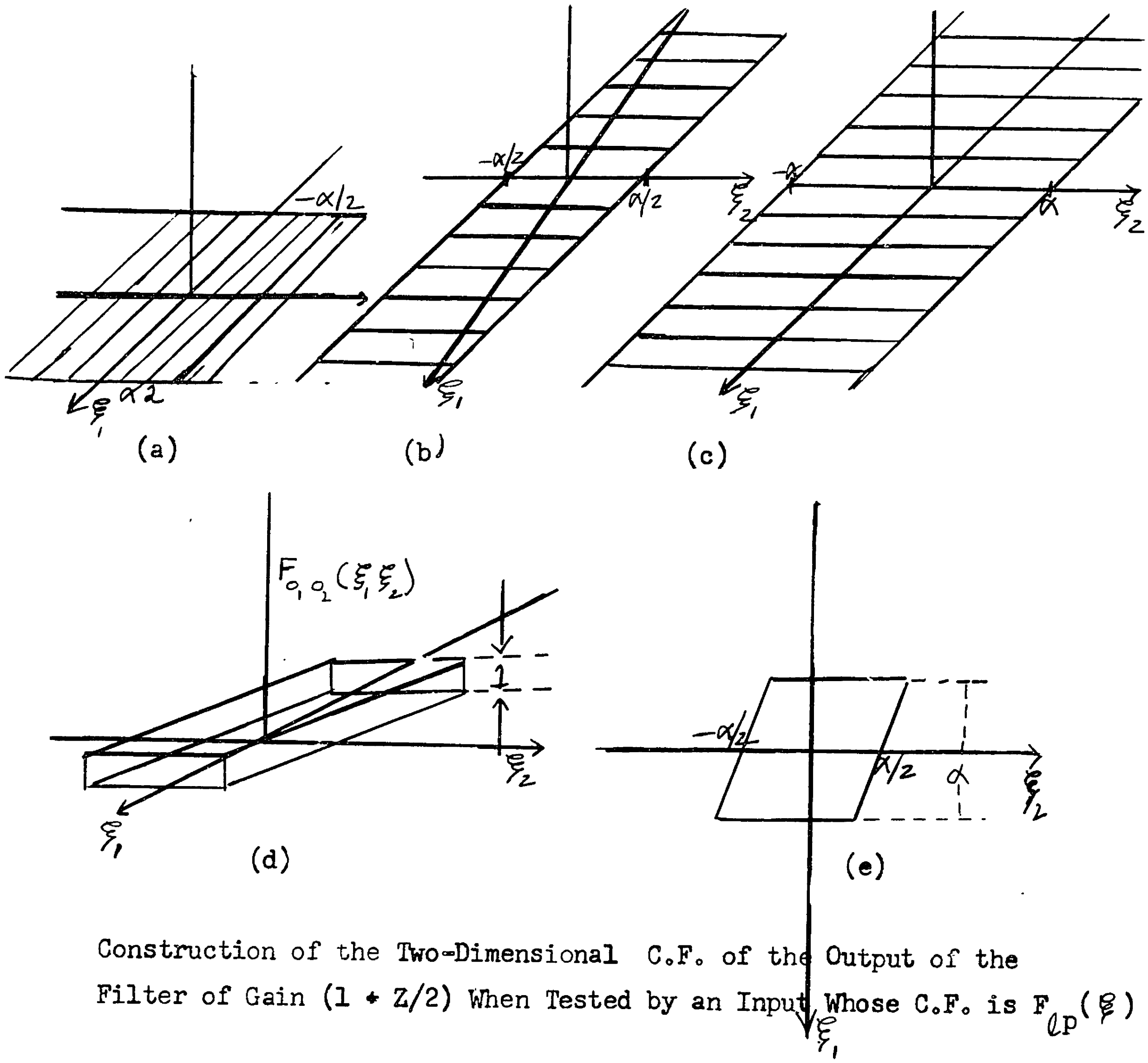
The shaded regions of Figure 47 (a), (b), and (c) show where on the ξ_1, ξ_2 plane the three factors of the output c.f. are finite. Figure 47 (e) is a region where all three factors are finite and have the value unity. An input whose c.f. is finite only when $-\alpha/2 < (\xi_1, \xi_2, \ldots \xi_n) < \alpha/2$ will give an output c.f. only within the enclosed region of Figure 47(e). The multidimensional Nyquist restriction would be satisfied by a quantizer at the filter output having a quantization frequency $\phi > 3/2\alpha$. Satisfaction of a first order restriction would be guaranteed of $\phi > \alpha$, but it would not be clear whether or not the resulting quantization noise would be uncorrelated although its first order distribution density would by Q(n).

A similar process could be used in exploring the extent of a c.f. after linear filtering for filters of a higher but finite number of memory states. However, feedback sampled-data systems of infinite memory are quite common, and cannot be dealt with practically as above. Consider the linear feedback system of Figure 48, having first order (uncorrelated) input to the filter of gain D_1 can be infinite in c.f. space only where $\alpha/2 < (\xi_1, \xi_2 \ldots\ldots \xi_n) < \alpha/2$. This signal is of a high order process because of the feedback signal, but the input, being first order and hence statistically independent of the feedback, sets a



Characteristic Function That Has Unit Value Within A Finite Region And Zero Elsewhere

FIGURE 46



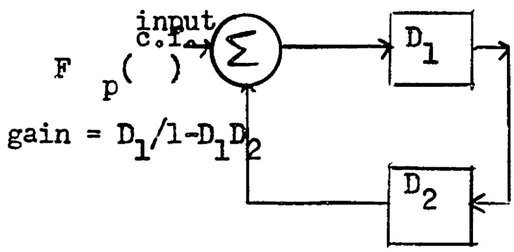
Construction of the Two-Dimensional C.F. of the Output of the Filter of Gain (1 + Z/2) When Tested by an Input Whose C.F. is $F_{\ell p}(\xi)$

FIGURE 47

limit on the spread of the c.f. into the filter D because the c.f.'s multiply. Thus the extent of the c.f. at the output of the system of Figure 48 is not increased by breaking the feedback link and the system may be conservatively checked by checking the filter of gain D_1 alone. The identical argument may be applied to the quantizer feedback systems shown in Figures 49 (a), (b), and (c). In (a), the restriction will be satisfied if the input itself has c.f. finite only in the region $-\phi/2 < (\xi_1, \xi_2, \xi_3 \ldots\ldots) < \phi/2$. In (b) only the filter of gain D need be tested. In (c) only the filter of gain D_1 need be tested. In general, single quantizer systems may be tested by testing the linear transmission from the input of the system to the input of the quantizer output link cut.

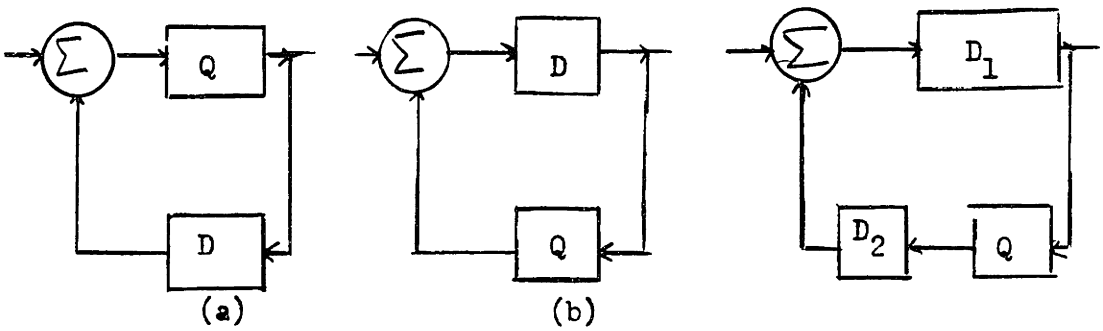
How to estimate the extent of the characteristic functions of the inputs in multiple-quantizer systems cannot be described in general because there are too many situations to consider. However, certain methods of approaching such problems can be illustrated in examples of two-quantizer systems.

Tests on the inputs to the quantizer Q_1 and Q_2 of Figure 50 are easy to make and are the same as for the systems of Figures 51(a) and (b) respectively. Fundamentally, the simplicity arises from the fact that the signal input to either quantizer does not come directly through the other. A system where an apprieciable portion of the input signal of quantizer Q_2 may come from the output of Q_1 is shown in Figure 52. The test for the input of Q_1 is the same as for the quantizer in Figure 53. The test for the input of Q_2 is not that simple, being the same as for the quantizer Q_2 in Figure 54. The feedback paths of Figure 52 which were broken to give Figure 54 carry signals that could only make characteristic functions narrower. Hence, the test on the system of Figure 54 is more conservative than is necessary, but simple enough to be made. First, calculate how narrow the width δ of the c.f. at the input to D_2 must be in order to satisfy the Nyquist restriction at Q_2. Then the problem is to determine how narrow the system input must be to insure that the c.f. at the output of the summer is narrower than δ in c.f. space. The c.f. of the sum (the c.f. of the input to D_2) is obtainable from the joint c.f. of the outputs of D_3 and Q_1. This joint c.f. is the same as would be if Q_1 were replaced by an independent quantization



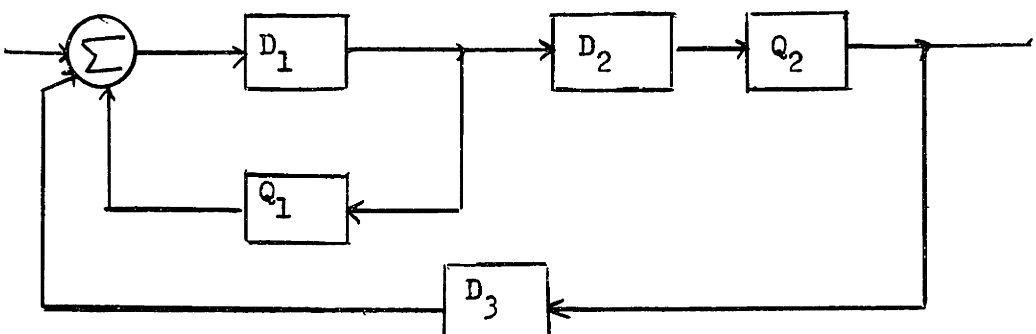
Linear Sampled-Data Feedback
System

FIGURE 48



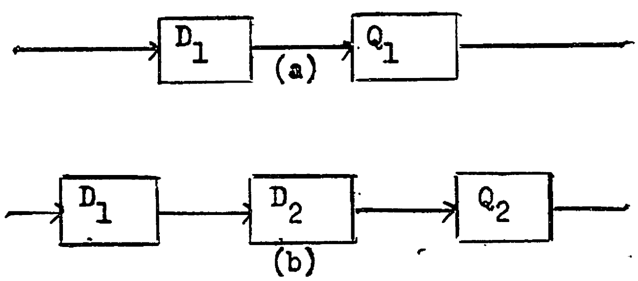
Quantizer Sampled-Data Feedback
Systems

FIGURE 49



A Simple Two-Quantizer Feedback System

FIGURE 50



Open-Loop Quantizer Systems That Are Related To Figure 50 For

Test Purposes

FIGURE 51

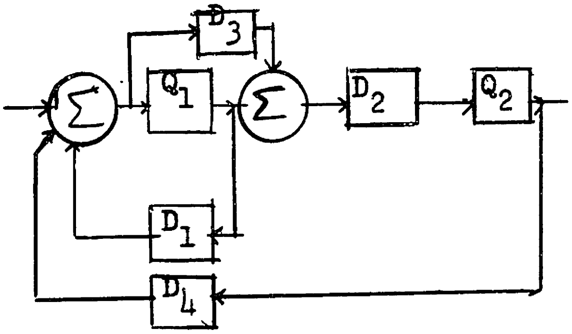
noise source, except that it is periodic along the coordinate of the c.f. variable corresponding to the Q_1 output.

Let the system input have a c.f. of $F_{\ell p}(\xi)$ and be uncorrelated. This is the usual test input. The output of D_3 has an order of one plus the number of memory states. The number of dimensions in c.f. space required of such a joint c.f., and how the c.f. of the sum is derived from the projection of the intersection of a 45^o cut plane and the c.f. surface for a filter D_3 having zero memory states.

Figure 55 is closely related to Figures 36 and 37. The c.f. of a sum is always given by the projections (either or any) of the 45^o cut going through the first quadrant. This c.f. is sketched in Figure 56. The periodicities of the joint c.f. cause the c.f. of the sum to be widened somewhat. The c.f. of the sum will surely be wider than that which results when the quantizer is replaced by an independent source of random quantization noise. It is necessary for the input of D_2 in the system of Figure 57 to have a c.f. of width less than δ calculated above. For simplicity, the random noise could be eliminated in this test and the results would be more conservative. The added noise makes the condition more easily satisfied because its c.f. $\sin(\pi \xi / \phi)/(\pi \xi / \phi)$ multiplies and could only make the result somewhat narrower.

