IX. STATISTICAL COMMUNICATION THEORY*

Prof. Y. W. Lee
Prof. A. G. Bose
Prof. D. J. Sakrison
Prof. M. Schetzen
Prof. H. L. Van Trees, Jr.
V. R. Algazi
R. Alter
D. S. Arnstein
M. E. Austin
R. F. Bauer
E. M. Bregstone
J. D. Bruce
A. M. Bush
J. K. Clemens
A. G. Gann
C. E. Gray
T. G. Kincaid
P. L. Konop
A. J. Kramer
D. E. Nelsen
J. K. Omura
A. V. Oppenheim
R. B. Parente
W. S. Smith
D. W. Steele

A. GENERALIZATION OF THE ERROR CRITERION IN NONLINEAR THEORY BASED ON THE USE OF GATE FUNCTIONS

In the Wiener theory of optimum nonlinear systems, the measure of performance is the mean-square error and the input is a Gaussian process. A reason for the choice of this error criterion and type of input is the resulting relative analytical and experimental simplicity by which a nonlinear system can be determined. Recently, experimental procedures have been studied by which one can determine optimum nonlinear systems for error criteria other than the mean-square error criterion and inputs other than a Gaussian process.[1,2] The basic procedure is to expand the class of systems of interest into a complete set of operators, $\mathcal{H}_n$, so that the output of any system of this class can be expressed as

$$y(t) = \sum_{n=1}^{N} a_n y_n(t) \tag{1}$$

in which

$$y_n(t) = \mathcal{H}_n[x(t)]. \tag{2}$$

Figure IX-1 is a schematic representation of Eq. 1. In this representation, the coefficients, a_n, are amplifier gains. The determination of an optimum system is thus reduced to the determination of these coefficients. The procedure is then to measure experimentally the desired function of the error and simply to adjust the amplifier gains in order to minimize this quantity. This procedure is guaranteed to result in the optimum system of the class being represented for convex error criteria, since we are then guaranteed that there is a unique minimum of the function of the error and that there are no local minima. A difficulty with this method is that, in general, the amplifier

*This work was supported in part by the National Institutes of Health (Grant MH-04737-02); and in part by the National Science Foundation (Grant G-16526).

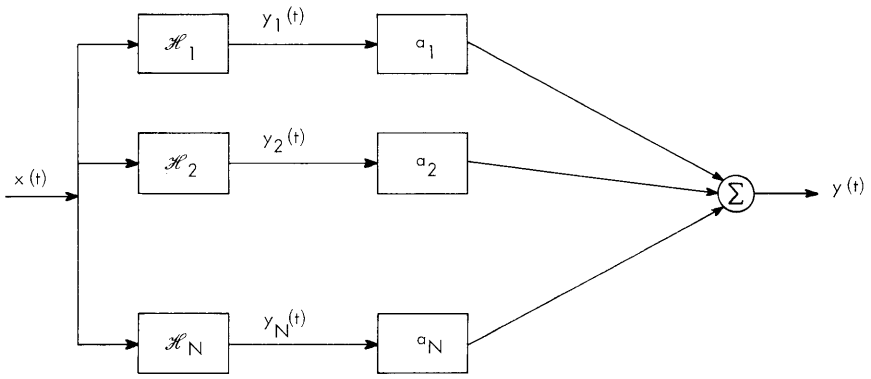

Fig. IX-1. Operator expansion of a system.

gain settings are not independent of one another and therefore an iterative adjustment procedure is required. This interdependence of the amplifier gain settings also makes any estimation of the optimum gain settings by analytical techniques too unwieldy to be obtained in any practical manner. Also, one generally has difficulty in obtaining a reasonable estimate of the class of nonlinear systems to represent. However, there is a set of operators based on the gate functions which can be used to expand the desired system with which, for any error criterion, the amplifier gain settings do not interact. In this report, we shall present some experimental and analytical techniques by which the amplifier gains can be determined when this set of operators is used. To explain the techniques, the determination of optimum nonlinear no-memory systems for various error criteria will be presented. The extension to nonlinear systems with memory will then be given.

1. The Gate Functions

For nonlinear no-memory systems, the set of operators that we shall use is the gate functions, Q_n. These functions were first introduced into the study of nonlinear systems by Bose.[3] For this set of operators and input, x(t), the outputs, $y_n(t)$, are

$$\left.\begin{aligned} y_o(t) = Q_o[x(t)] &= \begin{cases} 1 & \text{if } -\infty < x(t) < x_1 \\ 0 & \text{otherwise} \end{cases} \\ y_n(t) = Q_n[x(t)] &= \begin{cases} 1 & \text{if } x_n \leq x(t) < x_{n+1};\ n \neq 0, N \\ 0 & \text{otherwise} \end{cases} \\ y_N(t) = Q_N[x(t)] &= \begin{cases} 1 & \text{if } x_N \leq x(t) < \infty \\ 0 & \text{otherwise} \end{cases} \end{aligned}\right\} \tag{3}$$

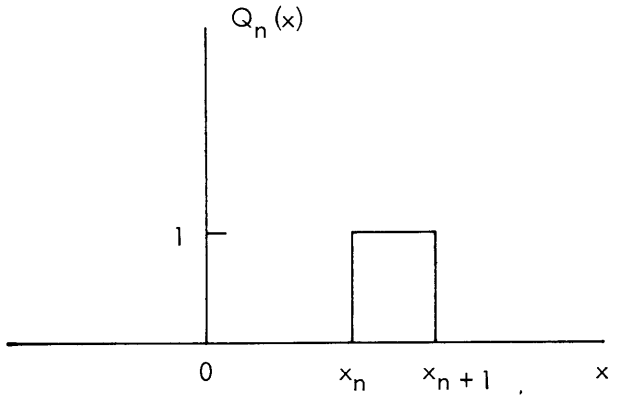

Fig. IX-2. Transfer characteristic of a gate function.

in which $x_m > x_n$ if $m > n$. Figure IX-2 is a plot of the n^{th} gate function. Any step approximation of a nonlinear no-memory system can be expressed in terms of this set as

$$y(t) = \sum_{n=0}^{N} a_n Q_n[x(t)], \tag{4}$$

in which we have not placed any restrictions on x(t).

Before proceeding, we note four important properties of the gate functions. The first is

$$\sum_{n=0}^{N} Q_n[x(t)] = 1. \tag{5}$$

The second is

$$Q_n^P[x(t)] = Q_n[x(t)] \qquad \text{for } P > 0. \tag{6}$$

The third is

$$\prod_{n=1}^{P} Q_{\alpha_n}[x(t)] = 0 \qquad \text{unless } \alpha_1 = \alpha_2 = \ldots = \alpha_P. \tag{7}$$

Then, by use of Eqs. 6 and 7, we obtain the fourth property, which is

$$\left[\sum_{n=0}^{N} a_n Q_n[x(t)]\right]^P = \sum_{n=0}^{N} a_n^P Q_n[x(t)]. \tag{8}$$

2. Optimum No-Memory Systems for Minimum $\overline{|E(t)|^P}$

In order to illustrate some of the properties of the gate functions and also methods for determining the coefficients, a_n, we first shall discuss the determination of optimum nonlinear no-memory systems for error criteria that are the mean P^{th} power of the magnitude of the error, E(t). That is, the error criteria that we first shall discuss are $\|E(t)\| = \overline{|E(t)|^P}$ for P = 1, 2, 3, ... We shall then generalize these results to error criteria that are arbitrary functions of the error.

