[10] H. Cramér and M. Leadbetter, "The moments of the number of crossings of a level by a stationary normal process," *Ann. Math. Stat.*, vol. 36, pp. 1656–1663, 1965.
[11] M. Loève, *Probability Theory*. New York: Van Nostrand, 1955, p. 125.
[12] D. J. Sakrison, "Efficient recursive estimation of the parameters of a radar or radio astronomy target," *IEEE Trans. Information Theory*, vol. IT-12, pp. 35–41, January 1966.
[13] L. Ehrman, "Analysis of a zero-crossing frequency discriminator with random inputs," *IEEE Trans. Aerospace and Navigational Electronics*, vol. ANE-12, pp. 113–119, June 1965.
[14] E. L. McMahon, "An extension of Price's theorem," *IEEE Trans. Information Theory (Correspondence)*, vol. IT-10, p. 168, April 1964.
[15] J. A. Mullen, "Optimal filtering for radar Doppler navigators," *1962 IRE Nat'l Conv. Rec.*, pt. 5, pp. 40–48.

Asymptotically Efficient Quantizing

HERBERT GISH, MEMBER, IEEE, AND JOHN N. PIERCE, SENIOR MEMBER, IEEE

Abstract—**It is shown, under weak assumptions on the density function of a random variable and under weak assumptions on the error criterion, that uniform quantizing yields an output entropy which asymptotically is smaller than that for any other quantizer, independent of the density function or the error criterion. The asymptotic behavior of the rate distortion function is determined for the class of vth law loss functions, and the entropy of the uniform quantizer is compared with the rate distortion function for this class of loss functions. The extension of these results to the quantizing of sequences is also given. It is shown that the discrepancy between the entropy of the uniform quantizer and the rate distortion function apparently lies with the inability of the optimal quantizing shapes to cover large dimensional spaces without overlap. A comparison of the entropies of the uniform quantizer and of the minimum-alphabet quantizer is also given.**

Introduction

THE ENTROPY H of the output of a quantizer is the minimum amount of information which must be transmitted in order to be able to determine the quantizer output with an arbitrarily small error. If we establish some mean error criterion E (e.g., mean square, mean absolute) between the quantizer input and output as a measure of quantizer reproduction fidelity, various types of quantizers can be ranked by comparisons of their $H(E)$ curves. In addition, the merits of quantization as opposed to other means of source encoding can be ascertained by comparing $H(E)$ to Shannon's [1], [2] rate distortion function $R(E)$. The function $R(E)$, which depends only on the distribution of the variable being transmitted, specifies the minimum amount of information which must be transmitted in order to reconstruct the variable with a mean error E.

In the following we investigate the relation between the entropy of the output of a quantizer and its mean error. In particular, we look at the asymptotic relation between the two quantities as the mean error is required to become very small and show that for a specified error uniform quantization yields minimum entropy. The result is shown to be valid under rather weak assumptions about the density function of the variable being quantized and the mean error criterion being used. The performance of the asymptotically optimum quantizer is compared to bounds on the rate distortion function.

Manuscript received June 23, 1967; revised February 5, 1968.
H. Gish is with SIGNATRON, Inc., Lexington, Mass. 02173
J. N. Pierce is with the AF Cambridge Research Laboratories, Bedford, Mass. 01730

Asymptotic Optimality of Uniform Quantizing

Let X be the random variable at the quantizer input we will assume that the density function $f(x)$ is reasonably smooth. The quantizer divides the range of X into a possibly infinite number of adjacent intervals I_n

$$I_n = (g_n, g_{n+1}) \tag{1}$$

and maps X into the discrete-valued random variable Y

$$Y = Y_n \quad \text{if} \quad X \in I_n. \tag{2}$$

The entropy of Y is

$$H = -\sum p_n \log p_n \tag{3a}$$

where

$$p_n = P(X \in I_n) = \int_{I_n} dx\, f(x). \tag{3b}$$

The mean-square error will be written as E; we will investigate more general loss functions subsequently. The value of E is

$$E = \sum \int_{I_n} dx\, f(x)(x - Y_n)^2. \tag{4}$$

If the lengths of all of the intervals are reasonably small, then $f(x)$ will be approximately constant over each interval, and we can write

$$p_n \approx f(g_n)(g_{n+1} - g_n). \tag{5}$$

Furthermore, under this condition, putting Y_n at the midpoints of the intervals will lead to approximately the

minimum mean-square error for the prescribed set of intervals

$$Y_n = (g_n + g_{n+1})/2, \tag{6}$$

$$E \approx \sum f(g_n)(g_{n+1} - g_n)^3/12. \tag{7}$$

Suppose now that we define the mesh points g_n by

$$g_n = g(n\delta) \tag{8}$$

where $g(t)$ is some suitably smooth monotone increasing function. Then we can inquire as to what choice of g leads to the slowest increase of H with decreasing E as δ is allowed to become small.