The effects of the periodicity of the joint c.f., i.e., the effects due to the fact that the quantization noise of Q_1 is causally related to the input of Q_1, are in question. Inspection of Figure 55 shows that, very conservatively, there would be no intersection of the 45^o cut with a c.f. surface of only the central c.f. section were written in the dotted square, i.e., if the radian fineness $\phi = 2\pi/q$ of Q_1 is greater than the width of the joint c.f. of the signals being added in Figure 57. This is simply the joint input-output c.f. of D_3. If this test is satisfied, then the causality of the noise of Q_1 has no effect on the c.f. of the input to D_2 of Figure 54; Q_1 is a source of independent noise. All arguments are general for a filter D_3 of any order. A filter of noise zero was used as an example because higher order c.f.'s could not be readily sketched.

The kind of testing just illustrated is applicable to most multiple quantizer situations where a quantizer input comes from a combination of another quantizer output and a signal of continuous

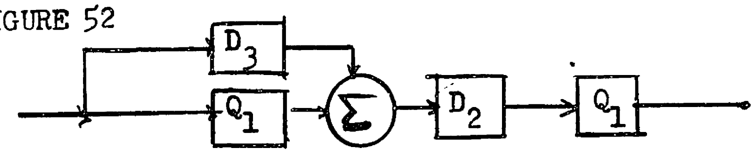


A Two-Quantizer System that Illustrates
The Basic Problem of Testing Quantizer
Inputs In Multiple Quantizer Systems.

FIGURE 52

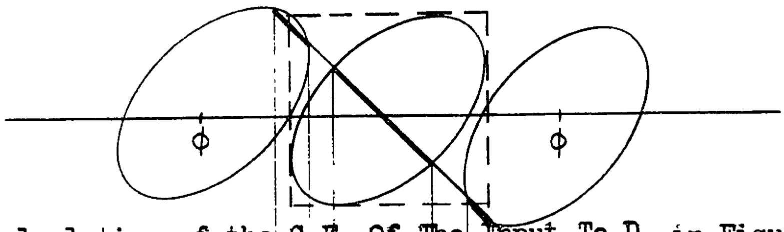
An Open Loop Quantizer System
Used For The Test Upon The Input
Signal of Q_1 (Figure 52)

FIGURE 53



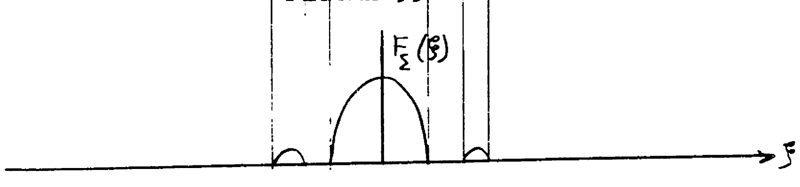
An Open Loop Quantizer System Used For The Test Upon The Input Signal
Of Q_1 (Figure 52)

FIGURE 54



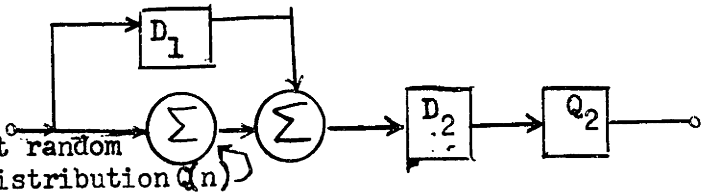
Calculation of the C.F. Of The Input To D_2 in Figure 54

FIGURE 55



The C.F. Of The Input To D_2 In Figure 54

FIGURE 56



independent random
noise of distribution Q(n)

Replacement Of Quantizer In Figure 54 By A Noisy Source

FIGURE 57

probability density. The other possibility, where the input to a quantizer has an impulse-type of probability density, can only be treated analytically in simple cases where levels are commensurate, as has been shown before.

3. Derivation of the moments and joint moments of the signals in linear quantizer systems.

It has been found that if the c.f. of a signal input to a quantizer is finite (significant) only within the region $-\phi/2 < (\xi_1, \xi_2, \xi_3, \ldots \xi_n) < \phi/2$. and zero elsewhere, all moments and joint moments are the same as if the quantizer were replaced by a source of additive independent first order noise of distribution Q(n). It will be shown that as far as all moments and joint moments in a linear system are concerned, the quantizers may be replaced by independent noisy sources having distribution Q(n) provided that this Nyquist condition is met at the inputs of each quantizer.

The output of a quantizer system consists primarily of two components, one due to the given input signal flowing through the linear equivalent system and the other a sum of noises that are generated at and propagate from the quantizers. The distributions and joint distributions of these two separate components are calculable, being descriptions of the propagation of random signals having known distributions through linear sampled-data systems. Most often, this is adequate for analysis and design. However, it may be desirable to have knowledge of the statistics of the sums of signals and noises at various system points. The issue is beclouded because the noise is causally related to the signal but it turns out that the moments of the output (signal plus noise) and joint input-output moments are unaffected by this causality and are the same as if the noises are completely unrelated to the signals.

There are three kinds of devices that appear in linear sampled-data quantizer systems. They are quantizers, linear filters, and adders. It is necessary to investigate the nature of changes in moments as signals propagate through these devices which are treated as "statistics changers".

A given probability density distribution has definite moments which may be derived from the derivatives of its c.f. at the origin. In reverse, only if the c.f. is everywhere representable by a Taylor's series about the origin do a given set of moments define the c.f. and distribution. As shown in Appendix II, the c.f.'s of the probability distributions and joint distributions throughout any linear sampled-data system may be calculated directly from the c.f. of the input signal. If moments of an input always define its c.f., then it could be stated that all moments and joint moments in a linear sampled-data system are calculable from the moments of the joint signal. This result turns out to be true, but needs to be proven.

The proof is simple: Consider as an example, the filter of gain (1 + AZ) (Figure 58) being excited by a signal having the first order c.f. of $F_i(\xi)$. The joint c.f. between output and input, according to Appendix II is (3.3).

$$F_{i_1,0_1,0_2}(\xi_{1a}, \xi_{1b}, \xi_{2b}) = F_i(\xi_{1a})F_i(\xi_{1b})F_i(A\xi_{1b} + \xi_{2b})F_i(A\xi_{2b}) \qquad (3.3)$$

The c.f. of the output signal is (3.4), the same as for $\xi_{1a} = 0$.

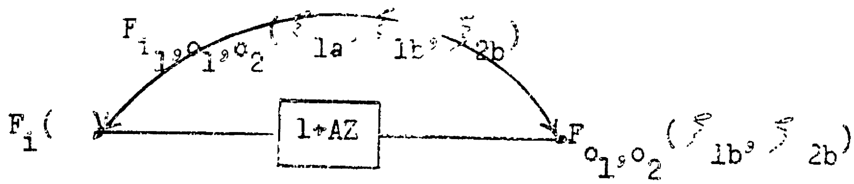
$$F_{0_1,0_2}(\xi_{1b}, \xi_{2b}) = F_i(\xi_{1b})F_i(A\xi_{1b} + \xi_{2b})F_i(A\xi_{2b}) \qquad (3.4)$$

Joint moments between input and output and moments of the output signal are obtainable from the respective c.f. expressions. These are to be differentiated partially and evaluated where all ξ's = 0. Differentiation of the product yields sums of derivatives multiplied by certain factors. When evaluated near the origin in c.f. space, all factors take the value unity (their arguments are zero), and the derivatives are simply proportional to the derivatives of the original c.f. of the input at the origin. For this example, it is thus proven that all moments and joint moments depend only upon the moments of the filter input. This is true in general for higher order inputs and linear filters of many memory states because characteristic functions will always be products of factors of the same general form as in the original (as shown in Appendix II).

Signals progagating through a linear quantizer system may be combined by addition. It will be shown next that moments of the sum and joint moments between the sum and the respective new constituents depend only upon the joint moments of the constituents. Consider the signals X and Y to be added. The nth moment of their sum is (3.5).

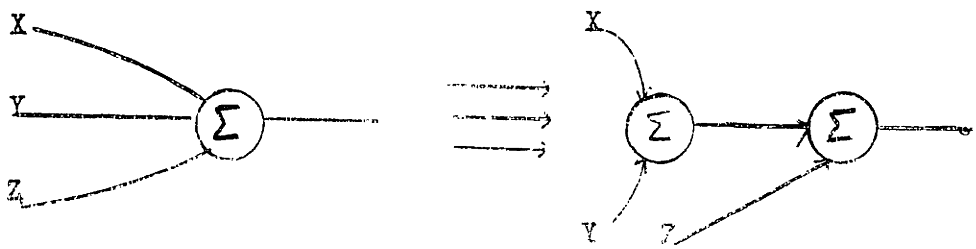
$$\overline{(X+Y)^n} = \overline{X^n} + n\overline{X^{n-1}} + n(n-1)/2!\,\overline{X^{n-2}Y^2} + \dots\dots + \overline{Y^n} \qquad (3.5)$$

Thus any moment of (X + Y) is a linear combination of certain of the moments and joint moments of X and Y. The m,nth joint moment between



Sampled-Data System Of One Memory State Excited By A Random Input Having C.F. of $F_i(\xi)$

FIGURE 58



Equivalence In Ways Of Adding Signals

FIGURE 59

X and the sum is (3.6)

$$\overline{X^m(X + Y)^n} = \overline{X^{m+n}} + \overline{X^{m+n-1}Y} + n(n-1)/2! \ \overline{X^{m+n-2}Y^2} + \ldots + \overline{X^mY^n} \qquad (3.6)$$

This expression is also a linear combination of certain of the moments and joint moments of X and Y. The addition of several signals at a point is equivalent to several separate additions, as illustrated in Figure 59. The theorem is proven and is general in the sense of Figure 59.

Consider a linear sampled-data quantizer system driven by an input of sufficient dynamic range to insure that the Nyquist restriction is satisfied at the input of each quantizer. If one of the quantizers is replaced by an additive first order noise of distribution density Q(n), the probability density at the quantizer output changes from an impulse distribution to a continuous distribution. Considerable changes in the distributions may take place at other points in the system, but nowhere do moments of any kind change. The c.f. at the former quantizer output is very much narrowed. As a result, c.f.'s at other points could only be narrowed and the Nyquist restrictions at the quantizer inputs are more easily met in general. Substitution of the noisy source for the quantizer does not change the "flow of moments" from the quantizer input to output points. The same moments in give the same moments out plus the same joint in-out momemts. The rest of the system components consisting of linear filters, adders, and "satisfied" quantizers, sense only moments and as far as moments are concerned, cannot detect the change. In like manner, each quantizer may be replaced by an independent first order source, having the distribution of quantization noise Q(n), without changing moments of any kind.

Since the characteristic functions when considering signals, noises, or their sums are obtainable by the methods of Appendix II, differentiation of these at their origins gives the moments of the equivalent quantizer system.

Distribution of the signal output component and joint input-output distributions are obtainable by considering the propagation of the actual input signal through the linear equivalent system and applying the methods of Appendix II. In like manner, the distribution of the noise component of the output signal is derived by cutting off the system input signal and replacing the quantizers by independent noisy sources. If the input is applied simultaneously with the independent noises, all moments that result are the same as for the signals plus quantization noises of the original system. The distributions of the combinations of signals and quantizer noises are often obtainable from these their moments, but they are rarely needed once the signals are known, the noises are known, and the moments of their sums are known.

B. Analysis Of Non-Linear Sampled-Data Quantizer Systems

Only certain cases of quantization in non-linear sampled-data systems may be treated by the methods presented. Limitations in application come from the inability to perform the Nyquist test on the probability density distributions of the quantization noise component in a system output signal. The basic difficulty lies in the determination of the distributions of signals resulting from propagation through non-linear systems.

Quantization noise distributions at system outputs may be calculated in cases where non-linear systems are "small-signal linear" where doubling the box size doubles the output noise component amplitude. The "macroscopic" desired signal propagates through the system, and as it swings the non-linear devices over their various dynamic ranges, it causes their incremental gains to vary. To the "microscopic" quantization noises, the system is linear and variable with time. These variations with time depend critically on the "macroscopic" signal in addition to the system characteristics. In other words, the time-varying impulse response from the various quantizer positions to the system output depend upon the system characteristics, the driving functions, and initial conditions. Generally, the statistical output of a time-varying system will be non-stationary, so that precise statistical descriptions will not be sought.