For the input x(t), let the desired output be z(t). The error is then

$$E(t) = z(t) - \sum_{n=0}^{N} a_n Q_n[x(t)]. \tag{9}$$

For convenience, we shall omit writing the argument t. Thus we shall write the error as given by Eq. 9

$$E = z - \sum_{n=0}^{N} a_n Q_n[x] \tag{10}$$

with the understanding that E, x, and z are functions of time. Then the P^{th} power of the error is

$$E^P = \left[z - \sum_{n=0}^{N} a_n Q_n(x) \right]^P. \tag{11}$$

We shall obtain a more convenient form of this expression. Equation 11 can be expanded as

$$\begin{aligned} E^P = z^P &+ c_1 z^{P-1} \left[\sum_{n=0}^{N} a_n Q_n(x) \right] \\ &+ c_2 z^{P-2} \left[\sum_{n=0}^{N} a_n Q_n(x) \right]^2 + \ldots \\ &+ c_P \left[\sum_{n=0}^{N} a_n Q_n(x) \right]^P \end{aligned} \tag{12}$$

in which the coefficients, c_n, are the binomial coefficients. Substituting Eq. 8 in Eq. 12, we then have

$$E^P = z^P + c_1 z^{P-1} \sum_{n=0}^{N} a_n Q_n(x)$$

$$+ c_2 z^{P-2} \sum_{n=0}^{N} a_n^2 Q_n(x) + \ldots$$

$$+ c_P \sum_{n=0}^{N} a_n^P Q_n(x). \tag{13}$$

Equation 13 can be written in the form

$$E^P = z^P + \sum_{n=0}^{N} \left[c_1 z^{P-1} a_n + c_2 z^{P-2} a_n^2 + \ldots + c_P a_n^P \right] Q_n(x)$$

$$= z^P + \sum_{n=0}^{N} \left[(z-a_n)^P - z^P \right]$$

$$= z^P \left[1 - \sum_{n=0}^{N} Q_n(x) \right] + \sum_{n=0}^{N} [z-a_n]^P Q_n(x). \tag{14}$$

By use of Eq. 5, we then have

$$E^P = \sum_{n=0}^{N} [z-a_n]^P Q_n(x). \tag{15}$$

The magnitude of the P^{th} power of the error can be expressed in the form

$$|E|^P = [E^{2P}]^{1/2}. \tag{16}$$

Thus, from Eq. 15 we have

$$\overline{|E|^P} = \overline{\left[\sum_{n=0}^{N} (z-a_n)^{2P} Q_n(x) \right]^{1/2}}, \tag{17}$$

in which the bar indicates the time average of the function.

The optimum set of coefficients, a_n, for which Eq. 17 is a minimum can be determined by setting the derivative with respect to a_j equal to zero. Thus, from Eq. 17 we have

$$\frac{\partial \overline{|E|^P}}{\partial a_j} = -P\left[\overline{\frac{(z-a_j)^{2P-1}Q_j(x)}{|E|^P}}\right]. \tag{18}$$

The numerator in Eq. 18 is zero except when the amplitude of x(t) falls in the interval of the j^{th} gate function. At such times, the denominator is equal to $\left[(z-a_j)^{2P}Q_j(x)\right]^{1/2}$. Thus we can rewrite Eq. 18 to obtain

$$\frac{\partial \overline{|E|^P}}{\partial a_j} = -P\left[\overline{\frac{(z-a_j)^{2P-1}}{|z-a_j|^P}Q_j(x)}\right]$$

$$= -P\left[\overline{\frac{|z-a_j|^P}{(z-a_j)}Q_j(x)}\right]. \tag{19}$$

However,

$$(z-a_j) = |z-a_j|\,\mathrm{Sgn}(z-a_j) \tag{20}$$

in which

$$\mathrm{Sgn}(z-a_j) = \begin{cases} 1 & \text{if } z > a_j \\ 0 & \text{if } z = a_j \\ -1 & \text{if } z < a_j. \end{cases} \tag{21}$$

Substituting Eq. 20 in Eq. 19, we find that the condition for $\overline{|E|^P}$ to be a minimum is

$$\overline{|z-a_j|^{P-1}\mathrm{Sgn}(z-a_j)\,Q_j(x)} = 0. \tag{22}$$

We can show that Eq. 22 is the condition for a minimum by differentiating Eq. 18 with respect to a_j and substituting Eq. 22 in the result. Then we obtain

$$\frac{\partial^2 \overline{|E|^P}}{\partial a_j^2} = (2P-1)\,P\left[\overline{\frac{(z-a_j)^{2(P-1)}}{|E|^P}Q_j(x)}\right] \geq 0. \tag{23}$$

Equation 23 is positive, since $P \geq 1$ and all terms being averaged are always positive. Thus Eq. 22 is the condition for a minimum. We note that, by using the gate functions, the amplifier gain settings do not interact and they can be determined individually by means of Eq. 22.

Equation 22 provides a convenient experimental method for determining the desired

coefficients, a_n. We note that if P is even, then Eq. 22 can be written in the form

$$\overline{(z-a_j)^{P-1}Q_j(x)} = 0 \qquad [P \text{ even}]. \tag{24}$$

A circuit for experimentally determining the value of a_j that satisfies Eq. 24 is depicted in Fig. IX-3. The battery voltage, V, is adjusted to make the meter read zero; the voltage, V, is then numerically equal to the optimum value of a_j. Another procedure is to expand Eq. 24 as

$$\overline{(z-a_j)^{P-1}Q_j(x)} = \sum_{n=0}^{P-1} c_n \overline{z^n Q_j(x)}\, a_j^{(P-1-n)} = 0, \tag{25}$$

in which the coeffieients, c_n, are the binomial coefficients. The averages, $\overline{z^n Q_j(x)}$, can be measured experimentally. The optimum value of a_j can then be determined by substituting the measured values of $\overline{z^n Q_j(x)}$ in Eq. 25 and solving for the root of the resulting (P-1)-degree polynomial in a_j.

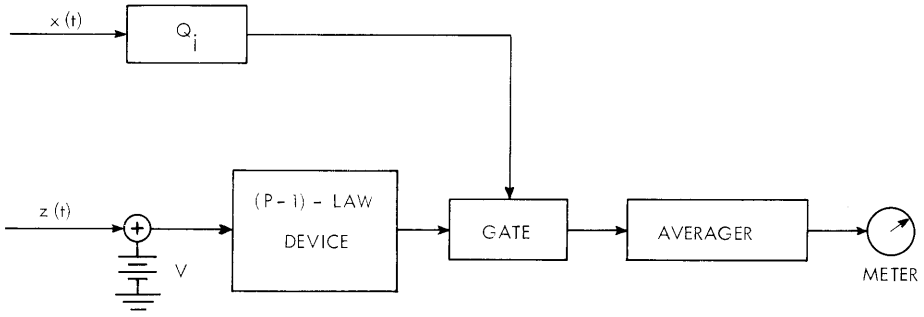

Fig. IX-3. A circuit for determining a_j for P even.