We first note that for small δ

$$g_{n+1} - g_n \approx \delta g'(n\delta) \tag{9}$$

where the prime indicates derivative, so that, from (3a) and (5)

$$H \approx -\sum \delta g'(n\delta) f(g(n\delta)) \log (\delta g'(n\delta) f(g(n\delta)))$$

which, as δ becomes arbitrarily small, goes over into the integral form

$$H \approx -\int dt\, g'(t) f(g(t)) \log (\delta g'(t) f(g(t))). \tag{10}$$

Similarly,

$$E \approx \sum f(g(n\delta))(\delta g'(n\delta))^3/12$$

approaches

$$E \approx \int dt\, g'(t) f(g(t))(\delta g'(t))^2/12. \tag{11}$$

Making the obvious substitution

$$s = g(t)$$

in these integrals then leads to

$$H \approx -\int ds\, f(s) \log (\delta f(s) g'(g^{-1}(s))) \tag{12}$$

and

$$E \approx (\delta^2/12) \int ds\, f(s)(g'(g^{-1}(s)))^2. \tag{13}$$

If we define H_0 to be the entropy of the continuous distribution

$$H_0 = -\int ds\, f(s) \log (f(s)) \tag{14}$$

and temporarily define

$$\gamma(s) = g'(g^{-1}(s)) \tag{15}$$

we then have

$$E \approx (\delta^2/12) \int ds\, f(s)(\gamma(s))^2 \tag{16}$$

and

$$H \approx H_0 - \log \delta - \int ds\, f(s) \log (\gamma(s)). \tag{17}$$

It is readily verified that for fixed δ and E, H is a minimum when $\gamma(s)$ is a constant, or equivalently, when $g'(t)$ is independent of t. We may conveniently take this constant to be unity and set

$$g(t) = t \tag{18}$$

which then leads to[1]

$$E \approx \delta^2/12 \tag{19}$$

$$H_{\min} \approx H_0 - \log \delta. \tag{20}$$

From this pair of equations we can write the equation relating H to E

$$H_{\min} \approx H_0 - (\tfrac{1}{2}) \log (12E) \quad \text{as} \quad E \to 0. \tag{21}$$

If we define V_0 to be entropy variance corresponding to H_0,[2] that is, the variance of a Gaussian distribution having the same entropy H_0, we have

$$H_0 = (\tfrac{1}{2}) \log (2\pi e V_0) \tag{22}$$

so that (21) can be rewritten as

$$H_{\min} \approx (\tfrac{1}{2}) \log (V_0/E) + (\tfrac{1}{2}) \log (2\pi e/12)$$

or as

$$H_{\min} \approx (\tfrac{1}{2}) \log_2 (V_0/E) + 0.255 \tag{23}$$

as $E \to 0$, in bits.

We now note that the result can be generalized somewhat. Suppose that in place of (4) we define an average error by

$$E_L = \sum \int_{I_n} dx\, f(x) L(x - Y_n) \tag{4'}$$

where

a) $L(0) = 0$.
b) L is an increasing function of the magnitude of its argument.
c) The function $M(x)$ defined from $L(x)$ by (7'b) below satisfies: $xM'(x)$ is monotone.

Then (7) becomes

$$E_L \approx \sum f(g_n) \int_{-(g_{n+1}-g_n)/2}^{(g_{n+1}-g_n)/2} du\, L(u)$$

which we rewrite in the form

$$E_L \approx \sum f(g_n)(g_{n+1} - g_n) M(g_{n+1} - g_n) \tag{7'a}$$

where the function M is defined by

$$M(v) = (1/v) \int_{-v/2}^{v/2} du\, L(u). \tag{7'b}$$

Then (11) is replaced by

$$E_L \approx \int dt\, g'(t) f(g(t)) M(\delta g'(t)) \tag{11'}$$

[1] Refer to Appendix I, for conditions on $f(x)$ which establish the validity of (19) and (20). Also, see Appendix II for the outline of a more rigorous but less intuitive approach to the derivation of (21).
[2] Analogous to the entropy power of a continuous process.