The quantization of Gaussian signals was found to produce first order quantization noise of distribution Q(n) as long as the granularity was finer than two standard deviations, i.e., as long as almost all of the probability density lies within 3 or 4 levels. If the dynamic range of a variable extends over 25 levels, it is safe to say that the quantization noise is first order and is distributed according to Q(n) almost irrespective of the statistical nature of the signal, whether stationary or non-stationary. In numerical analysis where accuracy is of interest, the granularity everywhere is at least that fine. This insures that the quantization noise is "small-signal", first order, and of distribution Q(n).

The origin and nature of quantization noise generated by round-off in numerical analysis being understood leads next to the problem of how it propagates. The useful statistical properites to be considered are again the moments, only here "average" moments will be used to describe non-stationary statistics. As far as moments are concerned, the quantizer may be replaced by adding independent quantization noise. The distribution of the noise component in the output signal may be calculated by making the same substitution just as in linear quantizer systems. A knowledge of the statistical relation between signal and noise is not necessary because the noise is small (always less than 20% of the signal) and is virtually impossible to obtain in non-linear systems. Noise dis-distributions may be estimated with sufficient accuracy from their moments and in many cases, it is sufficient to obtain first order noise distributions by Gaussian fit to match mean squares. The most useful and easiest moment to work with is the second as may be seen in the subsequent example. Consider the homogenous first order non-linear equation with its initial condition (3.7).

$$dy/dt + y^2 = 0$$
$$y(0) = 1.12 \qquad (3.7)$$

An associated difference equation is (3.8).

$$y_{k+1} = y_k - 1/10\, y_k^2$$
$$y_0 = 1.12 \qquad (3.8)$$

A block diagram of the exact numerical solution is shown in Figure 60(a). The squaring operation introduces numbers of double length. Quantization of the squared variable is very convenient for hand calculation and natural with crude 3 decimal digit machine calculation. Figure 60(b), a modification of 60(a), is the block diagram of the equivalent quantizer system. The quantization box size is unity, whence the scaling in 60(b). The quantization process in the point-by-point manual solution of the difference equation is illustrated in the table of Figure 61.

The error to be expected in such a solution due to round-off may be predicted by replacing the quantizer of Figure 60(b) by an independent noise source having a mean square of 1/12, and replacing the multiplier by a linear "time variable potentiometer". The linear incremental gain of the squaring box is $dy_k^2/dy_k = 2\,y_k$. Over the five time intervals of the example, y_k does not change radically, so that an average $2\,y_k$ would be $2(1.12 + .68/2) = 1.80$. The complication of time variable is not necessary. A small-signal equivalent of Figure 60(b) is given in Figure 62(a), and a reduction of this is 62(b). The response at y_k to a single unit negative pulse at the quantizer is shown in Figure 63. The quantization noise at the fourth sample time, for example, is the sum of four noises, due to four past quantizations. All noises are independent and had the same mean square when they originated. The noises are weighed according to the impulse response, and the expected mean square contribution is weighed with the square of the impulse response. Since mean squares of independent noises add, the mean square in the fifth output sample is

$$1/12\ \left[(1/100)^2 + (/82/100)^2 + (.672/100)^2 + (.552/100)^2\right] = 0.2(10)^{-4}.$$

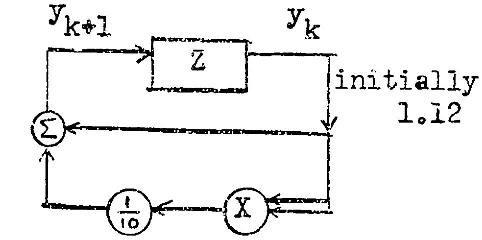
Thus the mean square error in the fifth sample is about one part in 10^4 of the mean square of the fifth sample of the actual signal. Round-off error is negligible. The quantization box size could be ten times as big, and the corresponding mean square error would be only 1% of the mean square output over many "runs" with various initial conditions in the vicinity of 1.12.

Non-linear systems that are "small-signal linear" differ from linear systems in that they contain multipliers, dividers, and non-linear devices whose input-output relations are representable by Taylor's series over the ranges used. All of these devices yield moments and joint moments which depend only upon input moments.

Consider first a device whose in-out relation is the Taylor's series (3.10).

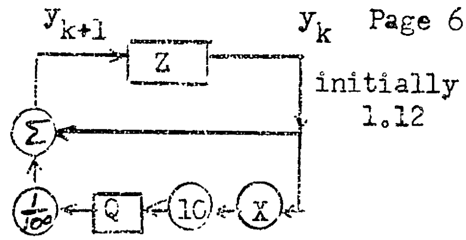
$$o = a_0 + a_1 i + a_2 i^2 \ldots \qquad (3.10)$$

the kth output moment is $\overline{o^k} = \overline{(a_0 + a_1 i + a_2 i^2 \ldots .)^k}$. This will give a sum of average powers of i, or a sum of the moments of i. The most general moments, the joint in-out moments, are $\overline{o^k i^l} = \overline{i^l (a_0 + a_1 i + a_2 i^2 \ldots)^k}$.



Block Diagram Of The Difference Equation (3.9)

FIGURE 60 (a)



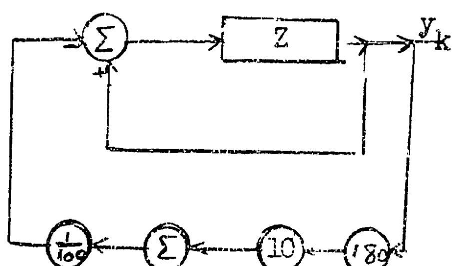
Block Diagram Scaled For Unit Quantization Box Size

FIGURE 60(b)

k	y_k	y_k^2	y_k^2(rounded)	$y_{k+1} = y_k - 1/10\ y_k^2$(rounded)
0	1.12	1.2544	1.3	$y_1 = 1.12 - 0.13 = 0.99$
1	0.99	0.9801	1.0	$y_2 = 0.99 - 0.10 = 0.89$
2	0.89	0.7921	0.8	$y_3 = 0.89 - 0.08 = 0.81$
3	0.81	0.6561	0.7	$y_4 = 0.81 - 0.07 = 0.73$
4	0.73	0.5329	0.5	$y_5 = 0.73 - 0.05 = 0.68$

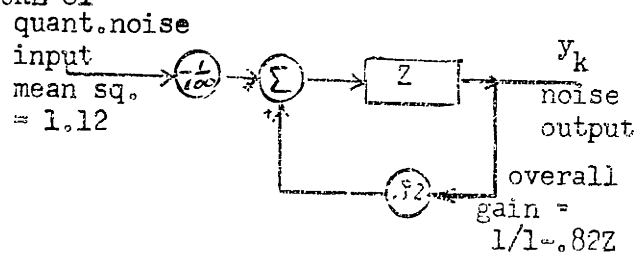
Point-By-Point Calculation of the Solution to (3.9)

FIGURE 61



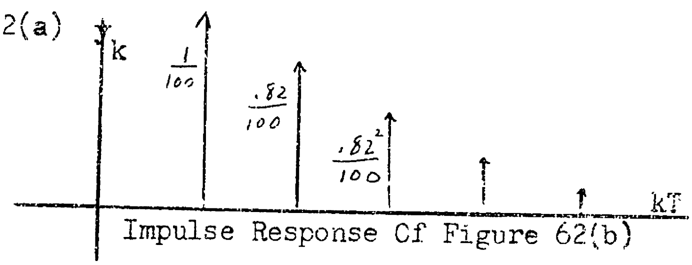
Replacement of Quantizer With A Noisy Source

FIGURE 62(a)



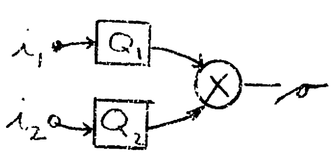
A Reduction Of Figure 62(a)

FIGURE 62(b)



Impulse Response Of Figure 62(b)

FIGURE 63



Formation Of The Product Of Two Quantized Variables

FIGURE 64

This, too, depends only upon the moments of i.

A device that multiplies two input variables i_1 and i_2 may be described by $o = i_1 i_2$.

The kth output moment is $\overline{o^k} = \overline{i_1^k i_2^k}$.

A general in-out moment is $\overline{o^k i_1^l} = \overline{i_1^{k+l} i_2^k}$

Both output and joint moments depend only on the joint moments between i_1 and i_2. If two quantized variables are multiplied as in Figure 64 there will be no change in the moment picture when the quantizers are replaced by independent added quantization noises as long as the high order joint distribution between i_1 and i_2 is wide enough along the i_1 and i_2 dimensions to satisfy the joint Nyquist restriction. In c.f. space, there will be periodicities in the joint c.f. along the ξ_1 dimensions of radian fineness ϕ_1, and periodicities of radian fineness ϕ_2 along the ξ_2 dimensions. ϕ_1 and ϕ_2 correspond to the quantization frequencies of Q_1 and Q_2, while ξ_1 and ξ_2 are the variables transformed from i_1 and i_2. No overlap means that the Nyquist restriction is satisfied.

C. Analysis Of Continuous and Continuous-Sampled Quantizer Systems.

A general discussion of how and where these techniques may be applied to continuous and part continuous sampled systems can best be induced from a study of a specific example.

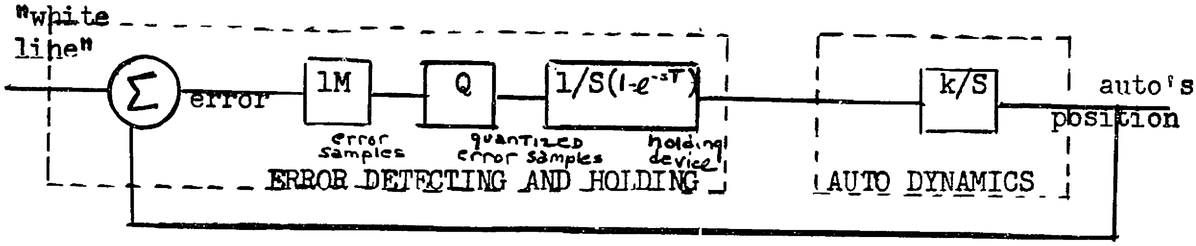
A crude and possibly cheap automatic guidance system for the steering of an automobile may be realized if the power source or controller is only required to produce several uniformly-spaced (angle-wise) headings. These headings are held for fixed intervals, whereupon they may be changed if necessary. The automobile is tracking an invisible "white line" painted on the highway, and is always trying to keep the error small, the error being the difference between the automobile position and the white line. Figure 65 is a system diagram including controller, the automobile dynamics, and the error detector. The problem is to find the effect of quantization on system performance and to estimate the largest tolerable quantization box size and thereby arrive at the required number of steering angles.

The system of Figure 65 is identical to one described by Linvill[1] expect for the quantizer. The holding device is shown as being linear and capable of holding a continuum of levels. Actually, in this application it holds only the discrete levels corresponding to quantizer outputs (steering wheel positions) and is a discrete-level device. The auto dynamics given by k/s is exact for small deflections. By conventional system reduction techniques, Figure 65 may be reduced to a fairly simple form. Figure 66(a), (b), and (c) show several stages of this development. Full justification may be found in Reference 5. The system of Figure 66(c) with the quantizer replaced by a box of gain unity is "perfectly" compensated when k is adjusted, so that $kT = 1$. Error sensed at a given sample time is brought to zero by the next sample time. Let k be adjusted as such in the quantizer system. This is a reasonable procedure, but it may not be precisely the best. If in addition, the substitution $e^{-sT} = Z$ is made, Figure 66(c) becomes Figure 67.

The Nyquist test upon the probability density at the quantizer input is easily made by cutting the feedback link (Figure 67). No matter what statistics are assumed for the system input (highway characteristics), the Nyquist restriction will be practically always satisfied. Recall that this condition was met for Gaussian inputs quantized with granularity as large as two standard deviations. In any event, there will always be a lot of arbitrariness in the assumption of input statistics so it is perfectly reasonable to replace the quantizer by a source of independent quantization noise. This is done in Figure 68(a), and a further reduction is Figure 68(b). Figure 68(b) is an accurate picture of the control system. A reasonable sampling rate might be once per second at 6.8 MPH, or a sampling every 10 feet. On this basis, the speed of the vehicle does not enter into the problem.

