THE BELL SYSTEM

Technical Journal

Volume 48 November 1969 Number 9

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, B. S. Gilmer, President, C. E. Wampler, Vice President and Secretary, J. J. Scanlon, Vice President and Treasurer. Checks for subscriptions should be made payable to American Telephone and Telegraph Company and should be addressed to the Treasury Department, Room 2312C, 195 Broadway, New York, N. Y. 10007. Subscriptions $7.00 per year; single copies $1.25 each. Foreign postage $1.00 per year; 15 cents per copy. Printed in U.S.A.

THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 48 November 1969 Number 9

Coupled Wave Theory for Thick Hologram Gratings

By HERWIG KOGELNIK

(Manuscript received May 23, 1969)

A coupled wave analysis is given of the Bragg diffraction of light by thick hologram gratings, which is analogous to Phariseau's treatment of acoustic gratings and to the "dynamical" theory of X-ray diffraction. The theory remains valid for large diffraction efficiencies where the incident wave is strongly depleted. It is applied to transmission holograms and to reflection holograms. Spatial modulations of both the refractive index and the absorption constant are allowed for. The effects of loss in the grating and of slanted fringes are also considered. Algebraic formulas and their numerical evaluations are given for the diffraction efficiencies and the angular and wavelength sensitivities of the various hologram types.

I. INTRODUCTION

Holographic recording in thick media ("volume recording") is of particular interest for high-capacity information storage,[1-3] for color holography[4] and for efficient white-light display of holograms.[5-9] The high efficiency of light conversion which is attainable with thick dielectric holograms is also important for microimaging, and it may make it practical to use holographic optical components (for example, gratings or fly's eye lenses) in a variety of optical systems.

In thick holograms it is light diffraction at or near the Bragg angle which leads to efficient wavefront reconstruction. This is true for both transmission and reflection holograms, and both types are considered in this paper. The (volume) record of the holographic interference pattern (fringe pattern) usually takes the form of a spatial modulation of the absorption constant or the refractive index of the medium, or both. Modulations of the absorption constant are produced in conventional photographic emulsions and in photochromics, while newer materials, like dichromated gelatin[10,11] lithium niobate,[12] or photopolymer materials[13] yield modulations of the refractive index.

This paper considers the properties of all these types of thick (or "deep") holograms. Of particular interest is their efficiency of converting light into the useful reconstructed wave (diffraction efficiency) and the angular dependence of this diffraction efficiency as the incident light deviates from the Bragg angle. We are also interested in the wavelength dependence and in the way the diffraction properties are changed in the presence of loss or a slant of the fringe pattern with respect to the surface of the recording medium.

Leith and his associates, and Gabor and Stroke have already considered some of the properties of thick holograms, in particular the angular and the wavelength dependence of the diffracted light.[14,15] Their theories are essentially linear or perturbational theories which use the Kirchhoff integral or the first Born approximation with the basic assumption that the incident light wave is not disturbed by the diffraction process. Their results are valid as long as this assumption is good. For high diffraction efficiencies (like 90 percent) the incident wave is strongly depleted and another approach is called for. One such approach is to use electronic computers to solve the relevant complicated electromagnetic problem accurately. Results of such computations are available for special cases. Klein, Tipnis, and Hiedemann have computed data for light diffraction by ultrasonic waves,[16,17] and Burckhardt has reported results for dielectric hologram gratings.[18,19] The method of Bathia and Noble[20] is another approach in which they employed integral equations to analyze acoustic diffraction of light.

Yet another approach is the use of a coupled wave theory, which is the subject of this paper. Such a theory can predict the maximum possible efficiencies of the various hologram types (results which one cannot hope to obtain from linear theories), and the angular and wavelength dependence at high diffraction efficiencies. Following Phariseau,[21] coupled wave theories have been successfully used in the treatment of light diffraction by acoustic waves[22] and by electrooptic gratings[23] where very similar diffraction processes are at work as in holography.

Closely related to the diffraction in thick holograms are also the diffraction of electrons in lattices and the diffraction of X-rays in crystals. The dynamical theory of X-ray diffraction[24] is also a theory of coupled waves and its application to holography has already been suggested.[25]

We have earlier reported some of the results and an outline of the coupled wave theory for hologram gratings.[26,27] Here we propose to give further results and a more detailed account of the basic assumptions and the analysis. We give analytic formulas for the various hologram types as well as numerical evaluations which include results on the angular sensitivities and the influence of loss and slant.

For simplicity the analysis is restricted to the holographic record of *sinusoidal* fringe patterns which we call hologram gratings. To some degree a more complicated hologram can be regarded as a multiplicity of such hologram gratings.

II. COUPLED WAVE ANALYSIS

2.1 *Derivation of the Coupled Wave Equations*

The coupled wave theory assumes monochromatic light incident on the hologram grating at or near the Bragg angle and polarized perpendicular to the plane of incidence.* Only two significant light waves are assumed to be present in the grating: the incoming "reference" wave R and the outgoing "signal" wave S. Only these two waves obey the Bragg condition at least approximately, the other diffraction orders violate the Bragg condition strongly and are neglected. They should be of little influence on the energy interchange between S and R. The last assumption limits the validity of the coupled wave theory to *thick* hologram gratings. Section 6 gives a more detailed discussion of this limitation.

Figure 1 shows the model of a hologram grating which is used for our analysis. The z-axis is chosen perpendicular to the surfaces of the medium, the x-axis in the plane of incidence and parallel to the medium boundaries and the y-axis perpendicular to the paper. The fringe planes are oriented perpendicular to the plane of incidence and slanted with respect to the medium boundaries at an angle ϕ. The fringes are shown dotted. The grating vector $\mathbf{K}$ is oriented perpendicular to the fringe planes and is of length $K = 2\pi/\Lambda$, where Λ is the period of the grating. The same average dielectric constant is assumed for the region inside and outside the grating boundaries. The angle of incidence measured *in* the medium is θ.

* A generalization to parallel polarization is given in the appendix.

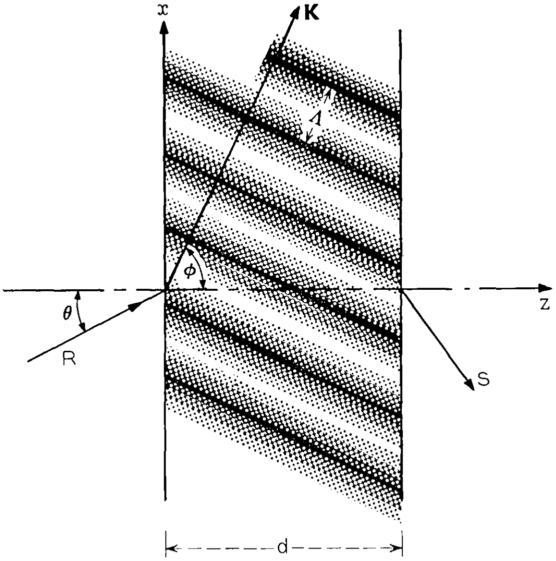


Fig. 1—Model of a thick hologram grating with slanted fringes. The spatial modulation of n or α is indicated by the dotted pattern. The grating parameters are: θ—angle of incidence in the medium, K—grating vector (perpendicular to the fringe planes), Λ—grating period, ϕ—slant angle, and d—grating thickness.

Wave propagation in the grating is described by the scalar wave equation

$$\nabla^2 E + k^2 E = 0, \tag{1}$$

where $E(x, z)$ is the complex amplitude of the y-component of the electric field, which is assumed to be independent of y and to oscillate with an angular frequency ω. The propagation constant $k(x, z)$ is spatially modulated and related to the *relative* dielectric constant $\epsilon(x, z)$ and the conductivity $\sigma(x, z)$ of the medium by

$$k^2 = \frac{\omega^2}{c^2} \epsilon - j\omega\mu\sigma \tag{2}$$

where c is the light velocity in free space and μ is the permeability of the medium which we assume to be equal to that of free space. In our model the constants of the medium are independent of y. The fringes of the hologram grating are represented by a spatial modulation of ϵ or σ:

$$\begin{aligned} \epsilon &= \epsilon_0 + \epsilon_1 \cos(\mathbf{K}\cdot\mathbf{x}) \\ \sigma &= \sigma_0 + \sigma_1 \cos(\mathbf{K}\cdot\mathbf{x}) \end{aligned} \tag{3}$$

where ϵ_1 and σ_1 are the amplitudes of the spatial modulation, ϵ_0 is the average dielectric constant and σ_0 the average conductivity. ϵ and σ are assumed to be modulated in phase. To simplify the notation we have used the radius vector $\mathbf{x}$ and the grating vector $\mathbf{K}$

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}; \quad \mathbf{K} = K \begin{bmatrix} \sin\phi \\ 0 \\ \cos\phi \end{bmatrix}; \quad K = 2\pi/\Lambda.$$

Equations (2) and (3) can be combined in the form

$$k^2 = \beta^2 - 2j\alpha\beta + 2\kappa\beta(e^{j\mathbf{K}.\mathbf{x}} + e^{-j\mathbf{K}.\mathbf{x}}) \tag{4}$$

where we have introduced the average propagation constant β and the average absorption constant α:

$$\beta = 2\pi(\epsilon_0)^{\frac{1}{2}}/\lambda; \qquad \alpha = \mu c\sigma_0/2(\epsilon_0)^{\frac{1}{2}}, \tag{5}$$

and the coupling constant κ was defined as

$$\kappa = \frac{1}{4}\left(\frac{2\pi}{\lambda}\epsilon_1/(\epsilon_0)^{\frac{1}{2}} - j\mu c\sigma_1/(\epsilon_0)^{\frac{1}{2}}\right). \tag{6}$$

This coupling constant describes the coupling between the reference wave R and the signal wave S. It is the central parameter in the coupled wave theory. For $\kappa = 0$ there is no coupling between R and S and, therefore, there is no diffraction.

Optical media are usually characterized by their refractive index and their absorption constant. We also find it convenient to use these parameters if the following conditions are met

$$2\pi n/\lambda \gg \alpha; \qquad 2\pi n/\lambda \gg \alpha_1, \qquad n \gg n_1, \tag{7}$$

which is true in almost every practical case. Here n is the average refractive index, and n_1 and α_1 are the amplitudes of the spatial modulation of the refractive index and the absorption constant, respectively [compare equation (3)]. λ is the wavelength in free space. Under the above conditions we can write with good accuracy

$$\beta = 2\pi n/\lambda \tag{8}$$

and for the coupling constant

$$\kappa = \pi n_1/\lambda - j\alpha_1/2. \tag{9}$$

The spatial modulation indicated by n_1 or α_1 forms a grating which couples the two waves R and S and leads to an exchange of energy between them. We describe these waves by complex amplitudes $R(z)$ and $S(z)$ which vary along z as a result of this energy interchange or because of an energy loss from absorption. The total electric field in the grating is the superposition of the two waves:

$$E = R(z)e^{-j\boldsymbol{\rho}\cdot\mathbf{x}} + S(z)e^{-j\boldsymbol{\sigma}\cdot\mathbf{x}}. \tag{10}$$

The propagation vectors $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ contain the information about the propagation constants and the directions of propagation of R and S. $\boldsymbol{\rho}$ is assumed to be equal to the propagation vector of the free reference wave in the absence of coupling. $\boldsymbol{\sigma}$ is forced by the grating and related to $\boldsymbol{\rho}$ and the grating vector by

$$\boldsymbol{\sigma} = \boldsymbol{\rho} - \mathbf{K} \tag{11}$$

which has the appearance of a conservation of momentum equation. $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ have been chosen to conform as closely as possible with our picture of the physical process of the diffraction in the grating. If the actual phase velocities differ somewhat from the assumed values, then these differences will appear in the complex amplitudes $R(z)$ and $S(z)$ as a result of the theory.

Figure 2 shows the vectors of interest and their orientation. The components of $\boldsymbol{\rho}$ are ρ_x and ρ_z which are given by

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_x \\ 0 \\ \rho_z \end{bmatrix} = \beta \begin{bmatrix} \sin\theta \\ 0 \\ \cos\theta \end{bmatrix}. \tag{12}$$

From this and equation (11) follow the $\boldsymbol{\sigma}$-components σ_x and σ_z

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_x \\ 0 \\ \sigma_z \end{bmatrix} = \beta \begin{bmatrix} \sin\theta - \dfrac{K}{\beta}\sin\phi \\ 0 \\ \cos\theta - \dfrac{K}{\beta}\cos\phi \end{bmatrix}. \tag{13}$$

The vector relation (11) is shown in Fig. 3 together with a circle of radius β. The general case is shown in Fig. 3a, where the Bragg

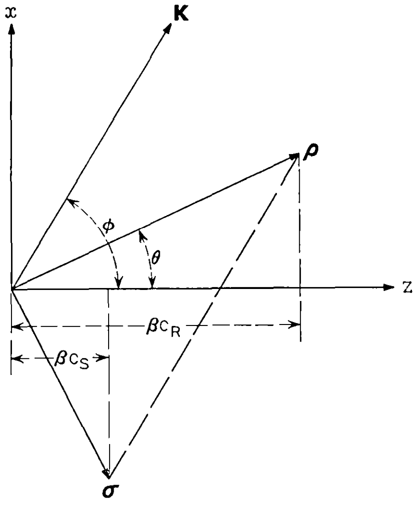


Fig. 2 — ϱ and σ, the propagation vectors of the reference wave R and the signal wave S, and their relation to the grating vector **K**. The obliquity factors c_R and c_S are indicated.

condition is not met and the length of σ differs from β. Figure 3b shows the same diagram for incidence at the Bragg angle θ_0 . In this special case the lengths of both, ϱ and σ are equal to the free propagation constant β, and the Bragg condition

$$\cos (\phi - \theta) = K/2\beta \tag{14}$$

is obeyed.

For a fixed wavelength the Bragg condition is violated by angular

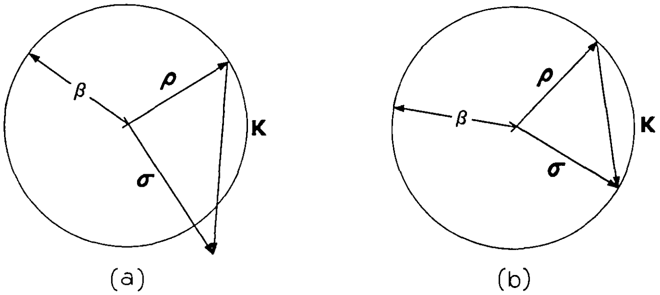


Fig. 3 — Vector diagram (conservation of momentum) for (a) near and (b) exact Bragg incidence.

deviations $\Delta\theta$ from the Bragg angle θ_0. For a fixed angle of incidence a similar violation takes place for changes $\Delta\lambda$ from the correct wavelength λ_0. We write

$$\theta = \theta_0 + \Delta\theta, \tag{15}$$

and

$$\lambda = \lambda_0 + \Delta\lambda,$$

and assume in the following that the deviations $\Delta\theta$ and $\Delta\lambda$ are small.

Angular changes $\Delta\theta$ have very similar effects on the behavior of the grating as wavelength changes $\Delta\lambda$, and there is a close relation between the angular sensitivity and the wavelength sensitivity of thick hologram gratings. We get an idea of this relationship by differentiating the Bragg condition (14), from which results

$$\frac{d\theta_0}{d\lambda_0} = K/4\pi n \sin(\phi - \theta_0). \tag{16}$$

The $\theta - \lambda$ connection shows up in the dephasing measure ϑ which appears in the coupled wave equations and which is defined by

$$\vartheta \equiv (\beta^2 - \sigma^2)/2\beta = K \cos(\phi - \theta) - \frac{K^2}{4\pi n}\lambda \tag{17}$$

and which has been expressed in this form using equation (13). A Taylor series expansion of equation (17) yields the following expression for ϑ which is correct to the first order in the deviations $\Delta\theta$ and $\Delta\lambda$:

$$\vartheta = \Delta\theta \cdot K \sin(\phi - \theta_0) - \Delta\lambda \cdot K^2/4\pi n. \tag{18}$$

Note that the deviations $\Delta\theta$ and $\Delta\lambda$ which produce equal dephasing ϑ are related by equation (16).

We are now ready to derive the coupled wave equations. We combine equations (1) and (4), and insert the expressions of (10) and (11). Then we compare the terms with equal exponentials ($e^{-j\boldsymbol{\rho}\cdot\mathbf{x}}$ and $e^{-j\boldsymbol{\sigma}\cdot\mathbf{x}}$) and arrive at

$$R'' - 2jR'\rho_z - 2j\alpha\beta R + 2\kappa\beta S = 0 \tag{19}$$

and

$$S'' - 2jS'\sigma_z - 2j\alpha\beta S + (\beta^2 - \sigma^2)S + 2\kappa\beta R = 0, \tag{20}$$

where the primes indicate differentiation with respect to z. The waves generated in the directions of $\boldsymbol{\rho} + \mathbf{K}$ and $\boldsymbol{\sigma} - \mathbf{K}$ are neglected, together with all other higher diffraction orders. In addition we assume that the energy interchange between S and R is slow and that energy is absorbed slowly, if at all. This allows us to neglect R'' and S''. We will check the results of the theory later for a more detailed justification of this last step. We can now introduce equation (18) and rewrite the above equations in the form

$$c_R R' + \alpha R = -j\kappa S \tag{21}$$

$$c_S S' + (\alpha + j\vartheta)S = -j\kappa R. \tag{22}$$

These are the coupled wave equations which are the basis for our analysis. The abbreviations c_R and c_S stand for the expressions

$$\begin{aligned} c_R &= \rho_z/\beta = \cos\theta \\ c_S &= \sigma_z/\beta = \cos\theta - \frac{K}{\beta}\cos\phi. \end{aligned} \tag{23}$$

Our physical picture of the diffraction process is reflected in the coupled wave equations. A wave changes in amplitude along z because of coupling to the other wave (κR, κS) or absorption (αR, αS). For deviations from the Bragg condition S is forced out of synchronism with R and the interaction decreases (ϑS).

The energy balance of the coupled-wave model is described by the relation

$$(c_R RR^* + c_S SS^*)' + 2\alpha(RR^* + SS^*) + j(\kappa - \kappa^*)(RS^* + R^*S) = 0 \tag{24}$$

where the asterisk denotes a complex conjugate. This is easily derived from equations (21) and (22) by multiplying them with R^* and S^*, respectively, and adding the results together with the complex conjugate results. The presence of the obliquity factors c_R and c_S in the first part of equation (24) indicates that it is the power flow of the two waves *in the z direction* that enters the energy balance. In the absence of ohmic loss this power flow is conserved. The second and the third part in the equation describe the energy loss resulting from absorption in the grating. They correspond to the relevant terms of σEE^*.

2.2 *Solution of the Coupled Wave Equations*

It is straight forward to obtain the general solution of the coupled wave equations, which is

$$R(z) = r_1 \exp(\gamma_1 z) + r_2 \exp(\gamma_2 z) \tag{25}$$

$$S(z) = s_1 \exp(\gamma_1 z) + s_2 \exp(\gamma_2 z) \tag{26}$$

where the r_i and s_i are constants which depend on the boundary conditions. To determine the constants γ_i we insert equations (25) and (26) into the coupled wave equations and obtain

$$(c_R\gamma_i + \alpha)r_i = -j\kappa s_i \tag{27}$$

$$i = 1, 2$$

$$(c_S\gamma_i + \alpha + j\vartheta)s_i = -j\kappa r_i \;. \tag{28}$$

After multiplying the equations with each other we get a quadratic equation for γ_i

$$(c_R\gamma_i + \alpha)(c_S\gamma_i + \alpha + j\vartheta) = -\kappa^2, \tag{29}$$

with the solution

$$\gamma_{1,2} = -\frac{1}{2}\left(\frac{\alpha}{c_R} + \frac{\alpha}{c_S} + j\frac{\vartheta}{c_S}\right) \pm \frac{1}{2}\left[\left(\frac{\alpha}{c_R} - \frac{\alpha}{c_S} - j\frac{\vartheta}{c_S}\right)^2 - 4\frac{\kappa^2}{c_Rc_S}\right]^{\frac{1}{2}}. \tag{30}$$

At this point we divert briefly from the main derivation, because now we have the means to check the validity of neglecting R'' and S'' in Section 2.1. This step is justified if the conditions $R'' \ll \zeta_z R'$, and $S'' \ll \sigma_z S'$ are obeyed. In view of equations (25) and (26) this will happen if $\gamma_i \ll \beta$. According to equation (30) the above requirement is met if $\Delta\theta \ll 1$ and if the inequalities of equation (7) are satisfied (which is usually the case).

Continuing the coupled wave analysis, we have to determine the constants r_i and s_i . To do this we have to introduce boundary conditions into our model. These are different for transmission holograms and for reflection holograms. Figure 4 gives an indication of this. For both hologram types the reference wave R is assumed to start with unit amplitude at $z = 0$. It decays as it propagates to the right and couples energy into S. In transmission holograms the signal S starts out with zero amplitude at $z = 0$ and propagates to the right ($c_S > 0$). In reflection holograms the signal travels to the left ($c_S < 0$) and it starts with zero amplitude at $z = d$.

Let us first analyze transmission holograms where $c_S > 0$. Here, the boundary conditions are

$$R(0) = 1, \qquad S(0) = 0 \tag{31}$$

as discussed before. If we insert these boundary conditions into equa-

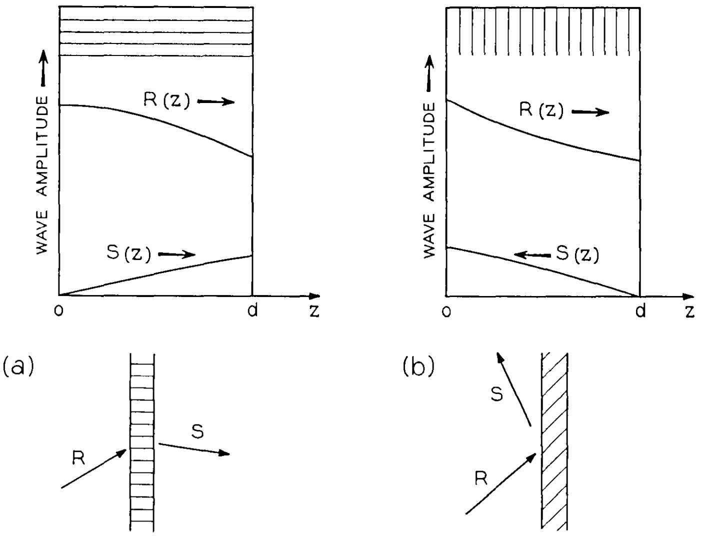


Fig. 4 — Wave propagation in (a) transmission and (b) reflection holograms. The reference wave R decays while it propagates to the right. In (a) the signal S travels to the right and gains with z, while in (b) S travels to the left and gains with decreasing z. The shading indicates the orientation of the fringes.

tions (25) and (26),it follows immediately that

$$r_1 + r_2 = 1,$$

and (32)

$$s_1 + s_2 = 0.$$

Combining these relations with equation (28) we obtain

$$s_1 = -s_2 = -j\kappa/c_S(\gamma_1 - \gamma_2). \tag{33}$$

Introducing these constants in equation (26) we arrive at an expression for the amplitude of the signal wave at the output of the grating

$$S(d) = j\frac{\kappa}{c_S\,(\gamma_1 - \gamma_2)}\,(\exp\,(\gamma_2 d) - \exp\,(\gamma_1 d)). \tag{34}$$

This is a general expression, which is valid for all types of thick transmission holograms including the cases of off-Bragg incidence, of lossy gratings and of slanted fringe planes.

The analysis of reflection holograms follows a pattern similar to the above. We have $c_S < 0$ and boundary conditions given by

$$R(0) = 1, \qquad S(d) = 0. \tag{35}$$

The output plane for the signal wave is, now, at $z = 0$, and $S(0)$ is the output amplitude of interest. Inserting the boundary conditions in equations (25) and (26) yields

$$r_1 + r_2 = 1$$

and (36)

$$s_1 \exp(\gamma_1 d) + s_2 \exp(\gamma_2 d) = 0.$$

To proceed with our derivation we rewrite the above relation for s_1 and s_2 in the form

$$\begin{aligned} s_1(\exp(\gamma_2 d) - \exp(\gamma_1 d)) &= (s_1 + s_2)\exp(\gamma_2 d) \\ s_2(\exp(\gamma_2 d) - \exp(\gamma_1 d)) &= -(s_1 + s_2)\exp(\gamma_1 d). \end{aligned} \tag{37}$$

Then we sum equation (28) for $i = 1$ and $i = 2$ and obtain the relation

$$-j\kappa(r_1 + r_2) = -j\kappa = (s_1 + s_2)(\alpha + j\vartheta) + c_S(\gamma_1 s_1 + \gamma_2 s_2). \tag{38}$$

Using the relations (37) to substitute the sum $(s_1 + s_2)$ for the terms s_1 and s_2 in this equation we finally arrive at the result for the amplitude $S(0)$ of the output signal of a reflection hologram

$$S(0) = s_1 + s_2 = -j\kappa \bigg/ \left\{\alpha + j\vartheta + c_S \frac{\gamma_1 \exp(\gamma_2 d) - \gamma_2 \exp(\gamma_1 d)}{\exp(\gamma_2 d) - \exp(\gamma_1 d)}\right\}. \tag{39}$$

This is, again, a formula of quite general validity, including off-Bragg incidence, loss, and slant.

In the following sections we discuss the behavior of transmission and reflection holograms in greater detail, using the general formulas derived above. In these discussions a parameter of prime interest is the diffraction efficiency η, which is defined as

$$\eta = \frac{|c_S|}{c_R} SS^* \tag{40}$$

where S is the (complex) amplitude of the output signal for a reference wave R incident with unit amplitude. η is the fraction of the incident light power which is diffracted into the signal wave. S is equal to $S(d)$ for transmission holograms and equal to $S(0)$ for reflection holograms in the notation of this section. But for reasons of simplicity we omit the arguments in the following sections. The obliquity factors c_R and c_S appear in the above definition for the same reason they have appeared in the energy balance of equation (24): in the absence of loss it is the power flow in the z direction which is conserved.

For slanted gratings another important parameter is the slant factor c which is defined as the ratio between the obliquity factors

$$c = c_R/c_S = -\cos\theta/\cos(\theta_0 - 2\phi)$$

which we have expressed here, for Bragg incidence, in terms of the angle of incidence θ_0 and the slant angle ϕ. Figure 5 indicates lines of constant c as a function of θ_0 and ϕ. For transmission holograms c is positive ($c > 0$), and for reflection holograms c is negative ($c < 0$). In the diagram transmission and reflection holograms are separated by the line for $c = \infty$.

III. TRANSMISSION HOLOGRAMS

In this section we discuss transmission holograms in greater detail. We give algebraic formulas and their numerical evaluations for the diffraction efficiencies and the angular and wavelength sensitivities of dielectric and of absorption gratings. This includes results on the influence of loss and slant.

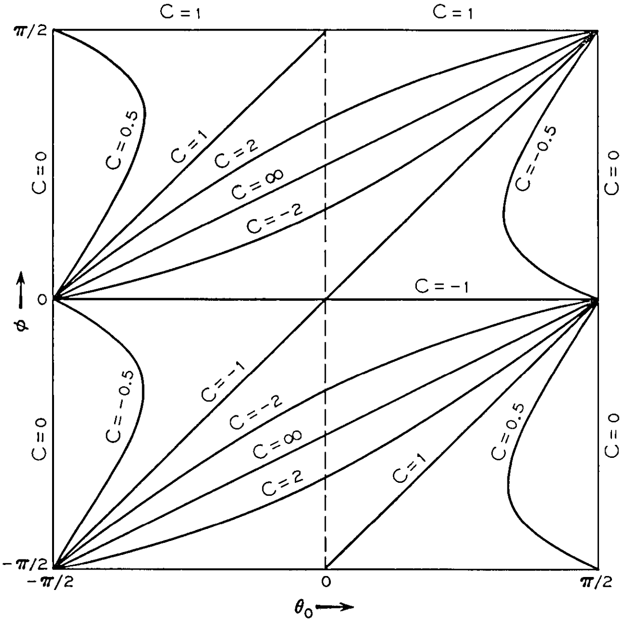


Fig. 5 — The slant factor c as a function of the angle of incidence θ_0 and the slant angle ϕ. c is positive for transmission holograms and negative for reflection holograms.

It is convenient to write the various diffraction formulas in terms of parameters ν and ξ, which are redefined for each grating type. In these parameters are lumped together the constants of the medium (n, α, n_1, α_1, κ), the obliquity factors (c_R, c_S), the wavelength, the grating thickness d, and the dephasing measure ϑ. By using ν and ξ, various trade-offs become immediately apparent.

We recall that, for transmission holograms, c_S is positive and the output signal appears at $z = d$. Combining equations (30) and (34) we obtain a general formula for the signal amplitude S of a transmission grating

$$\begin{aligned} S &= -j\left(\frac{c_R}{c_S}\right)^{\frac{1}{2}}\cdot\exp\left(-\alpha d/c_R\right)\cdot e^{\xi}\cdot\sin\left[\nu^2 - \xi^2\right]^{\frac{1}{2}}/\left[1 - \xi^2/\nu^2\right]^{\frac{1}{2}}, \\ \nu &= \kappa d/(c_R c_S)^{\frac{1}{2}}, \qquad\qquad (41) \\ \xi &= \tfrac{1}{2}d\left(\frac{\alpha}{c_R} - \frac{\alpha}{c_S} - j\frac{\vartheta}{c_S}\right), \end{aligned}$$

where κ is the coupling constant given in equation (9), ϑ the dephasing measure of equation (18), c_R and c_S are the obliquity factors of equation (23), α is the absorption constant and d the grating thickness. In the above form the parameters ν and ξ are, in general, of complex value.

3.1 *Lossless Dielectric Gratings*

For completeness we give the formulas for the lossless dielectric grating. For the unslanted case of this grating these formulas have been previously obtained by several workers whose prime interest was light diffraction by acoustic waves.[20,21,28] For this grating type it is easy to include the effect of slanted fringes.* For the lossless dielectric grating we have a coupling constant $\kappa = \pi n_1/\lambda$ and $\alpha = \alpha_1 = 0$. Equation (41) can be rewritten in the form

$$\begin{aligned} S &= -j\left(\frac{c_R}{c_S}\right)^{\frac{1}{2}} e^{-i\xi}\sin\left(\nu^2 + \xi^2\right)^{\frac{1}{2}}/\left(1 + \xi^2/\nu^2\right)^{\frac{1}{2}}, \\ \nu &= \pi n_1 d/\lambda(c_R c_S)^{\frac{1}{2}}, \qquad\qquad (42) \\ \xi &= \vartheta d/2c_S \end{aligned}$$

where ν and ξ have been redefined and are real-valued. The associated formula for the diffraction efficiency is

* Slant was also included in the treatment of dielectric gratings in Ref. 29.

$$\eta = \sin^2 (\nu^2 + \xi^2)^{\frac{1}{2}}/(1 + \xi^2/\nu^2). \tag{43}$$

For significant deviations from the Bragg condition the parameters ν and ξ are of equal order of magnitude, and we can take ν as independent of $\Delta\theta$ or $\Delta\lambda$ without causing an appreciable change in the predictions of equation (43). In this equation the angular and wavelength deviations are represented by the parameter ξ which can be written in the form

$$\begin{aligned}\xi &= \Delta\theta \cdot Kd \sin (\phi - \theta_0)/2c_S \\ &= -\Delta\lambda \cdot K^2 d/8\pi n c_S \end{aligned} \tag{44}$$

by using equation (18).

The angular and wavelength sensitivities of lossless dielectric gratings are shown in Fig. 6, where the efficiencies as given by equation (43) are plotted (normalized) as a function of ξ for three values of ν. The figure shows the sensitivity of gratings with $\nu = \pi/4$ and a peak diffraction efficiency of $\eta_0 = 0.5$, with $\nu = \pi/2$ and a peak efficiency of $\eta_0 = 1$, and with $\nu = 3\pi/4$ and $\eta_0 = 0.5$. We notice that the half-power points are reached for values near $\xi = 1.5$. There is some narrowing in the sensitivity curves for increasing values of ν, and a marked increase in the side lobe intensity.

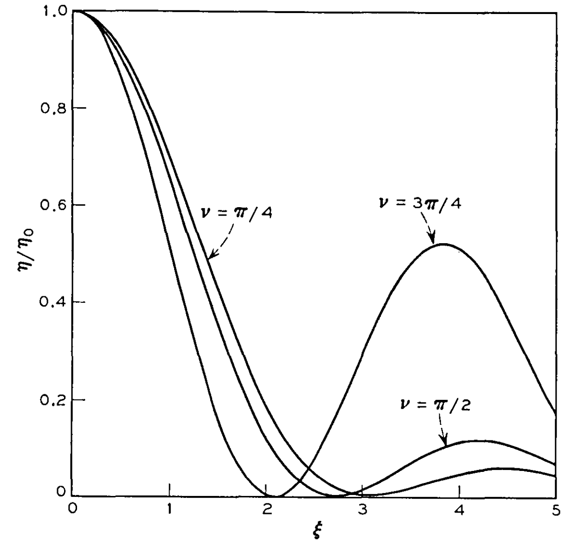


Fig. 6 — Transmission holograms—the angular and wavelength sensitivity of lossless dielectric gratings with the normalized efficiencies η/η_0 as a function of ξ.

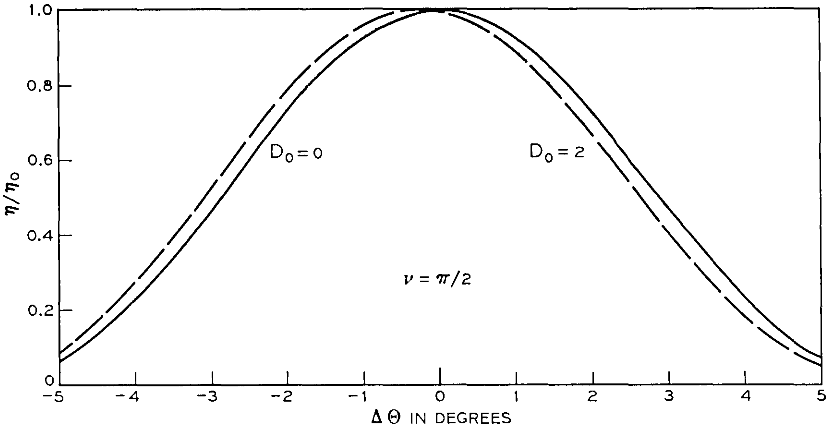


Fig. 7 — Transmission holograms—the angular sensitivity of a lossy dielectric grating with $\nu = \pi/2$ and $D_0 = 2$ compared with that of a lossless dielectric grating ($D_0 = 0$), for $\theta_0 = 30°$ and $\beta d = 50$.

The above formulas include the influence of slant through the obliquity factors c_R and c_S. If there is no slant ($\phi = \pi/2$) and if the Bragg condition is obeyed then $c_R = c_S = \cos\theta_0$ and equation (43) becomes the well known[20,21,30]

$$\eta = \sin^2 (\pi n_1 d/\lambda \cos\theta_0). \tag{45}$$

By inserting the above half-power values for ξ into equation (44) we obtain simple rules of thumb for the angular and spectral half-power bandwidths of unslanted gratings: $2\Delta\theta_{\frac{1}{2}} \approx \Lambda/d$, $2\Delta\lambda_{\frac{1}{2}}/\lambda \approx \cot\theta\cdot\Lambda/d$.