(13) is replaced by

$$E_L \approx \int ds\, f(s) M(g'(g^{-1}(s))) \tag{13'}$$

and, finally, the variational pair (16) and (17) is replaced by

$$E_L \approx \int ds\, f(s) M(\delta\gamma(s)) \tag{16'}$$

$$H \approx H_0 - \int ds\, f(s) \log (\delta\gamma(s)). \tag{17'}$$

The stationary solutions of the Euler–Lagrange equation for the system of (16') and (17') must satisfy

$$\gamma(s) M'(\gamma(s)) = \text{constant}. \tag{18'a}$$

The condition c) after (4') guarantees that $\gamma(s) = \text{constant}$ is the unique solution of the variational problem from which we are again led to the solution

$$g(t) = t \tag{18'b}$$

implying uniform quantizing and the relations

$$E_L \approx M(\delta) \tag{19'}$$

$$H_{\min} \approx H_0 - \log \delta. \tag{20'}$$

Finally, we arrive at the relation between $H_{\min}$ and E_L:

$$H_{\min} \approx H_0 - \log (M^{-1}(E_L)). \tag{21'}$$

We have thus shown that under rather weak assumptions about the density function of the random variable and about the error criterion, the uniform quantizer is asymptotically optimum.

Comparison with Rate Distortion Function

When E is mean-square error, it is well known[3] that the rate distortion satisfies

$$(\tfrac{1}{2}) \log (V_0/E) \leq R(E) \leq (\tfrac{1}{2}) \log (V/E) \tag{24}$$

where V and V_0 are the variance and entropy variance, respectively. Equation (23) shows that the uniform quantizer can always attain a performance asymptotically within approximately $\frac{1}{4}$ bit of the rate distortion lower bound.

When we use the more general error measure given by (4'), a lower bound on $R(E_L)$, which we will write as $r(E_L)$, is given by Shannon[4]

$$r(E_L) = H_0 - \phi(E_L) \tag{25}$$

where

$$\phi(E_L) = \sup_{\{p(u)\}} \left(-\int du\, p(u) \log (p(u)) \right) \tag{26}$$

subject to the constraints that $p(u)$ be a density and that

$$E_L = \int du\, L(u) p(u). \tag{27}$$

[3] See Shannon [1], p. 80.
[4] See Shannon [2], p. 120.

The solution of the variational problem specified by (26) and (27) is given by

$$p(u) = A \exp (-\lambda L(u)) \tag{28a}$$

where the constants A and λ are determined by the constraints. The corresponding value of $\phi(E_L)$ is

$$\phi(E_L) = -\log (A) - \lambda E_L \log (e). \tag{28b}$$

If we assume a loss function of the form

$$L(u) = |u|^a \tag{29a}$$

the corresponding value of ϕ is

$$\phi(E_L) = (1/a) \log (e \cdot a \cdot E_L) + \log (2\Gamma(1 + 1/a)). \tag{29b}$$

The same loss function substituted in (7'), (19'), and (21') gives

$$H_{\min} \approx H_0 - (1/a) \log ((1 + a)E_L) - \log (2), \quad \text{as} \quad E_L \to 0. \tag{30}$$

The asymptotic entropy of the uniform quantizer thus exceeds the lower bound on the rate distortion function by

$$H_{\min} - r(E_L) \approx (1/a) \log (a \cdot e/(1 + a)) + \log (\Gamma(1 + 1/a)). \tag{31}$$

Table I gives the value of the right-hand side of (31) for a few values of the loss exponent a. The last line of Table I indicates that for arbitrarily large loss exponents, the uniform quantizer achieves the rate distortion bound. This can be viewed as a verification of the intuitively satisfying notion of the optimality of uniform quantizing under a bounded error requirement.

We will now show that any difference between $H_{\min}$ and $r(E_L)$ is due to a limitation imposed by the quantization process rather than any inherent weakness in the lower bound $r(E_L)$. In fact, for a loss function which is positive and a monotonically increasing function of the magnitude of its argument, we will show

$$\lim_{E_L \to 0} [R(E_L) - r(E_L)] = 0. \tag{32}$$

First note that if X is the variable to be transmitted and Y its reconstruction, then by definition

$$R(E_L) = \min_{\{p(y|x)\}} \left[\iint dxdy\, p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \right] \tag{33}$$

where the expression in brackets is the mutual information between X and Y, and where the minimization is performed subject to the constraint

$$E_L = \iint dxdy\, L(x - y) p(x, y). \tag{34}$$

Thus, if we select some conditional density for which (34) is satisfied, the resulting mutual information between X and Y will provide an upper bound on $R(E_L)$.