If P is odd, then there are two basically different circuits by which the desired value of a_j can be determined. One is obtained by noting that

$$|z-a_j|^{P-1}\text{Sgn}(z-a_j) = (z-a_j)^{P-1}|z-a_j| \qquad [P \geq 3 \text{ and odd}]. \tag{26}$$

A circuit, based on Eq. 26, for experimentally determining the desired value of a_j is depicted in Fig. IX-4. The second circuit is obtained by noting that a system whose output is Sgn $(z-a_j)$ is a saturating amplifier. Thus, for P odd, a second circuit is shown in Fig. IX-5. We note that the multiplier in Fig. IX-5 need only be a polarity-reversing switch that is controlled by the output of the saturating amplifier. In the special case for P = 1, the coefficients for the system with a minimum mean magnitude

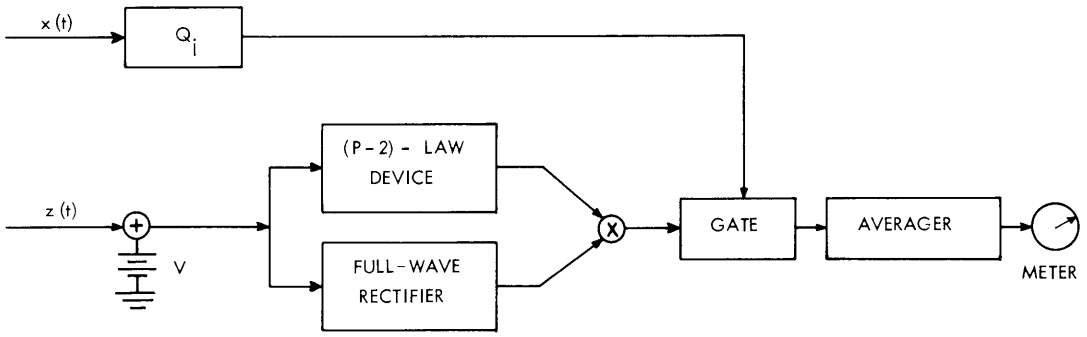

Fig. IX-4. A circuit for determining a_j for $P \geq 3$ and odd.

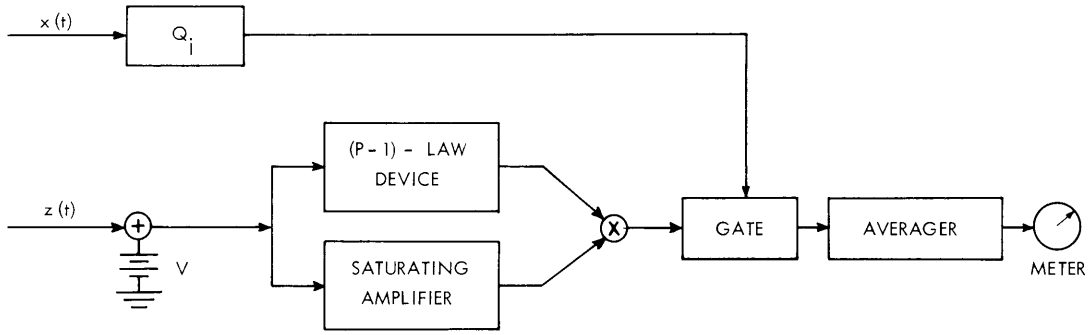

Fig. IX-5. A circuit for determining a_j for P odd.

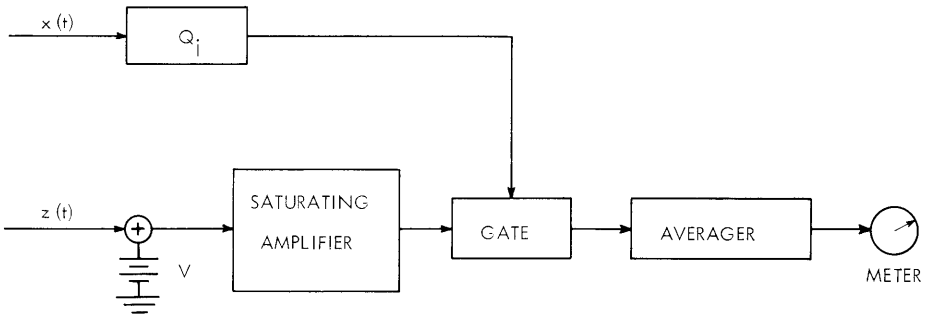

Fig. IX-6. A circuit for determining a_j for minimum $\overline{|E(t)|}$.

error are obtained. The circuit for determining these coefficients is relatively simple and is depicted in Fig. IX-6.

3. Optimum No-Memory Systems for Minimum $\overline{F[E(t)]}$

The procedure that we have just described can be extended to the determination of optimum nonlinear no-memory systems for error criteria that are arbitrary functions of the error. To extend this procedure, we have from Eq. 10

$$F[E] = F\left[z - \sum_{n=0}^{N} a_n Q_n(x)\right]. \tag{27}$$

According to the definition of the gate functions (Eqs. 3), at any instant only one term in the sum is nonzero. If, at that instant, x(t) is in the interval of the j^{th} gate function, then, at that instant, $F(E) = (z-a_j)\,Q_j(x)$. Thus, by the use of Eq. 5, we can express Eq. 27 as

$$F[E] = \sum_{n=0}^{N} F[z-a_n]\,Q_n(x). \tag{28}$$

Equation 28 is identical with Eq. 15 for $F(E) = E^P$. The set of coefficients for which the mean of Eq. 28 is stationary now can be determined by setting the derivative of $\overline{F(E)}$ with respect to a_j equal to zero. Thus, if we define

$$G(a) = \frac{d}{da} F(a), \tag{29}$$

then the condition that a_j be an extremal of the mean of Eq. 28 is

$$\overline{G(z-a_j)\,Q_j(x)} = 0. \tag{30}$$

A sufficient condition that assures us that the condition given by Eq. 30 yields a minimum of $\overline{F(E)}$ is that

$$\frac{d^2}{da^2} F(a) \geq 0. \tag{31}$$

Equation 31 is the statement that the error criterion is a convex function of its argument. For such cases, a circuit for experimentally determining the value of a_j that satisfies Eq. 30 is depicted in Fig. IX-7. The battery voltage, V, in the figure is adjusted to make the meter read zero; V is then numerically equal to the optimum value of a_j. If F(E) is not a convex function of its argument, this simple procedure is not sufficient, for local minima and maxima can then exist. However, the adjustment of any one coefficient affects only one term of the sum in Eq. 28. The minimum