3.2 *Lossy Dielectric Gratings*

Let us first study the influence of loss on the angular sensitivity of a dielectric grating. We assume that there is no slant ($\phi = \pi/2$) and therefore $c_R = c_S = \cos\theta$. With this and a coupling constant of $\kappa = \pi n_1/\lambda$ we obtain from equation (41) for the signal amplitude

$$\begin{aligned} S &= -j \exp(-\alpha d/\cos\theta)\cdot e^{-i\xi}\cdot\sin(\nu^2 + \xi^2)^{\frac{1}{2}}/(1 + \xi^2/\nu^2)^{\frac{1}{2}} \\ \nu &= \pi n_1 d/\lambda \cos\theta \qquad\qquad (46) \\ \xi &= \vartheta d/2\cos\theta = \Delta\theta\cdot\beta d \sin\theta_0 \end{aligned}$$

where ν and ξ have been redefined, and ξ has been expressed in the needed form with the use of equations (14) and (18).

Equation (46) has a form similar to that of equation (42) except for an additional exponential term containing the absorption constant α. This term decreases the peak efficiency and it changes the angular sensitivity of the grating. But this change is very small, even for high loss values, as illustrated in Fig. 7. This figure compares the angular sensitivitities of a lossless grating ($D_0 = 0$) with that of a grating of high loss ($D_0 = 2$) for a parameter value of $\nu = \pi/2$, a Bragg angle of $\theta_0 = 30°$, and an optical grating thickness of $\beta d = 2\pi n d/\lambda = 50$. The loss parameter D_0 was defined as

$$D_0 = \alpha d/\cos\theta_0 \tag{47}$$

which is closely related to the conventional photographic density D (except that D_0 is measured in the direction of the reference wave given by θ_0). A value of $D_0 = 2$, which is the parameter used for the dashed curve, represents very high loss, with a decrease of the peak efficiency by a factor of about 50. Still, the differences of the two sensitivity curves are very small and consist mostly of an angular shift. The differences are even smaller for larger values of βd (we checked up to $\beta d = 200$), and, of course, for smaller values of D_0. The main conclusion is that the presence of loss has very little influence on the angular sensitivity of a dielectric transmission grating. This is probably because absorption influences the phase relations between the waves R and S very little. It agrees with observations by Belvaux.[31]

Next let us consider the influence of loss on the efficiency of a slanted dielectric grating. For simplicity we assume Bragg incidence, that is, $\vartheta = 0$. The obliquity factors are positive and given by $c_R = \cos\theta_0$ and $c_S = -\cos(\theta_0 - 2\phi)$. For this case we can write equation (41) for the signal amplitude S in the form

$$\begin{aligned} S &= -j\left(\frac{c_R}{c_S}\right)^{\frac{1}{2}} \cdot \exp\left[-\tfrac{1}{2}D_0(1+c)\right] \sin(\nu^2 - \xi^2)^{\frac{1}{2}}/(1 - \xi^2/\nu^2)^{\frac{1}{2}} \\ \nu &= \pi n_1 d/\lambda(c_R c_S)^{\frac{1}{2}} \\ \xi &= \tfrac{1}{2}D_0(1-c) \end{aligned} \tag{48}$$

where we have used the loss parameter D_0 as above in equation (47), and the slant factor c

$$D_0 = \alpha d/c_R = \alpha d/\cos\theta_0$$

$$c = c_R/c_S = -\cos\theta_0/\cos(\theta_0 - 2\phi).$$

Figure 8 shows the diffraction efficiency of slanted grantings as calculated from equation (48). The efficiencies are plotted as a function

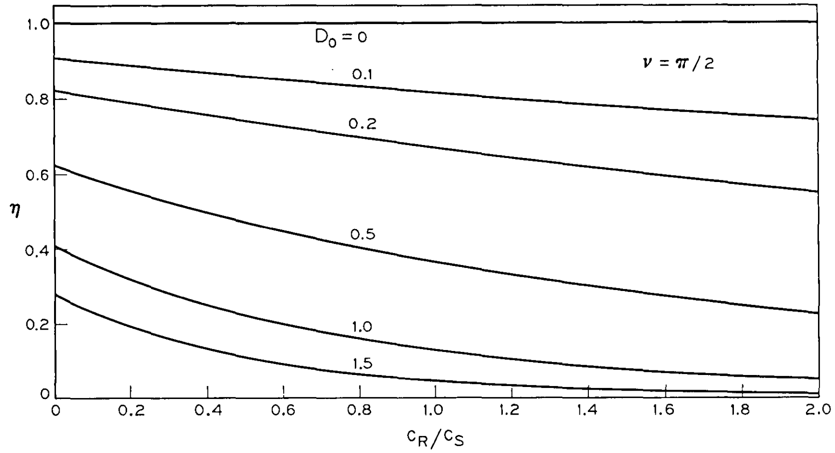


Fig. 8—Transmission holograms—the efficiency of lossy dielectric gratings as a function of slant for $\nu = \pi/2$. $c = c_R/c_S$ is the slant factor.

of the slant factor c for various values of D_0, and for a value of $\nu = \pi/2$ which corresponds to the maximum attainable efficiencies. Similar curves for $\nu = \pi/4$ and $\nu = 3\pi/4$ and the same D_0 values are almost identical to the curves of Fig. 8, except that the efficiency scale is reduced to a maximum efficiency of 0.5. This implies that for the range of chosen parameter values the exponential factor in equation (48) dominates in predicting the slant-dependence of the diffraction efficiency.

The results show that, for higher efficiencies, the grating prefers small c-values, assuming constant θ_0 and D_0. This is a preference of small exist angles for S which means that we get the best efficiency if the signal wave leaves the grating on the shortest possible path after it has been generated.

3.3 *Unslanted Absorption Gratings*

When one records holograms in conventional photographic emulsions one produces absorption gratings (bleaching can convert this into a dielectric grating). In an absorption grating there is no spatial modulation of the refractive index ($n_1 = 0$) and the coupling is provided by a modulation (α_1) of the absorption constant. We have, then, an imaginary coupling constant $\kappa = -j\alpha_1/2$. In this section we study the efficiencies and the angular and wavelength sensitivities of unslanted absorption gratings where $\phi = \pi/2$ and $c_R = c_S = \cos\theta$. From equation (41) we obtain for the signal amplitude

$$
\begin{aligned}
S &= -\exp(-\alpha d/c_R)\cdot e^{-i\xi}\cdot \mathrm{sh}\,(\nu^2 - \xi^2)^{\frac{1}{2}}/(1 - \xi^2/\nu^2)^{\frac{1}{2}} \\
\nu &= \alpha_1 d/2\cos\theta \qquad\qquad (49) \\
\xi &= \vartheta d/2\cos\theta \approx \Delta\theta\cdot\beta d\cdot\sin\theta_0 = -\tfrac{1}{2}(\Delta\lambda/\lambda)Kd\tan\theta_0
\end{aligned}
$$

where ν and ξ are real-valued, and equation (18) was used to express the parameter ξ again in various forms, showing explicitly the angular deviations $\Delta\theta$ and the wavelength deviation $\Delta\lambda$ from the Bragg condition.

For Bragg incidence we have $\xi = 0$, and obtain from the above a formula for the diffraction efficiency η of absorption gratings

$$\eta = \exp(-2\alpha d/\cos\theta_0)\cdot \mathrm{sh}^2\,(\alpha_1 d/2\cos\theta_0). \qquad (50)$$

As we exclude the presence of negative absorption (gain) in the medium, there is an upper limit for the amplitude α_1 of the assumed sinusoidal modulation, which is $\alpha_1 \leqq \alpha$. The highest diffraction efficiency possible for an absorption grating is reached in the limiting case $\alpha_1 = \alpha$ for a value of $\alpha d/\cos\theta_0 = \ln 3$. According to equation (50) this maximum efficiency has a value of $\eta_{\max} = 1/27$, or 3.7 percent.

Figure 9 shows values for the diffracted amplitude S of absorption gratings as computed from equation (50) as a function of the modulation amplitude α_1 and for various values of the depth of modulation. For convenience we have again used loss parameters, which are $D_0 = \alpha d/\cos\theta_0$ and $D_1 = \alpha_1 d/\cos\theta_0$. D_1 is a measure for the amplitude of the spatial modulation and $D_0/D_1 = \alpha/\alpha_1$ indicates the modulation depth. The dashed curves for constant D_0 show the grating behavior for constant background absorption. We have plotted S on a linear scale in order to identify the regions of linear grating response. Note that a good linear response and relatively good efficiency is obtained if the absorption background is held constant to a value of about $D_0 = 1$.

Equation (49) predicts also the angular sensitivity and the frequency sensitivity of absorption gratings. Such sensitivity curves are plotted in Fig. 10 for the special case of $\alpha_1 = \alpha_0$ and values of $\nu \equiv D_1/2 = 1$ (dashed) and $\nu = \frac{1}{2}\ln 3 = 0.55$. For the latter parameter value the peak efficiency of 3.7 percent is reached, and for $\nu = 1$ we have a peak efficiency of 2.5 percent. In the figure the relative efficiencies are plotted as functions of the parameter ξ. We note that there is very little difference between the sensitivity curves for the two ν-values chosen. We have also computed the sensitivity for smaller values of ν (0.2, 0.4), but the resulting curves differ so little from the ones shown that we have omitted them from the figure. The sensitivity curves are very

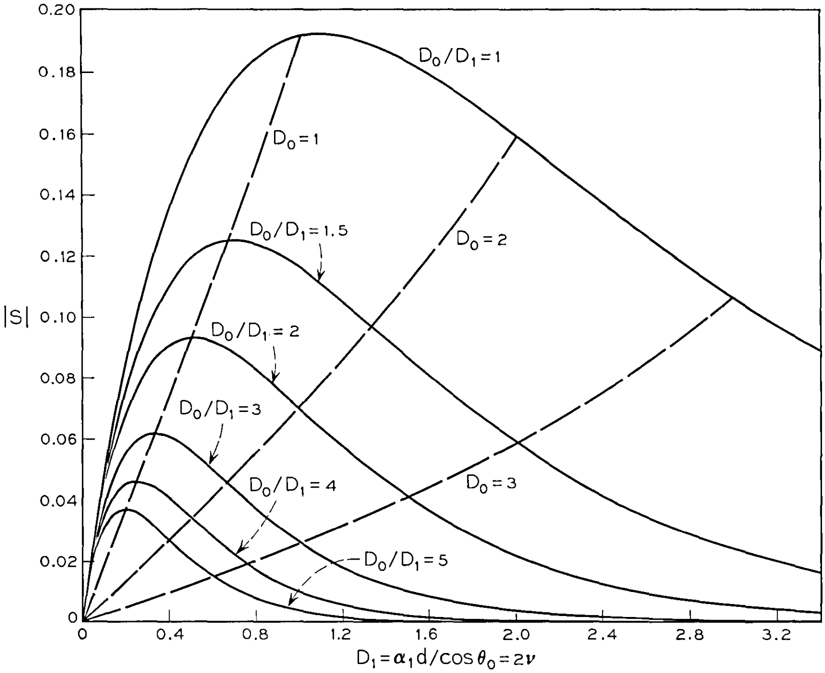


Fig. 9.—Transmission holograms—the diffracted amplitude of an absorption grating as a function of the modulation $D_1 = \alpha_1 d/\cos\theta = 2\nu$ for various modulation depths D_1/D_0 (solid curve) and various bias levels $D_0 = \alpha d/\cos\theta$ (dashed curve).

similar to those of the dielectric gratings with smaller ν-values which are shown in Fig. 6. Again, the half-power points are reached for about $\xi = 1.5$. But for absorption gratings there is no narrowing with increasing values of ν, and the side lobe intensity remains low.

3.4 *Slanted Absorption Gratings*

In this section we consider the influence of slant on the efficiency of an absorption grating. For simplicity we assume Bragg incidence ($\vartheta = 0$), and describe the slant by the obliquity factors $c_R = \cos\theta_0$ and $c_S = \cos(\theta_0 - 2\phi)$, as before. Using equation (41) we obtain, for this case, the following expression for the signal amplitude S

$$S = -\left(\frac{c_R}{c_S}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2}\alpha d\left(\frac{1}{c_R} + \frac{1}{c_S}\right)\right] \operatorname{sh}(\nu^2 + \xi^2)^{\frac{1}{2}}/(1 + \xi^2/\nu^2)^{\frac{1}{2}}$$

$$\nu = \alpha_1 d/2(c_R c_S)^{\frac{1}{2}} \tag{51}$$

$$\xi = \tfrac{1}{2}\alpha d\left(\frac{1}{c_R} - \frac{1}{c_S}\right),$$

where ν and ξ are redefined as real parameters. We have plotted the slant-dependence of absorption gratings in Fig. 11 for the special case of $\alpha_1 = \alpha$, that is, maximum depth of modulation. The diffraction efficiency η is shown as a function of the slant factor c for various values of the loss parameter D_0. These quantities are defined, as before, by

$$D_0 = \alpha d/c_R = \alpha d/\cos\theta_0$$

and (52)

$$c = c_R/c_S.$$

The efficiency is seen to reach its absolute maximum of 3.7 percent for the unslanted grating ($c = 1$) and for a loss parameter of $D_0 = \ln 3$. For larger values of D_0 the efficiencies reach relative maxima for exit

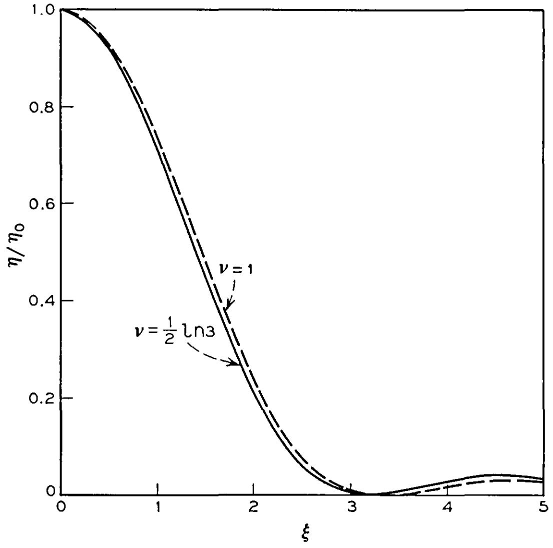


Fig. 10 — Transmission holograms—the angular and wavelength sensitivity of an absorption grating for $\alpha_1 = \alpha\,(D_1 = D_0)$ and values of $\nu = D_1/2 = 0.55$ ($\eta_0 = 0.037$) and $\nu = D_1/2 = 1$ ($\eta_0 = 0.025$).

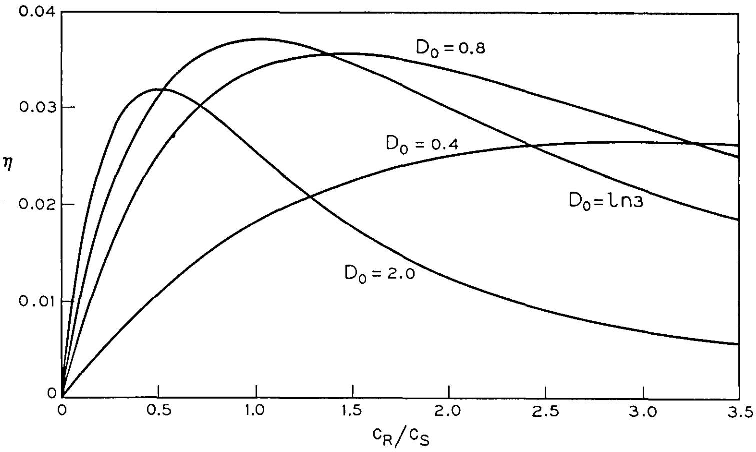


Fig. 11 — Transmission holograms—the efficiency of an absorption grating as a function of slant for $\alpha_1 = \alpha$ ($D_1 = D_0$). $c = c_R/c_S$ is the slant factor.

angles of the signal wave which are smaller than that of the reference wave ($c < 1$), while for smaller D_0-values the situation is reversed.

3.5 *Mixed Gratings*

Mixed gratings are those in which both the refractive index and the absorption constant are spatially modulated. This may occur in some recording materials (for example, as a result of incomplete bleaching, or in cases where strong absorption peaks are developed which cause refractive index changes according to the Kramers–Kronig relations).* Mixed gratings are described by a complex coupling constant, which is given in equation (9). For the special case of unslanted gratings ($\phi = \pi/2$) and Bragg incidence ($\vartheta = 0$) equation (41) simplifies to

$$S = -j \exp(-\alpha d/\cos\theta_0) \sin(\kappa d/\cos\theta_0) \tag{53}$$

where κ is complex. From this we obtain, after some algebra, an expression for the efficiency of mixed gratings

$$\eta = SS^* = [\sin^2(\pi n_1 d/\lambda \cos\theta_0) + \operatorname{sh}^2(\alpha_1/2\cos\theta_0)] \exp(-2\alpha d/\cos\theta_0), \tag{54}$$

where n_1 and α_1 are the amplitudes of the modulation of the refractive index and the absorption constant, and α is the average absorption

* Such effects have recently been observed by Nassenstein (see Ref. 32).

constant. We note that, at least for the special case considered here, there is a simple addition of the intensities diffracted by the dielectric grating and the absorption grating respectively [compare equations (46) and (50)!]. The exponential factor including α insures that the formula does not predict efficiencies larger than 1.

IV. REFLECTION HOLOGRAMS

In reflection holograms the recorded fringe-planes are of an orientation which is more or less parallel to the surfaces of the recording medium, and the signal appears as a "reflection" of the reference wave. We have illustrated this situation in Fig. 4b. It is expressed in the coupled wave analysis by negative values of the obliquity factor $c_S(c_S < 0)$. In addition, the signal amplitude S of interest is obtained by evaluating the signal wave in the plane $z = 0$, which is also the entrance plane for the reference wave R. For reflection holograms a slant angle $\phi = 0$ describes the case of unslanted gratings. Apart from these differences the following discussion of the detailed behavior of reflection holograms proceeds in a pattern similar to that of Section III, where we have discussed transmission holograms.

From equations (30) and (39) we obtain a general formula for the signal amplitude of reflection holograms which can be written in the form

$$\begin{aligned} S &= \left(\frac{c_R}{c_S}\right)^{\frac{1}{2}} \cdot \text{sh}\,(\nu\ \text{ch}\ a)/\text{ch}\,(a + \nu\ \text{ch}\ a) \\ \nu &= j\kappa d/\lambda(c_R c_S)^{\frac{1}{2}} \\ \xi &= \tfrac{1}{2}d\left(\frac{\alpha}{c_R} - \frac{\alpha}{c_S} - j\frac{\vartheta}{c_S}\right) \\ \text{sh}\ a &= \xi/\nu \end{aligned} \tag{55}$$

where we have again defined (complex) parameters ν, ξ and a, which lump together the constants of the medium $(n, \alpha, n_1, \alpha_1, \kappa)$, the obliquity factors c_R and c_S, the wavelength, the grating thickness d and the dephasing measure ϑ.

4.1 *Lossless Dielectric Gratings*

The lossless dielectric grating is characterized by a real-valued coupling constant $\kappa = \pi n_1/\lambda$, and by zero absorption $\alpha = \alpha_1 = 0$. As in the transmission-hologram counterpart, it is easy to include the case of slant in the analysis. For the present case we can rewrite equation (55) in the form

$$S = \left(\frac{c_R}{c_S}\right)^{\frac{1}{2}} / \{ j\xi/\nu + (1 - \xi^2/\nu^2)^{\frac{1}{2}} \cdot \coth (\nu^2 - \xi^2)^{\frac{1}{2}} \}$$

$$\nu = j\pi n_1 d/\lambda (c_R c_S)^{\frac{1}{2}} \tag{56}$$

$$\xi = -\vartheta d/2c_S$$

where ν and ξ have been redefined as real-valued parameters (c_S is negative!).

The associated formula for the diffraction efficiency of lossless dielectric gratings is

$$\eta = 1/\{1 + (1 - \xi^2/\nu^2)/\mathrm{sh}^2 (\nu^2 - \xi^2)^{\frac{1}{2}}\}, \tag{57}$$

which also provides a description of the angular and wavelength sensensitivities of the grating. For unslanted acoustic gratings this formula has been previously given by Quate and his associates.[22] Sensitivity curves calculated from equation (57) are shown in Fig. 12, where the normalized efficiencies are plotted as a function of ξ for various values of ν = const. The figure shows the sensitivity of a grating with $\nu = \pi/4$ and a peak efficiency of 43 percent, a grating with $\nu = \pi/2$ and $\eta_0 = 0.84$,

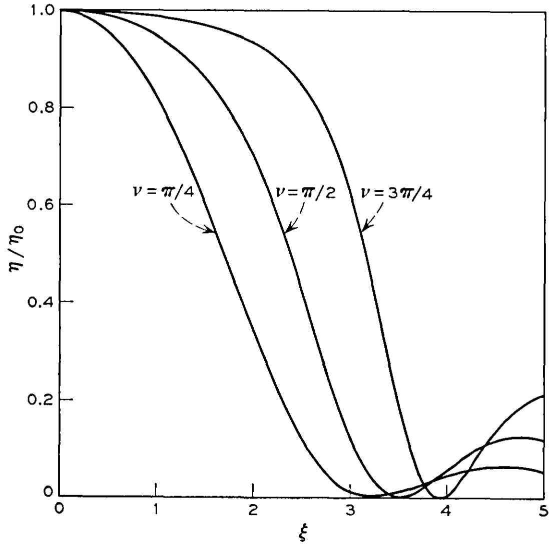


Fig. 12—Reflection holograms—the angular and wavelength sensitivity of a lossless dielectric grating with the normalized efficiency η/η_0 as a function of ξ.

and the corresponding values for $\nu = 3\pi/4$ and $\eta_0 = 0.96$. For $\nu = \pi/4$ the half-power points of the grating response are reached for values of approximately $\xi = 1.7$. But there is considerable broadening of the sensitivity curves for increasing values of ν, and an increase in the side-lobe level.

As in equation (44) for transmission holograms, we can express the parameter ξ directly in the angular deviation $\Delta\theta$ or the wavelength deviation $\Delta\lambda$ by using equation (18) to obtain

$$\begin{aligned} \xi &= \Delta\theta \cdot Kd \cdot \sin(\theta_0 - \phi)/2c_S \\ &= \Delta\lambda \cdot K^2 d/8\pi n c_S \ . \end{aligned} \tag{58}$$

These expressions can again be used to formulate rules for the angular bandwidth and the spectral bandwidth of the grating.

For an unslanted grating ($\phi = 0$) and Bragg incidence we have $c_R = -c_S = \cos\theta_0$, and equation (57) simplifies to

$$\eta = \text{th}^2\,(\pi n_1 d/\lambda \cos\theta_0). \tag{59}$$

This is a formula which has been obtained previously for light diffraction by acoustic waves.[33,22]

4.2 *Lossy Dielectric Gratings*

Let us first discuss the influence of loss on the angular and wavelength sensitivity of unslanted dielectric gratings. Here we have $\phi = 0$ and, to a good approximation

$$\begin{aligned} c_R &= \cos\theta_0(1 - \Delta\theta\tan\theta_0) = \cos\theta \\ c_S &= -\cos\theta_0(1 + \Delta\theta\tan\theta_0), \\ &= -\cos\theta(1 + 2\Delta\lambda/\lambda) \end{aligned} \tag{60}$$

at least as long $\tan\theta_0 \leqq 1$. One can show that the formula for the signal amplitude S, which we have given in equation (56), is still applicable for the present case of an unslanted lossy grating if we modify the parameters ν and ξ to

$$\begin{aligned} \nu &= \pi n_1 d/\lambda \cos\theta_0 \\ \xi &= \xi_0 - jD_0 \ , \\ \xi_0 &= -\Delta\theta \cdot \beta d \sin\theta_0 \\ D_0 &= \alpha d/\cos\theta_0 \end{aligned} \tag{61}$$

where ξ is now a complex parameter with ξ_0 describing the angular deviations and D_0 representing the loss. An evaluation of this formula is shown in Fig. 13, which shows the angular sensitivity of dielectric gratings for various values of the loss parameter D_0 and a grating parameter of $\nu = \pi/2$. In constrast to what we have observed in the case of dielectric transmission holograms (Fig. 7), we see here a quite noticeable effect of the grating loss on the sensitivity curves. With increasing loss values the curves broaden in the wings, sharpen somewhat in the center and the side-lobe level decreases.

To study the influence of loss on the diffraction efficiencies of dielectric gratings we rewrite equation (55) in the form

$$\begin{aligned} S &= \left(\frac{c_R}{c_S}\right)^{\frac{1}{2}} / \{\xi/\nu + (1 + \xi^2/\nu^2)^{\frac{1}{2}} \cdot \coth (\nu^2 + \xi^2)^{\frac{1}{2}}\} \\ \nu &= j\pi n_1 d/\lambda(c_R c_S)^{\frac{1}{2}} \qquad (62) \\ \xi &= \tfrac{1}{2} D_0 (1 - c) \end{aligned}$$

where we have written ν and ξ as real-valued parameters in a form which is valid for Bragg incidence and for slanted or unslanted gratings. Just as in the case of transmission holograms we have used the loss

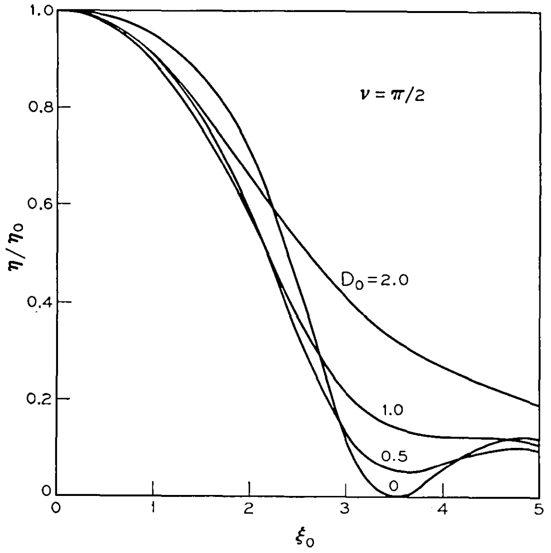


Fig. 13—Reflection holograms—the influence of loss on the angular and wavelength sensitivity of a dielectric grating for $\nu = \pi/2$. The normalized efficiencies η/η_0 are shown. The peak efficiencies are $\eta_0 = 0.84$ for $D_0 = 0$, $\eta_0 = 0.64$ for $D_0 = 0.5$, $\eta_0 = 0.28$ for $D_0 = 1$, and $\eta_0 = 0.12$ for $D_0 = 2$.

parameter D_0 and the slant factor c (which is now negative)

$$
\begin{aligned}
D_0 &= \alpha d/\cos\theta_0 \\
c &= c_R/c_S\ .
\end{aligned}
\tag{63}
$$

In the case of unslanted gratings the parameters ν and ξ simplify to

$$
\begin{aligned}
\nu &= \pi n_1 d/\lambda \cos\theta_0 \\
\xi &= D_0 = \alpha d/\cos\theta_0\ .
\end{aligned}
\tag{64}
$$

The results of a numerical evaluation for unslated gratings are shown in Fig. 14, where the signal amplitude is plotted as a function of ν for various values of the loss parameter D_0. The curve $D_0 = 0$ gives the values for lossless gratings, while the others indicate the influence of loss.

The behavior of slanted dielectric gratings in the presence of loss is shown in Fig. 15. The curves of this figure are also computed from equation (62) and show the diffraction efficiency as a function of the slant factor for $\nu = \pi/2$ and various values of the loss parameter D_0. For constant D_0 we notice an increase of the efficiency for decreasing values of the slant factor, as in the case of transmission holograms.

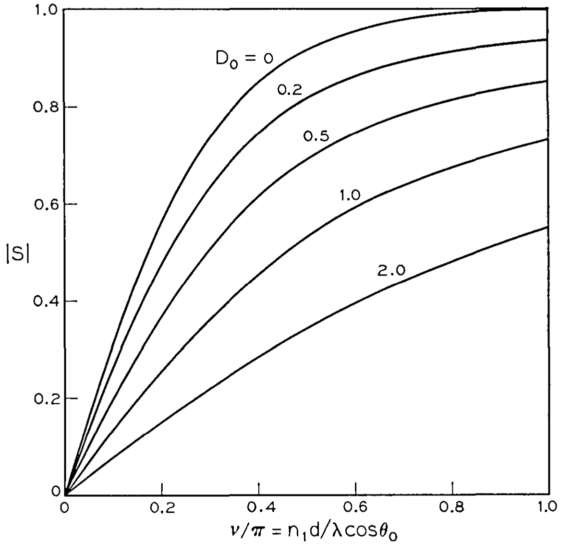


Fig. 14—Reflection holograms—the influence of loss on the diffracted amplitude S of an unslanted dielectric grating. $|S|$ is shown as a function of $\nu/\pi = n_1 d/\lambda \cos\theta_0$ for various loss parameters D_0.

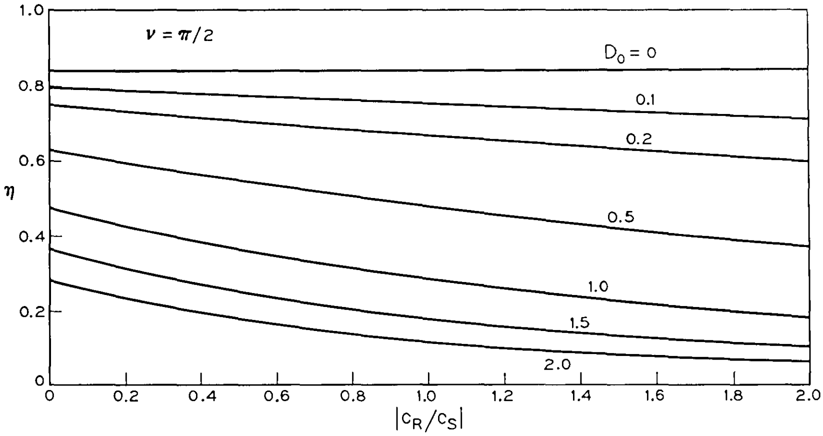


Fig. 15—Reflection holograms—the efficiency of a lossy dielectric grating as a function of slant for $\nu = \pi/2$. $c = c_R/c_S$ is the slant factor.

Again, for given loss and a given angle of incidence short signal paths through the grating (that is, small exit angles) are preferred for higher efficiencies.

4.3 *Unslanted Absorption Gratings*

Following the pattern set in the discussion of transmission holograms (Section III), we again describe an absorption grating by an imaginary coupling constant $\kappa = -j\alpha_1/2$, and proceed to study the diffraction efficiencies and the angular and wavelength sensitivities of unslanted ($\phi = 0$) gratings. In this case equation (55) simplifies to

$$\begin{aligned} S &= -j\left(\frac{c_R}{c_S}\right)^{\frac{1}{2}} \Big/ \{\xi/\nu + [(\xi/\nu)^2 - 1]^{\frac{1}{2}} \coth\,(\xi^2 - \nu^2)^{\frac{1}{2}}\} \\ \nu &= j\alpha_1 d/2(c_R c_S)^{\frac{1}{2}} \qquad (65) \\ \xi &= D_0 - j\xi_0 \end{aligned}$$

where the real-valued parameters D_0 and ξ_0 can be expressed to first order in the angular deviations $\Delta\theta$ and the wave-length deviations $\Delta\lambda$ by

$$\begin{aligned} D_0 &= \alpha d/\cos\,\theta_0 \\ \xi_0 &= \Delta\theta\cdot\beta d \sin\,\theta_0 = \tfrac{1}{2}(\Delta\lambda/\lambda)Kd. \end{aligned} \qquad (66)$$

D_0 is a loss parameter as before, and ξ_0 is a normalized measure for the angular or the wavelength deviations from the Bragg condition.

If the Bragg condition is obeyed equation (65) can be written in the form

$$S = -D_1/2[D_0 + (D_0^2 - D_1^2/4)^{\frac{1}{2}} \cdot \coth (D_0^2 - D_1^2/4)^{\frac{1}{2}}] \qquad (67)$$

where

$$D_1 = 2\nu_1 = \alpha_1 d/\cos \theta_0$$

measures the spatial modulation of the absorption constant (α_1).

For the deepest allowable modulation where we have $D_1 = D_0(\alpha_1 = \alpha_0)$, this equation predicts the maximum diffraction efficiency $\eta_{\max}$ which is possible for reflection holograms with a (sinusoidal) absorption modulation. We obtain $\eta_{\max} = 1/(2 + \sqrt{3})^2$, or a maximum efficiency of 7.2 percent for $D_0 = D_1 \rightarrow \infty$. The formula reflects the experimental fact that, for reflection holograms of the absorptive kind, one obtains the largest efficiencies for high photographic densities. Figure 16 shows a numerical evaluation of the above formula. Here the signal amplitude S is plotted as a function of the modulation amplitude D_1 for various levels of loss "bias" D_0 (dashed curves) and for various modulation depths D_0/D_1.

An evaluation of the grating sensitivity as predicted by equation (65) is shown in Fig. 17 for the special case of a maximum depth of modulation where $D_1 = D_0$. In this figure the (normalized) efficiency is plotted as a function of the parameter ξ_0 for various values of $D_1 = D_0$. As in the corresponding grating for the case of transmission holograms (Fig. 10) the sensitivity curves are seen to reach their half-power points for values of about $\xi_0 = 1.5$. But in the present case of reflection holograms there is a noticeable broadening of the curves with increasing loss values $D_1 = D_0$.

4.4 *Slanted Absorption Gratings*

In this section we consider the influence of slant on the diffraction efficiency of an absorption grating for reflection holograms. We assume Bragg incidence ($\vartheta = 0$) and again use the obliquity factors $c_R = \cos \theta_0$ and $c_S = -\cos (\theta_0 - 2\phi)$ to describe the slant (for reflection holograms we have $c_S < 0$!). We find that equation (65) can be used as a formula for the signal amplitude for the present case if we modify the parameters to

$$\nu = j\alpha_1 d/2(c_R c_S)^{\frac{1}{2}} = \frac{j}{2} D_1(c)^{\frac{1}{2}}$$

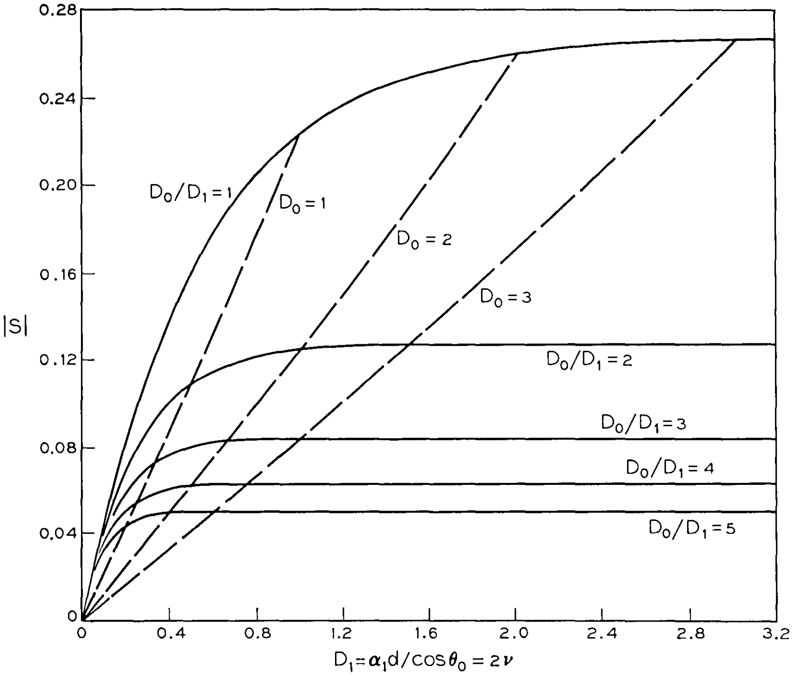


Fig. 16 — Reflection holograms—the diffracted amplitude of an absorption grating as a function of the modulation $D_1 = \alpha_1 d/\cos\theta = 2\nu$ for modulation depths D_1/D_0 (solid curve) and bias levels $D_0 = \alpha d/\cos\theta$ (dashed curve).

$$\begin{aligned} \xi &= \tfrac{1}{2}D_0(1 - c) \\ D_0 &= \alpha d/\cos\theta_0\,, \qquad D_1 = \alpha_1 d/\cos\theta_0 \\ c &= c_R/c_S \end{aligned} \tag{68}$$

where the slant factor c is negative. All these parameters are real-valued in the present case. For a maximum depth of modulation, that is, $\alpha_1 = \alpha$, there are further simplifications, and we obtain a simple expression for the slant-dependence of the diffraction efficiency

$$\eta = -c/\{1 - c + (1 - c + c^2)^{\frac{1}{2}}\cdot \coth \tfrac{1}{2}D_0(1 - c + c^2)^{\frac{1}{2}}\}^2. \tag{69}$$

Figure 18 shows a numerical evaluation of this formula for various values of $D_0 = D_1$. The slant factor value of $|\,c\,| = 1$ refers to unslanted gratings. In this case the maximum efficiency value $\eta_{\max} = 0.072$ is approached for large D_0. We note that for values of D_0 below unity the efficiencies increase for $|c|$-values larger than 1 and up to about 3, that is, for relatively large exit angles of the signal wave.