As a specific choice for the conditional density, consider

$$p(y \mid x) = e^{-\lambda L(y-x)} \Big/ \int dv\, e^{-\lambda L(v)} \tag{35}$$

TABLE I

a	$H_{\min} - r(E_L)$ (in bits)
0	∞
1/20	2.085
1/10	1.624
1/5	1.196
1/4	1.068
1/3	0.913
1/2	0.715
1	0.443
2	0.255
3	0.179
4	0.138
5	0.113
10	0.059
20	0.030
∞	0

where λ is determined by (34) and is not dependent on $p_1(x)$, the probability density of X. Letting $R_b(E_L)$ denote the mutual information between X and Y, we obtain

$$R_b(E_L) = -\int dy \log (p_2(y))p_2(y) - \phi(E_L) \geq R(E_L) \quad (36)$$

where p_2 is the density of y and is given by

$$p_2(y) = \left[\int dx\, e^{-\lambda L(y-x)} p_1(x)\right] \Big/ \int dv\, e^{-\lambda L(v)}. \quad (37)$$

Now as $E_L \to 0$, we have

$$p_2(y) \to p_1(y)$$

since $p(y \mid x)$ approaches a delta function (for as $E_L \to 0$, $\lambda \to \infty$). Hence, for $p_1(y)$ sufficiently well behaved

$$\lim_{E_L \to 0} R_b(E_L)$$

$$= \lim_{E_L \to 0} \left[-\int dy \log (p_2(y))p_2(y) - \phi(E_L)\right] \quad (38a)$$

$$= H_0(x) - \phi(E_L) = r(E_L) \quad (38b)$$

which establishes (32).[5]

Comparison with Minimum-Alphabet Quantizer

As a practical matter, it may be desirable to place a restriction on the number of quantizer output levels (particularly when uniform quantizing implies an infinite number of levels) even though this may be reflected in increased entropy. Now, for any specified average error there is a minimum number of quantizer levels which can yield an error that small. We will refer to such a quantizer as a minimum-alphabet quantizer. Its output entropy provides an upper bound on the minimum entropy attainable with consistent simultaneous constraints on average error and alphabet size.

We will restrict our attention here to the case of mean-square error, and compare the entropies of the minimum-alphabet and minimum-entropy quantizers.

[5] This is a generalization of the result of Gerrish and Schultheiss [3], who considered the case where $L(u) = u^2$.

Let N be the number of quantizer output levels, and let (A, B) be the range of the density function. Then from (8) we can take

$$g(0) = A, \quad g(N\delta) = B$$

or

$$N = (g^{-1}(B) - g^{-1}(A))/\delta.$$

By writing this as

$$N = (1/\delta) \int_{g^{-1}(A)}^{g^{-1}(B)} dt$$

and making the substitution $s = g(t)$, we get

$$N = (1/\delta) \int_A^B ds/g'(g^{-1}(s))$$

or, from (15),

$$N = (1/\delta) \int_A^B ds/\gamma(s). \quad (39)$$

The minimization of N subject to the constraint that E is fixed and given by (16) leads to the known results [4] that

$$\gamma(s) = (f(s))^{-1/3} \int dt (f(t))^{1/3} \quad (40a)$$

$$E \approx (\delta^2/12)\left(\int dt (f(t))^{1/3}\right)^3. \quad (40b)$$

Substitution of (40a) in (17) then gives the asymptotic entropy of the minimum-alphabet quantizer, which we will denote by H_{maq}

$$H_{\text{maq}} \approx H_0 - \log(\delta) + (\tfrac{1}{3}) \int dt\, f(t) \log (f(t)) - \log\left(\int dt (f(t))^{1/3}\right). \quad (40c)$$

Finally, expressing the entropy in terms of the ms error,

$$H_{\text{maq}} \approx H_0 - (\tfrac{1}{2}) \log (12E) + (\tfrac{1}{2}) \log \left(\int dt (f(t))^{1/3}\right) + (\tfrac{1}{3}) \int dt\, f(t) \log (f(t)), \quad \text{as} \quad E \to 0.$$

The increase in entropy relative to the uniform quantizer can then be written as

$$2(H_{\text{maq}} - H_{\min}) \approx \log \left(\int dt\, f(t)(f(t))^{-2/3}\right) - \int dt\, f(t) \log ((f(t))^{-2/3}). \quad (41)$$

(The somewhat awkward way of writing (41) was chosen to emphasize the positivity of the difference, which follows immediately from the convexity of the logarithm.) Table II gives the value of $H_{\text{maq}} - H_{\min}$ for a few common distributions. (Strictly speaking, the values for distributions defined on the entire line or half-line should be assumed to refer to distributions approximately equal to the