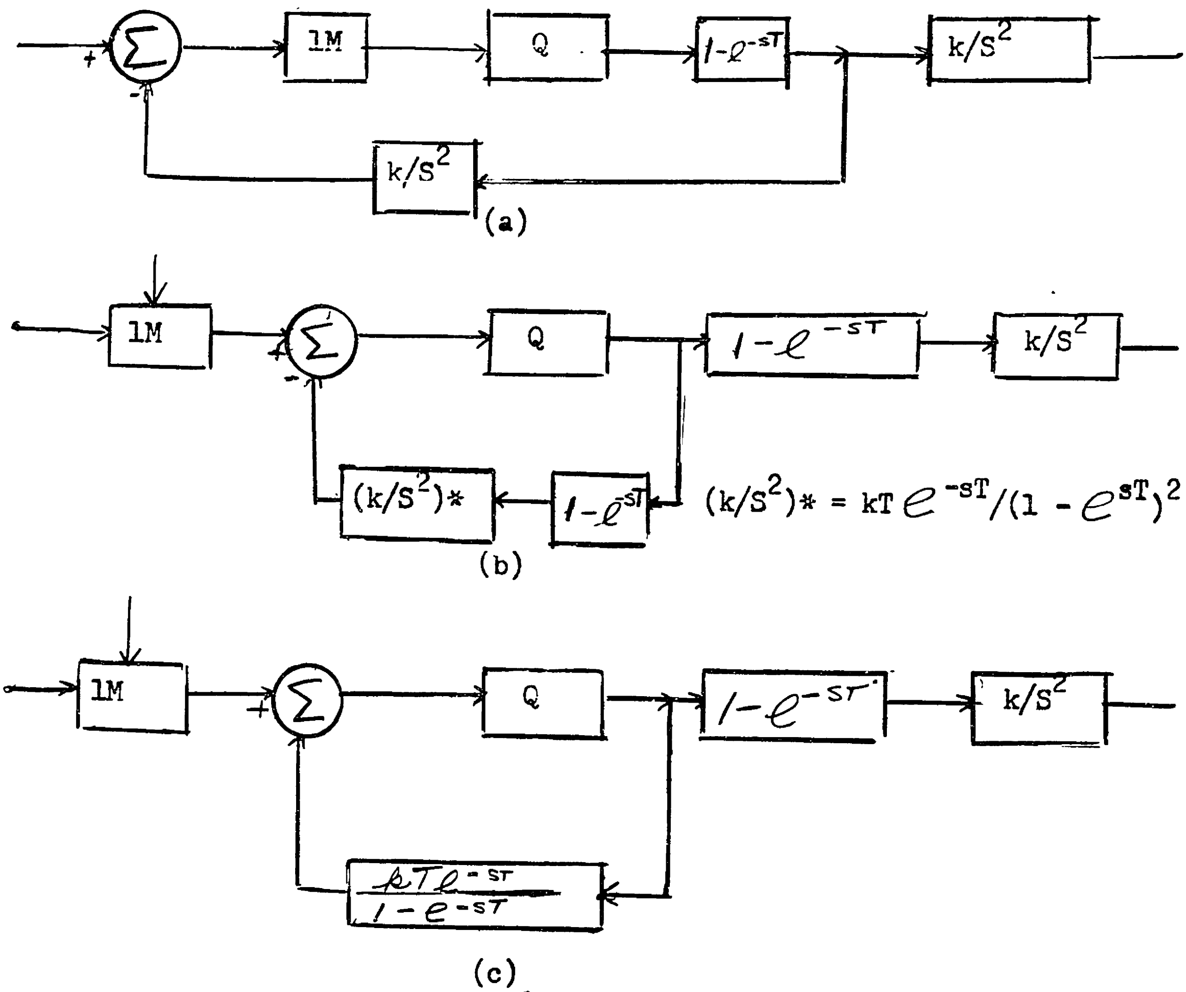
Instead of attempting analysis of the actual system output, consider only its samples. Formally, this is done by following the system with an impulse modulator. The analysis is greatly simplified, particularly if the second impulse modulator is synchronized with the first (Figure 69). Figure 69 exactly reduces to the very simple system of Figure 70.

Figure 70 shows that, except for the quantization noise, the samples of the output are identical with the samples of the input



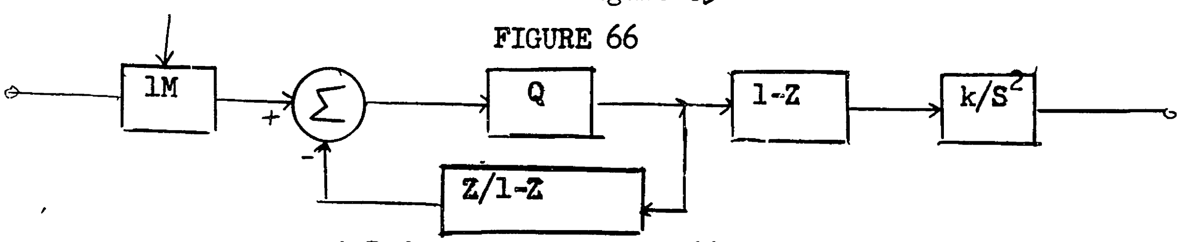
Automatic Guidance of An Automobile

FIGURE 65



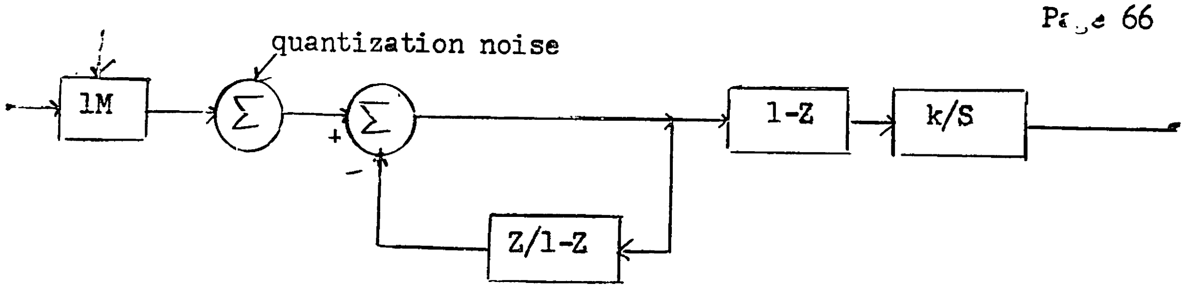
Reduction Of Figure 65

FIGURE 66



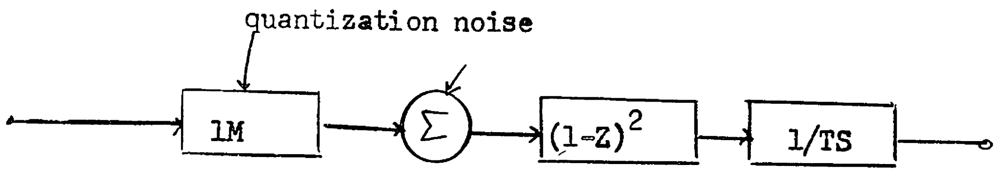
A Reduction of Figure 66

FIGURE 67



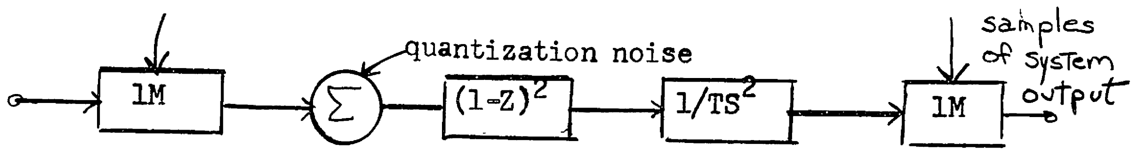
Reduction of Figure 67: Quantizer Exchanged For A Noisy Source

FIGURE 68(a)



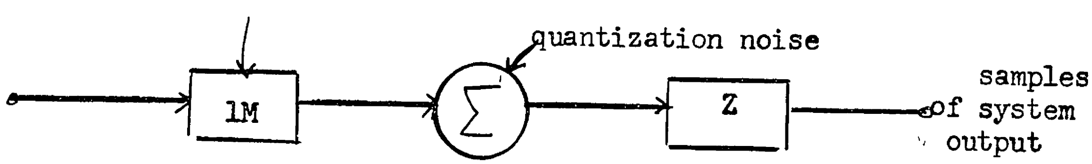
Reduction of Figure 68(a)

FIGURE 68(b)



Sampling The Output of Figure 68

FIGURE 69



A Reduction of Figure 69

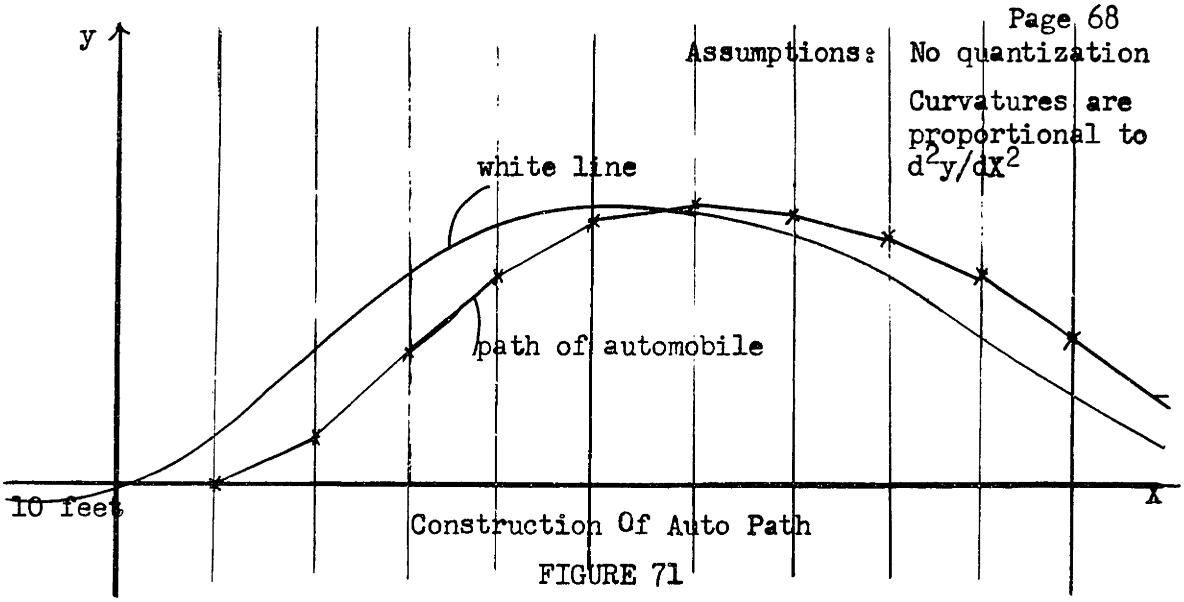
FIGURE 70

delayed by one sampling interval. Also, the path of the automobile is the same as the path of the white line shifted by the sampling interval, 10 feet, except for the quantization noise, and provided that the conventional Nyquist restriction on the bandwidth of the input is satisfied. See Figure 71.

It is now possible to attempt the question, how large may the quantization box be? Assume that the highway will never curve with a smaller radius than 25 feet, and that the automobile must always be guided to within ± 3 feet of the white line, Figure 72 shows a straight line path that suddenly takes on a curvature of 25 foot radius. The most adverse condition is that the last sample is taken just before the highway curves, as shown in the sketch. In one sampling interval, without quantization, a lateral error of two feet will develop. The quantization box size must therefore be less than 1 foot which, when added to the 2 foot "dynamic" error, will yield a net error within the 3 foot limit.

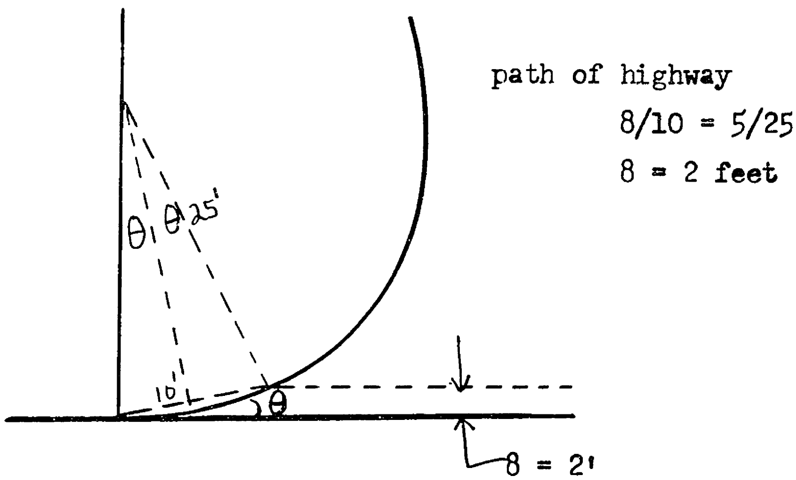
Only three quantization boxes would be necessary if the box size were set at 2 feet. One left, one right, and the third straight ahead. If the left state is selected, the system will turn left to give a lateral correction every sampling interval. This turns out to be just enough to allow turning within a 25 foot radius.