4.5 *Mixed Gratings*

Mixed gratings are described by a complex coupling constant $\kappa = \pi n_1/\lambda - j\alpha_1/2$ (see Section 3.5). For Bragg incidence ($\vartheta = 0$) and unslanted fringe-planes ($\phi = 0$) we can obtain from equation (55) a formula for the signal amplitude of mixed gratings, which is

$$S = -j\kappa \Big/ \left\{ \alpha + (\kappa^2 + \alpha^2)^{\frac{1}{2}} \cdot \coth \frac{d}{\cos \theta_0} (\kappa^2 + \alpha^2)^{\frac{1}{2}} \right\} \tag{70}$$

where κ is of complex value, α is the average absorption constant, d the grating thickness and θ_0 the angle of incidence.

V. AMPLITUDES OF THE DIRECT WAVES

For diagnostic purposes it is often of interest to monitor the change in amplitude of the direct reference wave R, which is depleted because of diffraction into S and absorption. The quantities of interest are the amplitudes $R(d)$ which can be obtained from the analysis of Section 2.2.

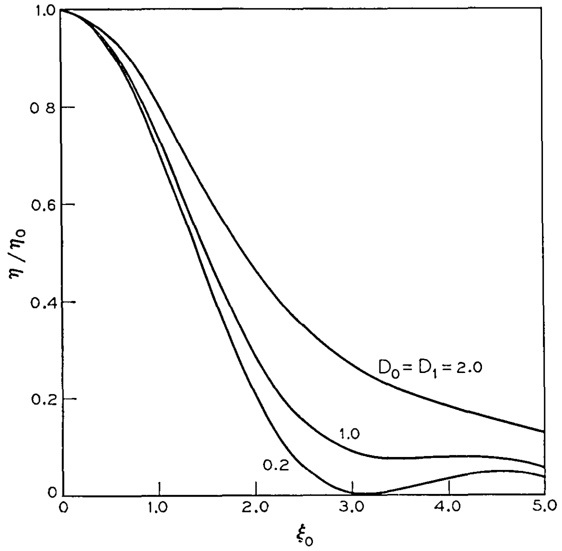


Fig. 17—Reflection holograms—the angular and wavelength sensitivity of an absorption grating for $\alpha_1 = \alpha$ ($D_1 = D_0$) and values of $D_1 = 2\nu = D_0 = 0.2$ ($\eta_0 = 0.007$), $D_1 = 1$ ($\eta_0 = 0.05$), and $D_1 = 2$ ($\eta_0 = 0.068$).

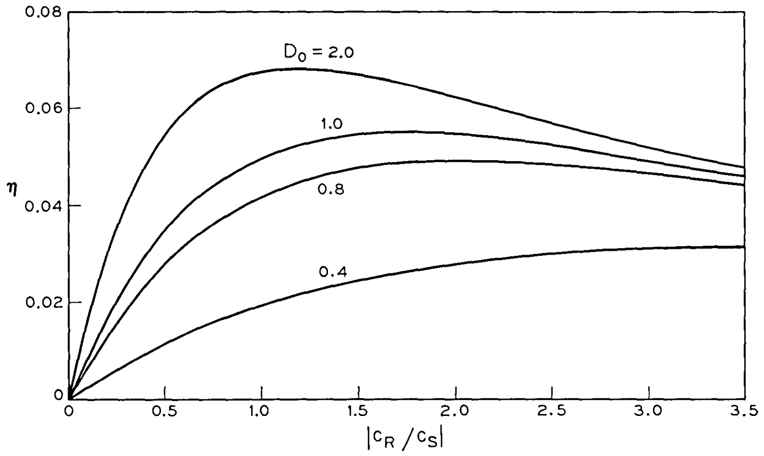


Fig. 18—Reflection holograms—the efficiency of an absorption grating as a function of slant for $\alpha_1 = \alpha$ $(D_1 = D_0)$. $c = c_R/c_S$ is the slant factor.

We will give here the general results for transmission and reflection holograms. The notation is that of Section 2.

5.1 *Transmission Holograms*

From equations (27) and (33) we get for the constants r_i of equation (25) the expressions

$$\begin{aligned} r_1 &= -\kappa^2/c_S(\gamma_1 - \gamma_2)(c_R\gamma_1 + \alpha) \\ r_2 &= \kappa^2/c_S(\gamma_1 - \gamma_2)(c_R\gamma_2 + \alpha). \end{aligned} \tag{71}$$

Using this we can write the output amplitude $R(d)$ of the reference wave in the form

$$R(d) = \frac{\kappa^2}{c_S(\gamma_1 - \gamma_2)} \left(\frac{\exp(\gamma_2 d)}{c_R\gamma_2 + \alpha} - \frac{\exp(\gamma_1 d)}{c_R\gamma_1 + \alpha} \right). \tag{72}$$

5.2 *Reflection Holograms*

For reflection holograms we use equations (27), (37), and (39) to express the constants r_i in the form

$$\begin{aligned} r_1 = {} & (c_S\gamma_1 + \alpha + j\vartheta) \exp(\gamma_2 d)/\{\exp(\gamma_2 d)(\alpha + j\vartheta + c_S\gamma_1) \\ & - \exp(\gamma_1 d)(\alpha + j\vartheta + c_S\gamma_2)\} \\ r_2 = {} & -(c_S\gamma_2 + \alpha + j\vartheta) \exp(\gamma_1 d)/\{\exp(\gamma_2 d)(\alpha + j\vartheta + c_S\gamma_1) \\ & - \exp(\gamma_1 d)(\alpha + j\vartheta + c_S\gamma_2)\}. \end{aligned} \tag{73}$$

The output amplitude $R(d)$ of the reference wave becomes

$$R(d) = c_S(\gamma_1 - \gamma_2)/\{(\alpha + j\vartheta + c_S\gamma_1) \exp(-\gamma_1 d) - (\alpha + j\vartheta + c_S\gamma_2) \exp(-\gamma_2 d)\}. \tag{74}$$

More detailed evaluations of the above formulas should follow the pattern prescribed in Sections III and IV. They can be undertaken for the specific case when the need arises.

VI. VALIDITY OF THE THEORY

We have tried to make our results as generally applicable as possible. We have allowed for the presence of absorption in the various hologram gratings and for a slant of the fringe planes. But a whole range of assumptions had to be made to make the simple coupled wave analysis possible. It seems appropriate to recount these assumptions to make clear the region of validity of the coupled wave theory. We have assumed that:

(*i*) The electric field of the light is polarized perpendicular to the plane of incidence. However, the appendix gives an extension of the theory to allow also for light of parallel polarization.

(*ii*) A slant of the fringe planes with respect to the z-axis is allowed, except that these planes are perpendicular to the plane of incidence. (This is reflected in the assumption $\epsilon(x, z)$, $\sigma(x, z)$.) But this assumption is not made in the generalization which we have given in the appendix.

(*iii*) The spatial modulation of the refractive index and the absorption constant is sinusoidal.

(*iv*) There is a small absorption loss per wavelength and a slow energy interchange (per wavelength) between the two coupled waves. This condition is stated mathematically in equation (7) and justifies neglecting the second derivatives R'' and S'' in the analysis.

(*v*) There is the same average refractive index n for the regions inside and outside the grating boundaries. If the grating has interfaces with air, then Snell's law has to be used to correct for the angular changes resulting from refraction.

(*vi*) Light incidence is at or near the Bragg angle and only the diffraction orders which obey the Bragg condition at least approximately are retained in the analysis. The other diffraction orders are neglected.

A detailed mathematical justification of assumption *vi* is outside the scope of our simple analysis. One can advance physical arguments to show that this step limits the validity of the theory to "thick" gratings, where the phase synchronism between the two coupled waves has enough time to develop a strong and dominating effect. Better definitions of a "thick grating" must come from more accurate theories which are available for special cases. A large amount of work has been done on acoustic diffraction gratings which correspond to the case of our unslanted, lossless, dielectric transmission-hologram gratings.[30] In acoustic diffraction one defines the parameter

$$Q = 2\pi\lambda d/n\Lambda^2 \tag{75}$$

as an appropriate measure of grating thickness. We can regard a grating as thick when the condition $Q \gg 1$ holds.[21,16] It appears that the coupled wave theory begins to give good results for values of $Q = 10$. This is particularly well demonstrated by Klein and his associates in theoretical and experimental work on acoustic gratings for the predictions of both the peak efficiencies and the angular sensitivities.[16,17,34] We hasten to add that for the majority of practical holograms the parameter Q is larger, and sometimes much larger, than 10.

Further checks of the validity of the coupled wave theory are provided by comparisons with accurate computer calculations and with experiments on special examples of gratings. Burckhardt has made computer calculations on unslanted, lossless, dielectric transmission holograms for selected values of grating parameters which are commonly encountered in holography.[18,19] Comparison with the results of the coupled wave theory shows very satisfactory agreement.[35] Measurements by Shankoff and Lin on dielectric transmission holograms prepared with dichromated gelatin yielded diffraction efficiencies approaching 100 percent, which agrees with the theory (even though there may be some uncertainty as to the exact nature of the refractive index variations).[10,11]

Efficiency measurements on thick absorption gratings for the case of transmission holograms were made by George, Mathews, and Latta.[36,37] Efficiencies approaching our predicted maximum value of 3.7 percent were observed.

Kiemle has studied unslanted ($\phi = 0$) reflection holograms for the special case of normal incidence ($\theta_0 = 0$) by analyzing equivalent four-terminal networks.[38] His treatment of absorption gratings corresponds to the material we discussed in Section 4.3 specialized to the case of $\theta_0 = 0$. But Kiemle's value of 2.8 percent for the maximum diffraction efficiency of absorptive reflection holograms does not agree with our prediction of 7.2 percent. This disagreement appears to derive from a set of restrictive assumptions made in Kiemle's work. Experimental observations on absorptive reflection holograms were made by

Lin and Lo Bianco.[9] Efficiency values as high as 3.8 percent were measured, which seems to support the predictions of the coupled wave theory. But further experiments are needed for a good confirmation.

VII. CONCLUSIONS

We have discussed a coupled-wave analysis of the Bragg diffraction of light by thick hologram gratings. This approach made it possible to derive simple algebraic formulas for the behavior of various types of holograms, even for the case of high diffraction efficiencies where the incident wave is strongly depleted. The treatment covers transmission holograms and reflection holograms, and it includes the spatial modulations of both the refractive index and the absorption constant. The influence of loss in the grating and of slanted fringes is also discussed. Formulas and their numerical evaluations are given for the diffraction efficiencies and the angular and wavelength sensitivities of various grating types.

For special cases we can compare the results of this theory with more accurate computations and with experimental observations. These comparisons give us the confidence to assume that the coupled wave predictions are good for a broad range of practical hologram types.

VIII. ACKNOWLEDGMENT

The author would like to acknowledge with thanks useful discussions with C. B. Burckhardt, and the patience of those who have waited so long for the promised completion of this article.

APPENDIX

Reduced Coupling for Light Polarized in the Plane of Incidence

In the body of this paper it was assumed that the incident light is polarized perpendicular to the plane of incidence. The purpose of this appendix is to show that we can use the results of the main paper also when the light is polarized in the plane of incidence, provided that we modify the coupling constant κ. Such a modification is suggested already by the dynamical theory of X-ray diffraction.

As in Section II we start with the wave equation

$$\nabla^2 \mathbf{E} - \nabla(\nabla \cdot \mathbf{E}) + k^2 \mathbf{E} = 0 \qquad (76)$$

for the electric field in the grating. Here, in contrast to equation (1), we have described the field by the *vector* quantity $\mathbf{E}$ and have included the term $\nabla(\nabla \cdot \mathbf{E})$, which is not necessarily zero. The constant k^2 is defined in equation (4). As in the main paper, we assume that only two waves are present in the grating, and put

$$\mathbf{E} = \mathbf{R}(z)e^{-i\boldsymbol{\rho}\cdot\mathbf{x}} + \mathbf{S}(z)e^{-i\boldsymbol{\sigma}\cdot\mathbf{x}} \tag{77}$$

using the vectors **R** and **S** to describe the amplitudes of the reference and signal waves. $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ are the propagation vectors (as in Section II) which point in the direction of the wavenormals. They are related by equation (11). In addition we assume that, both, R and S are transverse waves, that is, that the following conditions hold

$$\begin{aligned}(\boldsymbol{\rho}\cdot\mathbf{R}) &\equiv 0,\\ (\boldsymbol{\sigma}\cdot\mathbf{S}) &\equiv 0.\end{aligned} \tag{78}$$

Combining equations (76), (77), and (78) we get, after separating terms with equal exponentials and neglecting second derivatives $\partial^2/\partial z^2$

$$-2j\rho_z\mathbf{R}' + j\boldsymbol{\rho}R_z' - 2j\alpha\beta\mathbf{R} + 2\kappa\beta\mathbf{S} = 0 \tag{79}$$

$$-2j\sigma_z\mathbf{S}' + j\boldsymbol{\sigma}S_z' + (\beta^2 - \sigma^2 - 2j\alpha\beta)\mathbf{S} + 2\kappa\beta\mathbf{R} = 0 \tag{80}$$

where R_z and S_z are the z-components of R and S, and the notation of Section II is used.

We now make the additional assumption that the polarizations of R and S do not change in the grating and write

$$\begin{aligned}\mathbf{R}(z) &= R(z)\mathbf{r},\\ \mathbf{S}(z) &= S(z)\mathbf{s},\end{aligned} \tag{81}$$

where $R(z)$ and $S(z)$ are the scalar amplitudes of the two waves, and **r** and **s** are polarization vectors independent of z. These vectors are normalized so that

$$(\mathbf{r}\cdot\mathbf{r}) = 1, \qquad (\mathbf{s}\cdot\mathbf{s}) = 1. \tag{82}$$

Because of (78) we have

$$(\mathbf{r}\cdot\boldsymbol{\rho}) = 0, \qquad (\mathbf{s}\cdot\boldsymbol{\sigma}) = 0. \tag{83}$$

After forming the dot products of **r** with eq. (79) and of **s** with (80) we use equations (81), (82), and (83) to arrive at

$$-2j\rho_zR' - 2j\alpha\beta R + 2\kappa\beta S(\mathbf{r}\cdot\mathbf{s}) = 0 \tag{84}$$

$$-2j\sigma_zS' + (\beta^2 - \sigma^2 - 2j\alpha\beta)S + 2\kappa\beta R(\mathbf{r}\cdot\mathbf{s}) = 0. \tag{85}$$

As in Section II, we introduce the abbreviations

$$c_R = \rho_z/\beta, \qquad c_S = \sigma_z/\beta, \tag{86}$$

and

$$\vartheta = (\beta^2 - \sigma^2)/2\beta, \tag{87}$$

which allow us to write the above equations in the form

$$c_R R' + \alpha R = -j\kappa(\mathbf{r}\cdot\mathbf{s})S \tag{88}$$

$$c_S S' + (\alpha + j\vartheta)S = -j\kappa(\mathbf{r}\cdot\mathbf{s})R. \tag{89}$$

These are coupled wave equations which govern the Bragg diffraction of light polarized parallel to the plane of incidence, and indeed, of light of arbitrary polarization. They are similar in form to the coupled wave equations (21) and (22) which were derived for perpendicular polarization. The only difference is a reduction of the effective coupling constant by the dot product $(\mathbf{r}\cdot\mathbf{s})$ of the two polarization vectors.

Referring to the grating geometry of Fig. 1 we have $(\mathbf{r}\cdot\mathbf{s}) = 1$ for light polarized perpendicular to the plane of incidence. For parallel polarization the value of this dot-product depends on the inclination angles, and we have a reduced effective coupling constant $\kappa_{\parallel}$ given by

$$\kappa_{\parallel} = \kappa(\mathbf{r}\cdot\mathbf{s}) = -\kappa \cos 2(\theta_0 - \phi). \tag{90}$$

We can apply the results of the main paper for parallel polarization if we replace κ by $\kappa_{\parallel}$. For this polarization there is the trivial case of a Bragg angle of 45° (that is, diffraction angles of 90°) where $(\mathbf{r}\cdot\mathbf{s}) = 0$ and the intensity of the diffracted light goes to zero.

REFERENCES

1. van Heerden, P. J., "Theory of Optical Information Storage in Solids," Appl. Opt., *2*, No. 4 (April 1963), pp. 393–400.
2. Smits, F. M., and Gallaher, L. E., "Design Considerations for a Semipermanent Optical Memory," B.S.T.J., *46*, No. 6 (July–August 1967), pp. 1267–1278.
3. Vitols, V. A., "Hologram Memory for Storing Digital Data," IBM Technical Disclosure Bull., *8*, No. 11 (April 1966), p. 1581.
4. Pennington, K. S., and Lin, L. H., "Multicolor Wavefront Reconstruction," Appl. Phys. Letters, *7*, (August 1965), pp. 56–57.
5. Denisyuk, Y. N., "On the Reproduction of the Optical Properties of an Object by the Wave Field of Its Scattered Radiation," Opt. Spectroscopy, *15*, No. 4 (October 1963), pp. 279–284.
6. Stroke, G. W., and Labeyrie, A. E., "White-light Reconstruction of Holographic Images Using the Lippman-Bragg Diffraction Effect," Phys. Letters, *20*, (March 1966), pp. 368–370.
7. Lin, L. H., Pennington, K. S., Stroke, G. W., and Labeyrie, A. E., "Multicolor Holographic Image Reconstruction with White-light Illumination," B.S.T.J., *45*, No. 4 (April 1966), pp. 659–660.

8. Upatnieks, J., Marks, J., and Fedorwicz, R., "Color Holograms for White Light Reconstruction," Appl. Phys. Letters, *8*, No. 11 (June 1966), pp. 286–287.
9. Lin, L. H., and Lo Bianco, C. V., "Experimental Techniques in Making Multicolor White Light Reconstructed Holograms," Appl. Opt., *6*, No. 7 (July 1967), pp. 1255–1258.
10. Shankoff, T., "Phase Holograms in Dichromated Gelatin," Appl. Opt., *7*, No. 10 (October 1968), pp. 2101–2105.
11. Lin L. H., "Hologram Formation in Hardened Dichromated Gelatin Films," Appl. Opt., *8*, No. 5 (May 1969), pp. 963–966.
12. Chen, F. S., LaMacchia, J. T., and Fraser, D. B., "Holographic Storage in Lithium Niobate," Appl. Phys. Letters, *12*, No. 7 (October 1968), pp. 223–224.
13. Close, D. H., Jacobson, A. D., Margerum, J. D., Brault, R. G., and McClumg, F. J., "Hologram Recording in Photopolymer Materials," Appl. Phys. Letters *14*, No. 5, (March 1969), pp. 159–160.
14. Leith, E. N., Kozma, A., Upatnieks, J., Marks, J., and Massey, N., "Holographic Data Storage in Three-Dimensional Media," Appl. Opt., *5*, No. 8 (August 1966), pp. 1303–1311.
15. Gabor, D., and Stroke, G. W., "The Theory of Deep Holograms," Proc. Royal Soc. of London, *A. 304*, (February 1968), pp. 275–289.
16. Klein, W. R., Tipnis, C. B., and Hiedemann, E. A., "Experimental Study of Fraunhofer Light Diffraction by Ultrasonic Beams of Moderately High Frequency at Oblique Incidence," J. Acoust. Soc. Amer., *38*, No. 2 (August 1965), pp. 229–233.
17. Klein, W. R., "Theoretical Efficiency of Bragg Devices," Proc. IEEE, *54*, No. 5 (May 1966), pp. 803–804.
18. Burckhardt, C. B., "Diffraction of a Plane Wave at a Sinusoidally Stratified Dielectric Grating," J. Opt. Soc. Amer., *56*, No. 11 (November 1966), pp. 1502–1509.
19. Burckhardt, C. B., "Efficiency of a Dielectric Grating," J. Opt. Soc. Amer., *57*, No. 5 (May 1967), pp. 601–603.
20. Bathia, A. B., and Noble, W. J., "Diffraction of Light by Ultrasonic Waves II," Proc. Royal Soc. of London, *220A*, (1953), pp. 369–385.
21. Phariseau, P., "On the Diffraction of Light by Progressive Supersonic Waves," Proc. Ind. Acad. Sci., *44A*, (1956), pp. 165–170.
22. Quate, C. F., Wilkinson, C. D., and Winslow, D. K., "Interaction of Light and Microwave Sound," Proc. IEEE, *53*, No. 10 (October 1965), pp. 1604–1623. This paper includes a comprehensive bibliography of work on acoustic scattering of light.
23. Gordon, E. I., and Cohen, M. G., "Electro-Optic Diffraction Grating for Light Beam Modulation and Diffraction," IEEE J. Quantum Elec., *QE-1*, No. 5 (August 1965), pp. 191–198.
24. Batterman, B., and Cole, H., "Dynamical Diffraction of X-Rays by Perfect Crystals," Rev. Mod. Phys. *36*, No. 3 (July 1964), pp. 681–717. This paper contains a review of the work on the diffraction of X-rays.
25. Saccocio, E. J., "Application of the Dynamical Theory of X-Ray Diffraction to Holography," J. Appl. Phys., *38*, No. 10 (September 1967), pp. 3994–3998.
26. Kogelnik, H., "Reconstructing Response and Efficiency of Hologram Gratings," Proc. Symp. Modern Optics, Polytechnic Inst., Brooklyn, March 1967, pp. 605–617.
27. Kogelnik, H., "Hologram Efficiency and Response," Microwaves, *6*, No. 11 (November 1967), pp. 68–73.
28. Born, M., and Wolf, E., *Principles of Optics*, New York: Pergamon Press, 1959, Chapter 12.
29. Bergstein, L., and Kermisch, D., "Image Storage and Reconstruction in Volume Holography," Proc. Symp. Mod. Opt., Polytechnic Inst. Brooklyn, March 1967, pp. 655–680.
30. Gordon, E. I., "A Review of Acoustooptical Deflection and Modulation Devices," Proc. IEEE, *54*, No. 10 (October 1966), pp. 1391–1401.
31. Belvaux, Y., "Influence of Emulsion Thickness and Absorption in Hologram Reconstruction," Physical Letters, *26A*, No. 5 (January 1968), pp. 190–191.
32. Nassenstein, H., "A New Hologram with Wavelength-Selective Reconstruction" Physical Letters, *28A*, No. 2 (November 1968), pp. 141–142.

33. Tien, P. K., unpublished work.
34. Klein, W. R., "Light Diffraction by Ultrasonic Beams of High Frequency Near Bragg Incidence," Proc. 5th Congress Int. D'Acoustique, Liege, D24 (September 1965).
35. Kogelnik, H., "Bragg Diffraction in Hologram Gratings with Multiple Internal Reflections," J. Opt. Soc. Amer., *57*, No. 3, (March 1967), pp. 431–433.
36. George, N., and Mathews, J. W., "Holographic Diffraction Gratings," Appl. Phys. Lett., *9*, No. 5 (September 1966), pp. 212–125.
37. Latta, J. N., "The Bleaching of Holographic Diffraction Gratings for Maximum Efficiency," Appl. Opt., *7*, No. 12 (December 1968), pp. 2409–2416.
38. Kiemle, H., "Lippman-Bragg Holograms as Periodic Ladder Networks," Frequenz, *22*, No. 7 (July 1968), pp. 206–211.

Statistics on Attenuation of Microwaves by Intense Rain

By D. C. HOGG

(Manuscript received June 12, 1969)

Heavy rainfall and associated attenuation at centimeter and millimeter wavelengths are discussed. Measured attenuations are combined with path-rainfall statistics obtained from a rain-gauge network to produce plots of attenuation versus path length for a given probability of fading. Under the assumption that the spatial behavior of heavy rain is similar at various locations, the path-average rainfall statistics are combined with highly resolved point rain rates for geographically separated places to produce attenuation data appropriate to those places. Dual parallel-path-diversity is also evaluated; it is shown to be a very advantageous arrangement.

I. INTRODUCTION

An important problem in designing wide-band radio-relay systems at frequencies exceeding 10 GHz is reliability. Propagation through heavy rain is the significant factor in determining realiability of the medium. Thus it is important to examine the spatial and temporal behavior of heavy rain and the resultant attenuation.

Recent measurements of progagation at 18.5 GHz and 30.9 GHz, and analysis of rainfall data from the Crawford Hill rain-gauge network of Bell Telephone Laboratories at Holmdel, New Jersey, have led to an improved understanding of the rain environment.[1-4] Those data are used here to provide information on attenuation by rain for use in system design. In particular, the improvement in performance obtained by use of path diversity is evaluated.[5]

II. SINGLE-PATH STATISTICS (NEW JERSEY)

2.1 *The Magnitude of the Attenuation*

First, one must ask: What is the magnitude of the attenuation caused by heavy rain at frequencies exceeding 10 GHz? Figure 1 is a plot of

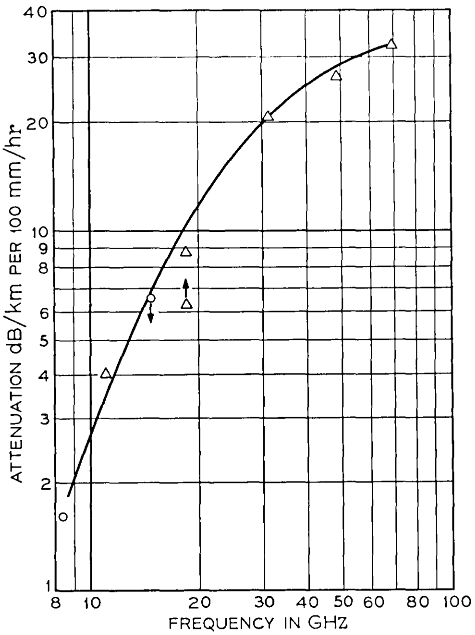


Fig. 1 — Attenuation measured during rain of rate 100 mm/hr (averaged over a 1 km path). The measurements at 8 and 15 GHz are from Ref. 6; 11 GHz from Ref. 7; 18 and 30 GHz from Refs. 1 and 2; and 50 and 70 GHz from Ref. 9. △ indicates Bell Telephone Laboratories data and ○ indicates DRB 1966 data—Canada (some extrapolation for both).

attenuation measured at a rain rate of 100 mm/hr (4 inches per hour) for a path length of 1 km. Data measured at 100 mm/hr, rather than at low rain rates, are used because path-average rain rates of this magnitude do indeed occur a significant percentage of the time in many places, including New Jersey.

Moreover, in the discussion that follows, we are concerned with attenuations caused by path-average rain rates of the order of 100 mm/hr, and the attenuations will be taken to be directly proportional to the path average rain rates; that is, proportional to the average density of rain along the path.* The curve in Fig. 1 serves as a benchmark by means of which attenuation is related to heavy path-average rain rates. Thus

* From theoretical considerations, the attenuation γ, at frequencies of the order 10 GHz is believed related to the rain rate R by $\gamma = \alpha R^{\beta}$. α is a function of frequency as indicated and β is also a mild function of frequency with values near unity. Here we use values of α measured at high rain rates (since β is taken to be unity) to minimize errors in the event β departs from unity.

for heavy rains one has:

$$\gamma = 0.04\mathrm{R}d; \qquad \gamma = 0.1\mathrm{R}d; \qquad \gamma = 0.2\mathrm{R}d$$

for frequencies of 11, 18, and 30 GHz, γ being the attenuation in decibels and R the average rain rate on a path of length d.

2.2 *Dependence of Path-Average Rain Rate on Path Length*

The path analysis of rain rate discussed in Ref. 3 encompasses the heavy rains of 1967 taken on 100 rain gauges forming a $130(\mathrm{km})^2$ grid in New Jersey. Obviously, there are many paths of various lengths in such a network and a relatively large amount of data is obtained for such paths from the several storms that occur during one year. Path average rain rates have been converted to a yearly base and are plotted in Fig. 2;* the curves show the probability of path-average rain rate with path length as parameter. At rain rates of the order 50 mm/hr, the probabilities are about the same for all path lengths, namely, about 0.01 percent; thus the probability of exceeding an average rate of 50 mm/hr on a 10.4 km path is about the same as at a point (path length—zero in Fig. 2). As the rain rate increases, the curves diverge. For example, the probability of a 100 mm/hr rain rate on a 10.4 km path is less by a factor of ten than that for a point; at 150 mm/hr, the factor is one hundred.

The data in Fig. 2 can be examined in another way. Consider a given probability, say, 0.001 percent (five minutes per year); the corresponding rain rate at a point is about 160 mm/hr, whereas for a 10.4 km path it is 80 mm/hr. This behavior tells us that heavy rains occur as localized showers. Of course, this behavior will show up in evaluating the attenuation on paths of various lengths.

2.3 *Dependence of Attenuation on Path Length*

The relationship between attenuation and path-average rain rate at various frequencies as given in Fig. 1, and the probability of occurrence of rain rates, Fig. 2, have been used to produce Figs. 3a and b. Two probability levels (0.01 percent, 50 min/yr, and 0.001 percent, 5 min/yr) and three frequencies (11, 18, and 30 GHz) have been chosen as representative of radio relay. These plots give computed attenuation that is exceeded for the percent of time indicated on the figure as a function of path length. Note that there is curvature in the plots. As one would expect, having looked at Fig. 2, the attenuation one obtains

* From curves A in Fig. 28 of Ref. 3.

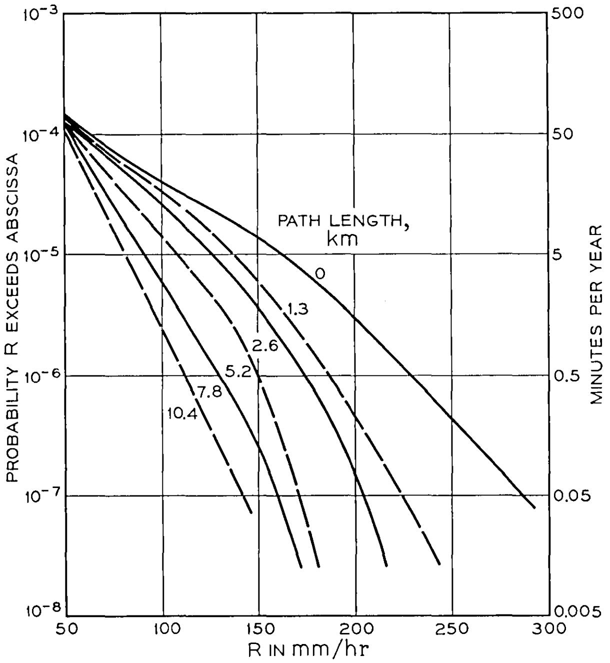


Fig. 2 — Probability of path-average rain rates for paths of various lengths; 1967 rain-gauge network data.

(for a low probability) on a 10 km path is less than one would expect by linearly extrapolating from the attenuation on a 1 km path.

The two propagation paths in operation within the Holmdel rain-gauge network are 1.9 and 6.4 km long at frequencies 30.9 and 18.5 GHz, respectively;[1,2] these lengths are indicated by arrows on the abscissas in Fig. 3.* Percent of time distributions of attenuation on these paths were measured throughout 1967 and 1968 and points taken at the indicated probability level are shown on the figures. For the 1.9 km path, the measured 30.9 GHz attenuations agree well with the computed curves for 30 GHz: somewhat higher in Fig. 3a and slightly lower in Fig. 3b.

Likewise, in Fig. 3a the points measured at 18.5 GHz (6.4 km) are in good agreement with, but are somewhat lower than, the computed curve for 18 GHz. In Fig. 3b the 18.5 GHz measurement for 1968 is somewhat below the computed curve; however, the 1967 measurement

* The 18.5 GHz signal is vertically polarized and the 30.9 GHz signal is polarized 45° from vertical.

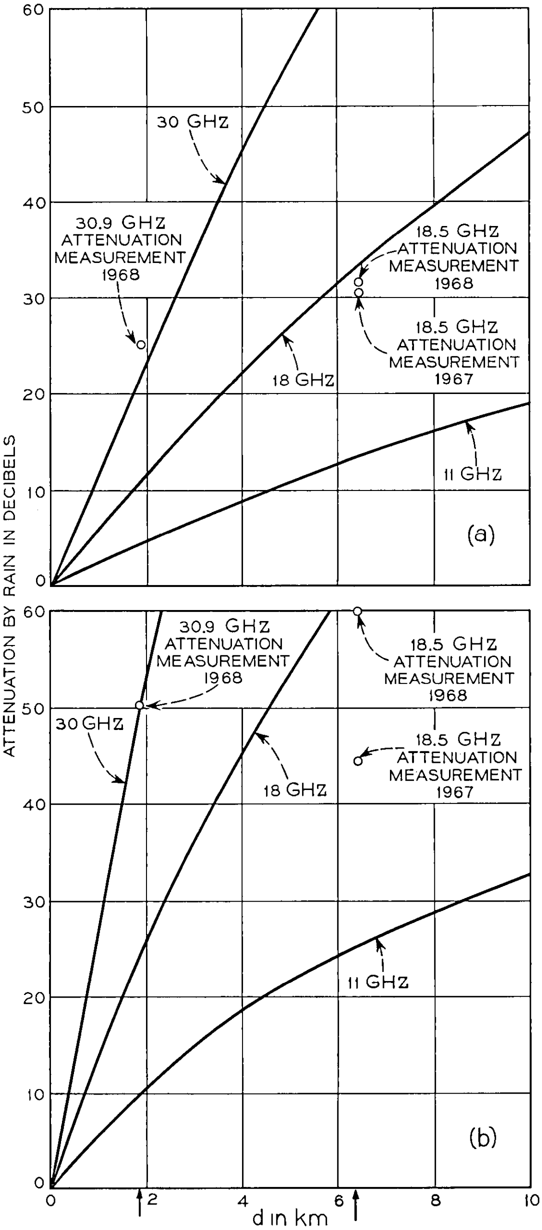


Fig. 3 — Attenuation as a function of path length at 11, 18 and 30 GHz (1967 network data) for (a) 0.01 percent probability (50 min/yr) and (b) 0.001 percent probability (5 min/yr) along with measurements on paths of length 1.9 km (30.9 GHz) and 6.4 km (18.5 GHz).

is considerably lower. Comparison of the 18.5 GHz attenuation distributions for 1967 and 1968 shows that heavy showers were more frequent on this path in 1968 than in 1967. The 18 GHz curves in Figs. 3a and b are apparently somewhat conservative.

Thus use of the attenuations in Fig. 1 to convert the pool of path-average rain rates from the rain-gauge network has led to a set of curves of attenuation versus path length that are consistent with independent measurements of attenuation. Accordingly, for design of a conventional tandem relay system at, say, 18 GHz, with a 30 dB margin, the repeater spacing is, from Fig. 3b, 2.5 km for 0.001 per cent probability on individual paths in coastal New Jersey.

III. SINGLE-PATH STATISTICS (OTHER LOCATIONS)

It is tempting to ask if the knowledge gained from the above studies can be used to say something about the attenuation environment in places other than coastal New Jersey. If certain assumptions are made concerning the spatial distribution in rain showers, that can be done.