TABLE II

$f(x)$	Range	$\Delta = H_{\max} - H_{\min}$ (in bits)
1	(0,1)	0
$2x$	(0,1)	0.052
$(n+1)x^n$, n large	(0,1)	0.312
$\log(1/x)$	(0,1)	0.122
$(2\pi)^{-1/2}\exp(-x^2/2)$	$(-\infty,\infty)$	0.156
$\exp(-x)$	$(0,\infty)$	0.312
$3x^{-4}$	$(1,\infty)$	0.944
$\alpha\exp(-x^\alpha)/\Gamma(1/\alpha)$	$(0,\infty)$	$0.312/\alpha$
$(2+\alpha)x^{-(3+\alpha)}$		0.571
(α very small)	$(1,\infty)$	$+(0.5)\log_2(1/\alpha)$
$\alpha x^{\alpha-1}$		$0.292 + 0.481/\alpha$
(α very small)	(0,1)	$-0.5\log_2(1/\alpha)$

tabulated ones, but truncated at some large absolute value, since the variational procedure which leads to (40) is based on boundedness of the random variable.) It can be seen that except for "peculiar" distributions, like the last three in Table II, or for distributions with infinite variance and finite entropy, the entropy of the minimum-alphabet quantizer is not strikingly larger than that of the uniform quantizer.

For completeness we include the asymptotic dependence of $\log N$ on E

$$\log N \approx (\tfrac{3}{2}) \log \left(\int ds (f(s))^{1/3} \right) - (\tfrac{1}{2}) \log (12E) \qquad (42a)$$

which can be found by substituting (40a) in (39). This relation indicates the transmitted digit rate required when entropy-reducing coding is not used. It can be verified by comparison with (41) that

$$\log N - H_{\min} = 3(H_{\max} - H_{\min}). \qquad (42b)$$

Quantizing of Sequences

Suppose that a stationary random process $X(t)$ is sampled every τ seconds to produce the finite sequence $X_1, X_2, \cdots, X_D$ with

$$X_n = X(n\tau) \qquad n = 1, \cdots, D. \qquad (43)$$

Then if each of these samples is quantized uniformly with step size δ, the average error is again

$$E_L \approx M(\delta) \qquad (44)$$

(ignoring any error introduced by the finite sampling frequency), whereas the average entropy per sample is now

$$H \approx -\log(\delta) + H_0/D \qquad (45)$$

where H_0 is the continuous entropy of the joint distribution of D consecutive samples

$$H_0 = -\int \cdots \int dx_1 \cdots dx_D f(\mathbf{x}) \log (f(\mathbf{x})). \qquad (46)$$

(We indicate vector-valued quantities here by bold-face notation.) For the case of independent samples, as from a band-limited, flat-spectrum Gaussian process sampled at twice its highest frequency, (45) of course yields the same entropy per sample as (20′).

We next observe that the uniform quantizing of the individual samples is equivalent to quantizing the D-dimensional vectors consisting of D consecutive samples by dividing D-space into D-cubes with sides of length δ parallel to the coordinate axes. A natural question to ask is whether any more efficient quantizing of D-space can be found. The answer to this question is yes, in general, but the concomitant problem of finding the optimum shapes in which to quantize D-space appears to be very difficult to solve.

Suppose that for the loss function we again adopt one which depends only on the pairwise coordinate differences; specifically, let

$$\Lambda(\mathbf{X} - \mathbf{Y}) = (1/D) \sum_{i=1}^{D} L(X_i - Y_i) \qquad (47)$$

be the error per coordinate incurred when the sequence $\mathbf{Y}$ is substituted for the sequence $\mathbf{X}$; correspondingly, E_L will be the expected value of $\Lambda(\mathbf{X} - \mathbf{Y})$. Then, if D-space is partitioned into translates of a region of shape S and volume V, it is easy to show, by the analog of the earlier derivation, that

$$H \approx H_0/D - \log (V)/D \quad \text{(entropy/coordinate)} \qquad (48)$$

and

$$E_L \approx M_S(V) \qquad (49a)$$

where

$$M_S(V) = (1/V) \int_S dx_1 \cdots dx_D \Lambda(\mathbf{x} - \mathbf{x}^0) \qquad (49b)$$

with $\mathbf{x}^0$ being that value which minimizes the integral. Using (49), the nonoptimality of D-cubes parallel to the coordinate axes can be verified for certain loss functions. For example, with the quadratic loss function $L(u) = u^2$, a covering of 2-space by hexagons is slightly better than a covering by squares. For another example, with the absolute loss function $L(u) = |u|$, a covering of 2-space by squares with diagonals parallel to the coordinate axes is slightly better than a covering with normally-oriented squares.