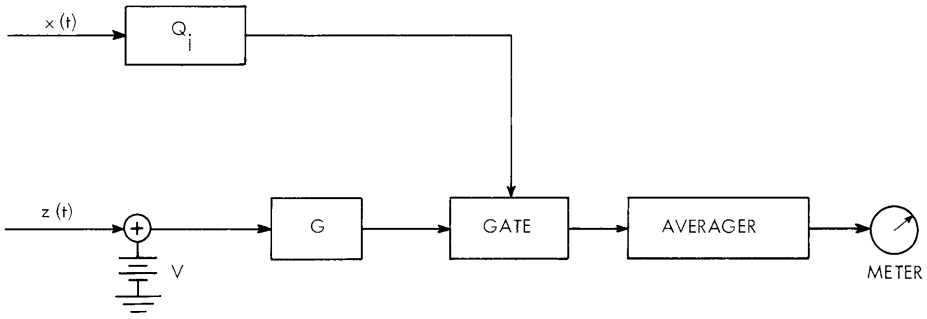

Fig. IX-7. A circuit for determining a_j for minimum $\overline{F[E(t)]}$.

of F(E) is thus obtained only if each term in Eq. 28 is a minimum. Therefore, because of the special properties of the gate functions, the amplifier gains, a_n, do not interact even for arbitrary functions of the error and thus they can be determined individually. This affords enormous simplification in search procedures for the determination of optimum systems for nonconvex functions of the error. However, if $G(a)$ can be expressed as a polynomial, then the values of a_j which satisfy Eq. 30 can be determined without a search procedure. For, if $G(a)$ can be expressed as a polynomial, then from Eq. 30 a polynomial expression in a_j can be obtained. The coefficients of the various powers of a_j will involve only the averages $\overline{z^n Q_j(x)}$ which can be measured experimentally. The values of a_j which satisfy Eq. 30 can then be determined by solving for the roots of the polynomial in a_j. Thus, we have the important result that for the cases in which F(E) can be expressed as a polynomial, no search procedure is required; then the optimum value of a_j can be determined analytically by solving for the minimum of the polynomial, $\overline{F(z-a_j)\, Q_j(x)}$.

4. An Example

An important consequence of the fact that the amplifier gains do not interact is that, in many cases, analytical estimates of the desired optimum nonlinear system can be obtained. We shall consider a simple example as an illustration. A received signal is the sum of a message and a noise that are statistically independent. The optimum nonlinear no-memory system is to be determined for an output that is the message. The error criterion to be used is $\overline{|E|^P}$ in which $P > 1$, but not necessarily an integer. The message is a binary signal that assumes the value 1 with probability 1/2 or the value -1 with probability 1/2. We shall determine the coefficients of the optimum system from Eq. 30. To do this, let the random variable of amplitude of the message be ζ which takes on values z; let the random variable of amplitude of the noise be η which takes on values y. Then, the random variable of amplitude of the received signal, ξ, is

$$\xi = \zeta + \eta. \tag{32}$$

On the ensemble basis, Eq. 30 for our example is

$$\overline{G(\zeta - a_j)\, Q_j(\xi)} = 0. \tag{33}$$

For our problem, $F(\alpha) = |\alpha|^P$ so that from Eq. 29

$$G(\alpha) = P|\alpha|^{P-1}\mathrm{Sgn}(\alpha). \tag{34}$$

Thus Eq. 33 becomes

$$0 = \overline{|\zeta - a_j|^{P-1}\mathrm{Sgn}(\zeta - a_j)\, Q_j(\zeta+\eta)}$$

$$= \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dz\, |z - a_j|^{P-1}\mathrm{Sgn}(z - a_j)\, Q_j(z+y)\, P_{\zeta,\eta}(z,y). \tag{35}$$

However, since the message is binary and independent of the noise, we have

$$P_{\zeta,\eta}(z,y) = P_\zeta(z)\, P_\eta(y)$$

$$= \frac{1}{2}[u(z-1)+u(z+1)]\, P_\eta(y), \tag{36}$$

in which u(z) is the unit impulse function. Substituting Eq. 36 in Eq. 35 and integrating with respect to z, we obtain

$$0 = \frac{1}{2}|1-a_j|^{P-1}\mathrm{Sgn}(1-a_j) \int_{-\infty}^{\infty} Q_j(y+1)\, P_\eta(y)\, dy$$

$$+ \frac{1}{2}|-1-a_j|^{P-1}\mathrm{Sgn}(-1-a_j) \int_{-\infty}^{\infty} Q_j(y-1)\, P_\eta(y)\, dy. \tag{37}$$

The equation that a_j must satisfy is thus

$$-\left|\frac{1-a_j}{1+a_j}\right|^{P-1} \frac{\mathrm{Sgn}(1-a_j)}{\mathrm{Sgn}(-1-a_j)} = \frac{\int_{-\infty}^{\infty} Q_j(y-1)\, P_\eta(y)\, dy}{\int_{-\infty}^{\infty} Q_j(y+1)\, P_\eta(y)\, dy}. \tag{38}$$

We note that the right-hand side of the equation is the ratio of the probability that $(\eta-1)$ is in the interval of the j^{th} gate function to the probability that $(\eta+1)$ is in the interval of the j^{th} gate function. Let us denote this ratio, which is positive, by β_j. We then have

$$-\left|\frac{1-a_j}{1+a_j}\right|^{P-1} \frac{\mathrm{Sgn}(1-a_j)}{\mathrm{Sgn}(-1-a_j)} = \beta_j. \tag{39}$$

Since $\beta \geq 0$, we require for a solution that

$$\frac{\mathrm{Sgn}(1-a_j)}{\mathrm{Sgn}(-1-a_j)} < 0.$$

Thus we require that $|a_j|$ be less than one. Equation 39 can thus be written as

$$\left(\frac{1-a_j}{1+a_j}\right)^{P-1} = \beta_j; \qquad |a_j| < 1. \tag{40}$$

For any set of gate functions, β_j can be determined; a_j can then be obtained by means of Eq. 40. If each gate function is chosen to be of infinitesimal width, then

$$\beta_j = \beta(y) = \frac{P_\eta(y-1)}{P_\eta(y+1)}, \tag{41}$$

in which $P_\eta(y)$ is the probability density of the amplitude of the noise. For such a case, $a_j = a(y)$ is the transfer characteristic of the no-memory system. Solving for a(y) from Eq. 40, we then have

$$a(y) = \frac{1 - [\beta(y)]^{1/(P-1)}}{1 + [\beta(y)]^{1/(P-1)}}; \qquad P > 1 \tag{42}$$

in which $\beta(y)$ is given by Eq. 41. We thus have obtained an explicit expression for the transfer characteristic of the no-memory system. The analogy between the approach taken in this example and the solution of a differential equation by means of difference equations is to be noted.