3.1 *Point Rainfall Rates of High Resolution*

Distribution of point rain rates with high resolution have been measured in a few places, shown in Fig. 4. Four of the full curves were measured in the United states by the Illinois State Water Survey using a photographic method measured over the best part of a year; they form a consistent set of data.[6] This method is capable of measuring drops in a small volume during a short interval every ten seconds. The solid line for Bedford, England is from a four-year sample;[5] gauges with two-minute resolution were used. The dashed curve is the distribution for the pool of data taken during 1967 on the rain-gauge network at Holmdel, New Jersey;[3] gauges with a time constant less than one second were sampled every ten seconds.

For a given probability of occurrence, how much heavier does it rain at other locations than in New Jersey? Table I shows the point rain-rate intensity in other places relative to New Jersey for the 0.01 and 0.001 per cent levels; the Illinois state survey set of curves and the data from England in Fig. 4 are used in this comparison.

Thus in the regime of low probability (high rain rate), the rain intensity in New Jersey is about one quarter that of Miami, Florida, and five times that of Corvallis, Oregon. These data must now be linked with the spatial distributions obtained in New Jersey in order to determine the attenuations.

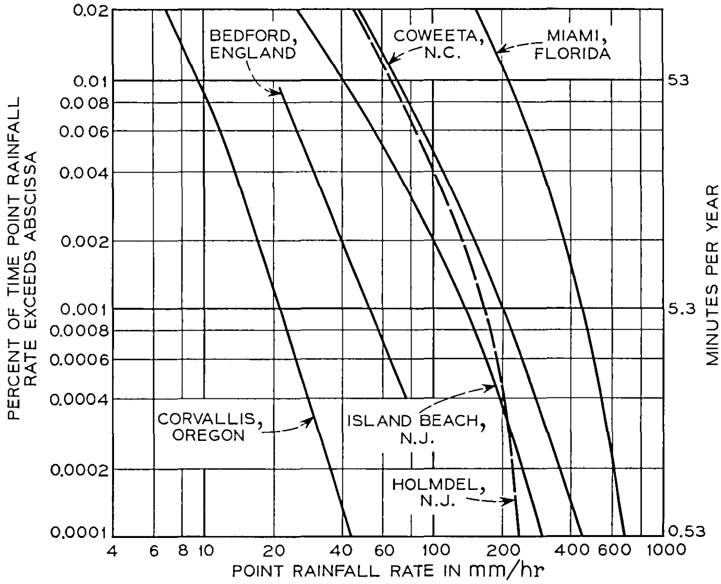


Fig. 4—Point rainfall rates measured in several places by instruments with rapid response.

3.2 *The Spatial Distribution of Rain Showers*

The data in Fig. 2 show that the probability of a given path-average rain rate decreases with increasing path length for heavy rains, a not too surprising result since one is dealing with rain cells of limited size. Likewise, for a given probability level, the path-average rain rate decreases with increasing path length as shown in Fig. 5. For relatively high probability (10^{-4}), this decrease does not amount to much; as shown by the lowest curve in Fig. 5, the average rain rate for a 10 km path is about the same as that for a point ($d = 0$). However, for example, on the upper-

TABLE I—RELATIVE INTENSITY OF POINT RAIN RATES

Probability Level	Miami Florida	Coweeta North Carolina	Island Beach New Jersey	Bedford England	Corvallis Oregon
(a) 10^{-4} (50 min/yr)	5	1.75	1	0.48	0.25
(b) 10^{-5} (5 min/yr)	3.5	1.55	1	0.42	0.15
AVERAGE of (a) & (b)	4.2	1.65	1	0.45	0.2

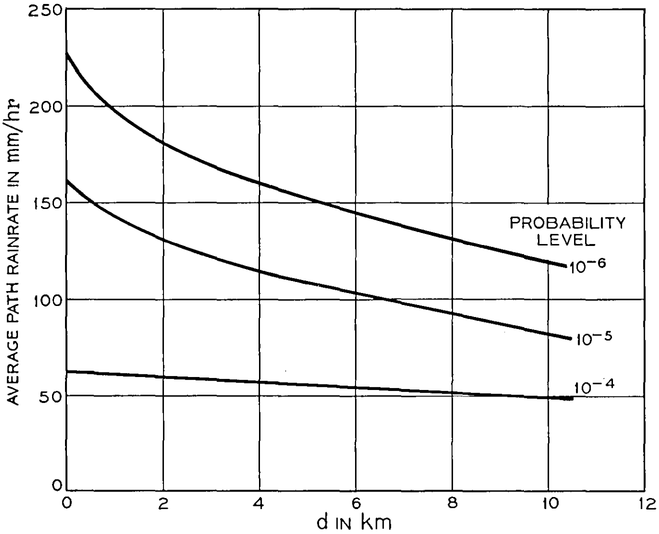


Fig. 5 — Average path rain rate in New Jersey versus path length for probability levels 10^{-4}, 10^{-5}, and 10^{-6}.

most curve in Fig. 2 (for 10^{-6} probability) the average rain rate on a 10 km path is only one half the rate at a point.

Assume that the spatial behavior of heavy rainfall is the same in other places as it is in New Jersey. This means that in a place with relatively low point rain rates (such as Oregon, Fig. 4), path-average rates are about the same as point rates (such as in the lowest curve in Fig. 5), that is, large-area rain. Whereas, where the point rain rates are very high (such as in Florida, Fig. 4), the path-average rates are much less than the point rates (such as in the uppermost curve in Fig. 5), that is, showers. To determine whether this assumption is warranted, one must await spatial measurements of rain rate in other places.

The data in Figs. 2 and 4 are used to construct Table II, a list of path-average rain rates for the various locations, as a function of path length, *d*. Some extrapolation of the curves in Fig. 2 was necessary to obtain the column for Miami, Florida.

Table II has been converted to attenuation at 18 GHz by way of the relationship discussed above as shown in Fig. 6. As one would expect, the attenuation for Oregon is linear with path length, whereas for places with heavy rain, there is considerable curvature. Figure 6 tells us that a single transmission path at 18 GHz with a 30 dB fading margin should not exceed 1, 2, 3, 6, and 15 km in Florida, North Carolina, New Jersey,

TABLE II—PATH-AVERAGE RAIN RATES IN MM/HR FOR THE 10^{-5} PROBABILITY LEVEL

d-km	Corvallis Oregon	Bedford England	Island Beach New Jersey	Coweeta North Carolina	Miami Florida
0	20	55	130	200	450
1.3	20	53	110	165	325
2.6	20	52	103	153	265
5.2	20	50	90	135	215
7.8	20	48	75	110	210
10.4	20	45	70	95	

Bedfordshire-England*, and Oregon, respectively, if a probability of 10^{-5} is stipulated.

One might argue that in Florida (for example) where the water vapor available for production of rain exceeds that of New Jersey, the dimension of a rain cell of given rain rate may exceed that of a cell of the same rain rate in New Jersey. If this were true, the attenuation for Florida and North Carolina in Fig. 6 would be somewhat higher than shown.

IV. PATH DIVERSITY (NEW JERSEY)

The analysis of the rain-gauge network data by Freeny and Gabbe[3] encompasses not only single paths of various lengths but also joint statistics for pairs of parallel paths separated by various distances.† These data are applicable to the design of path-diversity systems in that they are statistics of the percentage of time that the average rain rate on both paths exceeds given values. Of course, the idea in path diversity is to switch to the path with lowest attenuation.[5]

4.1 *Two Parallel Paths with a Given Separation*

An example of how path-average rain rates in the diversity arrangement convert to attenuation is given in Figs. 7a, b, and c for frequencies of 11, 18, and 30 GHz. The curves apply to the 0.001 percent probability level (5 min/yr) and a diversity separation of 5.2 km (3.25 miles). For comparison, the attenuation for a single path is shown by a dashed line

* In a recent Committee Consultatif Internationale Radio document (United Kingdom Document IX/164-E, May 9, 1969), attenuation distributions for a 24 km path, and for the worst year (1968) observed to date, indicate that the path length appropriate to 0.001 percent probability and 30 dB attenuation is something less than 12 km in Bedfordshire at 18 GHz. This presumably means that, even in a relatively low rain-rate environment, the heavier rains do indeed occur as showers of limited size (see also Ref. 5). That being the case, the curve for England in Fig. 6 would have more curvature than indicated, that is, the curve in Fig. 6 would be quite conservative.

† See Fig. 28 of Ref. 3.

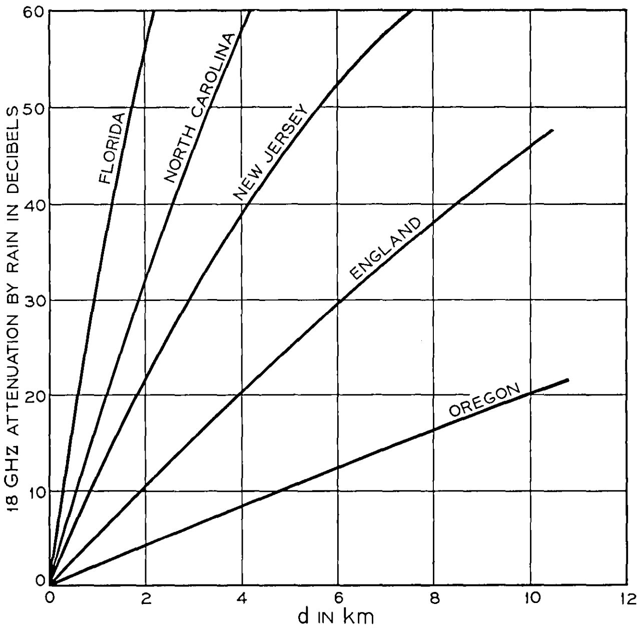


Fig. 6 — 18 GHz attenuation versus path length for various places; probability level, 10^{-5} (5 min/yr).

on each figure. If transmission paths at 18 GHz with a 30 dB fading margin are considered, Fig. 7b shows that the path length in the diversity arrangement with 5.2 km separation can be just over 5 km, compared with 2.5 km when no diversity is used.

4.2 *Relationship Between Interrepeater Path Length and Diversity Separation*

A somewhat more general question is: For a given attenuation margin and a given probability level, how does the inter-repeater path length change with diversity separation? As an example, 18 GHz, 30 dB, and 0.001 percent are chosen for the frequency, margin, and probability level; the data are plotted in Fig. 8. Note that the path length d for a diversity separation s of 7.5 km is 7 km, about thrice the path length (2.5 km), for the nondiversity arrangement ($d = 0$). Results such as these have considerable economic implications. The data can also be plotted as in Fig. 9 where 18 GHz attenuation is given as a function of path length with path separation a parameter. Apparently, for a given

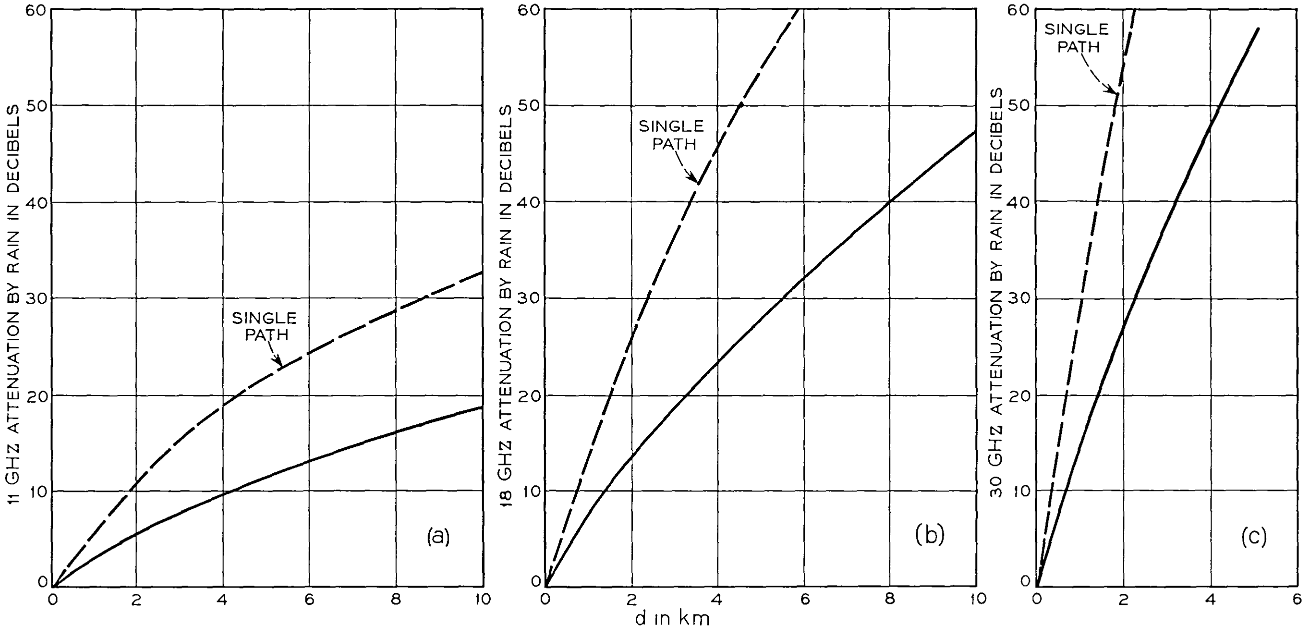


Fig. 7—Attenuation appropriate to a dual parallel-path-diversity separation of 5.2 km as a function of path length for frequencies (a) 11 GHz, (b) 18 GHz, and (c) 30 GHz. The dashed curves are for conventional (nondiversity) paths. (Attenuation exceeded 0.001 percent of the time jointly for two parallel paths spaced 5.2 km apart.)

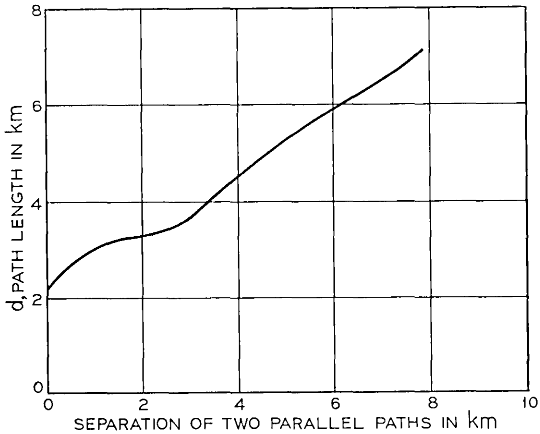


Fig. 8 — Path length as a function of dual parallel-path-diversity separation for a 0.001 percent probability level (5 min/yr) at 18 GHz with a 30 dB attenuation margin; 1967 network data.

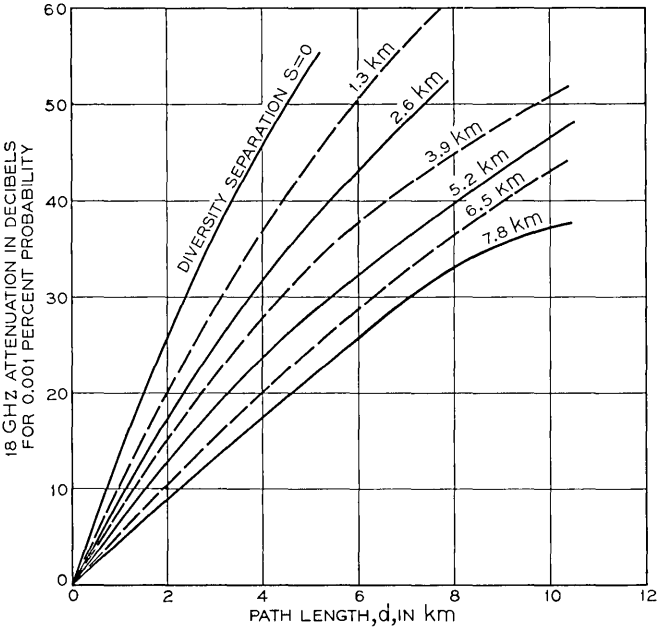


Fig. 9 — 18 GHz attenuation appropriate to 0.001 percent probability versus path length for various diversity separations; 1967 network data.

diversity separation, the advantage of diversity over nondiversity is not a strong function of the fading margin.

As yet we have no actual attenuation measurements on path diversity. However, the data of Figs. 8 and 9 are believed conservative in the same sense as those of Fig. 3; of course, they apply only to coastal New Jersey.

V. DISCUSSION

Although the data given here result in well-resolved design curves hopefully useful in design of radio systems, at least two important questions remain. A system is comprised of many paths in tandem forming a route of length l, whereas here only single paths have been discussed. If one has n such paths in tandem, is the probability P_l of attenuation by rain on the system simply nP_1 where P_1 is the probability for a single path? In other words, is there no correlation between heavy fades on tandem paths? Obviously, if a dense rain cell were centered on a repeater, there would be correlation of attenuation on the two paths associated with that repeater. From such considerations and examination of rainfall data, the relationship $P_l = nP_1$ is believed too conservative.

The other question is related to path diversity. We have only discussed the case of two (single) parallel paths separated by various distances. But in an actual system one deals with several paths in tandem on each leg of the route; these two legs must of course merge if one wishes to switch from one to the other. The path lengths for merge points lie between those given in Fig. 3 and those appropriate to a parallel path diversity arrangement.[5] Moreover, the diversity analysis here deals with two (single) parallel paths of given separation whereas in practice one would be dealing with a line of tandem paths parallel to, and displaced from, a second such set. In that case, the advantage gained by path diversity must be investigated beyond what we have done here.

Finally, it should be pointed out that the microwave systems to which the above discussion is pertinent would carry very wide bands of information. Clearly, the advantages of dual paths in providing equipment diversity (in addition to propagation reliability) would be considerable in such systems, especially from the viewpoint of maintenance.

VI. ACKNOWLEDGMENT

Discussion with R. A. Semplak, J. D. Gabbe, and Mrs. A. E. Freeny stimulated production of this paper.

REFERENCES

1. Semplak, R. A., and Turrin, R. H., "Some Measurements of Attenuation at 18.5 GHz," B.S.T.J., *48,* No. 6 (July-August 1969), pp. 1767–1788.
2. Semplak, R. A., unpublished work.
3. Freeny, A. E., and Gabbe, J. D., "A Statistical Description of Intense Rainfall," B.S.T.J., *48,* No. 6 (July-August 1969), pp. 1789–1852.
4. Semplak, R. A., and Keller, H. E., "A Dense Network for Rapid Measurement of Rainfall Rate," B.S.T.J., *48,* No. 6 (July-August 1969), pp. 1745–1756.
5. Hogg, D. C., "Path Diversity in Propagation of Millimeter Waves Through Rain," IEEE, *AP-15,* No. 3 (March 1967), p. 410.
6. Blevis, B. C., Dohoo, R. M., and McCormick, R. K., "Measurements of Rainfall Attenuation at 8 and 15 GHz," IEEE, *AP-15,* No. 5 (May 1967), pp. 394–403.
7. Hathaway, S. D., and Evans, H. E., "Radio Attenuation at 11 kmc and Implications Affecting Relay System Engineering," B.S.T.J., *38,* No. 1 (January 1959), pp. 73–97.
8. Mueller, E. A., and Sims, A. L., "Investigation of the Quantitative Determination of Point and Areal Precipitation by Radar Echo Measurements," Technical Rep. ECOM-00032-F, Illinois State Water Survey, December 1966.
9. Hogg, D. C., "Millimeter-Wave Communication Through the Atmosphere," Science, *159,* No. 3810 (January 5, 1968), pp. 39–46.

Work-Scheduling Algorithms: A Nonprobabilistic Queuing Study (with Possible Application to No. 1 ESS)

By JOSEPH B. KRUSKAL

(Manuscript received November 7, 1968)

In many large computer systems with real-time use (such as the No. 1 Electronic Switching System), the central processing unit handles much of its work through queues. It may spend much of its time cycling through the queues, performing the work requests it finds there. To accomodate varying degrees of urgency, the cycle may visit some hoppers more often than others. (No. 1 ESS strongly relies on this procedure.) This paper provides an approximate method for evaluating different cycles.

Using the evaluation method and some approximations, we obtain a formula for the optimum relative frequency with which different queues should be visited.

The model used is nonprobabilistic, and treats requests as continuous rather than discrete. The model also ignores certain interdependencies between queues. Despite these drastic simplifications, the results probably provide useful guidance, if interpreted cautiously.

I. INTRODUCTION

In many large computer systems, especially those with real-time use, the central processor handles much of its work through queues, which contain work requests. (The queues may also be called hoppers, buffers, waiting lines, files, and so forth. In this paper we call them hoppers.) The processor examines each hopper in turn, and performs some or all of the work requests if any, which it finds there.

Some work requests require processing more urgently than others. One method of providing appropriate response times is to examine more frequently hoppers which contain urgent work, and other hoppers less frequently. For example, the No. 1 ESS (Electronic Switching

System) has many hoppers which it groups into five different urgency classes.[1,2] The five classes are examined (or "visited") in a fixed recurring cycle, of length 30, during which the classes are visited 15, 8, 4, 2, and 1 times, respectively. (During a single visit to a single class, the individual hoppers are visited once each, in a fixed sequence.)

This paper contains a practical approximate model for evaluating various alternative cycles. The conceptual basis for the evaluation is the expected time each work request must wait in the hopper before being serviced by the central processor. (Such times depend not only on the cycle, but also on the times required to process requests, and on the rates at which new requests are initiated. These are all assumed given.) The expected waiting times for different hoppers are multiplied by frequencies and also by weights w_i, the "average penalty per second of delay," and added. The resulting sum is called P, the "expected total penalty per second." The weights w_i, which reflect the relative importance of delaying different work requests, are assumed given, and we seek to minimize P by choosing the cycle wisely. By way of illustration, the calculations required to evaluate any given cycle are given for two very simple cycles.

When applied to general cycles, our model yields the plausible conclusion that visits to the same hopper should be spaced as evenly as possible around the cycle (in terms of elapsed time between visits). Furthermore, the model permits us to estimate how sensitive P is to deviations from this ideal.

Our most important conclusion is an explicit formula for how frequently each hopper should be visited. To obtain this formula, we assume that visits to each hopper are evenly spaced around the cycle. Then P becomes a function of the visit frequency (and not of detailed visit pattern). We explicitly optimize this function, to obtain a formula for visit frequencies.

The time required to examine a hopper, whether or not it contains any work requests, is small but highly significant, and is an important consideration in the problem. Our model explicitly reflects this fact. (Indeed, it is known though sometimes overlooked that the No. 1 ESS central processor finds most hoppers empty on a majority of its visits, even when it is heavily loaded with work and operating near its capacity limits. This can occur because the number of hoppers is so large, and because each work request requires a relatively long time to service compared with the time to visit a single hopper.)

In this study, we assume that work enters the hoppers as a result of some outside process, which is independent of how the hoppers are being served. In No. 1 ESS, as in many other situations, much work does enter hoppers in this manner. However, it is also true that servicing a request from one hopper may place work, directly or indirectly, in another hopper. This interdependence may well be important in choice of a cycle. Nevertheless, the present model, which ignores such interdependence, is probably usable if we are suitably cautious about interpreting our results.

Service requests are discrete items and enter the hoppers according to an exceedingly complicated random process. Our model, however, assumes that each kind of request comes in at a constant rate, with no statistical fluctuation whatsoever. Furthermore, we treat the number of requests as a continuous quantity (so that requests keep trickling in like water) rather than a discrete quantity.

Despite the drastic nature of all these simplifications, we believe that this analysis is better than no analysis at all. Furthermore, we feel that our conclusions are probably valid approximations. It also seems plausible that our model could provide the jumping-off place for a more realistic study. Both interdependence and statistical fluctuation could be introduced in a limited way. (Since this was first written, R. W. Landgraff has done a study which extends this model to include interdependence.[3]) This might well permit their main effects to enter the model, without opening the Pandora's box of an extremely general stochastic process with one server and many interdependent queues.

II. SOME ASSUMPTIONS AND NOTATION

We suppose that there are I hoppers. For each hopper i we assume that we have three parameters:

s_i = service time = average time to service one request in this hopper,

r_i = request time = average time between occurrence of requests $\gg s_i$, and

w_i = weight = average penalty per second of delay for a single request of type i.

We also use

$$\lambda_i = \frac{s_i}{r_i} \ll 1, \qquad \Lambda = \sum_{i=1} \lambda_i \ .$$

(To permit a steady-state solution, we assume $\Lambda < 1$.) Note that the definition of w_i implies that *on the average* the penalty for delaying one kind of task is *proportional* to the delay time. The w_i are the proportionality constants. This simple assumption could be refined somewhat without too much trouble if desired.

In No. 1 ESS, one major penalty caused by hopper delays is the extra waiting time they cause to the telephone user at various stages of his call. For some hoppers, such as those involved during the process of dialing, undue delays can cause mishandling of the call. (Also, the delays tie up memory capacity and indirectly cause a need for extra memory equipment. However, this effect is probably minor.) By considering the loss incurred by the user due to various waiting periods, and the loss due to the probability of mishandled call, it would be possible to assign sensible values to the w_i . Although a truly realistic appraisal of the losses would require a quite elaborate study, some fairly reasonable simplifying assumptions which would make this study much simpler are available. Furthermore, assignment of the w_i on a direct intuitive basis would probably be adequate for many purposes.

To measure the total delay penalty paid by any work-scheduling algorithm, we combine the various penalties into a single number P:

d_i = expected delay for a request of type i,
p_i = expected penalty per request of type $i = w_i\, d_i$, and
P = expected total penalty per second

$$= \sum_{i=1}^{I} \frac{1}{r_i} p_i = \sum_{i=1}^{I} \frac{w_i}{r_i} d_i .$$

(Of course, $1/r_i$ is the expected number of requests of type i in one second.) We seek to minimize P by proper choice of a work-scheduling algorithm. Only the delays d_i may be influenced in this way, so we concentrate on evaluating the d_i .

A model which, like ours, treats requests as continuous has the danger of "discovering" that the hoppers are serviced infinitely fast, accumulating only an infinitesimal amount of work between visits. The following assumption, which in any case reflects an important reality, avoids this collapse.

To examine the ith hopper, whether or not it contains any work, requires a certain amount of time. We assume this amount of time is H_i . For simplicity we shall assume all the H_i are equal, and shall call their common value H, although it would be easy to work with unequal values if desired. Thus if x requests are serviced during one visit to hopper i, this visit requires $H_i + xs_i$ seconds.

It will turn out later that the value chosen for H is not very important in the context of this model. The comparison between different work-scheduling algorithms is unaffected by the (nonzero) value used.

III. WORK-SCHEDULING AND SERVICE POLICY

We suppose that the hoppers are visited in a fixed cycle of length N, namely,

$$(i_1\,,\, i_2\,,\, \cdots\,,\, i_N).$$

This means that hopper i_1 is visited first, then hopper i_2, and so on. After i_N is visited, the cycle starts over again with hopper i_1. One simple cycle with $I = 4$ and $N = 6$ is (1, 4, 2, 4, 3, 4). No. 1 ESS uses $I = 5$ hoppers (classes of hoppers, actually), and a cycle of length $N = 30$:

$$\overbrace{\overbrace{\overbrace{1\ 2\ 1}\ 3\ \overbrace{1\ 2\ 1}}\ 4\ \overbrace{\overbrace{1\ 2\ 1}\ 3\ \overbrace{1\ 2\ 1}}}\ 5\ \overbrace{\overbrace{\overbrace{1\ 2\ 1}\ 3\ \overbrace{1\ 2\ 1}}\ 4\ \overbrace{\overbrace{1\ 2\ 1}\ 3\ \overbrace{1\ 2}}}$$

If i is any given hopper, we shall let $V(i)$ indicate the set of all visits to hopper i. Thus for the cycle (1, 4, 2, 4, 3, 4), we have

$$V(1) = [1], \quad V(2) = [3], \quad V(3) = [5], \quad \text{and} \quad V(4) = [2, 4, 6].$$

In the No. 1 ESS cycle,

$$V(1) = [1, 3, 5, \cdots, 29], \qquad V(2) = [2, 6, 10, 14, 18, 22, 26, 30],$$
$$V(3) = [4, 12, 20, 28], \qquad V(4) = [8, 24], \qquad V(5) = [16].$$

For any visit n, the last previous visit to the *same* hopper is called $b(n)$ ("b" for before). Thus in the cycle (1, 4, 2, 4, 3, 4), visit 6 is to hopper 4, and the last previous visit to the same hopper is on visit 4. Thus $b(6) = 4$. Because "last previous" is understood in a cyclic sense, $b(2) = 6$. We have

$$b(1) = 1, \qquad b(2) = 6, \qquad b(3) = 3,$$
$$b(4) = 2, \qquad b(5) = 5, \qquad b(6) = 4.$$

Whenever a hopper is visited, we suppose that all work requests there are serviced. However, during the period when the hopper is being serviced new requests can enter it. What about these requests which enter the hopper while it is being serviced? These can either be handled when they are reached during the same visit, which we call the "come-right-in" policy, or they can be left for the next visit to the hopper, which we call the "please-wait" policy. We shall treat both of these hopper service policies, because their solutions are very similar.

IV. HOW TO EVALUATE P

As there is no statistical variation left in our model, it is easy to analyze. Let

t_n = time spent emptying the hopper i_n during visit n.

Let C be the time spent during an entire cycle, so that C consists of N hopper visits. Hopper visit n consists of time H to examine the hopper, and time t_n to service it. Thus

$$C = \sum_{n=1}^{N} (H + t_n) = NH + \sum_{n=1}^{N} t_n .$$

Now consider the requests which are serviced during t_n . Let

T_n = the interval during which they enter hopper i_n .

Recalling that $b(n)$ is the last prior visit to hopper i_n , we see from Fig. 1 that

$$T_n = \begin{cases} \sum_{p=b(n)+1}^{n} (H + t_p) = [n - b(n)]H + \sum_{b(n)+1}^{n} t_p , & \text{``come-right-in;''} \\ \sum_{p=b(n)}^{n-1} (t_p + H) = [n - b(n)]H + \sum_{b(n)}^{n-1} t_p , & \text{``please-wait.''} \end{cases} \tag{1}$$

Note that $b(n)$ and the summation indices must be understood in a suitable "cyclic" sense, so that (for example) if $b(n) = n$, then $n - b(n)$ means once around the cycle and hence equals N, not 0. Now it is easy to see that

$$\begin{aligned}(\text{the number of requests served during } t_n) &= t_n/s_{i_n} \\ &= (\text{the number of requests initiated during } T_n) = T_n/r_{i_n} ,\end{aligned}$$

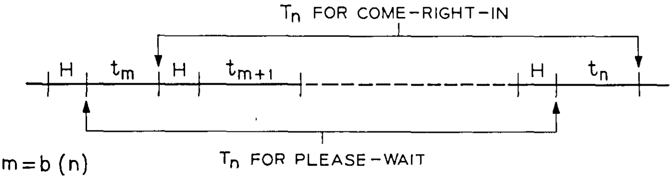


Fig. 1 —Time flow diagram illustrating "please-wait" and "come-right-in" policies.

so

$$t_n = \lambda_{i_n} T_n \,. \tag{2}$$

By using equation (2), we can eliminate either all T_n or all t_n from equations (1). This will leave us with N linear equations in N unknowns, which in fact turn out to be linearly independent. By solving these equations and using equation (2), we can find the T_n and the t_n , and from them all else will follow, as we show below. For convenient reference, we state the equations after eliminating the t_n :

$$T_n = \begin{cases} [n - b(n)]H + \sum\limits_{p=b(n)+1}^{n} \lambda_{i_p} T_p \,, & \text{“come-right-in;”} \\ [n - b(n)]H + \sum\limits_{p=b(n)}^{n-1} \lambda_{i_p} T_p \,, & \text{“please-wait.”} \end{cases} \tag{3}$$

Recall the special cyclic interpretation of $n - b(n)$ and the summations.

It is worth digressing briefly to derive an explicit formula for C, and to show how the N equations (3) can be reduced to $N - I$ equations in $N - I$ unknowns by using it. It is easy to see that if we sum T_n over all visits to some particular hopper j, the result must equal C:

$$\sum_{n \text{ in } V(j)} T_n = C \quad \text{for every } j. \tag{4}$$

Now sum equation (3) over all n in $V(j)$, and use equation (4) several times:

$$\begin{aligned} \sum_{n \text{ in } V(j)} T_n &= \sum_{n \text{ in } V(j)} \left\{ [n - b(n)]H + \sum_{p=b(n)+1}^{n} \lambda_{i_p} T_p \right\}, \\ C &= NH + \sum_{p=1}^{N} \lambda_{i_p} T_p \\ &= NH + \sum_{j=1}^{I} \lambda_j \Big[\sum_{p \text{ in } V(j)} T_p \Big] \\ &= NH + \Big(\sum_{j=1}^{I} \lambda_j \Big) C \\ &= NH + \Lambda C. \end{aligned}$$

This yields

$$C = \frac{NH}{1 - \Lambda}. \tag{5}$$

Since C is now given directly in terms of known quantities, we can use equation (4) to solve for one T_n in terms of others. We can do this separately for each $j = 1$ to I, and thereby reduce the number of unknowns and equations to N–I.

Once we have the values of T_n (and hence of t_n), we may easily evaluate e_n, the average delay for requests serviced during visit n. (Each delay is reckoned from occurrence of request to when its processing starts.) By elementary reasoning, we see that

$$\left.\begin{aligned} e_n &= \tfrac{1}{2}(T_n - t_n), \quad \text{“come-right-in,”} \\ e_n &= \tfrac{1}{2}(T_n + t_n), \quad \text{“please-wait.”} \end{aligned}\right\} \tag{6}$$

Of course T_n/r_{i_n} requests are serviced in visit n. Thus the average delay per request of type j is

$$d_j = \frac{\sum\limits_{n \text{ in } V(j)} \dfrac{T_n}{r_j} e_n}{\sum\limits_{n \text{ in } V(j)} \dfrac{T_n}{r_j}}. \tag{7}$$

Using equations (6), (2), and (4), we get

$$\left.\begin{aligned} d_i &= \frac{1 - \lambda_i}{2C} \sum_{n \text{ in } V(i)} T_n^2, \quad \text{“come-right-in,”} \\ d_i &= \frac{1 + \lambda_i}{2C} \sum_{n \text{ in } V(i)} T_n^2, \quad \text{“please-wait.”} \end{aligned}\right\} \tag{8}$$

Now let

$$F_n = T_n/C = \text{fraction of a cycle used by } T_n,$$

so that

$$\sum_{n \text{ in } V(i)} F_n = 1, \qquad \text{all } i. \tag{9}$$

Then

$$d_i = \begin{cases} \tfrac{1}{2}(1 - \lambda_i)C \sum\limits_{n \text{ in } V(i)} F_n^2, & \text{“come-right-in,”} \\ \text{same, but with } 1 + \lambda_i \text{ for } 1 - \lambda_i, & \text{“please-wait.”} \end{cases} \tag{10}$$

Using equation (5) and the definition of P, we now easily find a formula for the penalty P, which is the key quantity we use to evaluate work-scheduling algorithms:

$$P = \begin{cases} \dfrac{NH}{2(1 - \Lambda)} \left\{ \sum_i \dfrac{w_i}{r_i} (1 - \lambda_i) \sum_{n \text{ in } V(i)} F_n^2 \right\}, & \text{“come-right-in,”} \\ \text{same, but with } 1 + \lambda_i \text{ for } 1 - \lambda_i, & \\ 1 - \Lambda \text{ is unaffected} & \text{“please-wait.”} \end{cases} \tag{11}$$

(However, note that the values of the F_n may differ for the two policies.) We note that the work-scheduling algorithm influences equation (11) in only two ways: through N, and through the fractions F_n. From this formula we can evaluate and compare different work-scheduling algorithms. Also we can compare "come-right-in" with "please-wait."