The effectiveness of quantizing with D-cubes can be gauged, in a manner analogous to the one-dimensional case, by making a comparison with the lower bound on the rate distortion function for the D-dimensional sequence $\mathbf{X}$. The lower bound for the loss function $\Lambda(\mathbf{X} - \mathbf{Y})$ parallels the one-dimensional case and is given by

$$r(E_L) = H_0/D - \phi(E_L) \quad \text{(entropy/coordinate)} \qquad (50)$$

where $\phi(E_L)$ is given by (26). Thus, if we let $H_C(E_L)$ denote the entropy of the D-cube quantizer output and take $L(u) = |u|^\alpha$, then the difference $H_C(E_L) - r(E_L)$ is asymptotically equal to the right-hand side of (31), and the calculations of Table I apply.

We should further note, by an extension of the arguments used previously, that in D-space

$$\lim_{E_L \to 0} [R(E_L) - r(E_L)] = 0. \qquad (51)$$

In general, the shape which minimizes $M_S(V)$ for a fixed volume V is the shape

$$S = \{\mathbf{x}: \Lambda(\mathbf{x}) \leq b\} \tag{52a}$$

with the constant b chosen to yield the specified volume V; however, in most cases it is impossible to cover D-space with such shapes without overlap. For the loss function $L(u) = |u|^{1/c}$, the volume of the shape in (52a) is[6]

$$V = (2b^cD^c\Gamma(1 + c))^D/\Gamma(1 + cD) \tag{52b}$$

and the normalized loss is

$$M_S(V) = bcD/(1 + cD) \tag{52c}$$

so that substituting in (48) and (49) we have

$$H \approx H_0/D - \log (2\Gamma(1 + c)) - c \log (E_L) - c \log (D) - c \log (1 + 1/cD) + (1/D) \log (\Gamma(1 + cD)) \tag{52d}$$

which, when D is large, approaches

$$H \approx H_0/D - \log (2\Gamma(1 + c)) + c \log (c/eE_L). \tag{52e}$$

If we let $c = 1/a$ and compare (52e) with (29b), we see that the two expressions are identical. Thus, for this loss function the discrepancy between the entropy with quantizing and the rate distortion bound rests entirely on the inability of the optimum shapes to cover spaces of large dimensionality.

Appendix I

Convergence Implied by (19) and (20)

Suppose that the density $f(x)$ satisfies the following three conditions:

1) $f(x)$ is continuous except at finitely many points. (53a)

2) If x_0 is a point of discontinuity of f, then for some K and some $a < 1$,

$$|f(x)| < K |x - x_0|^{-a} \quad \text{as} \quad x \to x_0. \tag{53b}$$

3) For some K and some $a > 0$,

$$|f(x)| < K |x|^{-1-a} \quad \text{as} \quad |x| \to \infty. \tag{53c}$$

Then we will show that (19) and (20) can be rewritten as

$$E/\delta^2 \to 1/12 \quad \text{as} \quad \delta \to 0 \tag{54}$$

and

$$H_{\min} + \log \delta \to H_0 \quad \text{as} \quad \delta \to 0. \tag{55}$$

Let

$$f^*(x) = (1/\delta) \int_{n\delta}^{(n+1)\delta} dy\, f(y) \tag{56}$$

if $n\delta \leq x < (n + 1)\,\delta$, for all n. Then

$$H_{\min} + \log \delta = \int dx\, f^*(x) \log (f^*(x)). \tag{57}$$

[6] Refer to Appendix II for derivations of (52b) and (52c).

From (53a), $f^*(x) \to f(x)$ almost everywhere as $\delta \to 0$, from which it also follows that

$$f^*(x) \log (f^*(x)) \to f(x) \log (f(x)) \tag{58}$$

almost everywhere as $\delta \to 0$. Furthermore, the conditions (53b) and (53c) are sufficient to guarantee that the integrands of (57) are dominated in magnitude by an integrable function that is independent of δ. The result (54) follows by dominated convergence.

For the mean-square error we introduce the auxiliary function

$$\eta(x) = (x - n - \tfrac{1}{2})^2/\delta^2 \tag{59}$$

if $n\delta \leq x < (n + 1)\delta$, for all n, and write

$$E/\delta^2 = \int dx\, f(x)\eta(x). \tag{60}$$

Let $x_1, \cdots, x_m$ be the points of discontinuities of $f(x)$. Then, since f is integrable, given $\epsilon > 0$ there exists an A with $0 < A < 1$ such that

$$\int_S dx\, f(x) \leq \epsilon \tag{61a}$$

where

$$S = \{x: |x - x_m| < A \text{ for some } m\} \cup \{x: |x| > 1/A\}. \tag{61b}$$