Such a control system would no doubt give a rough ride at any apprieciable speed, and would place considerable strain upon the automobile. A quantitative way of determining the roughness of a ride is necessary in order to decide how much less than 2 feet the quantization box should be made. An arbitrary but reasonable procedure is to define a "roughness" of path in terms of the mean square of the lateral velocity. The first difference of the output samples is an approximate measure of the samples of lateral velocity. The roughness is then the mean square of the output of the system shown in Figure 73. Quantization with 2-foot boxes adds a roughness of $1/12\ q^2(2) = 1/12(4)(2) = 2/3\ ft^2$.
This may be compared with the roughness of a sine wave of amplitude A and a period of 360 feet. The output pulses are samples of $A \sin\alpha - A \sin(\alpha - 10^\circ) \cong 0.17\ A \sin(\alpha - 5^\circ)$, giving a roughness of $(0.17)^2 A^2/2\ ft^2$. Quantization as indicated above introduces as much roughness as would the addition of a sine wave of amplitude A = 6.8 feet and period of 360 feet. At a speed of 60 MPH, one cycle of the sine wave



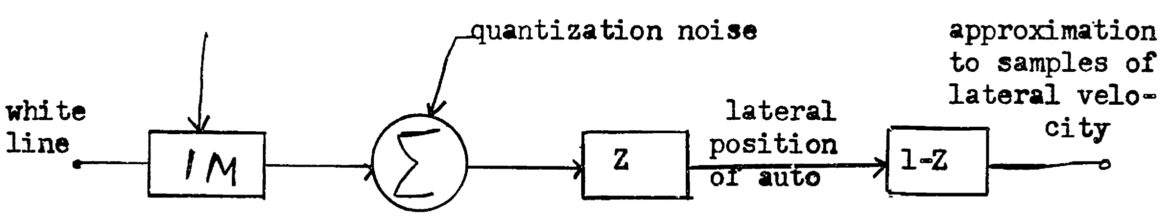
Construction Of Auto Path

FIGURE 71



A Sudden Sharp Highway Turn

FIGURE 72

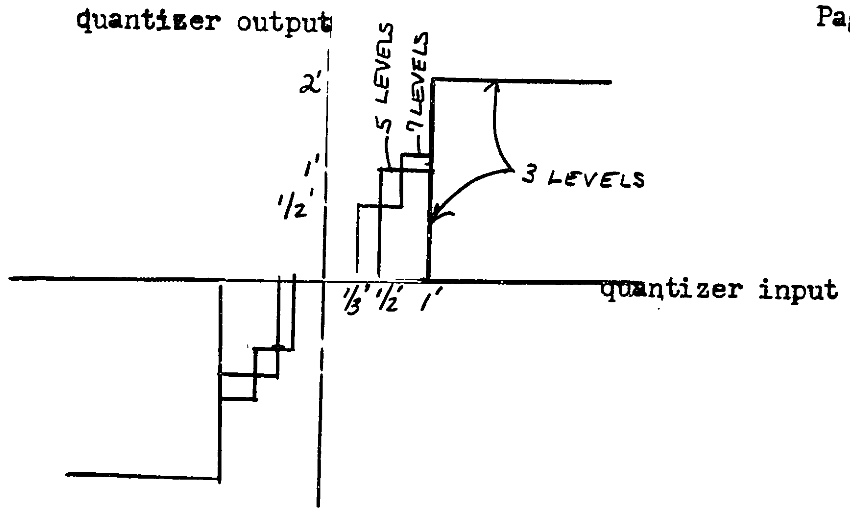


Derivation of Lateral Velocity

FIGURE 73

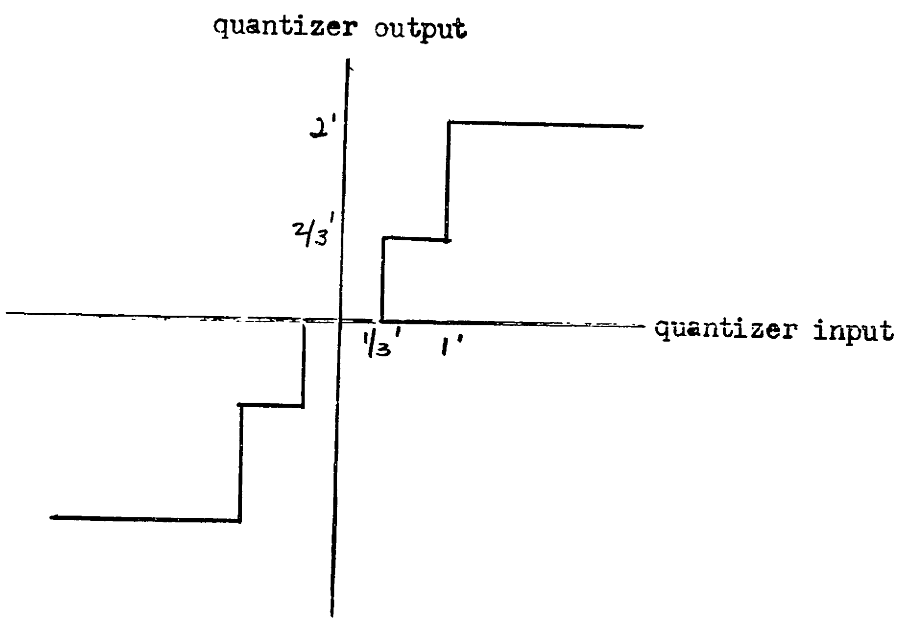
will be traversed in 4 seconds.

Instead of quantizing with 2 foot boxes, the same dynamic range could be covered by the addition of two more quantization levels with 1 foot quantization boxes. The addition of 2 more levels allows quantization boxes having widths of 2/3 feet (see Figure 74). The amplitudes of sine waves having equal roughness come down in the same proportion. 7-level quantization is probably fine enough so that its effects are negligible. Almost the same fineness of performance could be obtained from the quantizer described by Figure 75, at least for sections of highway having radius of curvature greater than 3(25) = 75 feet. Sharper curves will have quantization roughness corresponding to something between 5 and 3 levels, but never worse than 3 levels. Sharp curves present no serious problem, being relatively rare and always executed at low speeds. Roughness here will be small compared to the unavoidable roughness of the curve itself. The quantizer of Figure 75 is probably a good engineering device.



3, 5, and 7 Level Quantizers Covering Same
Dynamic Range

FIGURE 74



"Non-Linear" Quantizer

FIGURE 75

APPENDIX I

SHEPPARD'S CORRECTIONS FOR GROUPING

Quantization is known to statisticians as "grouping". Most of their statistical data is gathered by noting how many trials fall in each of a predetermined set of groups, giving either an impulse distribution or a histogram. Sheppard's corrections provide a means of modifying the moments of a given impulse distribution in order to arrive very much closer to the true moments of the unavailable ungrouped distribution. Sheppard first proposed his corrections in the Proceedings of the London Mathematical Society in 1898. His article was entitled "On the Calculation of the most Probable Values of Frequency-Constants, for Data arranged according to Equidistant Divisions of a Scale."

Before considering Sheppard's derivation, the corrections will be evaluated by assuming the quantization to be the same as the addition of an independent noise of distribution Q(n) as in Figure 16. It was shown that as far as moments are concerned, the above substitution is permissible when the Nyquist restriction is met.

Let the tth moment of the grouped data be μ_t, and let the tth moment of the true smooth distribution be m_t. Let the noise of distribution Q(n) be represented by n. The tth grouped moment is then given by equation (I.1). Since n is independent of X, the averages of (I.1)

$$\mu_t = \overline{(X+n)^t} = \overline{X^t} + \overline{tX^{t-1}n} + t(t-1)/2! \ \overline{X^{t-2}n^2} + \ldots + \overline{n^t} \qquad (I.1)$$

may be expressed as in equation (I.2). Note that all odd moments of Q(n) are zero because Q(n) is an even function.

$$\begin{aligned}\mu_t &= \overline{X^t} + t(t-1)/2! \ \overline{X^{t-2}}\ \overline{n^2} + t(t-1)(t-2)(t-3)/4! \ \overline{X^{t-4}}\ \overline{n^4} + \ldots \\ &= \overline{m_t} + t(t-1)/2! \ \overline{m_{t-2}}\ \overline{n^2} + t(t-1)(t-2)(t-3)/4! \ \overline{m_{t-4}}\ \overline{n^4} + \ldots\end{aligned} \qquad (I.2)$$

Notice that μ_t may be expressed in terms of some of the moments of X and n up to and possibly including the tth. The first four moments of the grouped distribution are as follows.

$$\begin{aligned}\mu_1 &= m_1 & \qquad \mu_3 &= m_3 + 3m_1\overline{n^2} \\ \mu_2 &= m_2 + \overline{n^2} & \qquad \mu_4 &= m_4 + 6m_2\overline{n^2} + \overline{n^4}\end{aligned} \qquad (I.3)$$

The moments of the rectangular distribution Q(n) are easily gotten to be

$$\overline{n^2} = 1/12\ q^2$$
$$\overline{n^4} = 1/80\ q^4 \qquad (I.4)$$

The equations (I.3) become (I.5).

$$\mu_1 = m_1$$
$$\mu_2 = m_2 + 1/12\ q^2$$
$$\mu_3 = m_3 + 1/4\ m_1 q^2 \qquad = m_3 + 1/4\ \mu_1 q^2$$
$$\mu_4 = m_4 + 1/2\ m_2 q^2 + 1/80\ q^4 = m_4 + 1/2\ \mu_2 q^2 - 7/240\ q^4 \qquad (I.5)$$

The moments of the exact distribution m_1, m_2, m_3, m_4 are given in terms of μ_1, μ_2, μ_3, μ_4 by equation (I.6).

$$m_1 = \mu$$
$$m_2 = \mu_2 - 1/12\ q^2$$
$$m_3 = \mu_3 - 1/4\ \mu_1 q^2$$
$$m_4 = \mu_4 - 1/2\ \mu_2 q^2 + 7/240 q^4 \qquad (I.6)$$

The Sheppard corrections are 0, $(1/12\ q^2)$, $(1/4\ \mu_1 q^2)$, $(1/2\ \mu_2 q^2 - 7/240\ q^4)$ for the first, second, third, and fourth moments respectively.

Dr. Sheppard's approach differs considerably. What follows is an abstraction of his derivation.

Let y = f(X) be a continuous probability density having area unity. Using the same notation, $m^t = \int_{-\infty}^{\infty} X^t\ f(X)\ dX$ and $\mu_t \equiv \sum_{a}^{b} X_p{}^t A_p$ where $A_p \equiv \int_{X_p - q/2}^{X_p + q/2} f(X)\ dX$. A_p is the area on the base $X_p \pm q/2$, p being integral, and (b-a) is the full range of the variable X.

$$A_p = \int_{-q/2}^{q/2} f(X_p + X)\ dX$$
$$= \int_{-q/2}^{q/2} f(X_p) + X\ f'(X_p) + X^2/2!\ \ f''(X_p) + \dots \ \ dX \qquad (I.7)$$
$$= \qquad q\ f(X_p) + q^3/24\ f''(X_p) + q^5/1920\ f''''(X_p) + \dots$$

μ_t is then given by (I.8).

$$\mu_t = \sum_a^b q\, X_p^{\,t}\, f(X_p) + \sum_a^b q^3/24\, X_p^{\,t}\, f''(X_p) + \sum_a^b q^5/1920\, f''''(X_p) \ldots \qquad (I.8)$$

The Euler-Maclaurin theorem, (I.9) which connects summation with integration, may be used to express the sums in terms of the true moments.

$$q \sum_a^b f(X_p) = \int_a^b f(X)dX + q/2 \left[f(b) + f(a) \right] + q^2/12 \left[f'(b) - f'(a) \right] + \ldots$$

Let f(X) be analytic over the full dynamic range of X, and let its value and the values of its derivatives be zero at the points a and b, the extremes of this range.

For all the substitutions of (I.9) into (I.8), the Euler-Maclaurin theorem becomes the very obvious equation (I.10).

$$q \sum_a^b f(X_p) = \int_a^b f(X)dX \qquad (I.10)$$

Substitution of (I.10) into (I.8) yields μ_t.

$$\mu_t = \int_a^b X^t f(X)dX + q^2/24 \int_a^b X^t f''(X)dX + q^4/1920 \int_a^b X^t f''''(X)dX + \ldots\ldots$$

$$= m_t + q^2/24 \int_a^b X^t f''(X)dX + q^4/1920 \int_a^b X^t f''''(X)dX + \ldots\ldots \qquad (I.11)$$

By continual integration by parts and applying the conditions at the extremities

$$\int_a^b X^t f''(X)dX = \Big| _a^b X^t f'(X) - t \int_a^b X^{t-1} f'(X)dX = t(t-1)\, m_{t-2} \qquad \text{and}$$

$$\int_a^b X^t f''''(X)dX = t(t-1)(t-2)(t-3)\, m_{t-4}$$

(I.11) then becomes

$$\mu_t = m_t + q^2/24\; t(t-1)\, m_{t-2} + q^4/1920\;\; t(t-1)(t-2)(t-3)\, m_{t-4} + \ldots \qquad (I.12)$$

Equation (I.12) is identical with equation (I.2) when the moments of quantization noise (I.4) are inserted. The remainder of Dr. Sheppard's derivation proceeds in the manner of equations (I.3) through (I.6).