V. SOME EXAMPLES

If there are $I = 3$ different hoppers, the simplest possible cycle is (1, 2, 3), for which $N = 3$. In this case we see trivially that $F_1 = F_2 = F_3 = 1$, for either "come-right-in "or "please-wait." Thus equation (10) for cycle (1, 2, 3) is:

$$P = \begin{cases} \dfrac{3H}{2(1 - \Lambda)} \sum_1^3 \dfrac{w_i}{r_i} (1 - \lambda_i), & \text{“come-right-in,”} \\ \text{same, but with } 1 + \lambda_i \text{ for } 1 - \lambda_i, & \\ 1 - \Lambda \text{ is unaffected} & \text{“please-wait.”} \end{cases}$$

Given the three input parameters s_i, r_i, and w_i for each hopper, this can be evaluated numerically.

Now suppose we use the cycle (1, 2, 1, 3), for which $N = 4$, with the "come-right-in" policy. Then equations (3) for cycle (1, 2, 1, 3) become the following four equations:

$$\begin{aligned} T_1 &= \lambda_1 T_1 + \lambda_3 T_4 + 2H, \\ T_2 &= \lambda_1(T_1 + T_3) + \lambda_2 T_2 + \lambda_3 T_4 + 4H, \\ T_3 &= \lambda_2 T_2 + \lambda_1 T_3 + 2H, \\ T_4 &= \lambda_1(T_1 + T_3) + \lambda_2 T_2 + \lambda_3 T_4 + 4H. \end{aligned}$$

However, taking C as known, and using equation (4) for cycle (1, 2, 1, 3) namely,

$$T_1 + T_3 = C, \qquad T_2 = C, \qquad T_4 = C,$$

we eliminate the unknowns T_2, T_3, and T_4, leaving one equation in one unknown, T_1 :

$$T_1 = \lambda_1 T_1 + \lambda_3 C + 2H.$$

We find

$$T_1 = \frac{1}{1 - \lambda_1} [2H + \lambda_3 C].$$

Dividing by C, and using $C = 4H/(1 - \Lambda)$ from equation (5), we see

$$F_1 = \frac{\frac{1}{2}(1 - \Lambda) + \lambda_3}{1 - \lambda_1} = \frac{1}{2} \frac{1 - \lambda_1 + \lambda_3 - \lambda_2}{1 - \lambda_1}$$

$$= \frac{1}{2}\left[1 + \frac{\lambda_3 - \lambda_2}{1 - \lambda_1}\right].$$

As $F_3 = 1 - F_1$, we find

$$F_1^2 + F_3^2 = \frac{1}{2}\left[1 + \left(\frac{\lambda_3 - \lambda_2}{1 - \lambda_1}\right)^2\right],$$

and also

$$F_2^2 = 1, \qquad F_4^2 = 1.$$

Thus equation (11) for cycle (1, 2, 1, 3) with the "come-right-in" policy is

$$P = \frac{4H}{2(1 - \Lambda)} \left\{ \frac{w_1}{r_1} (1 - \lambda_1) \frac{1}{2}\left[1 + \left(\frac{\lambda_3 - \lambda_2}{1 - \lambda_1}\right)^2\right] + \frac{w_2}{r_2}(1 - \lambda_2) + \frac{w_3}{r_3}(1 - \lambda_3) \right\}.$$

Through special circumstances which would not hold in general, the values for F_n using this cycle are all the same for "please-wait" as for "come-right-in," so P for "please-wait" is the same as the above but with $1 + \lambda_i$ substituted for $1 + \lambda_i$ in three places. Given the parameters s_i , r_i , and w_i for each hopper, this can be evaluated numerically.

VI. CONCLUSIONS

If we compare cycles of the same length and with the same number of visits to each hopper, then equation (11) yields the following conclusion: *The visits to a given hopper should be spaced as evenly around the cycle as possible.*

By this we mean that the values of T_n (and hence of F_n) pertaining to this hopper should be as equal as possible. This follows because the minimum of

$$\sum_{n \text{ in } V(i)} F_n^2 \quad \text{subject to} \quad \sum_{n \text{ in } V(i)} F_n = 1$$

occurs when the F_n with n in $V(i)$ are all equal. Furthermore, equation (11) can be used to estimate how serious any given deviation from equality is.

Suppose a cycle has N_i visits to hopper i, so that $N = \Sigma N_i$, and suppose that the N_i visits are spaced approximately evenly around the cycle for every i. Then for each visit n to hopper i,

$$F_n \approx \frac{1}{N_i}.$$

Thus

$$P \approx \frac{HN}{2(1 - \Lambda)} \sum_i \frac{w_i}{r_i} (1 - \lambda_i) \frac{1}{N_i}, \quad \text{"come-right-in."}$$

Either using a Lagrange multiplier to handle the constraint that $\Sigma N_i = N$, or by direct argument (see the appendix), it is easy to deduce that the values of N_i which minimize this satisfy

$$N_i \text{ proportional to } \left[\frac{w_i}{r_i}(1 - \lambda_i)\right]^{1/2}$$

so

$$\frac{N_i}{N} = \frac{\left[\frac{w_i}{r_i}(1 - \lambda_i)\right]^{1/2}}{\sum_i \left[\frac{w_i}{r_i}(1 - \lambda_i)\right]}.$$

This yields our most important conclusion: *The above approximate formula gives the optimum relative frequency of visits to each hopper in the cycle.*

By obtaining values for r_i, s_i, and less easily for w_i, it is possible to compare different work-scheduling algorithms with each other and with the "ideal" schedule with perfect spacing implied above. Notice that the actual value of H does not enter into this comparison. (If we had used unequal values for the H_i, only the ratios H_i/H_j would enter into the comparison, not the actual values of the H_i themselves.)

It would probably be worthwhile to analyze the actual work-scheduling algorithm used for ESS No. 1 in these terms. It would be interesting to compare this actual algorithm with the "ideal" algorithm.

Our model, with its highly simplified assumptions, cannot possibly provide the last word on work-scheduling evaluations, even with regard to delay times. However, this kind of approach is probably desirable. If greater realism is desired, the most important aspects are statistical variability and interdependence of hoppers.

APPENDIX

Direct Argument to Replace the LaGrange Multiplier Argument

Henry Pollak has pointed out a simple direct argument which shows that $\Sigma(a_i/N_i)$ is minimized, subject to the constraint $\Sigma N_i = N$, if N_i is proportional to $(a_i)^{\frac{1}{2}}$. Using $a_i = w_i(1 - \lambda_i)/r_i$, this yields the formula given above for N_i.

First, let $q = N/[\Sigma(a_i)^{\frac{1}{2}}]$. Now, we multiply the quantity to be minimized by q^2, and express it:

$$\Sigma \frac{q^2 a_i}{N_i} = \Sigma\left(q\left(\frac{a_i}{N_i}\right)^{1/2} - (N_i)^{1/2}\right)^2 + 2q\Sigma(a_i)^{1/2} - \Sigma N_i \,.$$

The middle term is constant by definition, and the last term is constant by constraint. The first term cannot be less than 0. The first term is 0 if

$$\frac{q^2 a_i}{N_i} = N_i \quad \text{or} \quad N_i = q(a_i)^{1/2}.$$

Since these values satisfy the constraint, we obtain the desired result.

REFERENCES

1. Keister, W., Ketchledge, R. W., and Vaughn, H. E., "No. 1 ESS: System Organization and Objectives," B.S.T.J., *43*, No. 5 (September 1964), pp. 1831–1844.
2. Harr, J. A., Hoover, Mrs. E. S., and Smith, R. B., "Organization of the No. 1 ESS Stored Program," B.S.T.J., *43*, No. 5 (September 1964), pp. 1923–1959.
3. Landgraff, R. W., unpublished work.

Some Properties of a Nonlinear Model of a System for Synchronizing Digital Transmission Networks*

By IRWIN W. SANDBERG

(Manuscript received April 22, 1969)

J. R. Pierce has recently proposed a system for synchronizing an arbitrary number of geographically separated oscillators, and, under the assumption of zero transmission delays between stations, has shown that a certain linear model of the system is stable in the sense that all of the station frequencies approach a common final value as $t \rightarrow \infty$.

The purpose of this paper is to report on some results concerning the dynamic behavior of a nonlinear version of an important special case of Pierce's model. The nonlinear model takes into account transmission delays.

It is proved under certain very general conditions that the nonlinear model possesses the stability property required of a synchronization system. More explicitly, it is proved that the model is stable for all nonnegative values of the delays. The results show that the model possesses some additional fundamental properties of engineering interest, and they provide an analytical basis for using a computer for further studies. In particular, a complete solution to the problem of determining the final frequency of the system and the final value of the content of an arbitrary buffer is presented, in the sense that it is shown that these quantities can be determined by solving a certain set of nonlinear equations which is proved to possess a unique solution.

I. INTRODUCTION

The purpose of this paper is to report on some results concerning properties of the solution $f_1(t)$, $f_2(t)$, $\cdots$, $f_n(t)$ of the set of equations

* This paper was presented as an invited contribution at both the Symposium on Mathematical Aspects of Electrical Networks (sponsored by the American Mathematical Society, New York City, April 1969) and the Joint Conference on Mathematical and Computer Aids to Design (sponsored by the ACM, SIAM, and IEEE; Anaheim, California; October 1969).

$$f_i(t) = \varphi_i\left\{\sum_{j\neq i}\varphi_{ij}\left\{\int_0^t [f_j(\tau - \tau_{ij}) - f_i(\tau)]\,d\tau + b_{ij}(0)\right\}\right\} + c_i$$
$$i = 1, 2, \cdots, n$$
$$t \geqq 0 \qquad (1)$$

in which n is an arbitrary positive integer such that $n \geqq 2$, the $\varphi_i(\cdot)$ and the $\varphi_{ij}(\cdot)$ are monotone functions that map the real interval $(-\infty, \infty)$ into itself, the τ_{ij} are nonnegative constants, and the c_i and the $b_{ij}(0)$ are real constants.

The set of equations (1) governs the behavior of a nonlinear model of the key part of a system for synchronizing digital transmission networks. Our main result is that synchronization is possible under very general conditions concerning the nonlinearities and the time delays τ_{ij}. In addition, an analytical basis for computing the final frequency of the system is presented; this involves proving that a certain set of nonlinear equations possesses a unique solution. Other results are presented concerning, for example, buffer requirements* and certain monotonicity properties of the frequency functions $f_i(\cdot)$.

1.1 *Pierce's Model*

When $\tau_{ij} = 0$ for all $i \neq j$, when $\varphi_i(x) = x$ for all i and all real x, and when $\varphi_{ij}(x) = a_{ij}x$ for all real x and all $i \neq j$, in which a_{ij} is a real constant for all $i \neq j$, we have

$$f_i(t) = \sum_{j\neq i} a_{ij}\left\{\int_0^t [f_j(\tau) - f_i(\tau)]\,d\tau + b_{ij}(0)\right\} + c_i$$
$$i = 1, 2, \cdots, n \qquad t \geqq 0. \qquad (2)$$

Equations (2) are the equations of a linear model of the principal part of a system for synchronizing digital transmission networks recently proposed by J. R. Pierce.[1] His system employs oscillators of adjustable frequency and buffers which accept pulses at an incoming rate and which produce corresponding output pulses at the local clock rate.

In Pierce's model the content b_{ij} of the buffer at station i which accepts pulses from station j is assumed to satisfy the equation†

$$\dot{b}_{ij}(t) = f_j(t) - f_i(t), \qquad t \geqq 0 \qquad (3)$$

in which $f_j(t)$ and $f_i(t)$ are the frequencies at time t at stations j and i,

* An explanation of the function of the device called a buffer is given in Section 1.1.

† As usual, a dot over a mathematical symbol denotes the derivative with respect to time.

respectively, and the overall system of coupled oscillators is assumed to satisfy equations (2) with $a_{ij} = a_{ji} \geqq 0$ for all $i \neq j$. Under the natural assumption that there is some path from each station to every other station, Pierce has shown, by directing attention to a passive RL network analog of equations (2), that the model is stable in the sense that each frequency f_i approaches the same final value as $t \rightarrow \infty$.*

1.2 *The Nonlinear Model*

Our interest in the properties of the solution of equations (1) arises as a consequence of Pierce's work as follows. First, we wish to take into account the time delay τ_{ij} associated with transmission to an arbitrary station i from an arbitrary station $j \neq i$. Thus we replace $f_j(t)$ by $f_j(t - \tau_{ij})$ in (3) and (2). The content $b_{ij}(t)$ of the ijth buffer is then

$$\int_0^t [f_j(\tau - \tau_{ij}) - f_i(\tau)]\, d\tau + b_{ij}(0) \tag{4}$$

for all $t \geqq 0$.

Our mathematical model of a buffer does not reflect the fact that the capacity of a real buffer is bounded; a real buffer is a device that can store at most some fixed finite number of pulses. Therefore it makes sense to study how a linear model of a synchronization system employing buffers, such as the one governed by (2), can be modified to reduce the possibility of occurrence of buffer overload (that is, the possibility that the capacity of the buffers will be exceeded). It is therefore reasonable to replace the expression (4) for the buffer content by some monotone nonlinear function $\varphi_{ij}(\cdot)$ of (4), with the idea in mind that $\varphi_{ij}(\cdot)$ is a function with moderate slope near the origin and very large slope corresponding to values of (4) that are in the neighborhoods of buffer overload. Similarly, in order to ease the requirements on the extent to which the frequencies of the adjustable oscillators must be variable, and in order to reduce the tendency of very large excursions in the frequencies f_i during a transient phase, it is reasonable to replace the sum

$$\sum_{j \neq i} \varphi_{ij}[b_{ij}(t)] \tag{5}$$

formed at the ith station by some monotone nonlinear function $\varphi_i(\cdot)$

* In Ref. 1 Pierce actually deals with a more general linear model than we have described here, but treats in most detail the important case described above. In connection with the more general model, Pierce has exploited the network analogy further in order to obtain an expression for the final frequency, and to make assertions concerning the behavior of the system when certain elements are nonlinear. For additional material dealing with various aspects of the problem of synchronizing geographically separated oscillators, see, for example, Refs. 2–7. In particular, Ref. 4 contains a short history of the problem.

of (5), in which $\varphi_i(\cdot)$ has moderate slope near the origin and very small slope far from the origin.

These considerations lead at once to the study of the properties of the set of equations (1). Of course the crucial question is: "Does the system governed by (1) possess the basic stability property required of a synchronization system?" Our main result concerning (1) is that, no matter what the values of the time delays τ_{ij}, under some conditions which are quite trivial from the engineering viewpoint (and rather weak from the mathematical viewpoint), it does.

II. SUMMARY OF RESULTS, AND SOME APPLICATIONS

2.1 *The Main Result Concerning (1)*

In order to describe the result, we first introduce some definitions and assumptions.

Definition 1: Let M denote an arbitrary $n \times n$ matrix with elements m_{ij}. Let the *graph of M* denote the graph containing n vertices (that is, n nodes), a directed edge (that is, a directed line segment or arc) from node j to node i for every pair i, j with $i \neq j$ and $m_{ij} \neq 0$, and no other directed edges.

Definition 2: Let M denote an arbitrary $n \times n$ matrix. Then we shall say that the graph of M is a *communicating graph* if and only if there is some path (not necessarily a direct path) from each node to every other node.

We assume throughout the paper that:

(*i*) τ_{ij} denotes an arbitrary nonnegative constant for all $i \neq j$.

(*ii*) For each i, $\varphi_i(\cdot)$ denotes a real-valued continuously differentiable function defined on $(-\infty, \infty)$ such that

$$\underline{k}_i \leqq \varphi_i'(x) \leqq \bar{k}_i \tag{6}$$

for all x, with $\underline{k}_i$ and $\bar{k}_i$ positive constants.

(*iii*) For each $i \neq j$, $\varphi_{ij}(\cdot)$ denotes a continuously differentiable real-valued function defined on $(-\infty, \infty)$ such that either $\varphi_{ij}(x) = 0$ for all x, or

$$\underline{k}_{ij} \leqq \varphi_{ij}'(x) \leqq \bar{k}_{ij} \tag{7}$$

for all x, with $\underline{k}_{ij}$ and $\bar{k}_{ij}$ positive constants.*

* At the price of some additional complication, we could have replaced assumptions (*ii*) and (*iii*) with assumptions concerning the behavior of the $\varphi_i(\cdot)$ and the $\varphi_{ij}(\cdot)$ on finite intervals. See Section 2.2.

(*iv*) The matrix M defined by

$$(M)_{ii} = 0 \qquad \text{for all} \quad i$$

$$(M)_{ij} = \varphi'_{ij}(0) \quad \text{for all} \quad i \neq j$$

is the matrix of a communicating graph.

(*v*) Each $f_i(\cdot)$ is defined and differentiable on $[-\bar{\tau}, \infty)$ in which $\bar{\tau} = \max_{i \neq j}\{\tau_{ij}\}$.

Assumption (*iv*) possesses a simple physical interpretation. It is a natural connectivity assumption of the type needed if synchronization is to be possible in the sense that all of the station frequencies approach a common final value as $t \to \infty$.

Our basic set of equations is

$$f_i(t) = \varphi_i\left\{\sum_{j \neq i} \varphi_{ij}\left\{\int_0^t [f_j(\tau - \tau_{ij}) - f_i(\tau)]\, d\tau + b_{ij}(0)\right\}\right\} + c_i \tag{8}$$

for all i and all $t \geqq 0$. By differentiating both sides of these equations with respect to t, we have

$$\dot{f}_i(t) = \varphi'_i[\xi_i(t)] \sum_{j \neq i} \varphi'_{ij}[\xi_{ij}(t)][f_j(t - \tau_{ij}) - f_i(t)], \qquad t \geqq 0 \tag{9}$$

for all i, in which of course

$$\xi_i(t) = \sum_{j \neq i} \varphi_{ij}\left\{\int_0^t [f_j(\tau - \tau_{ij}) - f_i(\tau)]\, d\tau + b_{ij}(0)\right\}$$

and

$$\xi_{ij}(t) = \int_0^t [f_j(\tau - \tau_{ij}) - f_i(\tau)]\, d\tau + b_{ij}(0).$$

Let $h_{ij}(t) = \varphi'_i[\xi_i(t)]\varphi'_{ij}[\xi_{ij}(t)]$ for all $t \geqq 0$ and all $j \neq i$. Then

$$\dot{f}_i(t) = \sum_{j \neq i} h_{ij}(t)[f_j(t - \tau_{ij}) - f_i(t)], \qquad t \geqq 0 \tag{10}$$

for all i. According to Theorem 1 (Section III) the coefficients $h_{ij}(\cdot)$ of (10) are such that there exists a real constant ρ with the property that for all i, $f_i(t) - \rho \to 0$ as $t \to \infty$. This means that the system is stable in the sense that all of the station frequencies approach a common final value. Note that this result does not involve assumptions concerning the values of the nonnegative delays τ_{ij}, that it is valid for monotone nonlinearities of a very general type, and that it does not involve symmetry assumptions such as $\varphi_{ij}(\cdot) = \varphi_{ji}(\cdot)$ for all $i \neq j$.

2.2. *A Monotonicity Property of the* $f_i(\cdot)$

The first of the two lemmas used in the proof of Theorem 1 asserts that the solution $f_1(\cdot), f_2(\cdot), \cdots, f_n(\cdot)$ of (10) possesses an interesting monotonicity property. Let T be an arbitrary nonnegative value of time t, and let the upper envelope and lower envelope $\bar{f}(t)$ and $\underline{f}(t)$, respectively, of the $f_i(t)$ be defined for each $t \geqq -\bar{\tau}$ by $\bar{f}(t) = \max_i f_i(t)$, $\underline{f}(t) = \min_i f_i(t)$. Let $\bar{f}_{\bar{\tau}}(T)$ and $\underline{f}_{\bar{\tau}}(T)$, respectively, denote the largest and smallest value of $\bar{f}(t)$ and $\underline{f}(t)$ for t belonging to the interval $[-\bar{\tau} + T, T]$. Then, according to the lemma just referred to, $\bar{f}(t) \leqq \bar{f}_{\bar{\tau}}(T)$ and $\underline{f}(t) \geqq \underline{f}_{\bar{\tau}}(T)$ for all $t \geqq T$. In particular, since the $f_i(t)$ approach a common final value, we see that the interval envelope functions $\bar{f}_{\bar{\tau}}(T)$ and $\underline{f}_{\bar{\tau}}(T)$ approach each other as $T \rightarrow \infty$.

Our assumptions (*ii*) and (*iii*) on the $\varphi_i(\cdot)$ and the $\varphi_{ij}(\cdot)$ concern the behavior of those functions for all, and in particular arbitrarily large, arguments. The upper and lower bounds just described show that it would have sufficed to have made similar assumptions on the behavior of the $\varphi_i(\cdot)$ on any finite interval $[-a, a]$ such that for all i

$$\varphi_i(x) \,\varepsilon\, [\underline{f}_{\bar{\tau}}(0) - \max_i c_i \,,\ \bar{f}_{\bar{\tau}}(0) - \min_i c_i]$$

for all $x \,\varepsilon\, [-a, a]$. On the basis of bounds of the type described in Section 2.4, similar statements can be made concerning the pertinent range of arguments of the $\varphi_{ij}(\cdot)$.

2.3. *Final-Frequency Determination*

We now turn our attention to the matter of determining the final frequency of the model governed by (1).

Let

$$p_i(t) = \int_0^t f_i(\tau)\, d\tau \tag{11}$$

for all $t \geqq 0$ and all i. Then, since for all $t \geqq 0$

$$\int_0^t f_j(\tau - \tau_{ij})\, d\tau = \int_0^{(t-\tau_{ij})} f_j(\tau)\, d\tau + \int_{-\tau_{ij}}^0 f_j(\tau)\, d\tau,$$

we have, using (1),

$$\dot{p}_i(t) = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[p_j(t - \tau_{ij}) - p_i(t) + \lambda_{ij}]\} + c_i \tag{12}$$

for all i and all $t \geqq 0$, in which

$$\lambda_{ij} = b_{ij}(0) + \int_{-\tau_{ij}}^0 f_j(\tau)\, d\tau. \tag{13}$$

According to Theorem 2 (Section III), there exists a unique real constant ρ and some real n-vector q such that

$$\rho = \varphi_i\{\sum_{j\neq i}\varphi_{ij}[-\rho\tau_{ij} + q_j - q_i + \lambda_{ij}]\} + c_i \quad \text{for all} \quad i. \tag{14}$$

With ρ and q such that (14) is satisfied, let

$$p_i(t) = \rho t + q_i + r_i(t), \qquad t \geqq -\bar{\tau} \tag{15}$$

for all i, in which the q_i are the components of q, and the $r_i(t)$ are some functions of t. Then, using (12),

$$\begin{aligned}\rho + \dot{r}_i(t) = \varphi_i\{\sum_{j\neq i}\varphi_{ij}[-\rho\tau_{ij} &+ q_j - q_i + \lambda_{ij} \\ &+ r_j(t - \tau_{ij}) - r_i(t)]\} + c_i\end{aligned} \tag{16}$$

for all i and all $t \geqq 0$. But, using (14) and (16),

$$\dot{r}_i(t) = \varphi_i\{\sum_{j\neq i}\varphi_{ij}[r_j(t - \tau_{ij}) - r_i(t) + s_{ij}]\} - \varphi_i\{\sum_{j\neq i}\varphi_{ij}[s_{ij}]\} \tag{17}$$

for all i and all $t \geqq 0$, in which $s_{ij} = -\rho\tau_{ij} + q_j - q_i + \lambda_{ij}$.

For each i and each $t \,\varepsilon\, [0, \infty)$, we have, by the mean-value theorem,

$$\begin{aligned}&\varphi_i\{\sum_{j\neq i}\varphi_{ij}[r_j(t - \tau_{ij}) - r_i(t) + s_{ij}]\} - \varphi_i\{\sum_{j\neq i}\varphi_{ij}[s_{ij}]\} \\ &\qquad = \varphi_i'[u_i(t)]\{\sum_{j\neq i}\varphi_{ij}[r_j(t - \tau_{ij}) - r_i(t) + s_{ij}] - \sum_{j\neq i}\varphi_{ij}[s_{ij}]\}\end{aligned}$$

for some $u_i(t)$ such that $u_i(t)$ lies within the closed interval with endpoints $\sum_{j\neq i}\varphi_{ij}[s_{ij}]$ and $\sum_{j\neq i}\varphi_{ij}[r_j(t - \tau_{ij}) - r_i(t) + s_{ij}]$. Similarly for each $j \neq i$ and each $t \,\varepsilon\, [0, \infty)$,

$$\begin{aligned}&\varphi_{ij}[r_j(t - \tau_{ij}) - r_i(t) + s_{ij}] - \varphi_{ij}[s_{ij}] \\ &\qquad = \varphi_{ij}'[w_{ij}(t)][r_j(t - \tau_{ij}) - r_i(t)]\end{aligned}$$

for a suitably chosen $w_{ij}(t)$. Therefore (17) can be written as

$$\dot{r}_i(t) = \sum_{j\neq i} c_{ij}(t)[r_j(t - \tau_{ij}) - r_i(t)] \tag{18}$$

for all i and all $t \geqq 0$, where $c_{ij}(t) = \varphi_i'[u_i(t)]\varphi_{ij}'[w_{ij}(t)]$. But, by Theorem 1, the coefficients $c_{ij}(\cdot)$ of (18) are such that there exists a constant σ with the property that for all i, $r_i(t) \rightarrow \sigma$ as $t \rightarrow \infty$. It follows [see (18)] that for all i, $\dot{r}_i(t) \rightarrow 0$ as $t \rightarrow \infty$. Since

$$\int_0^t f_i(\tau)\, d\tau = \rho t + q_i + r_i(t), \qquad t \geqq 0$$

for all i, it is clear that ρ is the final value of the $f_i(\cdot)$.

According to Theorem 2: there exists exactly one real n-vector q such that, with $U^{tr} = (1, 1, \cdots, 1)$,

$$U^{tr}q = \varphi_i\{\sum_{j\neq i}\varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}]\} + c_i$$

for all i, and $\rho = U^{tr}q$.

There are some simple special cases in which we can exhibit an explicit expression for ρ. Suppose, for example, that $\tau_{ij} = 0$ for all $i \neq j$, that $b_{ij}(0) = -b_{ji}(0)$ for all $i \neq j$, and that $\varphi_{ij}(x) = -\varphi_{ji}(-x)$ for all $i \neq j$ and all real x. Then, using (14), we have for all i

$$\varphi_i^{-1}(\rho - c_i) = \sum_{j\neq i}\varphi_{ij}[q_j - q_i + b_{ij}(0)]$$

in which $\varphi_i^{-1}(\cdot)$ is the inverse of $\varphi_i(\cdot)$, and

$$\sum_i \varphi_i^{-1}(\rho - c_i) = \sum_i \sum_{j\neq i}\varphi_{ij}[q_j - q_i - b_{ij}(0)] = 0.$$

Therefore, $n\rho = \sum_i c_i$ if $\varphi_i(x) = x$ for all real x and all i, or if $n = 2$ and $\varphi_1(x) = \varphi_2(x) = -\varphi_2(-x)$ for all real x.

Finally, as a relevant application of the material of Section 2.2, we have when $\tau_{ij} = 0$ for all $i \neq j$

$$\min_i (c_i + \varphi_i\{\sum_{j\neq i}\varphi_{ij}[b_{ij}(0)]\}) \leqq \rho \leqq \max_i (c_i + \varphi_i\{\sum_{j\neq i}\varphi_{ij}[b_{ij}(0)]\})$$

since $\bar{f}(t) \leqq \max_i f_i(0)$ and $\underline{f}(t) \geqq \min_i f_i(0)$ for all $t \geqq 0$, and, by (1),

$$f_i(0) = c_i + \varphi_i\{\sum_{j\neq i}\varphi_{ij}[b_{ij}(0)]\}$$

for all i.

2.4. *Bounds on Buffer Content*

In order to analytically formulate specifications to be met by real buffers such that buffer overload does not occur in a real synchronization system of the type under study, it is natural to consider the problem of obtaining useful upper bounds on the contents of the mathematical buffers of our model. We do not treat this entire problem in detail in this paper. However, we show here that under some strong assumptions, it is possible to exploit the material of Sections 2.2 and 2.3 to obtain a simple *uniform* bound on buffer content. In addition, in terms of the constant ρ and the vector q introduced in Section 2.3, we present a complete solution to the problem of evaluating the *final value* of the content of an arbitrary buffer.

According to Theorem 2, the vector q that satisfies (14) is unique to within an additive n-vector of the form αU, in which α is a real constant and U is the transpose of $(1, 1, \cdots, 1)$. In particular, the quantity $\Delta_q = (\max_i q_i - \min_i q_i)$ associated with any solution pair ρ, q of (14) is unique. In this section it is shown that when

$$\tau_{ij} = b_{ij}(0) = 0 \quad \text{for all} \quad i \neq j, \tag{19}$$

then the magnitude of the content

$$\int_0^t [f_j(\tau) - f_i(\tau)]\, d\tau \tag{20}$$

of an arbitrary buffer is bounded for all $t \geqq 0$ by $2\Delta_q$.

Let (19) be satisfied. As in Section 2.3, let

$$p_i(t) = \int_0^t f_i(\tau)\, d\tau, \qquad t \geqq 0$$

for all i. Then with $p_i(t) = \rho t + q_i + r_i(t)$, $t \geqq 0$ for all i, in which ρ and q satisfy (14), we find as in Section 2.3 that for suitably chosen functions $u_i(\cdot)$ and $w_{ij}(\cdot)$,

$$\dot{r}_i(t) = \sum_{j \neq i} c_{ij}(t)[r_j(t) - r_i(t)], \qquad t \geqq 0 \tag{21}$$

for all i, in which $c_{ij}(t) = \varphi_i'[u_i(t)]\varphi_{ij}'[w_{ij}(t)]$. Since (21) is an equation of the same type as (10) (more precisely, see Lemma 1 of the proof of Theorem 1), it follows that for all $t \geqq 0$, $r_i(t) \leqq \max_i r_i(0)$ and $r_i(t) \geqq \min_i r_i(0)$. But $r_i(0) = -q_i$ for all i. Thus, for any j and i with $j \neq i$

$$\begin{aligned} p_j(t) - p_i(t) &= q_j - q_i + r_j(t) - r_i(t), \qquad t \geqq 0 \\ &\leqq 2\Delta_q\,, \qquad t \geqq 0 \end{aligned}$$

and, similarly, $p_j(t) - p_i(t) \geqq -2\Delta_q$, $t \geqq 0$.

Concerning the problem of evaluating Δ_q, there are some cases in which it is possible to obtain simple and useful upper bounds. In one simple case we can obtain an explicit expression for Δ_q. For example, suppose that (19) is satisfied and that $n = 2$. Suppose also that $\varphi_1(x) = \varphi_2(x) = -\varphi_2(-x)$ for all x, and that $\varphi_{12}(x) = \varphi_{21}(x) = -\varphi_{21}(-x)$ for all x. Then $\rho = \varphi_1[\varphi_{12}(q_2 - q_1)] + c_1$, $\rho = \varphi_2[\varphi_{21}(q_1 - q_2)] + c_2$, and, using the fact that $\varphi_2(\cdot)$ and $\varphi_{21}(\cdot)$ are odd, $2\varphi_1[\varphi_{12}(q_2 - q_1)] = c_2 - c_1$. Therefore, in this case $\Delta_q = |\, q_2 - q_1 \,| = |\, \varphi_{12}^{-1}\{\varphi_1^{-1}[\frac{1}{2}(c_2 - c_1)]\} \,|$.

We now consider the matter of (proving the existence of and) evaluating the final value $\lim_{t \to \infty} b_{ij}(t)$ of the content of an arbitrary buffer. With ρ, q, the $r_i(\cdot)$, and the $p_i(\cdot)$ as defined in Section 2.3, we have for $t \geqq 0$ and any $i \neq j$

$$\begin{aligned} b_{ij}(t) &= \int_0^t [f_j(\tau - \tau_{ij}) - f_i(\tau)] \, d\tau + b_{ij}(0) \\ &= p_j(t - \tau_{ij}) - p_i(t) + b_{ij}(0) + \int_{-\tau_{ij}}^{0} f_j(\tau) \, d\tau \\ &= -\rho\tau_{ij} + q_j - q_i + r_j(t - \tau_{ij}) - r_i(t) + b_{ij}(0) + \int_{-\tau_{ij}}^{0} f_j(\tau) \, d\tau . \end{aligned}$$

Since $r_j(t - \tau_{ij}) - r_i(t) \to 0$ as $t \to \infty$, we have the result

$$\lim_{t \to \infty} b_{ij}(t) = -\rho\tau_{ij} + q_j - q_i + b_{ij}(0) + \int_{-\tau_{ij}}^{0} f_j(\tau) \, d\tau . \qquad (22)$$

Finally, if (19) is satisfied, then, using (22),

$$\max_{j \neq i} \left| \lim_{t \to \infty} b_{ij}(t) \right| = \max_{j \neq i} | q_j - q_i | = \Delta_q ,$$

which shows that our *uniform bound* $2\Delta_q$ is not unreasonable.

2.5 *Discussion*

The results presented in this paper are concerned with a reasonably realistic strongly-nonlinear model of an important type of synchronization system. They answer several key questions concerning the dynamic behavior of the system, and provide an analytical basis for using a computer for further studies in so far as we have proved, for example, that a solution pair ρ, q of the set of equations (14) exists, that this pair is unique in the sense indicated, and that it can be determined by computing the unique solution q of a related set of equations.

On the other hand, although we have proved that under very general conditions our nonlinear model possess the basic properties of a synchronization system, in this paper we have not considered the next natural problem, that of determining the extent to which the system performance can be improved as a result of the presence of the nonlinearities. There are several other important practical problems that are not considered here, such as the problem of predicting the effects of variable transmission delays (due to temperature changes). There is a clear need for much more work in this area, especially in connection with the problem of comparing the performance of alternative synchronization systems.

III. THEOREMS 1 AND 2

Throughout Sections III and IV:

(*i*) n denotes an arbitrary fixed positive integer such that $n \geqq 2$;

the statement "for all i" means for all $i = 1, 2, \cdots, n$, and "for all $j \neq i$" means for all $j \in \{1, 2, \cdots, n\}$ except $j = i$.

(*ii*) With v an arbitrary n-vector, v^{tr} denotes the transpose of v. The zero n-vector is denoted by θ.

(*iii*) If x denotes a differentiable function of t, then $\dot{x}$ indicates the derivative of x with respect to t.

(*iv*) All functions and constants considered are real valued.

The following two theorems are proved in Section IV.

Theorem 1: Suppose that the following conditions are satisfied:

(*i*) *For each $i \neq j$, $a_{ij}(\cdot)$ denotes a nonnegative bounded measurable function defined for all $t \in [0, \infty)$.*

(*ii*) *With $\underline{a}$ and $\bar{a}$ positive constants such that $\underline{a} \leqq \bar{a}$, for each $i \neq j$, $a_{ij}(\cdot)$ satisfies either $a_{ij}(t) = 0$ for all $t \in [0, \infty)$ or $\underline{a} \leqq a_{ij}(t) \leqq \bar{a}$ for all $t \in [0, \infty)$.*

(*iii*) *For $t \in [0, \infty)$, the $n \times n$ matrix A, with $(A)_{ij} = a_{ij}(t)$ for all $i \neq j$ and $(A)_{ii} = 0$ for all i, is the matrix of a communicating graph.*[*]

(*iv*) *For each $i \neq j$, τ_{ij} denotes a nonnegative constant and $\bar{\tau} = \max_{i \neq j} \tau_{ij}$.*

(*v*) *For each i, $x_i(\cdot)$ denotes a differentiable function defined on $[-\bar{\tau}, \infty)$ such that*

$$\dot{x}_i(t) = \sum_{j \neq i} a_{ij}(t)[x_j(t - \tau_{ij}) - x_i(t)], \qquad t \geqq 0$$

for all i.

Then there exists a constant ρ such that $x(t) - \rho U \rightarrow \theta$ as $t \rightarrow \infty$, in which $U = (1, 1, \cdots, 1)^{tr}$.