Restrict attention to $\delta < A/2$, and set

$$S_\delta = \text{union of those } \delta\text{-intervals wholly contained in } S. \tag{62}$$

Next, let

$$\eta^*(x) = \begin{cases} \eta(x) & \text{if} \quad x \in \bar{S}_\delta \\ 1/12 & \text{if} \quad x \in S_\delta \end{cases} \tag{63}$$

where the overbar means set complement. Then, we can approximate (60) by substituting η^* for η

$$\left| E/\delta^2 - \int dx\, f(x)\eta^*(x) \right| \leq \epsilon/6. \tag{64}$$

Now $\bar{S}_\delta$ contains intervals of total length less than $4/A$, and since f is continuous on its closure, it is uniformly continuous there, and we can take $\Delta > 0$ such that

$$|f(x) - f^*(x)| < 5A\epsilon/6 \quad \text{if} \quad \delta \leq \Delta. \tag{65}$$

Finally,

$$\int dx\, f(x)\eta^*(x) = \int_{S_\delta} dx\, f(x)\eta^*(x) + \int_{\bar{S}_\delta} dx\, f^*(x)\eta^*(x) + \int_{\bar{S}_\delta} dx(f(x) - f^*(x))\eta^*(x).$$

However, the sum of the first two integrals can easily be shown to be 1/12, whereas (65) provides a bound on the magnitude of the third integral, and we have

$$\left| \int dx\, f(x)\eta^*(x) - 1/12 \right| \leq 5\epsilon/6 \quad \text{if} \quad \delta \leq \Delta. \tag{66}$$

Combining (66) and (64) completes the proof of (54).

The derivations given here carry over directly to the analogous expressions for (19′) and (20′).

Appendix II

Outline of Rigorous Derivation of (21)

In Appendix I we were able to establish conditions under which it could be rigorously shown that if H_U is the entropy of the uniform quantizer, and E is its mean-square error, then

$$H_U - \log \delta \to H_0 \quad \text{as} \quad \delta \to 0$$

and

$$E/\delta^2 \to 1/12 \quad \text{as} \quad \delta \to 0$$

from which it can be deduced that

$$H_U - \log \sqrt{12E} \to H_0 \quad \text{as} \quad E \to 0.$$

It was not shown rigorously that this was indeed the best performance attainable, i.e., that

$$H_{\min} - \log \sqrt{12E} \to H_0 \quad \text{as} \quad E \to 0.$$

Unfortunately, the heuristic derivation given in the text is almost impossible to make rigorous. Here we give an outline of a derivation which can be made rigorous.

Starting from the definition of H_0, we have

$$H_0 = \sum_n \int_{I_n} dx\, f(x) \log (1/f(x)). \tag{67}$$

(The notation used here is consistent with that of (1) to (7), with the additional notation δ_n = length of I_n.) Noting that

$$(\partial^2/\partial f^2)(-f \log f) < 0 \tag{68}$$

it follows easily that

$$H_0 \leq \sum_n p_n \log (\delta_n/p_n) \tag{69}$$

or

$$H_0 \leq H + \sum_n p_n \log \delta_n. \tag{70}$$

Writing this as

$$H_0 \leq H + (\tfrac{1}{2}) \sum_n p_n \log (\delta_n^2) \tag{71}$$

and noting that $\{p_n\}$ is a discrete distribution and that

$$(\partial^2/\partial y^2) \log y < 0$$

we have

$$H_0 \leq H + (\tfrac{1}{2}) \log (\sum_n p_n \delta_n^2). \tag{72}$$

But

$$\sum_n p_n \delta_n^2 \approx 12E \tag{73}$$

so

$$H_0 \leq H + (\tfrac{1}{2}) \log (12E)$$

approximately. Since the result is independent of H, we have

$$H_{\min} + (\tfrac{1}{2}) \log (12E) \geq H_0 \tag{74}$$

approximately. This outline can serve as the basis of a rigorous proof. Specifically, it is possible to show that if

$$f(x) \text{ is uniformly continuous on its domain} \tag{75a}$$

and

$$f(x) \log f(x) \text{ is absolutely integrable} \tag{75b}$$

then

$$\liminf_{E \to 0} (H_{\min} + (\tfrac{1}{2}) \log (12E)) \geq H_0. \tag{75c}$$

Combining this with the result of Appendix I shows that indeed

$$\lim_{E \to 0} (H_{\min} + (\tfrac{1}{2}) \log (12E)) = H_0. \tag{76}$$

The complete proof is surprisingly long and will not be given here. The essential modification of the outline is to find a set S which is a finite disjoint union of bounded open sets, with f uniformly larger than some positive number on S and such that most of the contribution to both

$$\int f \log f \quad \text{and} \quad \int f$$

arises from the integral over S. Both the entropy and mean-square error for an arbitrary partition can then be bounded below by the entropy and mean-square error for the partition induced in S, with small corrections whose size is controllable by the uniform continuity.