APPENDIX II

THE PROPAGATION OF RANDOM STATIONARY SIGNALS THROUGH LINEAR SAMPLED-DATA SYSTEMS

A. First Order Inputs

(1) First Order Filters

A first order filter, having a single memory state, is shown in Figure II-1. This sampled-data filter has a gain 1 + az, which indicates that the present output equals the present input plus (a) times the previous input. A way of showing present and past inputs and outputs for the purpose of calculation of their joint probability density is Figure II-2. The input signal is stepped at each sampling time from input node to input node, entering at the "present" node at the top and proceeding downward. These nodes make up an "analogue stepping register." The filter input is first order, so that all the input nodes of Figure II-2 carry statistically independent signals at each sample time.

The cause of the statistical dependence among the output samples is clearly seen in Figure II-2. O_1 and O_2 have a common "source of supply" in the node X_2. Outputs O_1 and O_3, on the other hand, have no common supply node, being driven by statistically independent sources and are, therefore, statistically independent. The output is a second order process described by $W(O_1,O_2)$. The order of an output is always the sum of the order of the input statistics plus the order or number of memory states of the filter.

The two-dimensional characteristic function of the output signal is given by equation II.1.

$$F_{O_1,O_2}(\xi_1\xi_2) = \iint_{-\infty}^{\infty} W(O_1,O_2)\, e^{j(O_1\xi_1 + O_2\xi_2)}\, dO_1 dO_2 \qquad \text{(II.1)}$$

Formally, this is very similar to the characteristic function of the sum of O_1 and O_2, given by equation II.2. The two are identical for $\xi_1 = \xi_2 = \xi$.

$$F_{O_1+O_2}(\xi) = \iint_{-\infty}^{\infty} W(O_1,O_2)\, e^{j(O_1\xi + O_2\xi)}\, dO_1 dO_2 \qquad \text{(II.2)}$$

Thus the characteristic function of the sum of O_1 and O_2 allows calculation of their joint characteristic function, along a 45° cut in the ξ_1, ξ_2 plane. The entire $F_{O_1 O_2}(\xi_1, \xi_2)$ surface may be explored by finding the statistics of sums of O_1 and O_2 when multiplied by the constants k_1 and k_2 respectively. A flow graph of this process is given in Figure II-3). Notice that the arrows have been omitted for clarity. All links are unilateral with signals flowing from left to right.

From Figure II-3, it may be seen that $\sum$ is the sum (equation II.3) of three statistically independent signals.

$$k_1 O_1 + k_2 O_2 = \sum = k_1 X_1 + (ak_1 + k_2) X_2 + ak_2 X_3 \qquad \text{(II.3)}$$

The characteristic function of $\sum$ is equation (II.4), the product of three characteristic functions.

$$F_{\Sigma}(\xi) = F_x(k_1 \xi) F_x\left[(ak_1 + k_2)\xi\right] F_x(ak_2 \xi) \qquad \text{(II.4)}$$

The characteristic function of $\sum$ may also be expressed in terms of $F_{O_1 O_2}(\xi_1, \xi_2)$.

$$F_{\Sigma}(\ ) = \iint_{-\infty}^{\infty} W(O_1, O_2)\, e^{j(k_1 O_1 \xi + k_2 O_2 \xi)}\, dO_1 dO_2 \qquad \text{(II.5)}$$

Equation (II.5) is identical with equation (II.1) if the substitutions (II.6) are made.

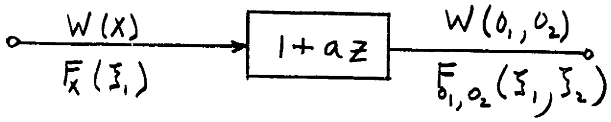
$$k_1 \xi = \xi_1$$
$$k_2 \xi = \xi_2 \qquad \text{(II.6)}$$

Therefore the second order output characteristic function is

$$F_{O_1, O_2}(\xi_1, \xi_2) = F_x(\xi_1) F_x(a\xi_1 + \xi_2) F_x(a\xi_2) \qquad \text{(II.7)}$$

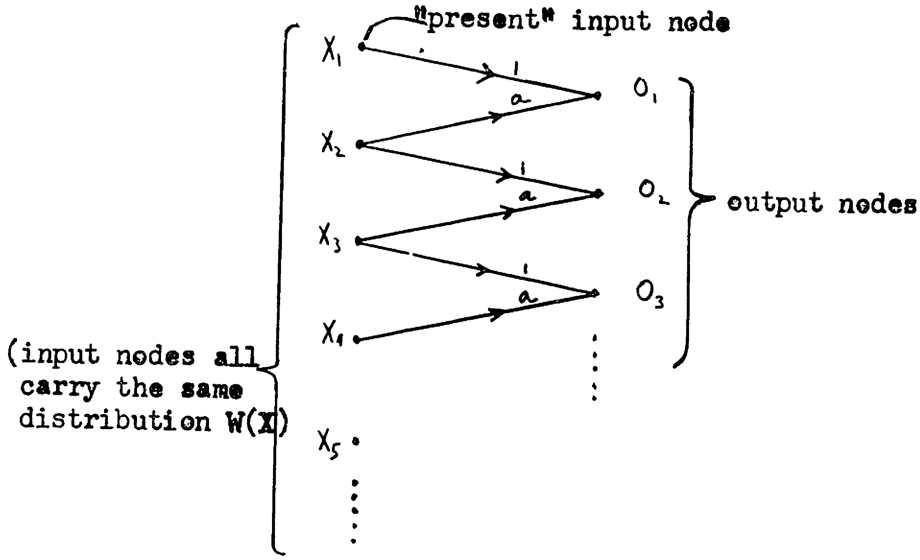
The output characteristic function and by inverse transformation $W(O_1, O_2)$, are completely determined by the characteristic function of the input signal $F_x(\xi)$ when the substitution into equation (II.7) is made. As an example, let the input be Gaussian with the characteristic function

$$F_x(\xi) = e^{\frac{-\xi^2 \sigma^2}{2}}$$



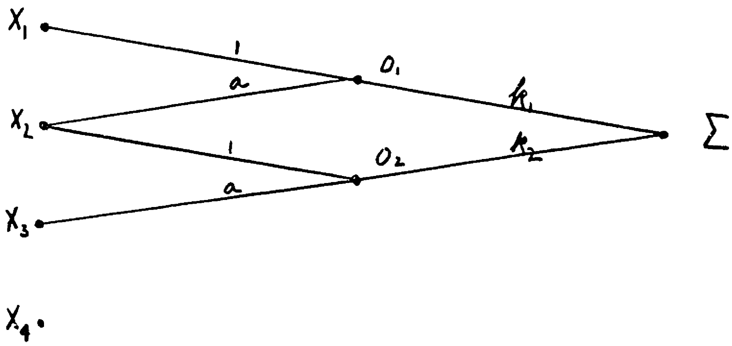
A First Order Filter Excited By A First Order Random Input

FIGURE II-1



Sequential Flow Graph Of The System Of Figure II-1

FIGURE II-2



Construction For The Calculation Of $F_{O_1,O_2}(\xi_1,\xi_2)$

FIGURE II-3

The second order output is, accordingly,

$$F_{0_1,0_2}(\xi_1,\xi_2) = e^{\frac{-\xi_1^2\sigma^2}{2}} e^{\frac{-(a^2\xi_1+\xi_2)^2\sigma^2}{2}} e^{\frac{-(a\xi_2)^2\sigma^2}{2}}$$

$$= e^{\frac{-\xi_1^2\sigma^2}{2}} e^{\frac{-(a^2\xi_1^2+\xi_2^2+2a\xi_1\xi_2)\sigma^2}{2}} e^{\frac{-a^2\xi_2^2\sigma^2}{2}}$$

$$= e^{\frac{-(1+a^2)(\xi_1^2+\xi_2^2)+2a\xi_1\xi_2\sigma^2}{2}}$$

Thus the output is a second order process having a variance of $(1+a^2)\sigma^2$ and a covariance of $a\sigma^2$.

Inspection of Figure II-3 shows that the two related variables, 0_1 and 0_2, are only related to the input variables X_1, X_2, and X_3. Therefore, the most general joint input-output characteristic function includes these five variables. It is necessary at this point to adopt a new, nore complete notation. X_1, X_2, and X_3 transform into ξ_{1a}, ξ_{2a} and ξ_{3a}, while 0_1 and 0_2 transform into ξ_{1b}, and ξ_{2b}. The desired joint characteristic function $F_{x_1,x_2,x_3,0_1,0_2}(\xi_{1a}, \xi_{2a}, \xi_{3a}, \xi_{1b}, \xi_{2b})$ will be derived like $F_{0_1,0_2}(\xi_1, \xi_2)$ from linear combinations of X_1, X_2, X_3, 0_1, and 0_2.

The formation of these linear combinations is shown in Figure II-4. Here,

$$\Sigma = k_{1a}X_1 + k_{2a}X_2 + k_{3a}X_3 + k_{1b}0_1 + k_{2b}0_2$$

Σ may be expressed in terms of the three independent variables as in equation (II.8).

$$\Sigma = X_1(k_{1a} + k_{1b}) + X_2(ak_{1b} + k_{2a} + k_{2b}) + X_3(ak_{2b} + k_{3a}) \quad (II.8)$$

The characteristic function of Σ is

$$F_\Sigma(\xi) = F_x(k_{1a}\xi + k_{1b}\xi)F_x(ak_{1b}\xi + k_{2a}\xi + k_{2b}\xi)F_x(ak_{2b}\xi + k_{3a}\xi) \quad (II.9)$$

Inserting the substitutions (II.10) into equation (II.9), the desired joint characteristic function appears as equation (II.11).

$$\begin{aligned} k_{1a}\xi &= \xi_{1a} \\ k_{2a}\xi &= \xi_{2a} \\ k_{3a}\xi &= \xi_{3a} \\ k_{1b}\xi &= \xi_{1b} \\ k_{2b}\xi &= \xi_{2b} \end{aligned} \qquad \text{(II.10)}$$

$$\begin{aligned} &F_{x_1,x_2,x_3,0_1,0_2}(\xi_{1a}, \xi_{2a}, \xi_{3a}, \xi_{1b}, \xi_{2b}) \\ &\quad = F_x(\xi_{1a} + \xi_{1b})F_x(a\xi_{1b} + \xi_{2a} + \xi_{2b})F_x(a\xi_{2b} + \xi_{3a}) \end{aligned} \qquad \text{(II.11)}$$

Equation (II.11) is the most general statistical description of the physical situation shown in Figure II.1. The joint output characteristic function of (II.7) may be derived from it by eliminating the input variables, i.e., $\xi_{1a} = \xi_{2a} = \xi_{3a} = 0$, and be reverting to the old simpler notation $\xi_{1b} = \xi$, and $\xi_{2b} = \xi_2$.

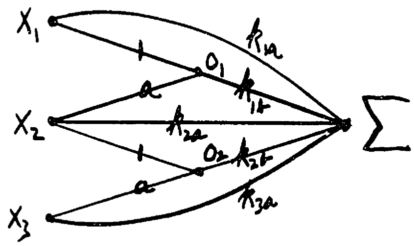
(2) Second Order Filters

A second order filter driven by a first order random input is shown in Figure II-5. The flow graph corresponding to the one of Figure II-2 is Figure II-6. Notice that the output is third order. Not only do 0_1 and 0_2 have common sources of supply in X_2 and X_3, but 0_1 and 0_3 also have X_3 in common. There is no statistical connection between 0_4 (not shown) and 0_1. The characteristic function of the output $F_{0_1,0_2,0_3}(\xi_{1b}, \xi_{2b}, \xi_{3b})$ will be derived before the joint input-output characteristic function.