Theorem 2: Suppose that assumptions (i) through (iv) in Section 2.1 are satisfied. Let U denote the n-vector $(1, 1, \cdots, 1)^{tr}$. Then (a) there exists a unique n-vector q such that

$$U^{tr}q = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}]\} + c_i \quad \text{for all} \quad i,$$

in which the λ_{ij} and the c_i are constants, and (b) concerning the solution ρ, q of

$$\rho = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho\tau_{ij} + q_j - q_i + \lambda_{ij}]\} + c_i \quad \text{for all} \quad i,$$

the value of ρ is unique, and q is unique to within an additive n-vector αU, in which α is an arbitrary real constant.

[*] See definitions 1 and 2 in Section 2.1.

IV. PROOF OF THEOREMS 1 AND 2

In this section:

(*i*) *1_n denotes the identity matrix of order n.*
(*ii*) *The transpose of any matrix M is denoted by M^{tr}.*
(*iii*) *If v is an n-vector, then $\| v \|$ denotes $(v^{tr}v)^{\frac{1}{2}}$.*
(*iv*) *If F denotes an n-vector-valued function, then $(F)_i$ denotes the i^{th} component of F.*

4.1. *Proof of Theorem 1*

We first prove the following lemma.

Lemma 1: Suppose that (i), (iv), and (v) of Theorem 1 are satisfied. For all $t \in [-\bar{\tau}, \infty)$, let $\bar{x}(t)$ and $\underline{x}(t)$ denote $\max_q x_q(t)$ *and* $\min_q x_q(t)$, *respectively. Let T be a nonnegative constant. Then, for all $t \geqq T$, $\bar{x}(t) \leqq$* $\sup_{[-\bar{\tau}+T, T]} \bar{x}(t)$ *and* $\underline{x}(t) \geqq \inf_{[-\bar{\tau}+T, T]} \underline{x}(t)$.

Proof: (upper bound) We have for all i

$$\dot{x}_i(t) = \sum_{j \neq i} a_{ij}(t)[x_j(t - \tau_{ij}) - x_i(t)], \qquad t \geqq 0. \tag{23}$$

Thus

$$\dot{x}_i(t) + x_i(t) \sum_{j \neq i} a_{ij}(t) = \sum_{j \neq i} a_{ij}(t)x_j(t - \tau_{ij}), \qquad t \geqq 0$$

and

$$x_i(t) = x_i(T) \exp\left[-\int_T^t \sum_{j \neq i} a_{ij}(t)\, dt\right]$$
$$+ \int_T^t \exp\left[-\int_\tau^t \sum_{j \neq i} a_{ij}(t)\, dt\right] \sum_{j \neq i} a_{ij}(\tau)x_j(\tau - \tau_{ij})\, d\tau, \quad t \geqq T \tag{24}$$

for all i.

It is convenient to introduce the function $I(\cdot, \cdot, \cdot)$ defined by

$$I(u, v, k) = \exp\left[-\int_u^v \sum_{j \neq k} a_{kj}(t)\, dt\right]$$

for all real $u \leqq v$ and all positive integer $k \leqq n$. Thus, for example, (24) is equivalent to

$$x_i(t) = x_i(T)I(T, t, i) + \int_T^t I(\tau, t, i) \sum_{j \neq i} a_{ij}(\tau)x_j(\tau - \tau_{ij})\, d\tau,$$
$$t \geqq T. \tag{25}$$

Let t_0 denote an arbitrary positive constant. There exist an index k and a $t_1 \in [T, T + t_0]$ such that

$$x_k(t_1) = \sup_{[T, T+t_0]} \bar{x}(t).$$

Clearly

$$x_k(t_1) = x_k(T)I(T, t_1, k) + \int_T^{t_1} I(\tau, t_1, k) \sum_{j \neq k} a_{kj} x_j(\tau - \tau_{kj}) \, d\tau.$$

Therefore, since the a_{kj} are nonnegative,

$$\begin{aligned} x_k(t_1) &\leqq x_k(T)I(T, t_1, k) + \int_T^{t_1} I(\tau, t_1, k) \sum_{j \neq k} a_{kj}(\tau) \, d\tau \\ &\qquad\qquad \cdot \max_{j \neq k} \sup_{[T-\tau_{kj}, t_1-\tau_{kj}]} x_j(t) \\ &\leqq x_k(T)I(T, t_1, k) + \int_T^{t_1} I(\tau, t_1, k) \sum_{j \neq k} a_{kj}(\tau) \, d\tau \sup_{[T-\bar{\tau}, t_1]} \bar{x}(t). \end{aligned}$$

But

$$\int_T^{t_1} I(\tau, t_1, k) \sum_{j \neq k} a_{kj}(\tau) \, d\tau = 1 - I(T, t_1, k).$$

Thus

$$x_k(t_1) \leqq x_k(T)I(T, t_1, k) + [1 - I(T, t_1, k)] \sup_{[T-\bar{\tau}, t_1]} \bar{x}(t).$$

Either

$$\sup_{[T-\bar{\tau}, T]} \bar{x}(t) \leqq \sup_{[T, t_1]} \bar{x}(t) \tag{26}$$

or

$$\sup_{[T-\bar{\tau}, T]} \bar{x}(t) > \sup_{[T, t_1]} \bar{x}(t). \tag{27}$$

If (26) holds, then

$$x_k(t_1) \leqq x_k(T)I(T, t_1, k) + [1 - I(T, t_1, k)]x_k(t_1)$$

[since $x_k(t_1) = \sup_{[T, t_1]} \bar{x}(t)$], and hence

$$x_k(t_1) \leqq x_k(T),$$

which implies that $x_k(t_1) \leqq \sup_{[T-\bar{\tau}, T]} \bar{x}(t)$. If (27) holds, then [since $x_k(T) \leqq \sup_{[T-\bar{\tau}, T]} \bar{x}(t)$]

$$x_k(t_1) \leqq I(T, t_1 , k) \sup_{[T-\bar{\tau},T]} \bar{x}(t) + [1 - I(T, t_1 , k)] \sup_{[T-\bar{\tau},T]} \bar{x}(t)$$
$$\leqq \sup_{[T-\bar{\tau},T]} \bar{x}(t).$$

We have shown that

$$\sup_{[T,T+t_0]} \bar{x}(t) \leqq \sup_{[T-\bar{\tau},T]} \bar{x}(t). \tag{28}$$

But t_0 is an arbitrary positive number. Therefore

$$\sup_{t \geqq T} \bar{x}(t) \leqq \sup_{[T-\bar{\tau},T]} \bar{x}(t).$$

(lower bound) Our proof of the inequality

$$\inf_{t \geqq T} \underline{x}(t) \geqq \inf_{[T-\bar{\tau},T]} \underline{x}(t) \tag{29}$$

parallels the derivation of the upper bound, and is outlined below.

There exists an index l and a $t_2 \,\varepsilon\, [T, T + t_0]$ such that

$$x_l(t_2) = \inf_{[T,T+t_0]} \underline{x}(t).$$

Thus

$$x_l(t_2) \geqq x_l(T)I(T, t_2 , l) + [1 - I(T, t_2 , l)] \inf_{[T-\bar{\tau},t_2]} \underline{x}(t). \tag{30}$$

Either

$$\inf_{[T-\bar{\tau},T]} \underline{x}(t) \geqq \inf_{[T,t_2]} \underline{x}(t)$$

or

$$\inf_{[T-\bar{\tau},T]} \underline{x}(t) < \inf_{[T,t_2]} \underline{x}(t).$$

In either case, we find using (30), that (29) is satisfied. □

We note that it is a consequence of Lemma 1 that the components of $x(\cdot)$ are bounded on $[0, \infty)$, and, since $x(\cdot)$ and $\dot{x}(\cdot)$ are related by (23), that the components of $\dot{x}(\cdot)$ are bounded on $[0, \infty)$.

Assume that

$$\sup_{[u-\bar{\tau},u]} \bar{x}(t) - \inf_{[u-\bar{\tau},u]} \underline{x}(t)$$

[$\bar{x}(\cdot)$ and $\underline{x}(\cdot)$ are defined in the statement of Lemma 1] *does not* approach zero as $u \to \infty$. We shall show that this assumption implies that the components of $x(\cdot)$ are *not bounded* on $[0, \infty)$, a contradiction.

Since, by assumption, $\sup_{[u-\bar{\tau},u]} \bar{x}(t) - \inf_{[u-\bar{\tau},u]} \underline{x}(t)$ does not approach zero as $u \to \infty$, there exist a positive constant ϵ and a set $\{u_q\}_0^\infty$

with $u_q \varepsilon [0, \infty)$ and $\sup_q u_q = \infty$ such that

$$\sup_{[u_q - \bar{\tau}, u_q]} \bar{x}(t) - \inf_{[u_q - \bar{\tau}, u_q]} \underline{x}(t) \geqq 2\epsilon$$

for all q. For each q let $t_q \varepsilon [u_q - \bar{\tau}, u_q]$ and $t_q' \varepsilon [u_q - \bar{\tau}, u_q]$ be such that

$$\underline{x}(t_q) = \inf_{[u_q - \bar{\tau}, u_q]} \underline{x}(t)$$

$$\bar{x}(t_q') = \sup_{[u_q - \bar{\tau}, u_q]} \bar{x}(t).$$

Of course $\sup_q t_q = \infty$ and $| t_q - t_q' | \leqq \bar{\tau}$. Thus there exists a set $\{\lambda_q\}_0^\infty$ of real constants such that $| \lambda_q | \leqq \bar{\tau}$ for all q, with the property that $\bar{x}(t_q + \lambda_q) - \underline{x}(t_q) \geqq 2\epsilon$ for all q. It follows from the definition of $\underline{x}(\cdot)$ and $\bar{x}(\cdot)$ that for each q there exist indices $l(q)$ and $s(q)$ such that $x_{l(q)}(t_q + \lambda_q) - x_{s(q)}(t_q) \geqq 2\epsilon$.

Finally, since the components of $\dot{x}(\cdot)$ are bounded on $[0, \infty)$, there exists a positive constant δ such that *for all* q $x_{l(q)}(t + \lambda_q) - x_{s(q)}(t) \geqq \epsilon$ for all $t \varepsilon [t_q - \frac{1}{2}\delta, t_q + \frac{1}{2}\delta]$.

At this point we need the following lemma.

Lemma 2: *If the hypotheses of Theorem 1 are satisfied, if T is a nonnegative constant, and if there exist three positive constants t_q, ϵ, and δ and indices $l(q)$ and $s(q)$ such that $t_q - \frac{1}{2}\delta > T + \bar{\tau}$ and $x_{l(q)}(t + \lambda_q) - x_{s(q)}(t) \geqq \epsilon$ for all $t \varepsilon [t_q - \frac{1}{2}\delta, t_q + \frac{1}{2}\delta]$, with λ_q a constant and $| \lambda_q | \leqq \bar{\tau}$, then there exist positive constants ξ and Δ such that, with $\bar{x}(t)$ as defined in the statement of Lemma 1,*

$$\sup_{t \geqq \xi} \bar{x}(t) \leqq \sup_{[T - \bar{\tau}, T]} \bar{x}(t) - \Delta$$

and Δ depends only on $\underline{a}$, $\bar{a}$, $\bar{\tau}$, ϵ, and δ.

Proof:

As in the proof of Lemma 1, it is convenient to introduce the function $I(\cdot, \cdot, \cdot)$ defined by

$$I(u, v, k) = \exp\left[-\int_u^v \sum_{j \neq k} a_{kj}(\tau)\, d\tau \right]$$

for all real $u \leqq v$ and all positive integer $k \leqq n$. The relation between T, $\bar{\tau}$, t_q, and δ is indicated in Fig. 1.

From (23)

$$x_i(t) = x_i(T) I(T, t, i) + \int_T^t I(\tau, t, i) \sum_{j \neq i} a_{ij}(\tau) x_j(\tau - \tau_{ij})\, d\tau$$

for all $t \geqq T$ and all i. By Lemma 1, $\bar{x}(t) \leqq \sup_{[T - \bar{\tau}, T]} \bar{x}(t)$ for all $t \geqq T$.

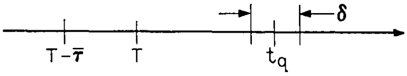


Fig. 1 — Relation between T, $\bar{\tau}$, t_q, δ.

Therefore

$$\sup_{[T-\bar{\tau},T]} \bar{x}(t) - x_{s(q)}(t) \geqq \epsilon \tag{31}$$

for all $t \in [t_q - \frac{1}{2}\delta, t_q + \frac{1}{2}\delta]$.

Let k_1 be an index such that $a_{k_1 s(q)}(t) \neq 0$ for all $t \geqq 0$. Then for $t \geqq t_q + \frac{1}{2}\delta + \bar{\tau}$

$$\begin{aligned} x_{k_1}(t) &= x_{k_1}(T)I(T, t, k_1) + \int_T^t I(\tau, t, k_1) \sum_{j \neq k_1} a_{k_1 j}(\tau) x_j(\tau - \tau_{k_1 j})\, d\tau \\ &\leqq I(T, t, k_1) \sup_{[T-\bar{\tau},T]} \bar{x}(t) \\ &\quad + \int_T^t I(\tau, t, k_1) \sum_{\substack{j \neq k_1 \\ j \neq s(q)}} a_{k_1 j}(\tau) x_j(\tau - \tau_{k_1 j})\, d\tau \\ &\quad + \int_T^{t_q - \frac{1}{2}\delta + \tau_{k_1 s(q)}} I(\tau, t, k_1) a_{k_1 s(q)}(\tau) x_{s(q)}(\tau - \tau_{k_1 s(q)})\, d\tau \\ &\quad + \int_{t_q - \frac{1}{2}\delta + \tau_{k_1 s(q)}}^{t_q + \frac{1}{2}\delta + \tau_{k_1 s(q)}} I(\tau, t, k_1) a_{k_1 s(q)}(\tau) x_{s(q)}(\tau - \tau_{k_1 s(q)})\, d\tau \\ &\quad + \int_{t_q + \frac{1}{2}\delta + \tau_{k_1 s(q)}}^{t} I(\tau, t, k_1) a_{k_1 s(q)}(\tau) x_{s(q)}(\tau - \tau_{k_1 s(q)})\, d\tau. \end{aligned}$$

By Lemma 1, for each j,

$$x_j(\tau - \tau_{k_1 j}) \leqq \sup_{[T-\bar{\tau},T]} \bar{x}(t) \tag{32}$$

for all $\tau \geqq T + \tau_{k_1 j}$. But (32) is obviously satisfied also for $\tau \in [T, T + \tau_{k_1 j}]$. That is, (32) holds for all $\tau \geqq T$ and all j. Thus, using (31),

$$\begin{aligned} x_{k_1}(t) &\leqq I(T, t, k_1) \sup_{[T-\bar{\tau},T]} \bar{x}(t) \\ &\qquad + \int_T^t I(\tau, t, k_1) \sum_{j \neq k_1} a_{k_1 j}(\tau)\, d\tau \sup_{[T-\bar{\tau},T]} \bar{x}(t) \\ &\qquad - \epsilon \int_{t_q - \frac{1}{2}\delta + \tau_{k_1 s(q)}}^{t_q + \frac{1}{2}\delta + \tau_{k_1 s(q)}} I(\tau, t, k_1) a_{k_1 s(q)}(\tau)\, d\tau \\ &\leqq \sup_{[T-\bar{\tau},T]} \bar{x}(t) - \epsilon \int_{t_q - \frac{1}{2}\delta + \tau_{k_1 s(q)}}^{t_q + \frac{1}{2}\delta + \tau_{k_1 s(q)}} I(\tau, t, k_1) a_{k_1 s(q)}\, d\tau \end{aligned}$$

for all $t \geqq t_q + \frac{1}{2}\delta + \bar{\tau}$, since

$$\int_T^t I(\tau, t, k_1) \sum_{j \neq k_1} a_{k_1 j}(\tau)\, d\tau = 1 - I(T, t, k_1).$$

But, for all k_1,

$$\int_{t_q - \frac{1}{2}\delta + \tau_{k_1 s(q)}}^{t_q + \frac{1}{2}\delta + \tau_{k_1 s(q)}} I(\tau, t, k_1) a_{k_1 s(q)}(\tau)\, d\tau \geqq \underline{a} \int_{t_q - \frac{1}{2}\delta + \tau_{k_1 s(q)}}^{t_q + \frac{1}{2}\delta + \tau_{k_1 s(q)}} e^{-(n-1)\bar{a}(t-\tau)}\, d\tau$$
$$\geqq K \exp\{-\beta[t - t_q - \tau_{k_1 s(q)} - \tfrac{1}{2}\delta]\}$$

in which $\beta = (n-1)\bar{a}$ and $K = \underline{a}\{1 - \exp[-(n-1)\bar{a}\delta]\}[(n-1)\bar{a}]^{-1}$. Therefore

$$x_{k_1}(t) \leqq \sup_{[T-\bar{\tau}, T]} \bar{x}(t) - \epsilon K \exp\{-\beta[t - t_q - \tau_{k_1 s(q)} - \tfrac{1}{2}\delta]\}$$

for all $t \geqq t_q + \frac{1}{2}\delta + \bar{\tau}$. In particular,

$$\sup_{[T-\bar{\tau}, T]} \bar{x}(t) - x_{k_1}(t) \geqq \epsilon K e^{-\beta(\delta+\bar{\tau})}$$

for all $t \,\varepsilon\, [t_q + \frac{1}{2}\delta + \bar{\tau}, t_q + \frac{3}{2}\delta + \bar{\tau}]$. Similarly, if the index k_2 is such that $a_{k_2 k_1}(t) \neq 0$ for all $t \geqq 0$, we have for all $t \geqq t_q + \frac{3}{2}\delta + 2\bar{\tau}$

$$\begin{aligned} x_{k_2}(t) &= x_{k_2}(T) I(T, t, k_2) \\ &\qquad + \int_T^t I(\tau, t, k_2) \sum_{j \neq k_2} a_{k_2 j}(\tau) x_j(\tau - \tau_{k_2 j})\, d\tau \\ &\leqq \sup_{[T-\bar{\tau}, T]} \bar{x}(t) - \epsilon K^2 e^{-\beta(\delta+\bar{\tau})} \\ &\qquad \cdot \exp\{-\beta[t - t_q - \tau_{k_2 k_1} - \tfrac{3}{2}\delta - \bar{\tau}]\}. \end{aligned} \tag{33}$$

In particular, for $t \,\varepsilon\, [t_q + \frac{3}{2}\delta + 2\bar{\tau}, t_q + \frac{5}{2}\delta + 2\bar{\tau}]$

$$\sup_{[T-\bar{\tau}, T]} \bar{x}(t) - x_{k_2}(t) \geqq \epsilon K^2 e^{-2\beta(\delta+\bar{\tau})}.$$

Since the graph of A is a communicating graph, we may continue in this manner to obtain an upper bound of the type (33) for *all* of the $x_k(\cdot)$. More explicitly, for each $k_1 \,\varepsilon\, \{1, 2, \cdots, n\}$, let $\{k_1, k_2, \cdots, k_p\}$ denote a finite set of positive integers, with the integer p dependent on k_1, such that $\{k_1, k_2, \cdots, k_p\} \supset \{1, 2, \cdots, n\}$ and

$$a_{k_2 k_1} a_{k_3 k_2} \cdots a_{k_p k(p-1)} \neq 0, \qquad t \geqq 0.$$

Then, with $B = \sup_{[T-\bar{\tau}, T]} \bar{x}(t)$, $u = e^{-\beta(\delta+\bar{\tau})}$, and

$$T_r = t_q + \tfrac{1}{2}\delta + \bar{\tau} + (r-1)(\delta + \bar{\tau}) \quad \text{for all} \quad r = 1, 2, \cdots, p,$$

we have

$$\begin{aligned} x_{k_1}(t) &\leqq B - \epsilon K e^{-\beta\bar{\tau}} e^{-\beta(t-T_1)}, && t \geqq T_1 \\ x_{k_2}(t) &\leqq B - \epsilon K^2 u e^{-\beta\bar{\tau}} e^{-\beta(t-T_2)}, && t \geqq T_2 \\ &\vdots \\ x_{k_p}(t) &\leqq B - \epsilon K^p u^{p-1} e^{-\beta\bar{\tau}} e^{-\beta(t-T_p)}, && t \geqq T_p\,. \end{aligned}$$

Now let $t = T_p + \eta$ with $\eta \varepsilon [0, \bar{\tau}]$. Then

$$\begin{aligned} x_{k_1}(t) &\leqq B - \epsilon K e^{-\beta\bar{\tau}} e^{-\beta(p-1)(\delta+\bar{\tau})} e^{-\beta\eta} \\ x_{k_2}(t) &\leqq B - \epsilon K^2 u e^{-\beta\bar{\tau}} e^{-\beta(p-2)(\delta+\bar{\tau})} e^{-\beta\eta} \\ &\vdots \\ x_{k_p}(t) &\leqq B - \epsilon K^p u^{p-1} e^{-\beta\bar{\tau}} e^{-\beta\eta}. \end{aligned}$$

Thus, for all $t \varepsilon [T_p, T_p + \bar{\tau}]$,

$$x_{k_r}(t) \leqq B - \Delta_{k_1}$$

for all $r = 1, 2, \cdots, p$, in which

$$\Delta_{k_1} = \min_r \{\epsilon K^r u^{r-1} e^{-2\beta\bar{\tau}} e^{-\beta(p-r)(\delta+\bar{\tau})}\}.$$

Let $\Delta = \min_{k_1} \Delta_{k_1}$, and observe that Δ depends only on $\underline{a}$, $\bar{a}$, $\bar{\tau}$, ϵ, and δ. By Lemma 1,

$$\bar{x}(t) \leqq B - \Delta$$

for all $t \geqq T_p + \bar{\tau}$. □

Since as indicated earlier, there are an infinite number of δ-intervals with centers t_q such that $\sup \{t_q\} = \infty$, and such that there exist indices $l(q)$ and $s(q)$ with the property that

$$x_{l(q)}(t + \lambda_q) - x_{s(q)}(t) \geqq \epsilon \tag{34}$$

for all $t \varepsilon [t_q - \frac{1}{2}\delta, t_q + \frac{1}{2}\delta]$, with the constants λ_q such that $|\lambda_q| \leqq \bar{\tau}$ we see that Lemma 2 and the assumption that

$$\sup_{[u-\bar{\tau},u]} \bar{x}(t) - \inf_{[u-\bar{\tau},u]} \underline{x}(t) \tag{35}$$

does not approach zero as $u \to \infty$ imply that $\bar{x}(t) \to -\infty$ as $t \to \infty$, which contradicts the fact that $\bar{x}(\cdot)$ is bounded on $[0, \infty)$. Therefore (35) approaches zero as $u \to \infty$. But, by Lemma 1, $\sup_{[u-\bar{\tau},u]} \bar{x}(t)$ is monotone nonincreasing in u and bounded from below. Thus there is a constant $\bar{L}$ such that

$$\sup_{[u-\bar{\tau},u]} \bar{x}(t) \rightarrow \bar{L}$$

as $u \rightarrow \infty$. Similarly, by Lemma 1, $\inf_{[u-\bar{\tau},u]} \underset{\sim}{x}(t)$ is monotone nondecreasing in u and bounded from above. Thus there is a constant $\underset{\sim}{L}$ such that

$$\inf_{[u-\bar{\tau},u]} \underset{\sim}{x}(t) \rightarrow \underset{\sim}{L}$$

as $u \rightarrow \infty$. But we have proved that $\underset{\sim}{L} = \bar{L}$. Therefore $\bar{x}(t)$ and $\underset{\sim}{x}(t)$ both approach $\underset{\sim}{L}$ as $t \rightarrow \infty$, which means that there is a constant ρ such that

$$x(t) - \rho U \rightarrow \theta \quad \text{as} \quad t \rightarrow \infty. \quad \square$$

4.2. *Proof of Theorem 2*

In part (*a*) of this proof we employ a theorem of R. S. Palais* according to which: if $F(\cdot)$ is a continuously-differentiable mapping of real Euclidean n-space E^n into itself with values $F(q)$ for $q \varepsilon E^n$ such that

(*i*) $\det J_q \neq 0$ for all $q \varepsilon E^n$, in which J_q is the Jacobian matrix of $F(\cdot)$ with respect to q, and
(*ii*) $\lim_{\|q\| \rightarrow \infty} \| F(q) \| = \infty$,

then $F(\cdot)$ is an invertible mapping of E^n onto itself and $F(\cdot)^{-1}$ is continually differentiable on E^n.

We have

$$U^{tr}q = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}]\} + c_i \quad \text{for all } i.$$

Let $F(\cdot)$ denote the mapping of E^n into itself defined by the condition that for all i and all $q \varepsilon E^n$:

$$[F(q)]_i = U^{tr}q - \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}]\}.$$

Our objective is to show that $F(\cdot)$ satisfies conditions (*i*) and (*ii*) of Palais' theorem.

We have, with F_i denoting $[F(\cdot)]_i$,

$$\frac{\partial F_i}{\partial q_i} = 1 + \varphi_i' \sum_{j \neq i} (1 + \tau_{ij})\varphi_{ij}' \quad \text{for all } i$$

* See Ref. 8 and the appendix of Ref. 9.

and

$$\frac{\partial F_i}{\partial q_k} = 1 + \varphi_i' \sum_{j \neq i} \tau_{ij}\varphi_{ij}' - \varphi_i'\varphi_{ik}' \quad \text{for all} \quad k \neq i$$

in which

$$\varphi_i' = \varphi_i'\{\sum_{j \neq i} \varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}]\}$$

and

$$\varphi_{ij}' = \varphi_{ij}'[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}].$$

Let $\beta_{ij} = \varphi_i'\varphi_{ij}'$ for all $i \neq j$, let V be the n-vector defined by

$$V^{tr} = (1 + \sum_{j \neq 1} \beta_{1j}\tau_{1j}\ , 1 + \sum_{j \neq 2} \beta_{2j}\tau_{2j}\ , \cdots , 1 + \sum_{j \neq n} \beta_{nj}\tau_{nj}),$$

and let B denote the $n \times n$ matrix defined by

$$(B)_{ii} = \sum_{j \neq i} \beta_{ij} \quad \text{for all } i, \qquad (B)_{ij} = -\beta_{ij} \quad \text{for all} \quad i \neq j.$$

Then $J_q = B + VU^{tr}$.

Suppose now that $\det J_q = 0$ for some n-vector q. For that q, there would exist an n-vector $x \neq \theta$ such that $J_q^{tr}x = (B^{tr} + UV^{tr})x = \theta$. Since the column space of B^{tr} is orthogonal to U, we must have $B^{tr}x = \theta$ and $V^{tr}x = 0$. But B is of rank $(n - 1)$ and the cofactors of B are nonnegative.*

Thus $B^{tr}x = \theta$ implies that $x = \xi y$, in which y is any column of the matrix of cofactors of B and ξ is some real nonzero constant.† But we must have $V^{tr}x = \xi V^{tr}y = 0$, which is a contradiction, since at least one element of y and all of the elements of V are positive. Therefore $F(\cdot)$ meets condition (i) of Palais' theorem.

We now show that $F(\cdot)$ satisfies condition (ii) of the theorem of Palais. It is a simple matter to verify that for all i

$$F_i = U^{tr}q - r_i \sum_{j \neq i} r_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i] - \varphi_i[\sum_{j \neq i} \varphi_{ij}(\lambda_{ij})]$$

in which

$$r_i = \frac{\varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}]\} - \varphi_i\{\sum_{j \neq i} \varphi_{ij}[\lambda_{ij}]\}}{\sum_{j \neq i} \varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}] - \sum_{j \neq i} \varphi_{ij}[\lambda_{ij}]}$$

* See Ref. 10 for a proof that B is of rank $(n - 1)$ and that the cofactors of all of the $(B)_{ii}$ elements of B are positive. Since $BU = \theta$, each of the columns of the transposed matrix of cofactors of B is proportional to the vector U (see the footnote that follows). Therefore *all* of the cofactors of B are positive.

† This follows from the well-known proposition that $M^{tr}C = 1_n \det M$, in which M is any square matrix and C is the matrix of cofactors of M.

and

$$r_{ij} = \frac{\varphi_{ij}[-\tau_{ij}U^{tr}q + q_j - q_i + \lambda_{ij}] - \varphi_{ij}[\lambda_{ij}]}{-\tau_{ij}U^{tr}q + q_j - q_i},$$

with the understanding that r_i is unity when the corresponding numerator is zero, r_{ij} is zero for all q and all $i \neq j$ for which φ_{ij} is identically zero, and r_{ij} is unity for all $i \neq j$ for which φ_{ij} is not identically zero and for all q for which the corresponding numerator is zero. Therefore, for all $q \, \varepsilon \, E^n$, $F(q) = Mq + s$, in which the $n \times n$ matrix M is obtained from J_q by replacing φ_i' by r_i and φ_{ij}' by r_{ij} for all i and all $i \neq j$, respectively, and the ith component of s is $-\varphi_i[\sum_{j \neq i} \varphi_{ij}(\lambda_{ij})]$ for all i. In particular, det $M \neq 0$ for all q. Therefore det $(M^{tr}M) > 0$ for all q. Since all of the r_i as well as all of the nonidentically zero r_{ij} are bounded above and below by positive constants uniformly for $q \, \varepsilon \, E^n$, there exists a positive constant ϵ such that det $(M^{tr}M) \geqq \epsilon$ for all $q \, \varepsilon \, E^n$.

Let $\lambda_1, \lambda_2, \cdots, \lambda_n$ denote the eigenvalues of $M^{tr}M$. Then $\lambda_1\lambda_2 \cdots \lambda_n \geqq \epsilon$ for all $q \, \varepsilon \, E^n$. Assume that $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$. Since all of the r_i and all of the r_{ij} are bounded from above uniformly for $q \, \varepsilon \, E^n$, there exists a positive constant λ such that $\lambda_n \leqq \lambda$ for all $q \, \varepsilon \, E^n$. Thus, for all $q \, \varepsilon \, E^n$ we have $\lambda_1 \geqq \epsilon\lambda^{-(n-1)}$. Therefore,

$$\begin{aligned} \| F(q) \| = \| Mq + s \| &\geqq \| Mq \| - \| s \| \\ &\geqq \epsilon^{\frac{1}{2}}\lambda^{-\frac{1}{2}(n-1)} \| q \| - \| s \| \end{aligned}$$

for all $q \, \varepsilon \, E^n$, from which it is clear that $\| F(q) \| \to \infty$ as $\| q \| \to \infty$ This completes the proof of part (*a*) of our theorem.

Next we show that there is *at most* one ρ with the property that there exists a $q \, \varepsilon \, E^n$ such that for all i

$$\rho = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho\tau_{ij} + q_j - q_i + \lambda_{ij}]\} + c_i \, .$$

Let $\rho^{(a)}$ and $\rho^{(b)}$ be two constants, and $q^{(a)}$ and $q^{(b)}$ two n-vectors, such that for all i

$$\rho^{(a)} = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho^{(a)}\tau_{ij} + q_j^{(a)} - q_i^{(a)} + \lambda_{ij}]\} + c_i$$

$$\rho^{(b)} = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho^{(b)}\tau_{ij} + q_j^{(b)} - q_i^{(b)} + \lambda_{ij}]\} + c_i \, .$$

Then with $q^{(c)} = q^{(b)} + \alpha U$, in which the constant α is chosen so that $\rho^{(a)} - \rho^{(b)} = U^{tr}(q^{(a)} - q^{(c)})$, we have for all i

$$\rho^{(c)} = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho^{(c)}\tau_{ij} + q_j^{(c)} - q_i^{(c)} + \lambda_{ij}]\} + c_i$$

where $\rho^{(c)} = \rho^{(b)}$. Thus

$$\rho^{(a)} - \rho^{(c)} = \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho^{(a)}\tau_{ij} + q_j^{(a)} - q_i^{(a)} + \lambda_{ij}]\} - \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho^{(c)}\tau_{ij} + q_j^{(c)} - q_i^{(c)} + \lambda_{ij}]\}$$

and

$$\rho^{(a)} - \rho^{(c)} = U^{tr}(q^{(a)} - q^{(c)}).$$

Therefore we can define nonnegative ratios p_i and p_{ij} similar to the r_i and r_{ij} above, such that

$$U^{tr}(q^{(a)} - q^{(c)}) = p_i \sum_{j \neq i} p_{ij}[-U^{tr}(q^{(a)} - q^{(c)})\tau_{ij} + (q_j^{(a)} - q_j^{(c)}) - (q_i^{(a)} - q_i^{(c)})] \quad \text{for all } i,$$

and such that these equations are equivalent to $M'(q^{(a)} - q^{(c)}) = \theta$ in which the $n \times n$ matrix M' is obtained from J_q by replacing φ'_i by p_i and φ'_{ij} by p_{ij} for all i and all $i \neq j$, respectively, so that det $M' \neq 0$. But this implies that $q^{(a)} = q^{(c)}$ and hence that $\rho^{(a)} = \rho^{(b)}$.

We shall how prove that q is specified to within an additive vector of the form αU in which α is a real constant.

Suppose that, with $q^{(a)}$ and $q^{(b)}$ two n-vectors,

$$\rho - \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho\tau_{ij} + q_j^{(a)} - q_i^{(a)} + \lambda_{ij}]\} = \rho - \varphi_i\{\sum_{j \neq i} \varphi_{ij}[-\rho\tau_{ij} + q_j^{(b)} - q_i^{(b)} + \lambda_{ij}]\}$$

for all i. Then, with the p_i and p_{ij} as introduced above,

$$p_i \sum_{j \neq i} p_{ij}[(q_j^{(a)} - q_j^{(b)}) - (q_i^{(a)} - q_i^{(b)})] = 0$$

for all i. Thus, since $p_i \neq 0$ for all i, and the $n \times n$ matrix P defined by

$$(P)_{ii} = \sum_{j \neq i} p_{ij}, \quad \text{for all } i$$

$$(P)_{ij} = -p_{ij}, \quad \text{for all} \quad i \neq j$$

is of rank* $(n - 1)$, and $PU = \theta$, we have $q^{(a)} - q^{(b)} = \alpha U$ for some real constant α. □

* See Ref. 10.

V. ACKNOWLEDGMENT

The writer is indebted to A. N. Willson, Jr., for carefully reading the draft.

REFERENCES

1. Pierce, J. R., "Synchronizing Digital Networks," B.S.T.J., *48*, No. 3 (March 1969), pp. 615–636.
2. Karnaugh, M., "A Model for the Organic Synchronization of Communications Systems," B.S.T.J., *45*, No. 10 (December 1966), pp. 1705–1735.
3. Gersho, A. and Karafin, B. J., "Mutual Synchronization of Geographically Separated Oscillators," B.S.T.J., *45*, No. 10 (December 1966), pp. 1689–1704.
4. Brilliant, M. B., "The Determination of Frequency in Systems of Mutually Synchronized Oscillators," B.S.T.J., *45*, No. 10 (December 1966), pp. 1737–1748.
5. Brilliant, M. B., "Dynamic Response of Systems of Mutually Synchronized Oscillators," B.S.T.J., *46*, No. 2 (February 1967), pp. 319–356.
6. Candy, J. C. and Karnaugh, M., "Organic Synchronization: Design of the Controls and Some Simulation Results," B.S.T.J., *47*, No. 2 (February 1968), pp. 227–259.
7. Sandberg, I. W., "On Conditions Under Which It Is Possible to Synchronize Digital Transmission Systems," B.S.T.J., *48*, No.6 (July-August 1969), pp. 1999–2022.
8. Palais, R. S., "Natural Operations on Differential Forms," Trans. Amer. Math. Soc., *92*, No. 1 (July 1959), pp. 125–141.
9. Holtzmann, C. A. and Liu, R., "On the Dynamical Equations of Nonlinear Networks with n-Coupled Elements," Proc. Third Annual Allerton Conf. on Circuit and System Theory, 1965, pp. 536–545.
10. Fiedler, M. and Pták, V., "On Matrices with Non-Positive Off-Diagonal Elements and Positive Principal Minors," Czech. Math. J., *12*, No. 3 (September 1962), pp. 391–392.