Appendix III

Derviation of (52b) and (52c)

From (52a) and (47)

$$S = \left\{ \mathbf{x} : \sum_{j=1}^{D} |x_j|^{1/c} \leq bD \right\} \tag{77}$$

so that

$$V = \int_S \cdots \int dx_1 \cdots dx_D. \tag{78}$$

Rewrite (78) as

$$V = 2^D \int_0^\infty \cdots \int dx_1 \cdots dx_D g(\mathbf{x}) \tag{79a}$$

where

$$g(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{D} x_j^{1/c} \leq bD \\ 0 & \text{otherwise} \end{cases}. \tag{79b}$$

Now, using the standard inverse Laplace transform relation

$$\int_{\sigma-i\infty}^{\sigma+i\infty} dw\, w^{-1} \exp(-wt)(\exp(w\mu) - 1) = \begin{cases} 1 & \text{if } t \in (0, \mu) \\ 0 & \text{otherwise} \end{cases} \quad (80)$$

which holds for any real σ, we can write $g(\mathbf{x})$ as

$$g(\mathbf{x}) = \int_{\sigma-i\infty}^{\sigma+i\infty} dw\, w^{-1}(\exp(wbD) - 1) \cdot \exp\left(-w \sum_{i=1}^{D} x_i^{1/c}\right). \quad (81)$$

Substituting this in (79b) and changing the order of integration (which is permissible if $\sigma > 0$), we have

$$V = 2^D \int_{\sigma-i\infty}^{\sigma+i\infty} dw\, w^{-1}(\exp(wb\,D) - 1) \cdot \prod_{i=i}^{D} \int_0^\infty dx_i \exp(-wx_i^{-1/c}) \quad \text{if} \quad \sigma > 0. \quad (82)$$

The integrations in the product can be accomplished by a substitution of the form $x = y^c$ to obtain

$$\int_0^\infty dx \exp(-wx^{-1/c}) = w^{-c}\Gamma(1 + c).$$

We then have

$$V = (2\Gamma(1 + c))^D \int_{\sigma-i\infty}^{\sigma+i\infty} dw\, w^{-cD-1} \cdot \exp((wb\,D) - 1) \quad \text{with} \quad \sigma > 0. \quad (83)$$

This again is recognizable as an inverse Laplace transform which can be evaluated as

$$V = (2\Gamma(1 + c))^D(bD)^{cD}/\Gamma(cD + 1) \quad (84)$$

which, with a trivial rewriting, is (52b).

Now, to arrive at (52c), let

$$S_\beta = \left\{\mathbf{x} : \sum_{i=1}^{D} |x_i|^{1/c} \leq \beta D\right\} \quad (85a)$$

and let

$$V(\beta) = \text{volume of } S_\beta. \quad (85b)$$

The $V(\beta)$ is given by (84) with the trivial substitution of β for b. But (49b) can be rewritten as

$$M_s(V) = (1/V) \int_0^b \beta\, dV(\beta). \quad (86)$$

If we rewrite (84) as

$$V = K\, b^{cD} \quad (87a)$$

where K includes all the multiplicative constants, then

$$V(\beta) = K\, \beta^{cD} \quad (87b)$$

and

$$M_s(V) = b^{-cD} \int_0^b \beta\, cD\beta^{cD-1}\, d\beta \quad (87c)$$

which yields (52c).

References

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, Ill.: University of Illinois Press, 1959.

[2] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Information and Decision Processes*, R. E. Machol, Ed. New York: McGraw-Hill, 1960.

[3] A. M. Gerrish and P. M. Schultheiss, "Information rates of non-Gaussian processes," *IEEE Trans. Information Theory*, vol. IT-10, pp. 265–271, October 1964.

[4] J. Max, "Quantizing for minimum distortion," *IRE Trans. Information Theory*, vol. IT-6, pp. 7–12, March 1960.

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

[6] T. J. Goblick, Jr. and J. L. Holsinger, "Analog source digitization: A comparison of theory and practice," *IEEE Trans. Information Theory (Correspondence)*, vol. IT-13, pp. 323–326, April 1967.

[7] J. T. Pinkston, III, "Encoding independent sample information sources," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, September 1967.

[8] F. Jelinek, "Evaluation of distortion rate functions for low distortions," *Proc. IEEE*, vol. 55, pp. 2067–2068, November 1967.