5. Probability of a Function of the Error

Since an arbitrary function of the error can be written in the form of Eq. 28, we can also determine the coefficients to minimize $\mathrm{Prob}\{F[E] > A\}$. We have from Eq. 28 the condition

$$\mathrm{Prob}\left\{\sum_{n=0}^{N} F[z-a_n]\, Q_n(x) > A\right\} = \text{minimum}. \tag{43}$$

But since, at any instant, only one term in the sum is nonzero, Eq. 43 can be written

$$\sum_{n=0}^{N} \mathrm{Prob}\{F[z-a_n]\, Q_n(x) > A\} = \text{minimum}. \tag{44}$$

However, the adjustment of any one coefficient affects only one term of the sum in Eq. 44. The minimum of Eq. 44 is thus obtained only if each term of the sum is a minimum. Thus the optimum value of a_j is that for which

$$\text{Prob}\left\{F[z-a_j]\, Q_j(x) > A\right\} = \text{minimum}. \tag{45}$$

A circuit by means of which the optimum value of a_j can be determined is depicted in Fig. IX-8. The voltage, V, is adjusted to minimize the meter reading; V is then

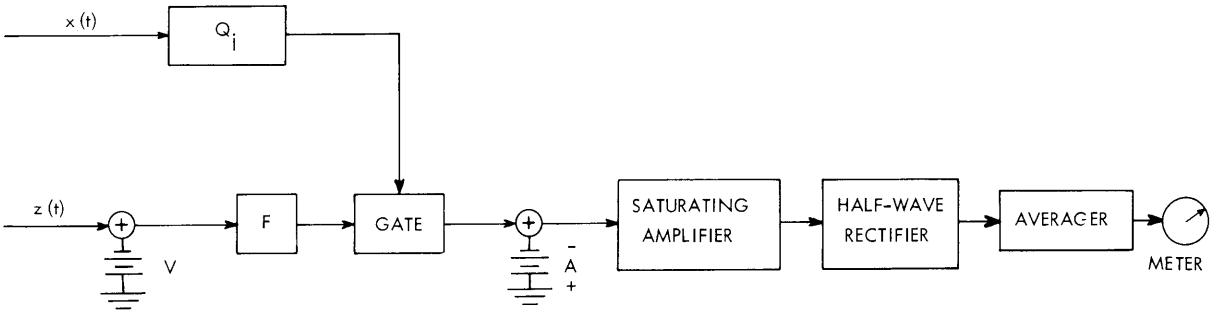

Fig. IX-8. A circuit for determining a_j that satisfies Eq. 44.

numerically equal to the desired value of a_j and the meter reading is the probability that $F[z-a_j]\, Q_j(x) > A$.

6. Weighted-Mean Function of the Error

All of the results that we have obtained can be extended to weighted functions of the error. By a weighted function of the error we mean $W(t)\, F[E(t)]$ in which $W(t)$ is an arbitrary function of time. For example, $W(t)$ can be a function of $[z(t)-x(t)]$. By use of Eq. 28, the weighted function of the error can be written

$$WF[E] = \sum_{n=0}^{N} F[z-a_n]\, WQ_n(x). \tag{46}$$

Thus all of our results apply if we replace $Q_n(x)$ by $WQ_n(x)$. Thus, to minimize a weighted mean of a convex function of the error, we have from Eq. 30 the result that the optimum value of a_j is that for which

$$\overline{G[z-a_j]\, WQ_j(x)} = 0. \tag{47}$$

7. Systems with Memory

We shall now present an extension of our previous results to nonlinear systems with memory. A complete set of operators that could be used for the synthesis of nonlinear

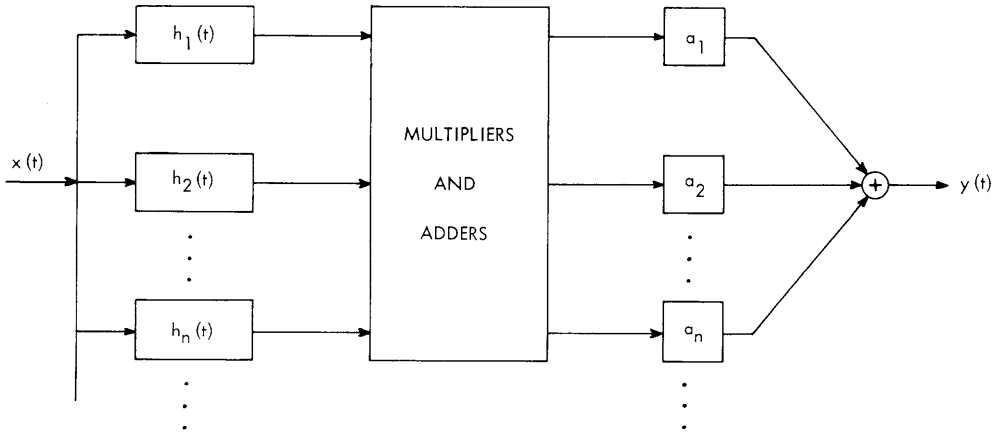

Fig. IX-9. Nonlinear system.

systems with memory is Wiener's G-functionals.[4] These operators can be synthesized in the form shown schematically in Fig. IX-9. As shown, the nonlinear system can be divided into three sections that are connected in tandem: a linear system whose impulse responses, $h_n(t)$, form a complete set of functions; a nonlinear no-memory section that consists of just multipliers and adders; and a section consisting of amplifiers whose outputs are then summed. In this model of a nonlinear system, the first section is the only one that has memory, since its outputs, $v_n(t)$, are the result of linear operations on the past of the input, x(t). We note from this model that the output, y(t), of a nonlinear system of the Wiener class is just a linear combination of various products of the outputs, $v_n(t)$, of the linear section. If the linear section consists of only K linear systems, then, according to the model, the output, y(t), can be expressed as

$$y = \sum_{k=1}^{\infty} \sum_{i_1=1}^{K} \cdots \sum_{i_k=1}^{K} A_{i_1,\ldots,i_k} v_{i_1} \cdots v_{i_k}. \tag{48}$$

By use of the gate functions, any step approximation to $v_i(t)$ can be given by

$$v_i \approx \sum_{n=0}^{N} B_n Q_n[v_i]. \tag{49}$$

By substituting Eq. 49 in Eq. 48, we note that we can express any step approximation to y(t) as

$$y \approx \sum_{k=1}^{K} \sum_{n_1=0}^{N} \cdots \sum_{n_k=0}^{N} C_{n_1,\ldots,n_k} Q_{n_1}(v_1) \ldots Q_{n_k}(v_k). \tag{50}$$

We note that by use of Eq. 5 we can write

$$\sum_{n_{k+1}=0}^{N} \cdots \sum_{n_K=0}^{N} Q_{n_{k+1}}(v_{k+1}) \cdots Q_{n_K}(v_K) = 1. \tag{51}$$

Thus Eq. 50 can be written in the form

$$y \approx \sum_{n_1=0}^{N} \cdots \sum_{n_K=0}^{N} D_{n_1,\ldots,n_K} Q_{n_1}(v_1) \cdots Q_{n_K}(v_K). \tag{52}$$