Overflow Oscillations in Digital Filters

By P. M. EBERT, JAMES E. MAZO,
AND MICHAEL G. TAYLOR

(Manuscript received May 9, 1969)

The cascade and parallel realizations of an arbitrary digital filter are both formed using second order sections as building blocks. This simple recursive filter is commonly implemented using 2's complement arithmetic for the addition operation. Overflow can then occur at the adder and the resulting nonlinearity causes self-oscillations in the filter. The character of the resulting oscillations for the second order section are here analyzed in some detail. A simple necessary and sufficient condition on the feedback tap gains to insure stability, even with the presence of the nonlinearity, is given although for many desired designs this will be too restrictive. A second question studied is the effect of modifying the "arithmetic" in order to quench the oscillations. In particular it is proven that if the 2's complement adder is modified so that it "saturates" when overflow occurs, then no self-oscillations will be present.

I. INTRODUCTION

A digital filter using idealized operations can easily be designed to be stable.[1] Nevertheless, in actual implementations, the output of such a stable filter can display large oscillations even when no input is present.* A known cause of this phenomenon is the fact that the digital filter realization of the required addition operation can cause overflow, thereby creating a severe nonlinearity.† Our purpose here is twofold. The first is to give a somewhat detailed analysis of the character of the oscillations when the filter is a simple second order recursive section with two feedback taps. This unit is the fundamental building block for the cascade and the parallel realization of digital filters, and as such is worthy of some scrutiny.[2] A simple conclusion which one can draw from

* To the best of our knowledge, these oscillations were first observed and diagnosed by L. B. Jackson of Bell Telephone Laboratories.

† In the present work rounding errors in multiplication or storage are neglected and therefore so are the little-understood oscillations attendant upon these nonlinearities.

the analysis is that the design of many useful filters requires using values of feedback coefficients such that the threat of oscillations is always present (with 2's complement arithmetic). Optimum solutions that cope with this state of affairs are still unknown. Some recent proposals include observing when overflow at the adder is to occur and then taking appropriate action. Our second purpose, then, is to discuss the effectiveness of some of these ideas, and to give a proof that modifying 2's complement arithmetic so that the adder "saturates" is an effective way to eliminate the oscillations. Questions of how this nonlinearity will affect the desired outputs from a particular ensemble of input signals are not yet answered however, and perhaps for some applications other solutions need be considered.

II. PROBLEM FORMULATION AND GENERAL DISCUSSION

As explained in the introduction, this paper deals primarily with the simple structure shown in Fig. 1. The outputs of the registers, which are storage elements with one unit of delay, are multiplied by coefficients a and b respectively, fed back, and "added" to the input in the accumulator. No round-off error is considered either in multiplication or storage, but overflow of the accumulator is not neglected. In other words, the accumulator will perform as a true adder if the sum of its inputs is in some range; otherwise a nonlinear behavior is observed.

Figure 2 shows the instantaneous input-output characteristic $f(v)$ of the device motivated by using 2's complement arithmetic. It is also important to note that there is no memory of the accumulator for past outputs; that is, the device is zeroed after the generation of each output.

If we let $x(t)$ be the input signal to the device, $y(t)$ the output, and

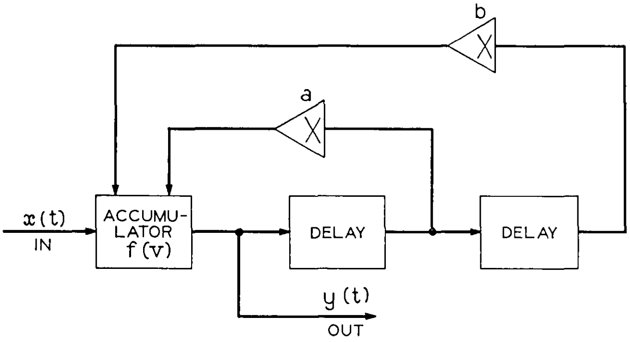


Fig. 1 — Basic configuration for the digital filter $\cdot y_{k+2} = f[ay_{k+1} + by_k + x_{k+2}]$.

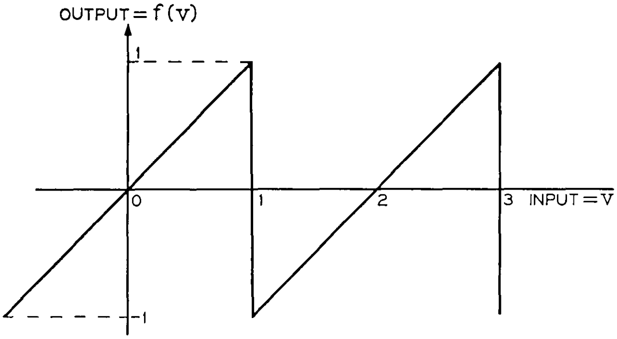


Fig. 2 — Instantaneous transfer function of the accumulator.

$f(\cdot)$ the nonlinear characteristic of the accumulator, we have the basic equation

$$y(t + 2) = f[ay(t + 1) + by(t) + x(t + 2)]. \tag{1}$$

We shall be concerned with the self-sustaining oscillations of the device that are observed even when no input is present $[x(t) = 0]$, and when linear theory would predict the device to be stable.

By making this linear approximation $f(v) = v$, the linearized version of equation (1) becomes, with no driving term in the equation,

$$y(t + 2) - ay(t + 1) - by(t) = 0. \tag{2}$$

The roots of the characteristic equation for equation (2) are

$$\rho_{1,2} = \frac{a \pm (a^2 + 4b)^{\frac{1}{2}}}{2} \tag{3}$$

and the region of linear stability corresponds to the requirement that $| \rho_i | < 1$. This region is depicted as a subset of the a–b plane in Fig. 3. One has $| \rho_i | < 1$ if and only if one is within the large triangle shown in Fig. 3. For this situation any solution of (2) will damp out to zero after a sufficient period of time. Now note that (2) is not necessarily a valid reduction of (1) even when $x(t) = 0$. The output, by choice of f, has been assumed to be constrained to be less than unity, but this is not sufficient to guarantee that the argument of the function f is less than unity. For this to be the case we require

$$| ay(t + 1) + by(t) | < 1. \tag{4}$$

Since $| y(t) | < 1$, equation (4) will always be satisfied provided that

$$| a | + | b | < 1. \tag{5}$$

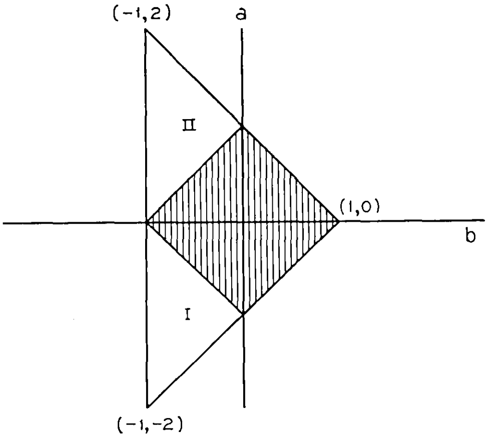


Fig. 3—Some interesting regions in the "space" of feedback tap weights. The hatching indicates stability even with the nonlinearity.

The subset of the a–b plane for which (5) is true is shown in Fig. 3 with vertical hatching, and is a subset of the region of linear stability. It is shown in this Section that if (5) is not satisfied there always exist self-sustained oscillations of the digital filter and hence (5) is both a necessary and sufficient condition for absence of self-sustained oscillations.* One way to avoid the oscillations in question is simply to impose the requirement (5). This trick has its limitations, however, for it clearly restricts design capabilities. The region of the s-plane which is shaded in Fig. 4 shows the allowable pole positions. Roughly speaking, one concludes that there are desirable filter characteristics that can be realized with this restriction and there are desirable characteristics that cannot.

It is not our purpose here to outline those applications for which (5) will not be restrictive; we proceed to sketch the situation when $| a | + | b | > 1$ and the threat of oscillation is present. Sections III and IV contain, we believe, a novel and interesting mathematical treatment of the general problem of classifying the self-oscillations of the nonlinear difference equation (1). However, for the user of digital filters a simple proof of the $| a | + | b | > 1$ being sufficient for threat of oscillations is of more immediate interest. After reading the simple proof of this fact given next in the present section, such a reader may wish to proceed directly to Section V.

Consider the possibility of undriven nonlinear operation giving a dc

* I. W. Sandberg has informed the authors that the necessity and sufficiency of (5) holding for absence of oscillations has also been obtained jointly by him and L. B. Jackson.

output, that is, $y_k \equiv y$ for all k. Equation (1), with $x(t) = 0$ becomes $y = f[(a + b)y]$. Assuming for definitness that $y > 0$, we can easily see from Fig. 2 that the above equation will be true if $(a + b)y = y - 2$, which implies $y = 2/(1 - a - b)$. One can show (see discussion following equation 17), that this y will have magnitude < 1 provided only that the tap values a and b lie in the region labeled I in Fig. 3. Thus a consistent dc oscillation is always possible for all (a, b) pairs in this region. Next consider the possibility of a period 2 oscillation. This amounts to finding a consistent solution to $y = f[(b - a)y]$. Proceeding as before we obtain

$$y = \frac{2}{1 + a - b}.$$

Thus y_k will be given by $(-1)^k y$, and will have magnitude less than unity if the (a, b) pair lies anywhere in region II of Fig. 3.

III. FURTHER ANALYSIS OF THE OSCILLATIONS

To analyze equation (1) in greater detail, it is very convenient to write it in the form similar to (2),

$$y(t + 2) - ay(t + 1) - by(t) = \sum_n a_n u(t + 2 - n), \qquad (6)$$

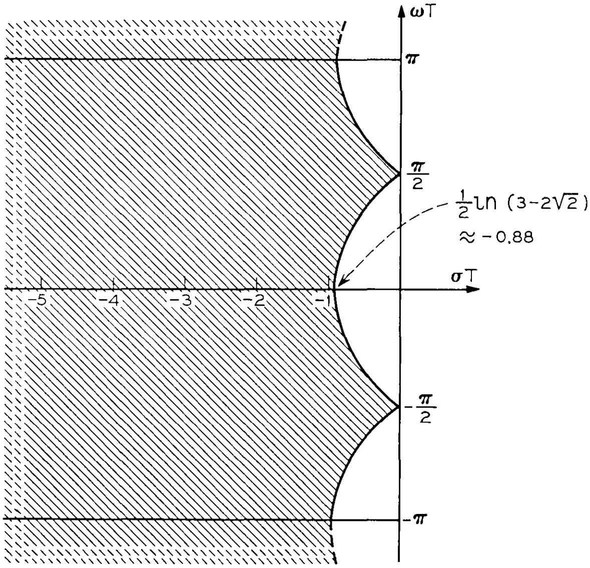


Fig. 4 — Pole locations in the s-plane (shaded region) realizable under the constraint that $|a| + |b| < 1$.

where $u(t)$ is a square pulse of unit height that one may conveniently think of as lasting from $t = 0$ until $t = 1$. This, of course, means that one interprets the solution of (6) to be a piecewise constant function like the actual output of the digital filter. For mathematical manipulations it is sometimes desirable to also interpret (6) as a difference equation, defined only for integer t. In this case one would write that $u(t - n) = \delta_{tn}$ where δ_{tn} is the familiar Kroneker symbol.

The point of the right side of (6) is simply to keep $| f(v) | < 1$ regardless of what value v has. From Fig. 2 we see that if $| v | < 1$, this added term is not needed and we take $a_n = 0$. If $1 < v < 3$ then we take $a_n = -2$, and if $-3 < v < -1$ we take $a_n = +2$. Since we have that $| y(t) | < 1$ and that linear stability (see Fig. 3) implies $| a | < 2$, $| b | < 1$, we need not consider further values of $| v |$. Thus in (6) $a_n = 0, \pm 2$ depending on whether or not $v(t) \equiv ay(t + 1) + by(t)$ crosses the lines $v = \pm 1$. It will be convenient to have a word for such crossings; we shall call them "clicks", borrowing a favorite word from FM theory. Then $a_n = 0, \pm 2$ depending on whether or not a click does not, or does, occur.

Note if one knew what the click sequence $\{a_n\}$ was, one could solve (6) simply by using the clicks to be the driving term for a linear equation. We are mainly interested in describing the self-sustained steady state oscillations of arbitrary period N. Hence initial conditions will play no essential role for us, for while they determine which oscillating mode appears as $t \to \infty$, they play no role in describing the modes. Our procedure will be as follows:

(*i*) Assume a click sequence of period N;

$$\begin{aligned} &a_0 , a_1 , a_2 , \cdots , a_{N-1} \\ &a_{lN+k} = a_k . \qquad l = 0, 1, \cdots \\ &\qquad\qquad\qquad 0 \leqq k < N - 1. \end{aligned} \tag{7}$$

(*ii*) Using the assumed $\{a_n\}$, find the steady state solution of (6). However, only solutions that have $| y(t) | < 1$ for all t are allowed.
(*iii*) Check that this steady state solution actually generates the assumed click sequence.

In carrying out the above program for some simple cases we observed that step *iii* never seemed to yield anything new. Indeed, surprising as it seems at first glance, step *iii* never has to be carried out. If one obtains a solution with $| y(t) | < 1$, this solution is consistent. That is, it automatically generates the assumed click sequence. The proof is simple.

One calculates the argument of the function f from (6):

$$ay(t + 1) + by(t) = y(t + 2) - \sum a_n u(t + 2 - n). \tag{8}$$

We have a click at time $t + 2 = m$ if $| ay(m - 2) + by(m - 1) | > 1$. From (8),

$$| ay(m - 2) + by(m - 1) | = | y(m) - a_m |. \tag{9}$$

Note then if in (9) $a_m = 0$, then $| ay(m - 2) + by(m - 1) | = | y(m) | < 1$; thus if there is no click at a particular time in the assumed click sequence the "solution" will not generate one. Next assume $a_m = +2$; then

$$ay(m - 2) + by(m - 1) = y(m) - 2 < -1, \tag{10}$$

where we use $| y(t) | < 1$ again. Equation (10) says if a positive click is present in the assumed click sequence then the solution obtained from the linear equation (6), given by this click sequence, will reproduce the positive click. Obviously the same argument holds for a negative click, $a_m = -2$, and the proof of this point is complete.

The steady-state solution of our fundamental equation (6) for an arbitrary click sequence $\{a_m\}$ of period N is derived in the appendix. If we define

$$A_{N-1}\left(\frac{1}{z}\right) \equiv \sum_{n=0}^{N-1} a_n z^{-n} \tag{11}$$

and

$$D(z) \equiv z^2 - az - b, \tag{12}$$

and let r_i, $i = 1, \cdots, N$, be the N Nth roots of unity, then the (periodic) output values are given by

$$y_k = \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right)}{D(r_i)} r_i^k . \tag{13}$$

The above expression gives the $\{y_k\}$ output sequence for any click sequence. We emphasize, however, that it is only a solution corresponding to a self-sustained oscillation of the digital filter if we have $| y_k | < 1$, all k. Whether or not this is true depends on the particular click sequence assumed.

Another form of the solution can be obtained by manipulation of (13). To write this down, define

$$b_n^{(k)} \equiv (\bar{a}_{k-1-n} + \bar{a}_{k-1-n+N})/2, \tag{14}$$

where we understand $\bar{a}_j \equiv 0$ if j does not lie between 0 and $N - 1$, inclusive, and $\bar{a}_j \equiv a_j$ if it does. One of the $\bar{a}$'s in (14) will thus always be zero and $b_n^{(k)}$ has values of $\pm 1, 0$. The other form of the solution is then

$$y_k = \frac{2}{\rho_1 - \rho_2} \sum_{n=0}^{N-1} b_n^{(k)} \left[\frac{\rho_1^n}{1 - \rho_1^N} - \frac{\rho_2^n}{1 - \rho_2^N} \right] \qquad k = 0, 1, \cdots, N - 1 \tag{15}$$

where ρ_i are given in (3).

In (15) we have N vectors of dimension N, namely the $\{b_n^{(k)}\}$ $k = 0, 1, 2, \cdots, N - 1$. Note from (14), however, that they are all cyclic permutations of one another. Hence we may refer to the b vector, **b**, of a solution, understanding that the **b** and all its cyclic permutations generate a solution in the sense of (15). Note that a cyclic permutation of the y_k has no real significance here; it simply changes the origin of time.

An interesting property of the solutions which we have written down follows from the fact that if we transform the point (a, b) in the ab-plane into another point by

$$\begin{aligned} a &\rightarrow a' = -a \\ b &\rightarrow b' = b \end{aligned} \tag{16a}$$

then under this transformation

$$\begin{aligned} \rho_1 &\rightarrow \rho_1' = -\rho_2 \\ \rho_2 &\rightarrow \rho_2' = -\rho_1 \,. \end{aligned} \tag{16b}$$

The property is this: Let N be an even integer and let $\mathbf{b} = (b_0, b_1, \cdots, b_{N-1})$ be a click vector generating a solution at point (a, b). Then the vector $\mathbf{b}' = (b_0, -b_1, b_2, -b_3, \cdots, b_{N-1})$ generates a solution at reflected point $(-a, b)$. The proof is simple. Note from (15),

$$\begin{aligned} y'^{(k)} &= \frac{2}{\rho_1' - \rho_2'} \sum_n b_n'^{(k)} \left[\frac{\rho_1'^n}{1 - \rho_1'^N} - \frac{\rho_2'^n}{1 - \rho_2'^N} \right] \\ &= \frac{2}{-\rho_2 + \rho_1} \sum_n (-1)^{k+n} b_n \left[\frac{(-\rho_2)^n}{1 - \rho_2^N} - \frac{(-\rho_1)^n}{1 - \rho_1^N} \right] = (-1)^k y^{(k)}. \end{aligned}$$

Hence if $| y^{(k)} | < 1$ then $| y'^{(k)} | < 1$. Note that the proof also supplies the value for $y'^{(k)}$ in terms of $y^{(k)}$. This theorem will be used later to generate new solutions from old ones.

Before leaving this general discussion in favor of exhibiting some solutions in the next section, we list a few more observations related to the click vector **b**. The click vector **b**, whose only allowed component values are ±1, 0, completely characterizes the associated oscillation. Clearly there can then only be a finite number of oscillations of given period N. This number is upper bounded by 3^N, but will generally be much less. Also note that a cyclic permutation of the components of b cyclically permutates the output values y^k, and this latter is merely a shift in time. The permutated values are not physically distinct.

Also note that if we perform $\mathbf{b} \rightarrow -\mathbf{b}$ then $\mathbf{y} \rightarrow -\mathbf{y}$, and a solution of opposite sign is obtained. While this may often be distinguishable from the first solution, it is trivially related to it. Finally if one were to count the number **b** vectors of dimension N that yield new information, one would wish to exclude subperiods of N. Thus if $(+, 0, 0)$ is an generating **b** vector for period 3, $(+, 0, 0, +, 0, 0)$ generates a period 6 oscillation but this is not new information. We have not solved the problem of counting how many of the 3^N vectors are left after we impose the requirements of cyclic shifts, sign changes, and subperiods. At any rate, it is essential to test the ones that remain to check that they generate allowed solutions, $| y^k | < 1$.

IV. SOME EXPLICIT PERIODS AND REGIONS OF OSCILLATION

Now for a few explicit solutions. Consider the possibility of a dc "oscillation", namely, set $N = 1$. The only nontrivial click vector is $\mathbf{b} = (+)$. The solution is more immediate if we use (13). We have

$$y = \frac{2}{1 - a - b} \tag{17}$$

for the dc value of output. For what values of a and b within the triangle of Fig. 3 will we have $| y | < 1$? We require

$$| 1 - a - b | > 2 \tag{18}$$

which is equivalent to either

$$1 - a - b > 2 \tag{19a}$$

or

$$-1 + a + b > 2. \tag{19b}$$

Inequality (19a) (coupled with the linear stability requirement) defines the triangle labeled "I" in Fig. 3, while (19b) is outside the stability region and needs no further consideration. Thus any portion of the region $a < 0$ that we have not excluded from oscillations has now been shown to have them. They are of period 1; other period oscillations may (and do) occur in this region.

At this point it is amusing to use an earlier remark on the possibility of generating new solutions from an even period one by "reflection". Letting $N = 2$, the click vector $\mathbf{b} = (+, +)$ certainly generates a period 2 oscillation (albeit one with subperiods) in region I. Then the click vector $\mathbf{b} = (+, -)$ generates something really new: a period 2 oscillation in the region labeled II in Fig. 3. The amplitudes of the output are

$$y^{(k)} = (-1)^k \frac{2}{1 + a - b}, \qquad a > 0. \tag{20}$$

One more possibility of a click vector exists for period 2, and that is $\mathbf{b} = (+, 0)$. From (13) we write for possible output values

$$\begin{aligned} y_0 &= \frac{1}{1 - a - b} + \frac{1}{1 + a - b} \\ y_1 &= \frac{1}{1 - a - b} - \frac{1}{1 + a - b}. \end{aligned} \tag{21}$$

After a little uninteresting analysis one can conclude that we cannot have $| y_0 | < 1$, $| y_1 | < 1$ in (21) for any allowed values of a and b. Thus there are no other period 2 oscillations.

On to period 3. Now there are four click vectors which must be considered. These are $(+00)$, $(++0)$, $(+-0)$, $(++-)$. Even in this case an exhaustive check that the "solutions" generated are legitimate ones is trying. Therefore, we resort to a trick; we look for periods which may exist in the immediate neighborhood of the point $(a = 0, b = 1)$. This means $\rho_1 = i$, $\rho_2 = -i$. In this immediate neighborhood $\rho_2 = \rho_1^*$, and (15) reads

$$y = \frac{2}{\operatorname{Im} z} \operatorname{Im} \sum_{n=0}^{N-1} \frac{b_n z^n}{1 - z^N}, \tag{22}$$

where we have let $z = \rho_1$. Letting $N = 3$, $z = i$ gives

$$\begin{aligned} y_0 &= -b_0 + b_1 + b_2 \\ y_1 &= -b_1 + b_2 + b_0 \\ y_2 &= -b_2 + b_0 + b_1 . \end{aligned} \tag{23}$$

We now require $y_k = \pm 1$ as a test for the click vector $\mathbf{b}$. We see that only $(+00)$ qualifies as possibly yielding a solution in the neighborhood of $(a = 0, b = -1)$. A computer study shows that indeed the solution extends into the interior of the triangle and the region found is shown in Fig. 5. This immediately implies existence of the period 6 oscillation generated by $(+00-00)$ in the reflected region. Similarly, a period 5 oscillation region (with the concomitant period 10) generated by $(+0000)$ is shown in Fig. 6.

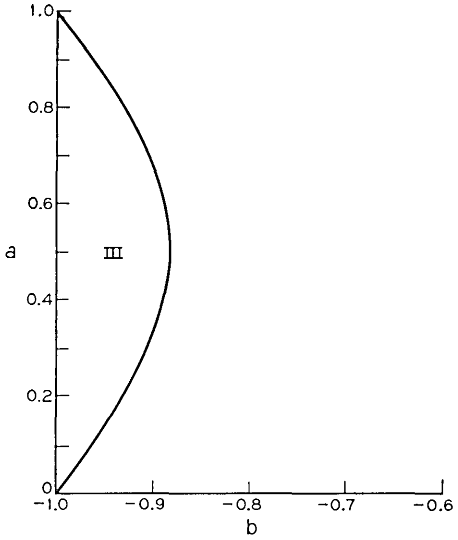


Fig. 5 — A region for period 3 oscillations.

It is very tempting to conjecture that the point ($a = 0, b = -1$) is a boundary point of any allowed region of oscillation. If this is true, a procedure like that used above may eliminate some otherwise very respectable **b** vectors from consideration. Note that for $N = 2$, $b = (+, 0)$ satisfies the required condition at $\rho_1 = i$, but we have shown this

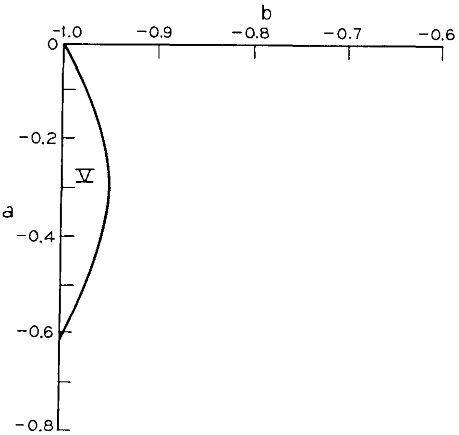


Fig. 6 — A region for period 5 oscillations.

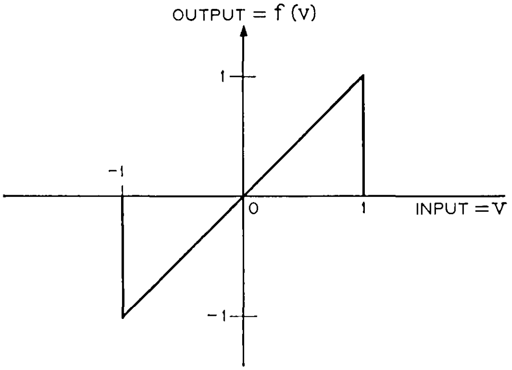


Fig. 7 — Zeroing arithmetic, shown above, also gives rise to oscillations.

is not extendable into the interior of the triangle. Hence existence at $z = i$ does not guarantee an allowed solution.

V. STABILITY WITH A MODIFIED ARITHMETIC

In an attempt to eliminate these oscillations, proposals have been made which rely on detecting overflow. One such suggestion dictates that when overflow occurs, the adder is directed to shift out zero. For refereace we call this zeroing arithmetic. The effective transfer function of the adder for zeroing arithmetic is given in Fig. 7. However, it can be shown by numerical example that such a procedure still leads to oscillations. Another possibility, "saturation arithmetic," is displayed in Fig. 8. Here a one (with the appropriate sign) is put out when overflow is detected. The remaining portion of this paper is devoted to proving that saturation arithmetic leads to stable operation whenever linear theory would predict it to be so.

To begin, we suppose for the moment that we ignore the fact that the digitally implemented adder is nonlinear. Then the second-order linear difference equation which governs the behavior of the undriven system has solutions y_k which may be described as follows:

Case 1: Complex roots for characteristic equation

$$y_k = \operatorname{Re} K_0 \exp(-\alpha k), \quad K_0 \text{ and } \alpha \text{ complex}, \quad \operatorname{Re} \alpha > 0. \qquad k = 0, 1, 2, \cdots . \tag{24}$$

Case 2: Real but unequal roots

$$y_k = K_1 \exp(-\alpha k) + K_2 \exp(-\beta k). \quad K_i \text{ real}; \quad \alpha > 0, \quad \beta > 0. \tag{25}$$

Case 3: Real and equal roots

$$y_k = [K_1 + K_2k] \exp(-\alpha k). \quad K_i \text{ real}; \quad \alpha > 0. \tag{26}$$

Using this information, coupled with knowledge of y_j and y_{j+1} for some j, it is easy to give a bound on the magnitudes of all future ($k \geqq j$) values of the output and to show this value goes to zero with increasing j. This is just another way to say that the solutions go to zero for the linear case. In the nonlinear case we cannot exclude the situation that some y_{k+1} will exceed unity and the nonlinearity will be operative. For saturation arithmetic the offending value must be set to unity if, for example, $y_{k+1} > +1$. We can, for conceptual purposes, regard this as a "squeezing" of the output from a value greater than unity down to the value one which is performed in a continuous fashion. The crux of the proof now comes in showing that the partial derivative of our bound (on future outputs) with respect to the most recent output y_{k+1} has, for saturation arithmetic, the same sign as y_{k+1}. Hence decreasing a value that is too large in magnitude will decrease the bound as well, and it will go to zero at least as fast as it does for the linear case.

To show how the above outline works, consider first the linear case with complex roots. From the form of the solution

$$y_k = \operatorname{Re} K_0 \exp(-\alpha k), \quad \operatorname{Re} \alpha > 0, \quad k = 0, 1, 2, \cdots,$$

it is clear that if we define

$$B_0 = |K_0|^2 \tag{27}$$

then $y_k^2 \leqq B_0$ for all $k \geqq 0$. We now express B_0 in terms of the values y_0, y_1 which are initially stored in the shift registers to yield

$$B_0 = y_0^2 + \frac{[y_1 - y_0 \operatorname{Re} \exp(-\alpha)]^2}{[\operatorname{Im} \exp(-\alpha)]^2}. \tag{28}$$

This suggests that one define the more general set of numbers

$$B_j = y_j^2 + \frac{[y_{j+1} - y_j \operatorname{Re} \exp(-\alpha)]^2}{[\operatorname{Im} \exp(-\alpha)]^2}. \tag{29}$$

Clearly, from the way that B_j is defined, we have that

$$y_k = \operatorname{Re} K_j \exp[-\alpha(k - j)], \qquad k \geqq j \tag{30}$$

where K_j is some appropriate complex number that satisfies

$$B_j = |K_j|^2. \tag{31}$$

From (30), the additional inequality that $y_k^2 \leqq B_j$ for all $k \geqq j$ follows.

Furthermore, one can see by comparing (30) and (24) that

$$| K_j |^2 = | K_0 |^2 \, | \exp(-\alpha j) |^2. \tag{32}$$

Hence, since the real part of α is positive, B_j goes monotonically to zero with increasing j.

To generalize the above arguments to a nonlinear situation of interest,* consider the following equation which follows from (29):

$$\frac{\partial B_j}{\partial y_{j+1}} = \frac{2}{[\operatorname{Im} \exp(-\alpha)]^2} [y_{j+1} - y_j \operatorname{Re} \exp(-\alpha)]. \tag{33}$$

Now imagine B_{j-1} has been calculated from values stored in the registers. From *linear* theory we predict $y_{j+1}^{(L)}$ and $B_j^{(L)} \leqq B_{j-1} \exp(-2\alpha)$, by (32). Now if the $y_{j+1}^{(L)}$ generated by the linear equation were too large, say, then decreasing it to unity would, according to (33), *decrease* the bound B_j if we knew that

$$y_{j+1} - y_j \operatorname{Re} [\exp(-\alpha)] \geqq 0 \quad \text{for} \quad y_{j+1}^{(L)} \geqq y_{j+1} \geqq y_{j+1}^{(C)} \tag{34}$$

where $y_{j+1}^{(L)}$ is the linear prediction for y_{j+1} and $y_{j+1}^{(C)}$ is the correct value for the nonlinear circuit resulting from "squeezing" $y_{j+1}^{(L)}$ down. Since $| y_j | \leqq 1$ and $\operatorname{Re} \exp(-\alpha) < 1$, (34) is always true for saturation arithmetic (see Fig. 8) because $y_{j+1}^{(C)} = +1$ (assuming $y_{j+1}^{(L)} > +1$) and (34) can never swing negative. Similar things happen, of course, if $y_{j+1} < -1$. Thus the bound decreases at least as fast as for the linear case (which is exponential) and stability is assured. For zeroing arithmetic $y_{j+1}^{(C)} = 0$, and thus the appropriate sign for (34) cannot be guaranteed which is in satisfying agreement with the known instability for this case.

For the next case of real but unequal roots, we now have reference to equation (25) and define our initial bound as

$$\begin{aligned} B_0 &= 2(K_1^2 + K_2^2) \\ &= 2 \frac{[y_1 - \exp(-\alpha) y_0]^2 + [y_1 - \exp(-\beta) y_0]^2}{[\exp(-\alpha) - \exp(-\beta)]^2}. \end{aligned} \tag{35}$$

The remaining details are too similar to those of the preceding case to warrant recording again; stability for saturation arithmetic holds here as well.

The last case to discuss occurs when we have real and equal roots.

* B_j calculated from (29) is a bound on future outputs for the nonlinear as well as the linear case. If $B_j \leq 1$ the two cases coincide, while of $B_j > 1$ the conclusion follows equally trivially since $|y_k| \leq 1$ for the nonlinear situation.

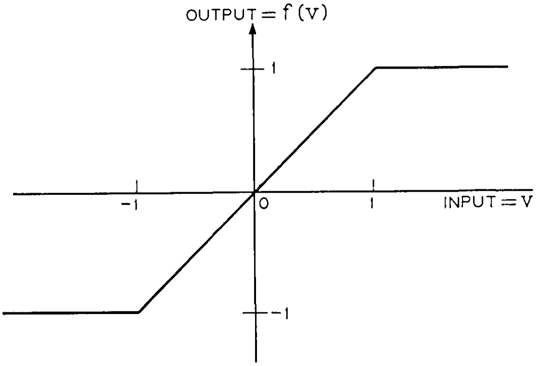


Fig. 8 — The above nonlinearity corresponds to saturation arithmetic and leads to stable behavior.

This situation, represented for the linear equation by equation (26), is more difficult to treat than the previous ones. The analog of (27) and (35) now is

$$B_0 = \max \begin{cases} 4K_1^2 \\ \dfrac{4K_2^2}{\alpha^2} \end{cases}. \tag{36}$$

That (36) yields a bound follows from the facts that (for $t \geqq 0$)

$$\begin{aligned} y_k^2 &\leqq \max_t \, [(K_1 + K_2 t) \exp(-\alpha t)]^2 \\ &\leqq 2 \max_t \, [K_1^2 + K_2^2 t^2] \exp(-2\alpha t) \\ &\leqq 4 \max \begin{cases} \max\limits_t K_1^2 \exp(-2\alpha t) \\ \max\limits_t K_2^2 t^2 \exp(-2\alpha t) \end{cases} \\ &= 4 \max \begin{cases} K_1^2 \\ \dfrac{K_2^2 \exp(-2)}{\alpha^2} \end{cases} \\ &\leqq 4 \max \begin{cases} K_1^2 \\ \dfrac{K_2^2}{\alpha^2} \end{cases}. \end{aligned}$$

Since

$$K_1^2 = y_0^2$$
$$\frac{K_2^2}{\alpha^2} = \frac{(y_1 \exp \alpha - y_0)^2}{\alpha^2}, \tag{37}$$

we define our general bound as

$$B_j = 4 \max \begin{cases} y_j^2 \\ \dfrac{(y_{j+1} \exp \alpha - y_j)^2}{\alpha^2}. \end{cases} \tag{38}$$

Using the solution $y_j = (K_1 + K_2 j) \exp(-\alpha j)$, we see that

$$\theta_j \equiv \frac{(y_{j+1} \exp \alpha - y_j)^2}{\alpha^2} \tag{39}$$

decreases by the multiplicative factor $\exp(-2\alpha)$ for every unit increase of j. Further, suppose that $B_j = 4y_j^2$ for some j. That is, suppose

$$\frac{(y_{j+1} \exp \alpha - y_j)^2}{\alpha^2} < y_j^2 . \tag{40}$$

This implies

$$y_{j+1}^2 < y_j^2(1 + \alpha)^2 \exp(-2\alpha), \tag{41}$$

and so if next time $B_{j+1} = 4y_{j+1}^2$, then we have decreased by $(1 + \alpha)^2 \exp(-2\alpha) < 1$. On the other hand, if at the next step we have to choose $B_{j+1} = 4\theta_{j+1}$, we see

$$\frac{B_{j+1}}{B_j} = \frac{\theta_{j+1}}{y_j^2} \leqq \frac{\theta_{j+1}}{\theta_j} \leqq \exp(-2\alpha). \tag{42}$$

Likewise if we go from $4\theta_j$ to $4\theta_{j+1}$ we decrease by $\exp(-2\alpha)$. Finally, a "transition" from $4\theta_j$ as a bound to $4y_{j+1}^2$ decreases the bound by a multiplicative factor of $(1 + \alpha)^2 \exp(-2\alpha)$. To see this we note that, by assumption,

$$B_j = \frac{4[y_{j+1} \exp \alpha - y_j]^2}{\alpha^2} \geqq 4y_j^2 . \tag{43}$$

Using the left-hand equality in (43) implies

$$| y_{j+1} | \exp \alpha \leqq \frac{\alpha(B_j)^{\frac{1}{2}}}{2} + | y_j |. \tag{44}$$

while $B_i \geqq 4y_i^2$ yields

$$| y_i | \leqq \frac{(B_i)^{\frac{1}{2}}}{2}. \tag{45}$$

Using (45) in (44) then allows us to deduce that

$$B_{i+1} = 4y_{i+1}^2 \leqq (1 + \alpha)^2 \exp(-2\alpha)B_i \tag{46}$$

as was claimed. To extend these arguments to the nonlinear case we again observe that

$$\frac{\partial B_i}{\partial y_{i+1}} \geqq 0 \tag{47}$$

for saturation arithmetic.