A flow graph showing the formation of a $\sum$ appropriate for the calculation of $F_{0_1,0_2,0_3}(\xi_{1b}, \xi_{2b}, \xi_{3b})$ is given in Figure II-7.

$$\sum = k_{1b}0_1 + k_{2b}0_2 + k_{3b}0_3$$

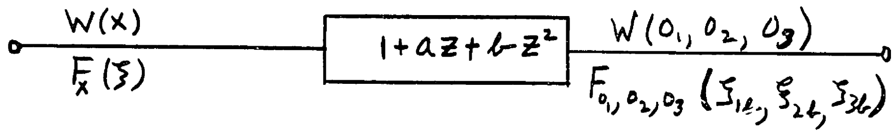
may be expressed in terms of independent input variables as equation (II.12) from inspection of Figure II-7.



Construction For The Calculation Of

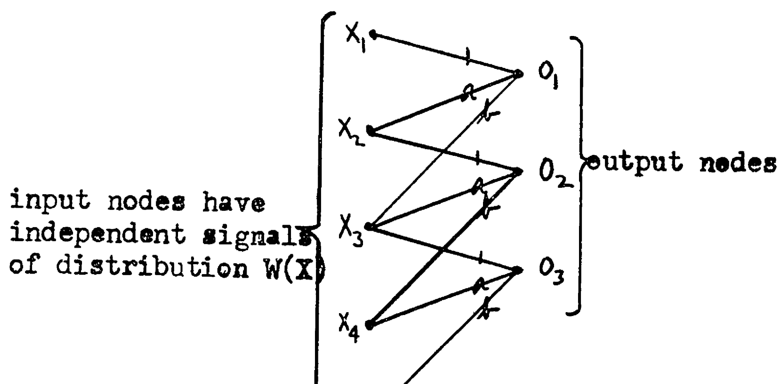
$F_{x_1 x_2 x_3 O_1 O_2}(\xi_{1a}, \xi_{2a}, \xi_{3a}, \xi_{1b}, \xi_{2b})$

FIGURE II-4



A Second Order Filter Excited By A First Order Random Input

FIGURE II-5



Sequential Flow Graph Of The System Of Figure II-5

FIGURE II-6

$$\sum = X_1(k_{1b}) + X_2(ak_{1b} + k_{2b}) + X_3(bk_{1b} + ak_{2b} + k_{3b})$$
$$+ X_4(bk_{2b} + ak_{3b}) + X_5(bk_{3b}) \qquad (II.12)$$

The characteristic function of $\sum$ is

$$F_{\Sigma}(\xi) = F_x(k_{1b}\xi)F_x(a\xi_{1b} + k_{2b}\xi)F_x(bk_{1b}\xi + ak_{2b}\xi + k_{3b}\xi)$$
$$F_x(bk_{2b}\xi + ak_{3b}\xi)F_x(bk_{3b}\xi) \qquad (II.13)$$

By substituting (II.14) into (II.13), the output characteristic function is derived.

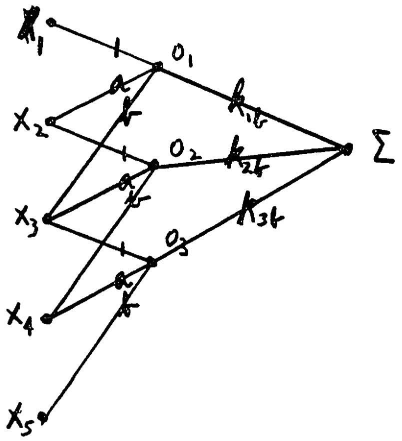
$$k_{1b}\xi = \xi_{1b}$$
$$k_{2b}\xi = \xi_{2b}$$
$$k_{3b}\xi = \xi_{3b} \qquad (II.14)$$

$$F_{O_1,O_2,O_3}(\xi_{1b}, \xi_{2b}, \xi_{3b}) = F_x(\xi_{1b})F_x(a\xi_{1b} + \xi_{2b})$$
$$F_x(b\xi_{1b} + a\xi_{2b} + \xi_{3b})F_x(b\xi_{2b} + a\xi_{3b})F_x(b\xi_{3b}) \qquad (II.15)$$

The flow graph showing the formation of a $\sum$ for the derivation of the most general joint characteristic function $F_{x_1,x_2,x_3,x_4,x_5,O_1,O_2,O_3}(\xi_{1a}\xi_{2a}\xi_{3a}\xi_{4a}\xi_{5a}\xi_{1b}\xi_{2b}\xi_{3b})$ is in Figure II-8. Commas separating the variables will be omitted for this many variables.

By inspection of Figure II-8, the characteristic function of $\sum$ is

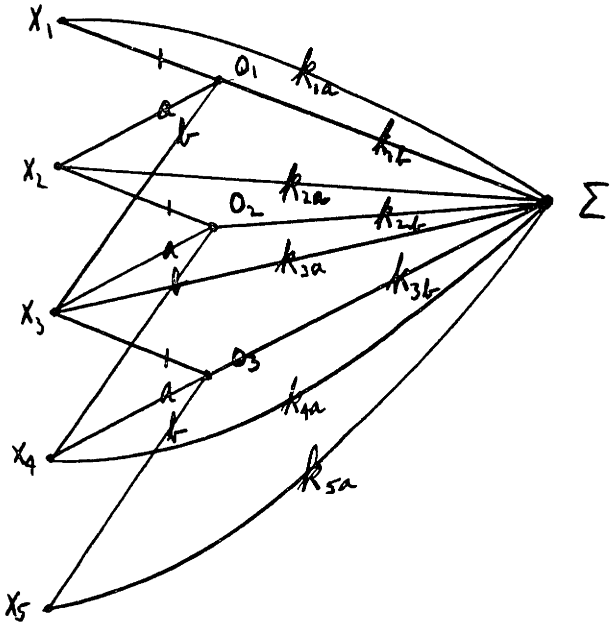
$$F_{\Sigma}(\xi) = F_x(k_{1a}\xi + k_{1b}\xi)F_x(ak_{1b}\xi + k_{2a}\xi + k_{2b}\xi)$$
$$F_x(bk_{1b}\xi + ak_{2b}\xi + k_{3a}\xi + k_{3b}\xi)F_x(bk_{2b}\xi + ak_{3b}\xi + k_{4a}\xi)$$
$$F_x(bk_{3b}\xi + k_{5a}\xi) \qquad (II.16)$$



Construction For Calculation Of

F_{O_1,O_2,O_3}(1b, 2b, 3b)

FIGURE II-7



Construction For The Calculation Of

$F_{x_1x_2x_3x_4x_5O_1O_2O_3}$(1a 2a 3a 4a 5a 1b 2b 3b)

FIGURE II-8

Substitution of (II.17) into (II.16) yields the desired result (II.18).

$$k_{1a}\xi = \xi_{1a} \qquad k_{1b} = \xi_{1b}$$
$$k_{2a}\xi = \xi_{2a} \qquad k_{2b} = \xi_{2b}$$
$$k_{3a}\xi = \xi_{3a} \qquad k_{3b} = \xi_{3b}$$
$$k_{4a}\xi = \xi_{4a}$$
$$k_{5a}\xi = \xi_{5a} \qquad \text{(II.17)}$$

$$F_{x_1x_2x_3x_4x_5O_1O_2O_3}(\xi_{1a}\xi_{2a}\xi_{3a}\xi_{4a}\xi_{5a}\xi_{1b}\xi_{2b}\xi_{3b}) = F_x(\xi_{1a}+\xi_{1b})$$
$$\cdot F_x(a\xi_{1b}+\xi_{2a}+\xi_{2b})F_x(b\xi_{1b}+a\xi_{2b}+\xi_{3a}+\xi_{3b})$$
$$\cdot F_x(b\xi_{2b}+a\xi_{3b}+\xi_{4a})F_x(b\xi_{3b}+\xi_{5a}) \qquad \text{(II.18)}$$

(3) Higher Order Filters

It is now possible to induce the forms that the output and joint input-output characteristic function will take when a filter of degree n is driven by a first order random input as in Figure II-9.

The general output characteristic function is equation (II.19), containing 2n + 1 factors.

$$F_{O_1\ldots O_{n+1}}(\xi_{1b}\ldots\xi_{n+1b}) = F_x(\xi_{1b})F_x(a\xi_{1b}+\xi_{2b})\ldots\ldots$$
$$\ldots\ldots F_x(p\xi_{1b}+\ldots+a\xi_{n-1b}+\xi_{nb})F_x(q\xi_{1b}+p\xi_{2b}+\ldots\ldots$$
$$\ldots\ldots+a\xi_{nb}+\xi_{n+1b})F_x(q\xi_{2b}+p\xi_{3b}+\ldots\ldots+a\xi_{n+1b})\ldots$$
$$\ldots\ldots F_x(q\xi_{nb}+p\xi_{n+1b})F_x(q\xi_{n+1b}) \qquad \text{(II.19)}$$

The general joint characteristic function is equation (II.20), consisting of 2n + 1 factors.

$$F_{x_1 x_2 \cdots x_{2n+1} 0_1 0_2 \cdots 0_{n+1}}(\xi_{1a} \cdots \xi_{2n+1a}\, \xi_{1b} \cdots \xi_{n+1b})$$

$$= F_x(\xi_{1a} + \xi_{1b}) F_x(\xi_{2a} + a\xi_{1b} + \xi_{2b}) \cdots\cdots$$

$$\cdots\cdots F_x(\xi_{na} + p\xi_{1b} + \cdots + a\xi_{n-1b} + {}_{nb})$$

$$F_x(\xi_{n+1a} + q\xi_{1b} + p\xi_{2b} + \cdots + a\xi_{nb} + \xi_{n+1b})$$

$$F_x(\xi_{n+2a} + q\xi_{2b} + p\xi_{3b} + \cdots + a\xi_{n+1b}) \cdots\cdots$$

$$\cdots\cdots F_x(\xi_{2na} + q\xi_{nb} + p\xi_{n+1b}) F_x(\xi_{2n+1a} + q\xi_{n+1b}) \qquad \text{(II.20)}$$

(II.20) is a complete statistical description of the propagation of first order random signals through linear sampled-data systems of nth order.

B. Second Order Inputs

(1) First Order Filters

A first order filter driven by a second order input is shown in Figure II.10. The flow graph that is useful in the determination of the characteristic function of the filter output is Figure II.11. Again, the output characteristic function is determined via the characteristic functions of linear combinations of output samples. The output signal is a third order process, where three successive output nodes are driven by four successive input nodes. Difficulty arises here because these four sources are not statistically independent as they would be if the input were first order. These four sources are more clearly shown in the modified flow graph of Figure II.12.

Although the input is a second order process, it may be described for convenience as a degenerate fourth order process having a characteristic function $F_{x_1x_2x_3x_4}\left[\xi_{1a}\ \xi_{2a}\ \xi_{3a}\ \xi_{4a}\right]$. Thus, the characteristic function of the variable in Figure II.12 just before they are combined is

$$F_{x_1x_2x_3x_4}\left[k_{1b}\xi_{1a},\ (ak_{1b}+k_{2b})\xi_{2a},\ (ak_{2b}+k_{3b})\xi_{3a}.\ ak_{3b}\xi_{4b}\right]$$

It follows that the characteristic function of $\sum$ is (II.21)

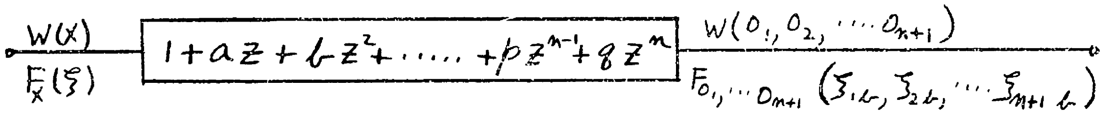
$$F_{\sum}(\xi) = F_{x_1x_2x_3x_4}\left[k_{1b}\xi\ ,\ (ak_{1b}+k_{2b})\xi,\ (ak_{2b}+k_{3b})\xi,\ ak_{3b}\xi\right] \tag{II.21}$$

Making the usual substitutions (II.22) into (II.21)

$$\begin{aligned} k_{1b}\xi &= \xi_{1b} \\ k_{2b}\xi &= \xi_{2b} \\ k_{3b}\xi &= \xi_{3b} \\ k_{4b}\xi &= \xi_{4b} \end{aligned} \tag{II.22}$$

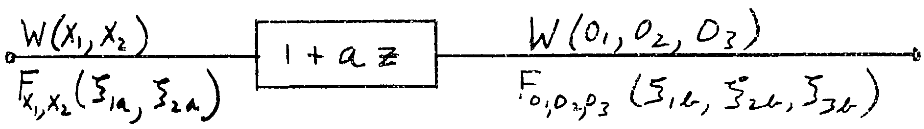
yields the desired result (II.23).