Define

$$\Phi_\alpha(\underline{v}) = Q_{n_1}(v_1) \cdots Q_{n_K}(v_K), \tag{53}$$

in which $\alpha = (n_1, n_2, \ldots, n_K)$. Then Eq. 52 can be written in the form

$$y \approx \sum_\alpha D_\alpha \Phi_\alpha(\underline{v}). \tag{54}$$

We note that the functions, $\Phi_\alpha(\underline{v})$, are K-dimensional gate functions, since $\Phi_\alpha(\underline{v})$ is nonzero and equal to one only when the amplitude of $v_1(t)$ is in the interval of $Q_{n_1}(v_1)$, and the amplitude of $v_2(t)$ is in the interval of $Q_{n_2}(v_2)$, and so on for each $v_n(t)$. That is, $\Phi_\alpha(\underline{v})$ is nonzero and equal to one only when $\underline{v}$ is in the α^{th} cell. We also note that

$$\sum_\alpha \Phi_\alpha(\underline{v}) = 1. \tag{55}$$

By use of the K-dimensional gate functions, all of our results can be extended to nonlinear systems that have memory. For we note that if the desired output is z(t), then any function of the error is

$$F[E] = F\left[z - \sum_\alpha A_\alpha \Phi_\alpha(\underline{v})\right]. \tag{56}$$

According to the definition of the K-dimensional gate functions, only one term in the sum is nonzero at any instant. If, at that instant, $\underline{v}$ is in the α^{th} cell, then at that instant $F[E] = F(z-A_\alpha)\,\Phi_\alpha(\underline{v})$. Thus, by use of Eq. 55, we can express Eq. 56 as

$$F[E] = \sum_\alpha F[z-A_\alpha]\,\Phi_\alpha(\underline{v}). \tag{57}$$

This equation is identical in form with Eq. 28 and thus all of the results that we have obtained for the one-dimensional case also apply to the K-dimensional case. For example, from Eq. 30 the condition that A_a be an extremal of $\overline{F(E)}$ is

$$\overline{G[z-A_a]\,\Phi_a(\underline{v})} = 0. \tag{58}$$

Experimentally, the values of A_a that satisfy Eq. 58 can be obtained by means of the circuit depicted in Fig. IX-10. The procedure is to adjust the battery voltage, V, until the meter reads zero; V is then numerically equal to the desired value of A_a for which

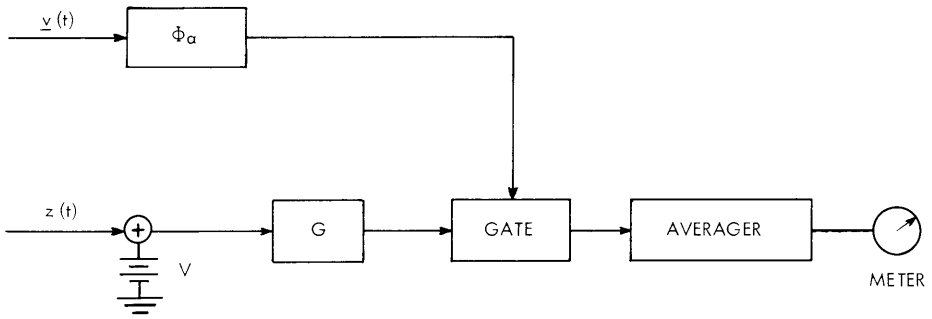

Fig. IX-10. A circuit for determining A_a for which $\overline{F[E]}$ is a minimum.

Eq. 58 is satisfied. However, if $G(a)$ can be expressed as a polynomial, then a polynomial expression in A_a can be obtained from Eq. 58. The coefficients of the various powers of A_a will involve only the averages $\overline{z^n\Phi_a(\underline{v})}$ which can be measured experimentally. The values of A_a that satisfy Eq. 58 can then be determined by solving for the roots of the polynomial. For example, if $F(E) = E^2$, then Eq. 58 becomes

$$\overline{(z-A_a)\,\Phi_a(\underline{v})} = 0. \tag{59}$$

Solving for A_a, we have

$$A_a = \frac{\overline{z\Phi_a(\underline{v})}}{\overline{\Phi_a(\underline{v})}}. \tag{60}$$

Equation 60 is the result obtained by Bose.[3] We thus have the important result that, as in the case of no-memory systems, no adjustment procedure of any sort is required in those cases in which F(E) can be expressed as a polynomial. If F(E) is not a convex function, the optimum value of A_a can still be determined analytically by solving for the minimum of the polynomial, $\overline{F[z-A_a]\,\Phi_a(\underline{v})}$, and thus no search procedure is required. Also, the set of coefficients $\{A_a\}$ can be determined for which the mean of

an arbitrary weighted function of the error or the probability of an arbitrary function of the error is a minimum. Each coefficient of this set can be determined individually in a manner similar to that used for the no-memory case.

M. Schetzen

References

1. M. Schetzen, Some Problems in Nonlinear Theory, Technical Report 390, Research Laboratory of Electronics, M. I. T., July 6, 1962.

2. D. Sakrison, Application of Stochastic Approximation Methods to System Optimization, Technical Report 391, Research Laboratory of Electronics, M. I. T., July 10, 1962.

3. A. Bose, A Theory of Nonlinear Systems, Technical Report 309, Research Laboratory of Electronics, M. I. T., May 15, 1956.

4. N. Wiener, Nonlinear Problems in Random Theory (The Technology Press of Massachusetts Institute of Technology, Cambridge, Mass., and John Wiley and Sons, Inc., New York, 1958).

B. OPTIMUM QUANTIZATION FOR A GENERAL ERROR CRITERION

1. Introduction

Quantization is the nonlinear, no-memory operation of converting a continuous signal to a discrete signal that assumes only a finite number of levels (N). Quantization occurs whenever it is necessary to represent physical quantities numerically. The primary concern in quantization is faithful reproduction, with respect to some fidelity criterion, of the input at the output. Thus, it is not necessary to require uniform quantization. In fact, since the number of output levels, N, will be specified, the "error" in the output will be minimized by adjusting the quantizer characteristic. Figure IX-11 illustrates the input-output characteristic of a quantizer.

Early investigations into the process of quantization considered speech to be the quantizer input signal. One of the first investigators was Bennett[1] who concluded that, with speech, it would be advantageous to taper the steps of the quantizer in such a manner that finer steps would be available for weak signals. Following Bennett, other investigators such as Smith,[2] Lozovoy,[3] and Lloyd[4] worked toward the characterization of optimum quantizers by assuming that a large number of levels would be used in the quantizer. Max[5] formulated the general problem of selecting the parameters of the optimum quantizer for a wide class of error criteria irrespective of the number of levels in the device. He also was able to determine a partial solution to the problem by paying particular attention to the mean-square error criteria. Bluestein[6] derived some extensions to Max's work for the special case of the mean-absolute error criteria.