$$\therefore F_{0_10_20_3}(\xi_{1b},\xi_{2b},\xi_{3b}) = F_{x_1x_2x_3x_4}\left[\xi_{1b},(a\xi_{1b}+\xi_{2b}),\ldots\ldots \ldots\ldots\ldots (a\xi_{2b}+\xi_{3b}),\ a\xi_{3b}\right] \tag{II.23}$$



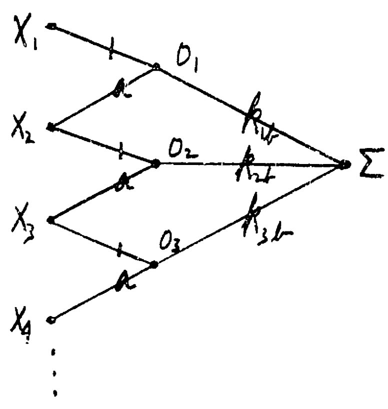
An nth Order Filter Excited By A First Order Random Input

FIGURE II-9



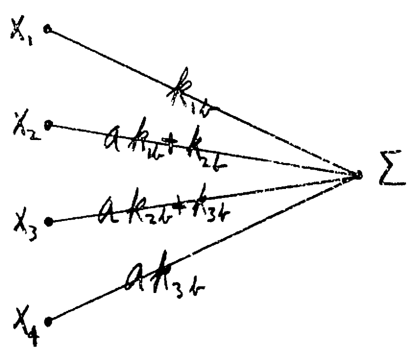
A First Order Filter Excited By A Second Order Random Input

FIGURE II-10



Construction for the Calculation of
$F_{O_1 O_2 O_3}($ 1b 2b 3b$)$

FIGURE II-11



Modification Of Figure II-11

FIGURE II-12

The general joint input-output characteristic function of which (II.23) is a special situation will be derived from the flow graph of Figure II.13. The input variables are introduced to by the links of gain k_{1a}, k_{2a}, k_{3a}, k_{4a} which are added to those of Figure II.12. The characteristic function of $\sum$ is (II.24).

$$F_{\Sigma}(\xi) = F_{x_1 x_2 x_3 x_4}\left[(k_{1b} + k_{1a})\xi,\ (ak_{1b} + k_{2b} + k_{2a})\xi,\ (ak_{2b} + k_{3b} + k_{3a})\xi,\ (ak_{3b} + k_{4a})\xi\right] \qquad \text{(II.24)}$$

Equation (II.25) gives the desired joint input-output characteristic function.

$$F_{x_1 x_2 x_3 x_4 0_1 0_2 0_3}\left[\xi_{1a}, \xi_{2a}, \xi_{3a}, \xi_{4a}, \xi_{1b}, \xi_{2b}, \xi_{3b}\right]$$
$$= F_{x_1 x_2 x_3 x_4}\left[(\xi_{1a} + \xi_{1b}),\ (a\xi_{1b} + \xi_{2b} + \xi_{2a}),\ (a\xi_{2b} + \xi_{3b} + \xi_{3a}),\ (a\xi_{3b} + \xi_{4a})\right] \qquad \text{(II.25)}$$

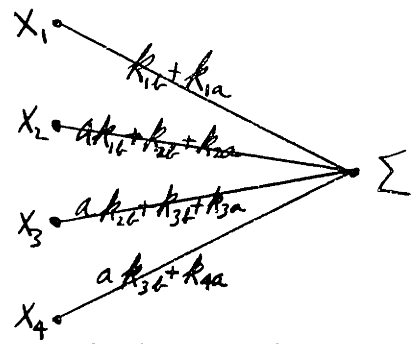
(2) Second Order Filters

A second order filter driven by a second order (Markov) input is shown in Figure II.14. The flow graph required in the calculation of the most general joint input-output characteristic function is Figure II.15.

The second order input is expressed as a degenerate sixth order process by its characteristic function

$$F_{x_1 x_2 x_3 x_4 x_5 x_6}\left[\xi_{1a}\, \xi_{2a}\, \xi_{3a}\, \xi_{4a}\, \xi_{5a}\, \xi_{6a}\right]$$

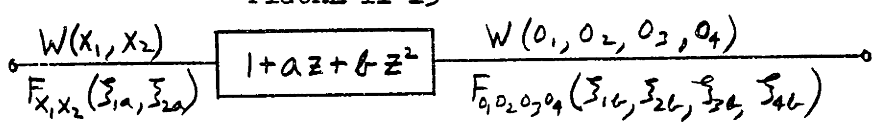
The characteristic function of $\sum$ is easily derived from Figure II.15 and from this the joint input-output characteristic function is calculated.



Flow Graph Useful In The Calculation Of

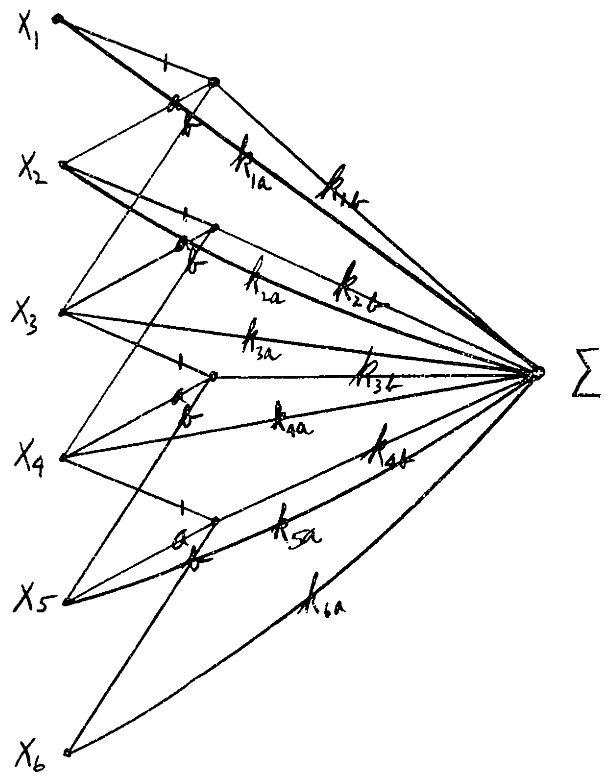
$F_{x_1 x_2 x_3 x_4 0_1 0_2 0_3}\left[\xi_{1a}\xi_{2a}\xi_{3a}\xi_{4a}\xi_{1b}\xi_{2b}\xi_{3b}\right]$

FIGURE II-13



A Second Order Filter Excited By A Second Order Input Signal

FIGURE II-14



Construction Of

$F_{x_1 x_2 x_3 x_4 x_5 x_6 0_1 0_2 0_3 0_4}\left[\xi_{1a}\xi_{2a}\xi_{3a}\xi_{4a}\xi_{5a}\xi_{6a}\xi_{1b}\xi_{2b}\xi_{3b}\xi_{4b}\right]$

FIGURE II-15

$$F_{x_1x_2x_3x_4x_5x_6\theta_1\theta_2\theta_3\theta_4}\left[\xi_{1a}\,\xi_{2a}\,\xi_{3a}\,\xi_{4a}\,\xi_{5a}\,\xi_{6a}\,\xi_{1b}\,\xi_{2b}\,\xi_{3b}\,\xi_{4b}\right]$$

$$= F_{x_1x_2x_3x_4x_5x_6}\Big[(\xi_{1b}+\xi_{1a}),(a\,\xi_{1b}+\xi_{2b}+\xi_{2a}),$$

$$(b\,\xi_{1b}+a\,\xi_{2b}+\xi_{3a}+\xi_{3b}),(b\,\xi_{2b}+a\,\xi_{3b}+\xi_{4a}+\xi_{4b}),$$

$$(b\,\xi_{3b}+\xi_{5a}+a\,\xi_{4b}),(b\,\xi_{4b}+\xi_{6a})\Big] \qquad \text{(II.26)}$$

C. <u>Higher Order Inputs, Higher Order Filters</u>

The same techniques may be generalized to derive output characteristic functions and joint input-output characteristic functions for any filter of finite order being driven by a random stationary input of finite order. A general literal expression for these comparable to (II.20) for first order inputs is too complex to serve useful purpose.

APPENDIX III

EXPERIMENTAL VERIFICATIONS

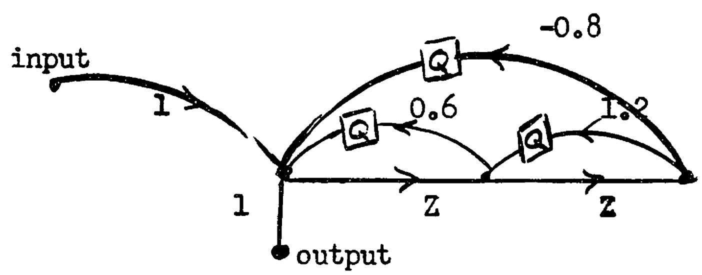
Two Bachelor's theses were done during the Spring term 1956 at MIT for the purpose of checking certain of the results of this thesis. They were done by Harry E. Pople under the supervision of the author, and by Alan I. Green under the supervision of R. F. Jenney. The results of both investigations came in too late to be included here in any detail.

A. Harry Pople; SB Thesis

A series of fine-grained correlated numbers were obtained from a turbulence noise study group in the MIT Hydrodynamics Laboratory. The autocorrelation of these numbers was calculated by the Whirlwind computer according to program by Douglas Ross. A modification in this program by Mr. Pople allowed rough quantization of the input numbers to take place before correlation. It was found, as predicted, that quantization noise is essentially uncorrelated and contributes to the mean square an amount equal to one twelfth the square of the box size. This held true until the entire range of the input was broken up into four levels, when the input numbers were rounded to two binary digits.

B. Alan Green; SB Thesis

The purpose of this thesis was to measure roundoff error propagation in a fairly complicated linear sampled-data feedback system. A flow graph of the system and its internal quantizers is shown in Figure III-1.



A Linear Flow Graph Containing Several Quantizerz

FIGURE III-1

The input was excited by random uncorrelated numbers that were uniformly distributed over a fixed range. The operation of the system as described by simultaneous difference equations was simulated by Whirlwind. Initial conditions for each run were zero. A run consisted of calculating the output over fifty sample points, and comparing these results with the precise solution. One hundred runs were made for each quantization box size, and the box sizes were adjusted by programming from 16 bit accuracy down to 3 bit accuracy.

By considering the quantizers as independent sources of random uniformly distributed noise, it was possible to calculate the mean square error of the net propagated noise. This agreed with the experimental results to well within the limits of the experiment (always within 30%) and departed radically when the Nyquist probability condition was violated.

BIOGRAPHICAL NOTE

Bernard Widrow was born in Norwich, Connecticut on December 24, 1929. He entered the Massachusetts Institute of Technology in September 1947, and received an S.B. in Electrical Engineering in 1951. At that time, he joined the staff of the Digital Computer Laboratory, which later became Division VI of Lincoln Laboratory, as a Research Assistant. Magnetic-core Memory development was his chief area of interest, and this led to a Master's thesis research. After receiving the S.M. degree in 1953, he joined the Lincoln Laboratory System Analysis Group under Prof. W. K. Linvill. From that time to the present, he has been engaged jointly in teaching Radar and System Analysis courses in the Electrical Engineering Department and in the work described by this thesis under the sponsorship of Lincoln Laboratory.

BIBLIOGRAPHY

1. Linvill, W.K., Class Notes for 6.54 (Pulsed-Data Systems), M.I.T., 1953.

2. Kochenburger, R.J., Analysis and Synthesis of Contactor Servomechanisms, Sc.D. Thesis, M.I.T., 1949.

3. Chow, C.K., Contactor Servomechanisms Employing Sampled-Data, Ph.D. Thesis, Cornell University, 1951.

4. Middleton, D., Class Notes for AM 215-216 (Applied Mathematics), Harvard University.

5. Linvill, W.K., and Sittler, R.W., Design of Sampled-Data Systems by Extension of Conventional Techniques, Lincoln Lab. Division 6 Report #R-222, M.I.T., July 1953

6. Vinogradov, I.M., Elements of Number Theory, Dover Publications, 5th edition, 1954.

7. Cramer, H., Mathematical Methods of Statistics, Princeton Univ. Press, Princeton, New Jersey, 1946.

8. Gardner, M. F., and Barnes, J.L., Transients in Linear Systems, John Wiley and Sons, New York, N.Y., 9th edition, 1952.

9. Sheppard, W.F., On the Calculation of the most Probable Values of Frequency-Constants, for Data arranged according to Equidistant Divisions of a Scale, Proc. of the London Math. Soc., Vol.29 (1898), P. 353.

10. Bowley, A.L., Elements of Statistics, P.S. King and Son, Ltd., London, 1937.

11. Bennett, W. R., Spectra of Quantized Signals, B.S.T.J., Vol.27, P. 446, July 1948.