In this report, the expression for the quantization error as a function of the quantizer

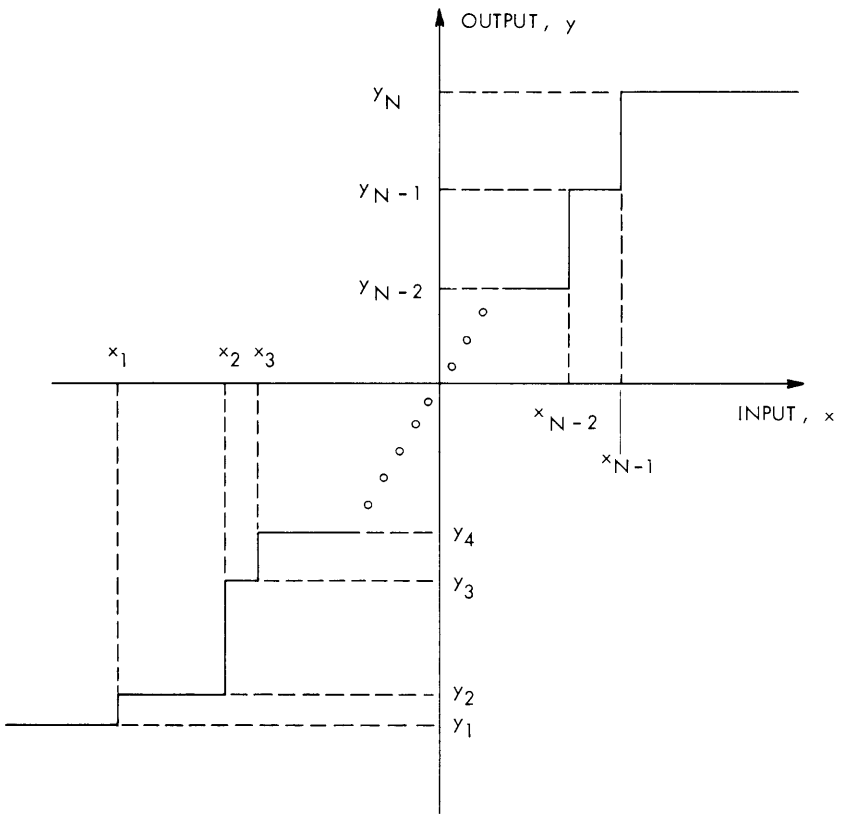

Fig. IX-11. Input-output relationship of the N-level quantizer.

parameters is presented and a method for determining these parameters by utilizing the technique of dynamic programming is developed. This method has two advantages over the previous ones. First, the method "searches" for the absolute minimum within the region of variation rather than the relative minima. Second, the method is guaranteed to converge to the absolute minimum in a specified number of steps.

2. Formulation of the Problem

Figure IX-11 shows the input-output relationship for the general quantizer. The output is y_k for $x_{k-1} \leq x < x_k$. The x_k are called the transition values; that is, x_k is the value of the input variable x for which there is a transition in the output from y_k to y_{k+1}. The y_k are called the representation values.

In the general quantizer problem under consideration here, the x_i, as well as the y_k (a total of 2N-1 quantities), are variables. These 2N-1 variables are to be chosen in such a manner that the error with respect to a specified error criterion is minimized.

In order to determine the specific values of x_i and y_k which specify the optimum quantizer, that is, the quantities that we shall call X_i and Y_k, we must first derive an

expression for the quantization error criterion. Since by optimum we shall mean minimum quantization error with respect to the specified error criterion, the X_i and the Y_k are the specific values of x_i and y_k which yield the absolute minimum value for the quantization error.

There is one set of constraints which must be imposed upon solutions to this problem. For purposes of organization as indicated in Fig. IX-11 we shall require that

$$\left.\begin{aligned} x_1 &< x_2 \\ x_2 &< x_3 \\ &\vdots \\ x_{N-2} &< x_{N-1} \end{aligned}\right\}. \tag{1}$$

With respect to the quantizer of Fig. IX-11, the error signal when the input signal satisfies the inequality

$$x_i \leq x < x_{i+1} \tag{2}$$

is

$$x - y_{i+1}. \tag{3}$$

We shall choose the error criterion to be the expected value of the function

$$g(x,y_{i+1}).$$

Then the measure of the error in this interval is

$$\int_{x_i}^{x_{i+1}} g(x,y_{i+1})\, p(x)\, dx, \tag{4}$$

where p(x) is the amplitude probability density of the input variable x. Since the error in each of the quantization intervals is independent of the error in the other intervals, the total quantization error is

$$\mathcal{E}(x_1, x_2, \ldots, x_{N-1}; y_1, y_2, \ldots, y_N) = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} g(x,y_{i+1})\, p(x)\, dx. \tag{5}$$

In writing this equation we included the two parameters x_o and x_N for convenience. If x is a bounded function with lower bound X_ℓ and upper bound X_u, then, by definition, x_o will equal X_ℓ and x_N will equal X_u.

Upon first observation it appears that the coordinates of the absolute minimum value of (5) can be determined by use of the methods of calculus. Indeed, if there is only a single critical point of Eq. 5 within the region of variation (the region specified by Eq. 1)

and if this critical point is a relative minimum, then it is also the absolute minimum of the function. However, if there is more than a single critical point within the region of variation, then one of these critical points may be the absolute minimum of (5). Or it might be that none of these critical points is the absolute minimum. Thus the method of calculus is not a satisfactory technique for solving this problem. What is needed is a method that yields the coordinates of the absolute minimum whether or not the absolute minimum is at a critical point. The method of dynamic programming[7,8] is such a technique.

3. Determining the Optimum Quantizer by Using Dynamic Programming

In order to determine the coordinates of the absolute minimum value of (5) it is necessary to define a sequence of error functionals

$$\{f_i(x_i)\} \qquad i = 1, 2, \ldots, N$$

as follows.

$$\left.\begin{aligned}
f_1(x_1) &= \min_{\substack{y_1 \\ X_\ell \le x_1 \le X_u}} \left[\int_{X_\ell}^{x_1} g(x,y_1)\, p(x)\, dx\right] \\
f_2(x_2) &= \min_{\substack{y_2,x_1 \\ X_\ell \le x_1 \le x_2 \le X_u}} \left[\int_{x_1}^{x_2} g(x,y_2)\, p(x)\, dx + f_1(x_1)\right] \\
f_3(x_3) &= \min_{\substack{y_3,x_2 \\ X_\ell \le x_2 \le x_3 \le X_u}} \left[\int_{x_2}^{x_3} g(x,y_3)\, p(x)\, dx + f_2(x_2)\right] \\
&\vdots \\
f_i(x_i) &= \min_{\substack{y_i,x_{i-1} \\ X_\ell \le x_{i-1} \le x_i \le X_u}} \left[\int_{x_{i-1}}^{x_i} g(x,y_i)\, p(x)\, dx + f_{i-1}(x_{i-1})\right] \\
&\vdots \\
f_{N-1}(x_{N-1}) &= \min_{\substack{y_{N-1},x_{N-2} \\ X_\ell \le x_{N-2} \le x_{N-1} \le X_u}} \left[\int_{x_{N-2}}^{x_{N-1}} g(x,y_{N-1})\, p(x)\, dx + f_{N-2}(x_{N-2})\right] \\
f_N(x_N) &= \min_{\substack{y_N,x_{N-1} \\ X_\ell \le x_{N-1} \le x_N \le X_u}} \left[\int_{x_{N-1}}^{x_N} g(x,y_N)\, p(x)\, dx + f_{N-1}(x_{N-1})\right]
\end{aligned}\right\} \qquad (6)$$

It should be noted that in the last member of Eq. 6 we have taken the liberty of permitting x_N to take on variation from X_ℓ to X_u in deference to our previous assumption (see explanation immediately following Eq. 5) that $x_N = X_u$. This variation is necessary if $f_N(x_N)$ is to be completely determined.

Also, from Eq. 6 we can show that the last member of this set of equations, when evaluated at $x_N = X_u$, can be written

$$f_N(X_u) = \min_{\substack{y_i, x_{i-1} \\ i=1,2,\ldots N \\ X_\ell = x_o \leq x_1 \leq \ldots \leq x_N = X_u}} \left[\sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} g(x, y_{i+1})\, p(x)\, dx \right]. \tag{7}$$

Therefore, $f_N(X_u)$ is the error for the optimum quantizer with N levels. (Note that, in general, $f_i(X_u)$ is the error for the optimum quantizer with i levels, $i \leq N$.)

It is possible to simplify the minimization process used to determine the error functionals, Eq. 6, when we realize that the minimum, with respect to y_i, occurs for a value of y_i which satisfies the equation

$$0 = \int_{x_{i-1}}^{x_i} \frac{\partial}{\partial y_i} [g(x, y_i)]\, p(x)\, dx. \tag{8}$$

If, as will usually be the case, $g(x, y_i)$ is greater than or equal to zero and is concave upward, then (8) has only a single solution. This solution will be denoted by $\overline{y_i}$. (Should (8) have more than one solution, the solution that minimizes (6) will be called $\overline{y_i}$.) Equation 8 then allows us to rewrite the general term of (6) and to eliminate the formal minimization with respect to y_i, since the value of y_i which minimizes $f_i(x_i)$ for specific values of x_i and x_{i-1} is $\overline{y_i}$. Thus (6) becomes

$$f_i(x_i) = \min_{\substack{x_{i-1} \\ X_\ell \leq x_{i-1} \leq x_i \leq X_u}} \left[\int_{x_{i-1}}^{x_i} g(x, \overline{y_i})\, p(x)\, dx + f_{i-1}(x_{i-1}) \right] \qquad i = 1, 2, \ldots, N \tag{9}$$

where $x_0 = X_\ell$, a constant, and $f_o(X_\ell) \equiv 0$. With reference to Eq. 7 we see that (8) has reduced the number of variables over which the formal minimization must be performed from 2N-1 to N.

In order to determine the optimum quantizer parameters it is necessary to define two auxiliary sets of functionals; first, the transition-value decision functionals

$$\{X_i(x)\} \qquad i = 1, 2, \ldots, N;$$

and second, the representation-value decision functionals

$$\{Y_i(x)\} \qquad i = 1, 2, \ldots, N.$$

The transition-value decision functionals are defined as follows

$$\left.\begin{aligned} &X_1(x) = x; \\ &X_2(x) = \text{the value of } x_1 \text{ which minimizes } f_2(x_2) \text{ for a specified } x = x_2,\ y_2 = \overline{y_2}; \\ &\quad\vdots \\ &X_N(x) = \text{the value of } x_{N-1} \text{ which minimizes } f_N(x_N) \text{ for a specified } x = x_N,\ y_N = \overline{y_N} \end{aligned}\right\}. \tag{10}$$

In a similar manner the representation-value decision functionals are defined as

$$\left.\begin{aligned} &Y_1(x) = \overline{y_1} \text{ for a specified } x = x_1; \\ &Y_2(x) = \overline{y_2} \text{ when } x_1 \text{ is that value [i.e., } x_1 = X_2(x)] \text{ which minimizes } f_2(x_2) \text{ for a specified } x = x_2; \\ &\quad\vdots \\ &Y_N(x) = \overline{y_N} \text{ when } x_{N-1} \text{ is that value [i.e., } x_{N-1} = X_N(x)] \text{ which minimizes } f_N(x_N) \text{ for a specified } x = x_N \end{aligned}\right\}. \tag{11}$$

Each of the functionals, Eqs. 6, 10, and 11, will be represented by a tabulation of its value at a number of points taken along an equispaced grid. In general, the same set of points will be used for each of the functionals.

As we have previously indicated, $f_N(X_u)$ is the measure of error for the optimum quantizer with N levels with the specified error criterion used. Therefore, the location of the transition value between levels (N-1) and (N) is specified by (10), that is,

$$X_{N-1} = X_N(X_u).$$

This level will be represented by the representation value

$$Y_N = Y_N(X_u),$$

specified by Eq. 11. At this point in the decision process, x that is such that $X_{N-1} \leq x \leq X_u$ has been allocated to the N^{th} level. The remaining values of x, $X_\ell \leq x \leq X_{N-1}$, remain to be quantized into N-1 levels. From (10) the transition value between levels (N-2) and (N-1) is given by

$$X_{N-2} = X_{N-1}(X_{N-1}),$$

and from (11) the representation value is

$$Y_{N-1} = Y_{N-1}(X_{N-1}).$$

This decision process is continued until the first representation value is found from (11) to be

$$Y_1 = Y_1(X_1).$$

Once the decision process is completed, the optimum quantizer with respect to the specified error criterion is determined.

As a first example, this method of selecting the optimum quantizer is being applied in the case in which speech is the input signal. In this application the error criterion will be the expected value of the function

$$g(x,y_{i+1}) = |x-y_{i+1}|^{\nu} W(x),$$

where

$$W(x) \geq 0 \qquad -\infty < x < \infty.$$

J. D. Bruce

References

1. W. R. Bennett, Spectra of quantized signals, Bell System Tech. J. 27, 446-472 (1948).

2. B. Smith, Instantaneous companding of quantized signals, Bell System Tech. J. 36, 653-709 (1957).

3. I. A. Lozovoy, Regarding the computation of the characteristics of compression in systems with pulse code modulation, Telecommunications (published by AIEE), No. 10, pp. 18-25, 1961.

4. S. P. Lloyd, Least Squares Quantization in PCM (unpublished manuscript, Bell Telephone Laboratories, Inc., 1958), reported by D. S. Ruchkin, Optimum Reconstruction of Sampled and Quantized Stochastic Signals, Doctor of Engineering Dissertation, Yale University, May 1960.

5. J. Max, Quantizing for minimum distortion, IRE Trans., Vol. IT-6, pp. 7-12, March 1960.

6. L. I. Bluestein, A Hierarchy of Quantizers, Ph.D. Thesis, Columbia University, May 1962.

7. R. Bellman and J. Dreyfus, Applied Dynamic Programming (Princeton University Press, Princeton, N. J., 1962).

8. R. Bellman and B. Kotkin, On the Approximation of Curves by Linear Segments Using Dynamic Programming—II, Memorandum RM-2978-PR, The Rand Corporation, Santa Monica, California, February 1962.