VI. GENERALIZATIONS TO OTHER STABLE NONLINEARITIES

Aside from the three nonlinearities already mentioned, there does not appear to be immediate engineering interest in seeing which other nonlinearities will or will not give rise to stable behavior of the filter. Having come this far, however, it is hard to resist asking if the method of proof we have used, or some slight extension of it, does suggest other nonlinearities for which stability will hold. The extension we consider is not to require

$$\frac{\partial B_i}{\partial y^{i+1}} \geqq 0$$

all during the "squeezing" operation, but merely that

$$B_i^L - B_i^C \geqq 0, \tag{48}$$

where B_i^L is the value of the bound using linear theory and B_i^C is the "correct" value. An inspection of the previous proofs shows that this is equivalent to

$$(y_{i+1}^L - ay_i)^2 - (y_{i+1}^C - ay_i)^2 > 0 \tag{49}$$

for all real a such that $| a | < 1$.

A little manipulation reduces (49) to

$$(y_{k+1}^L - y_{k+1}^C)(y_{k+1}^L + y_{k+1}^C - 2ay_k) \geqq 0. \tag{50}$$

Assuming $y_{k+1}^L > 0$, the first term in (50) to be nonnegative, and $| y_k | \leqq 1$, makes it apparent that

$$y_{k+1}^L + y_{k+1}^C \geqq 2 \tag{51}$$

is sufficient. The "stable nonlinearities" deduced from this kind of reasoning are outlined in Fig. 9. Thus any nonlinearity whose graph coincides with the identity function on the interval $[-1, 1]$ and whose remaining portions lie in the closed shaded region of Fig. 9 will be stable. The function in these regions need not be continuous and need not obey $f(-u) = -f(u)$.

An even higher degree of generality is achieved when we realize that nothing in our proofs required the nonlinearity $f(u)$ to be the same for successive values of the parameter k. This is tantamount to allowing the nonlinearity to be random in the following manner. Suppose a value of $y_{k+1}^L > 1$ has been predicted from linear theory (see Fig. 9). The perpendicular P to the v axis through y_{k+1}^L intersects the shaded region shown in Fig. 9 along a line segment. Choose randomly from this line segment the "value" of the nonlinearity to give y_{k+1}^C. The discussion in this Section shows that the solutions of the difference equation

$$y_{k+2} = f[ay_{k+1} + by_k] \tag{52}$$

which has the stochastic nonlinearity just described will be stable whenever the linear version has stable solutions.

APPENDIX

Derivation of the Steady-State Solution

We obtain the steady-state solution of our fundamental equation (6) using z-transforms. Recall that if one has a bounded sequence of number $\{a_n\}$, the z-transform is defined by

$$f(z) = \sum_{n=0}^{\infty} a_n z^{-n} \tag{53}$$

where (53) converges and is analytic outside the unit circle, $|z| > 1$. It is easy to show that if $\{a_n\}$ is periodic of period N, that is if $a_{N+n} = a_n$, then (53) becomes

$$f(z) = \frac{A_{N-1}\left(\frac{1}{z}\right)}{1 - z^{-N}} \tag{54}$$

where A_{N-1} is the polynomial of degree $(N - 1)$ in $1/z$ given by

$$A_{N-1}\left(\frac{1}{z}\right) = \sum_{n=0}^{N-1} a_n z^{-n}. \tag{55}$$

The N poles of $f(z)$ at the N roots of unity are apparent from (12), and there are no other poles.

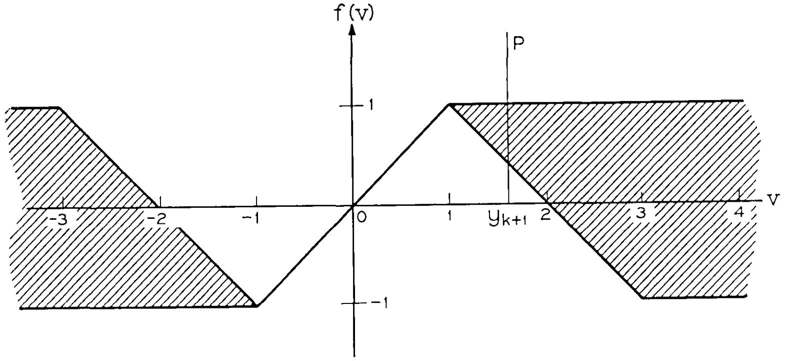


Fig. 9 — Any nonlinearity whose graph coincides with the identity function on the interval [−1, +1] and whose remaining portions lie in the (closed) shaded region will be stable. The possibility of generalizing this to a stochastic nonlinearity is also noted in the text.

Denoting by $Y(z)$ the z-transform of $y(t)$ *excluding* the additive terms involving initial conditions (since these will damp out because of linear stability) we have from (6) that

$$Y(z) = \frac{A_{N-1}\left(\frac{1}{z}\right)}{(z^2 - az - b)(1 - z^{-N})}. \tag{56}$$

The z-transform of the steady-state solution $\hat{Y}(z)$ must still be extracted from $Y(z)$. Since the unit circle $|z| = 1$ corresponds to the frequency axis if one were using Fourier transforms, we know, by analogy, the state steady-state portion of (56) will be the pole-terms. Let r_i, $i = 1, \cdots, N$ be the N Nth roots of unity and define

$$Q_i^{N-1}\left(\frac{1}{z}\right) \equiv \sum_{k=0}^{N-1} \left(\frac{1}{r_i}\right)^{N-1-k} \left(\frac{1}{z}\right)^{k} = \frac{1 - z^{-N}}{\frac{1}{r_i} - \frac{1}{z}}. \tag{57}$$

Note (57) implies

$$Q_i^{N-1}\left(\frac{1}{r_i}\right) = N r_i . \tag{58}$$

Then from (56)–(58) we have

$$\hat{Y}(z) = \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right)}{\left(\frac{1}{r_i} - \frac{1}{z}\right) \cdot N r_i \cdot D(r_i)}, \tag{59}$$

where we have let

$$D(z) = z^2 - az - b. \tag{60}$$

Using (57) once more, the steady-state solution (59) may be written

$$\hat{Y}(z) = \frac{1}{1 - z^{-N}} \cdot \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right) Q_i^{N-1}\left(\frac{1}{z}\right)}{r_i D(r_i)}. \tag{61}$$

Referring back to the discussion at the beginning of this section, we see that (61) is the z-transform of a sequence $\{y_k\}$ of period N where

$$y_k = \text{coefficient of } z^{-k} \text{ in } \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right) Q_i^{N-1}\left(\frac{1}{z}\right)}{r_i D(r_i)} \right\}$$

$$k = 0, 1, \cdots, N - 1. \tag{62}$$

Using (57) in (62) we obtain

$$y_k = \frac{1}{N} \sum_{i=1}^{N} \frac{A_{N-1}\left(\frac{1}{r_i}\right)}{D(r_i)} r_i^k, \tag{63}$$

where, in writing (63), we have used the fact that $r_i^N = 1$. Expression (63) thus gives the $\{y_k\}$ sequence for any click sequence. It is a solution corresponding to a self-sustained oscillation of the digital filter only if we have $| y_k | < 1$, all k.

Two sums appear in (63). The explicit one shown is the sum over the roots of unity; the hidden one is the polynomial $A_{N-1}(1/r_i)$. We will exhibit another form of solution (63) by explicitly doing the sum over the N roots. We begin by writing

$$A_{N-1}\left(\frac{1}{r_i}\right) = 2 \sum_{l=0}^{N-1} \frac{p_l}{r_i^l}, \qquad p_l = \pm 1, 0. \tag{64}$$

Thus p_l are the coefficients, except for the factor of 2, of the polynomial $A_{N-1}(z)$. We also write, by factoring $D(z)$ and expanding in partial fractions,

$$\frac{1}{D(z)} = \frac{1}{(z - \rho_1)(z - \rho_2)} = \frac{1}{\rho_1 - \rho_2} \left[\frac{1}{z - \rho_1} - \frac{1}{z - \rho_2} \right]. \tag{65}$$

Now note that if z is such a number than $z^N = 1$, we have (since $| \rho | < 1$ and $| z | = 1$)

$$\frac{1}{z-\rho}=\frac{1}{z}\sum_{n=0}^{\infty}\left(\frac{\rho}{z}\right)^{n}. \tag{66}$$

Let us look at the sum of the $n = 0, N, 2N$, etc., terms in the right side of (66), that is

$$1+\frac{\rho^{N}}{z^{N}}+\frac{\rho^{2N}}{z^{2N}}+\frac{\rho^{3N}}{z^{3N}}+\cdots$$

$$=1+\rho^{N}+\rho^{2N}+\rho^{3N}+\cdots=\frac{1}{1-\rho^{N}}. \tag{67}$$

Treating the sum of terms

$$\begin{aligned}
n &= 1, N+1, 2N+1, \cdots \\
n &= 2, N+2, 2N+2, \cdots \\
&\vdots \\
n &= N-1, N+(N-1), 2N+(N-1), \cdots
\end{aligned}$$

similarly, we have

$$\frac{1}{z-\rho}=\frac{1}{z}\cdot\frac{1}{1-\rho^{N}}\left[1+\frac{\rho}{z}+\frac{\rho^{2}}{z^{2}}+\cdots+\frac{\rho^{N-1}}{z^{N-1}}\right]. \tag{68}$$

Finally letting $z = 1/r_i$ gives

$$\frac{1}{\dfrac{1}{r_i}-\rho}=\frac{r_i}{1-\rho^{N}}\sum_{n=0}^{N-1}[\rho r_i]^{n}. \tag{69}$$

Using (65) and (64) in (63) yields

$$\begin{aligned}
y_k = \frac{1}{\rho_1-\rho_2}\cdot\frac{2}{N}\sum_{i} r_i^{k}\left(\sum_{l=0}^{N-1}\frac{p_l}{r_i^{l}}\right) \\
\cdot\left[\frac{1}{r_i}\sum_{n=0}^{N-1}\frac{1}{r_i^{n}}\left(\frac{\rho_1^{n}}{1-\rho_1^{n}}-\frac{\rho_2^{n}}{1-\rho_2^{n}}\right)\right].
\end{aligned} \tag{70}$$

Two sums in (70) are immediately done. First look at the sum over the roots of unity. This involves observing that

$$\sum_{i} r_i^{k-l-1-n}=\begin{cases} N & \text{if } \ k-l-1-n\equiv 0 \mod N, \\ 0 & \text{otherwise.}\end{cases} \tag{71}$$

The congruence indicated in (71) can only be satisfied here if $l = k -$

$1 - n$ or if $l = k - 1 - n + N$. Thus it is useful to define

$$2b_n^{(k)} \equiv \bar{a}_{k-1-n} + \bar{a}_{k-1-n+N} , \tag{72}$$

where we understand $\bar{a}_j \equiv 0$ if j does not lie between 0 and $N - 1$, inclusive, and $\bar{a}_j \equiv a_j$ if it does. One of the $\bar{a}$'s in (72) will thus always be zero and $b_n^{(k)}$ has values, like the p's, of $\pm 1, 0$. Using the discussion above surrounding equations (71) and (72) we perform next the sum over l and write another form of the solution:

$$y_k = \frac{2}{\rho_1 - \rho_2} \sum_{n=0}^{N-1} b_n^{(k)} \left[\frac{\rho_1^n}{1 - \rho_1^N} - \frac{\rho_2^n}{1 - \rho_2^N} \right]$$

$$k = 0, 1, \cdots , N - 1. \tag{73}$$

REFERENCES

1. Rader, C. M., Gold, B., "Digital Filter Design Techniques in the Frequency Domain," Proc. IEEE, *55*, No. 2 (February 1967), pp. 149–171.
2. Jackson, L. B., Kaiser, J. F., and McDonald, H. S., "An Approach to the Implementation of Digital Filters," IEEE Trans. Audio and Electroacoustics, *AV-16*, No. 3 (September 1968), pp. 413–421.

Rate Optimization for Digital Frequency Modulation

By J. E. MAZO, HARRISON E. ROWE, and J. SALZ

(Manuscript received June 12, 1969)

The data rate of a multilevel digital FM system is optimized subject to fixed RF bandwidth, signal-to-noise ratio, and output error rate. The possibility of optimizing such a system was first considered by J. R. Pierce at Bell Telephone Laboratories. He made the observation that it is possible to send many levels slowly or fewer levels rapidly for an FM wave of fixed RF bandwidth and error rate, and that there must be a choice of signaling rate and number of levels that optimize the data rate. The rigorous treatment of this problem is the subject of this paper. The mathematical model we analyze uses frequency-shift keying at the transmitter and ideal discrimination detection with an integrate-and-dump circuit as the post-detection filter. Our results are exhibited graphically showing the various dependencies among the pertinent system parameters.

I. INTRODUCTION

In this paper we optimize the information rate (subject to certain constraints) of a multilevel digital FM system. This problem of delivering the maximum information through an FM system has recently been formulated by J. R. Pierce.[1] Specifically, he considered how one should choose the baseband signaling rate and the number of levels to get the most information through the channel, subject to fixed bandwidth, fixed RF signal-to-noise ratio, and fixed output error rate. This optimization has recently been carried out under the assumption that the conventional FM receiver can be linearized.[2] Small-noise linear FM theory is satisfactory when analyzing analog systems, but has its well known pitfalls in digital applications.

The purpose of this paper is to reexamine this problem more rigorously, paying particular attention to the anomalies (clicks) which can result from the nonlinear character of the receiver. In order to do this we must choose a particular mathematical model for digital

FM which is amenable to analysis. Such a model uses frequency-shift keying (FSK) at the transmitter and ideal discrimination detection with an integrate-and-dump circuit as the postdetection filter. The noise at RF is assumed to possess gaussian statistics. Although realizable FM systems do not exactly conform to this ideal mathematical model, we feel that the results predicted with the use of this model are applic-cable to real FM systems. In any case, the numerical results agree well with those derived from the linear theory. According to our present calculations, this is due to the circumstance that the optimum number of levels leads to small enough deviations so that the contribution of the clicks to the error rate can be neglected.

II. ANALYSIS

Consider an n-level FSK communication system with a sample rate $N = 1/T$, square-wave modulation, and a level separation (in frequency) Δf. Such a system would yield a data rate R given by

$$R = N \log_2 n = 1.443\, N \ln n \text{ bits/s}, \tag{1}$$

and, according to Carson's rule, occupy a bandwidth*

$$B = N + (n - 1)\Delta f. \tag{2}$$

The FM signal plus gaussian noise enters a receiver consisting of an ideal RF filter (bandwidth B), limiter, discriminator, integrator (integration time T), and sampler (sampling rate N). The sampler outputs are simply the successive values of the instantaneous phase of the modulated wave following each (rectangular) modulation pulse, and would be separated by multiples of

$$\Delta\phi = 2\pi \frac{\Delta f}{N} \text{ radians} \tag{3}$$

in the absence of noise.

The simplicity of the present system (that is, the finite-time integrator post-detection filter) has permitted a fairly rigorous determination of the probability of error for high RF signal-to-noise ratio.[4] It is shown in Ref. 4 that the parameter $\Delta\phi$ given in equation (3) plays a very important role in the theory of error rates for digital FM. In particular, it is known that if $\Delta\phi < \pi$ (or equivalently, $\Delta f/N < \frac{1}{2}$), then it is the smooth noise at the baseband output which determines the error

* Comparison with the exact FSK spectra for $n = 2, 4, 8$ suggests that this approximation is valid for present purposes.[3]

rate; while if $\Delta\phi > \pi$ ($\Delta f/N > \frac{1}{2}$), then the clicks dominate, which is the basic reason for the probability of error taking on different forms in these two cases.

The optimum systems considered here are shown to correspond to the $\Delta f/N < \frac{1}{2}$ case, for which clicks are unimportant. Therefore we take the probability of error* P as given by twice equation (17a) of Ref. 4, with $\phi \to \Delta\phi/2 = \pi\,\Delta f/N$;

$$P \sim \frac{1}{(2\pi\rho)^{\frac{1}{2}}} \frac{\cot\left(\frac{\pi}{2}\frac{\Delta f}{N}\right)}{\left(\cos\left(\pi\frac{\Delta f}{N}\right)\right)^{\frac{1}{2}}} \exp\left[-2\rho\sin^2\left(\frac{\pi}{2}\frac{\Delta f}{N}\right)\right],$$

$$\rho \gg 1, \quad \frac{\Delta f}{N} < \frac{1}{2}, \tag{4}$$

and subsequently verify that $\Delta f/N$ is indeed less than $\frac{1}{2}$ for the resulting optimum systems. Here ρ is the RF signal-to-noise ratio in the frequency band B. We treat the asymptotic approximation (for large ρ) of equation (4) as an equality in the following.

For fixed error rate P and RF signal-to-noise ratio ρ, equation (4) determines $\Delta f/N$. Rewriting equation (2),

$$\frac{B}{N} = 1 + (n-1)\frac{\Delta f}{N}; \tag{5}$$

substituting equation (5) into equation (1),

$$\frac{R}{B} = \frac{1.443 \ln n}{1 + (n-1)\frac{\Delta f}{N}} \text{ bits/cycle.} \tag{6}$$

We set the derivative of equation (6) equal to zero, determining the optimum number of levels n_0 and maximum rate R_0.

$$n_0(\ln n_0 - 1) = \frac{1}{\Delta f/N} - 1. \tag{7}$$

$$\frac{R_0}{B} = \frac{1.443}{n_0(\Delta f/N)}. \tag{8}$$

Alternatively, once the optimum number of levels n_0 has been de-

* For multilevel output samples, most errors will be to adjacent levels. Assuming that something like the Gray code is used, the symbol probability of error P of equation (4) will be approximately the bit probability of error for the final reconstructed binary signal.

termined via equations (4) and (7), we may express the other parameters of the (optimum) system in terms of n_0 only:

$$\frac{\Delta f}{N} = \frac{1}{n_0(\ln n_0 - 1) + 1}, \tag{9}$$

$$\frac{R_0}{B} = 1.443 \left[\ln n_0 + \frac{1}{n_0} - 1 \right] \text{bits/cycle}, \tag{10}$$

$$\frac{B}{N} = \frac{n_0 \ln n_0}{n_0(\ln n_0 - 1) + 1}. \tag{11}$$

Note that the restriction $\Delta f/N < \frac{1}{2}$ implies via equation (7) that

$$n_0 \geqq 4. \tag{12}$$

Finally, the Shannon capacity for the RF channel is

$$\frac{C}{B} = 1.443 \ln (1 + \rho) \text{ bits/cycle}. \tag{13}$$

III. RESULTS

Figures 1 to 7 illustrate the parameters of optimum multilevel FM systems using a finite-time integrator as a post-detection filter for two representative error rates ($P = 10^{-6}, 10^{-8}$).

The solid curves of Fig. 1 show the optimum number of levels n_0 versus the RF signal-to-noise ratio in dB, 10 $\log_{10} \rho$, for the two values of P. The curves terminate at $n_0 = 4$, according to equation (12).

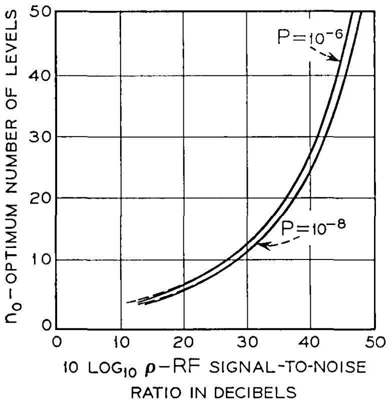


Fig. 1 — Number of levels for maximum data rate versus RF signal-to-noise ratio. Dashed lines indicate small-angle approximations.

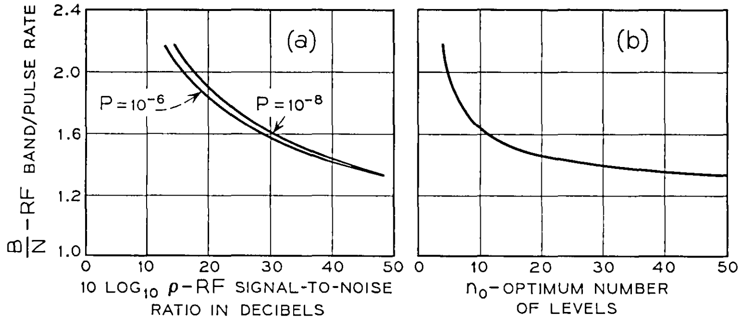


Fig. 2 — Bandwidth expansion factor for maximum data rate.

n_0 increases rapidly as ρ increases, for fixed P. The small-angle approximation for the trigonometric functions in equation (4) is shown by the dashed curves of Fig. 1; in this approximation changing P simply translates the curves of Fig. 1 horizontally. This is a reasonable approximation for the smallest n_0 permitted [by equation (12)], for the values of P of interest here.

Figures 2, 3, 4, and 5 show optimum system parameters plotted against two horizontal scales:

(*i*) 10 $\log_{10}\rho$—the RF signal-to-noise ratio in dB. Two plots are shown, for $P = 10^{-6}, 10^{-8}$. Using the small-angle approximation in equation (4), changing P translates these curves horizontally. This horizontal axis is the parameter of most direct physical interest.

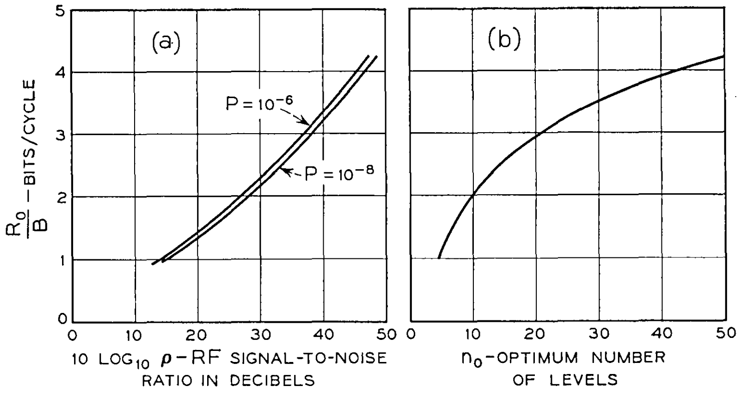


Fig. 3 — Maximum data rate per unit RF bandwidth.

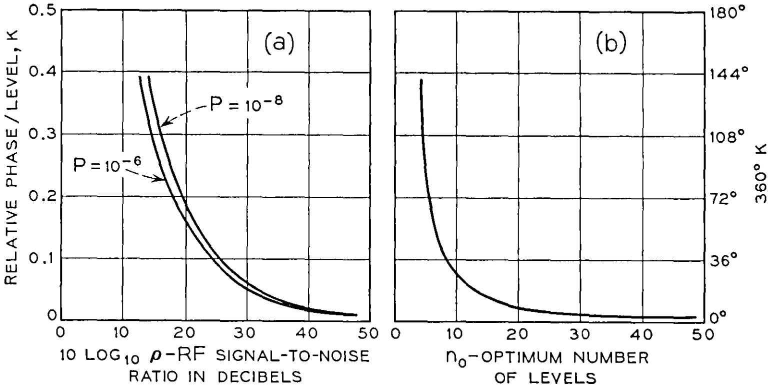


Fig. 4 — Relative phase shift per level in one sample interval for optimum systems.

(*ii*) n_0—the optimum number of levels, determined from Fig. 1. Here a single universal plot suffices rigorously for all P [That is, without small-angle approximations in equation (4)].

The vertical axes show:

Figure 2—B/N, the bandwidth expansion factor, roughly* one-half the ratio of RF to base-bandwidth. This factor varies from about 2 at small ρ or n_0, to an asymptotic limit of 1 as ρ, $n_0 \rightarrow \infty$. For large ρ, n_0 we have small-index phase modulation, with only the first sideband significant. Even for the smallest ρ, n_0 considered here the bandwidth expansion is moderate.

Figure 3—R_0/B, the normalized maximum rate in bits per cycle. This quantity increases monotonically with ρ, n_0.

Figure 4—$360 \cdot \Delta f/N$ represents the relative phase change in degrees corresponding to a change in modulation of one level.

Figure 5—$360\ (n - 1)\ \Delta f/N$ represents the maximum relative phase change in degrees in one sampling interval, corresponding to a change in modulation from the lowest to the highest level. The maximum value for this quantity, occurring for the smallest ρ, n_0 (that is, $n_0 = 4$) is not far from 360°. As ρ, n_0 increase, the maximum phase change becomes small for optimum systems.

Within the small-angle approximation, discussed in connection with Fig. 1, changing P merely shifts the horizontal (dB) axes of Fig. 1 and Figs. 2(a) to 5(a). Let us adopt the $P = 10^{-6}$ curves as standard,

* This is because the square-wave modulation assumed here is not strictly band-limited; in fact, its spectrum falls off so slowly that its rms bandwidth is infinite.

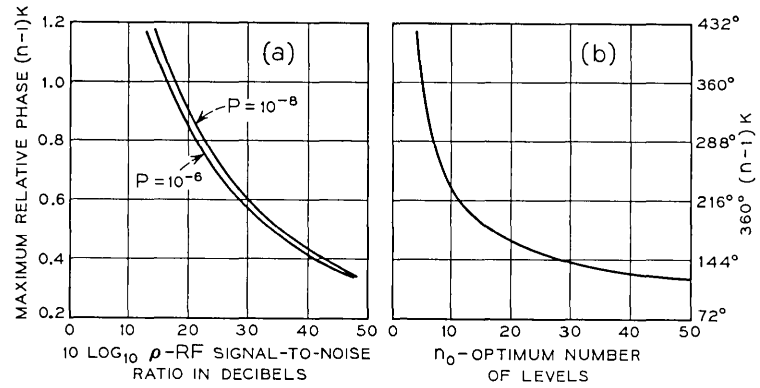


Fig. 5 — Maximum relative phase shift in one sample interval for optimum systems.

and plot the number of dB to be added to the 10 $\log_{10} \rho$ axes a sa function of P. This is shown in Fig. 6. We remark that this is only an approximation, and will begin to fail sooner as P decreases.

Finally, Fig. 7 compares the maximum data rate for the multilevel FM system with the Shannon capacity of the RF channel. The optimum data rate ranges from about 19 to 27 percent of the ideal RF channel capacity, for error probabilities P between 10^{-6} and 10^{-8}.

We have so far dealt with optimum systems. However, the number of levels may be fixed by other constraints, so that suboptimum systems are of interest. For example, it may not be practical to have the large number of levels required for optimum systems at large RF signal-to-

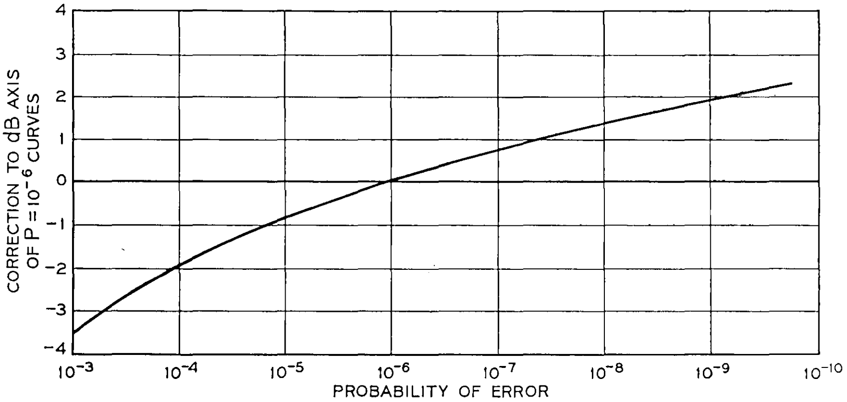


Fig. 6 — Correction for modifying $P = 10^{-6}$ curves to other error probabilities.

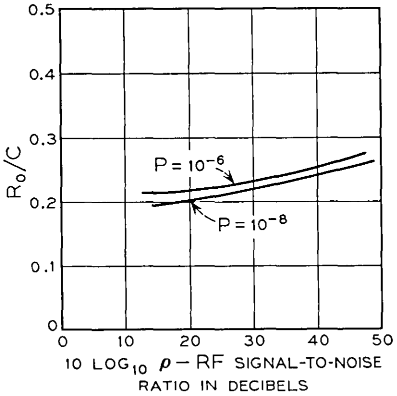


Fig. 7 — Ratio of maximum data rate to Shannon capacity.

noise ratios ρ; we may be restricted to 8 (or 16) levels, and it is necessary to determine how much the data rate will be reduced. Now rather than maximizing R by varying N and n in equation (1) subject to the constraints of equations (2) and (4), we fix n in equations (5) and (6). Figures 8 and 9 show the optimum rate R_0/B versus 10 $\log_{10}$ ρ [given

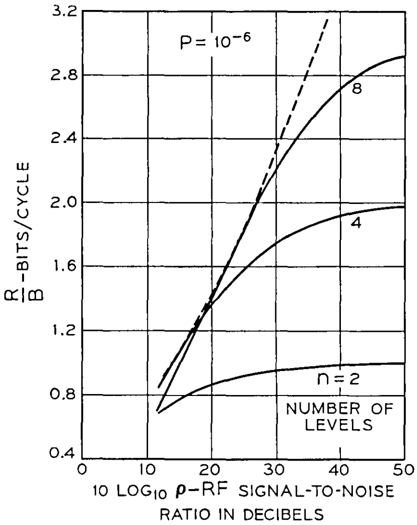


Fig. 8 — Best data rate for suboptimum systems with two, four, and eight levels compared to maximum data rate for optimum system. Dashed line—maximum data rate for optimum system, R_0/B (see Fig. 3).

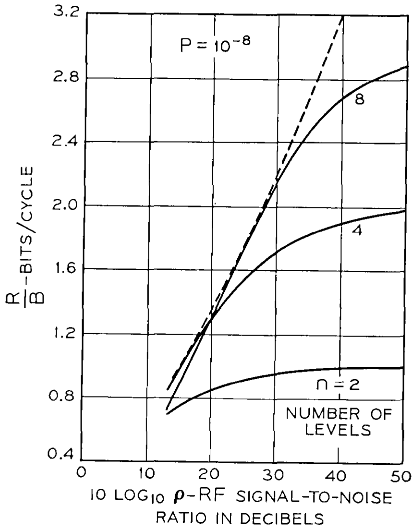


Fig. 9 — Best data rate for suboptimum systems with two, four, and eight levels compared to maximum data rate for optimum system. Dashed lined—maximum data rate for optimum system, R_0/B (see Fig. 3).

also in Fig. 3(a)], together with the rates for two, four, and eight levels, determined from equation (6) with $n = 2$, 4, and 8 for $P = 10^{-6}$, 10^{-8} in equation (4). While eight levels is strictly optimum only at the point of tangency between the R_8 and the R_0 curves, we see that the optimum is fairly broad. The corresponding bandwidth expansion factors are found from equation (5).

IV. DISCUSSION

We have presented the results of Figs. 1 through 9 as continuous curves. Actually, only isolated points of these curves are significant, since the number of levels must be integral. These continuous curves should consequently be replaced by appropriate "staircase" functions, but the difference will be significant only for small numbers of levels (that is, at low RF signal-to-noise ratios).

The present theory excludes two- and three-level systems. Naively, one might try to extend the present results to these cases by equation (17c) and Fig. 5 of Ref. 4. This may not be accurate for the error rates considered here ($P = 10^{-6}$, 10^{-8}), because the RF signal-to-noise ratio ρ becomes small, and the basic results of Ref. 4, that is, equations (17), (26), and (27), are asymptotic as ρ becomes large. However, for very much smaller error rates, for example, $P \approx 10^{-30}$, it is possible that this approach would be productive.

It would be desirable to extend the present results to binary and ternary systems; this will require a different or improved approach from the asymptotic evaluation of Ref. 4 for the error probability. It seems likely that clicks will dominate the error behavior for optimum two- and three-level systems.

The principal limitation in the present treatment (aside from the assumptions of the model, such as a finite-time integrator post-detection filter) lies in our lack of knowledge of the precise way in which the basic result for the probability of error P (equation (4) above) fails. We have merely assumed that this result holds for signal-to-noise ratios down to about 10 dB, independently of P or $\Delta f/N$. This provides additional motivation for further study of the asymptotic theory of Ref. 4.

REFERENCES

1. Pierce, J. R., unpublished work.
2. Rowe, H. E., unpublished work.
3. Lucky, R. W., Salz, J., and Weldon, E. J., *Principles of Data Communication*, New York: McGraw-Hill, 1968.
4. Mazo, J. E. and Salz, J., "Theory of Error Rates for Digital FM," B.S.T.J., *45*, No. 9 (November 1966), pp. 1511–1535.

Power Spectrum of Hard-Limited Gaussian Processes

By HARRY M. HALL

(Manuscript received September 10, 1968)

The power spectral density at the output of an ideal hard limiter (one-bit quantizer) is examined when the input is driven by a narrowband gaussian signal plus an additive gaussian noise that consists of a broadband background component plus narrowband interference. Assuming that the input signal-to-noise power ratio is small by virtue of the large bandwidth of the observed broadband noise, calculations are made of the average output signal power, the average output noise power in the signal band, and the average power of the strongest intermodulation product. The results support the intuitive conclusion that spectrum analyzer performance is degraded by the presence of the limiter and that this degradation is more pronounced when a strong narrowband interfering signal is present. They also indicate that the degradation can be minimized by making the bandwidth observed by the limiter sufficiently wide that the broadband noise power dominates both the signal and interference powers. In particular, for a typical example, the signal-to-noise power ratio measured in the signal band is degraded by less than about 1.3 dB by the presence of the limiter and the ratio of output signal power to power of the strongest intermodulation product is greater than about 14.5 dB as long as the broadband noise power exceeds the interfering-signal power.

I. INTRODUCTION

In this paper we examine the power spectral density at the output of an ideal hard limiter when the input is driven by a collection of independent gaussian processes. This work is motivated by the fact that in spectrum analysis, it is often convenient from the point of view of signal processing to precede the analyzer with a hard limiter. In order to determine the effect of the limiter on analyzer performance, it is of interest to compare the power spectral density at the limiter output with that at the limiter input. With this goal in mind, the ideal limiter to be analyzed is shown in Fig. 1. It is assumed that the limiter input is driven by the signal

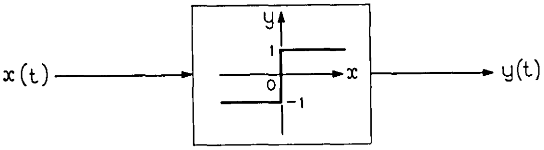


Fig. 1 — Ideal hard limiter.

$$x(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ is a sample function of the gaussian "signal" process $S(t)$ and $n(t)$ is a sample function of the gaussian "noise" process $N(t)$. More precisely, it is assumed that $S(t)$ and $N(t)$ are statistically independent, zero-mean, stationary, real, gaussian processes having continuous covariance functions $R_S(\tau)$ and $R_N(\tau)$ respectively. Further, motivated by the spectrum analysis application, the covariance functions $R_S(\tau)$ and $R_N(\tau)$ are specified: the signal process $S(t)$ is assumed to be a narrowband process with covariance function

$$R_S(\tau) = R_0(\tau) \cos \omega_0 \tau \tag{2}$$

where $S_0(f)$, the Fourier transform of $R_0(\tau)$, occupies a narrow band centered at zero frequency. The noise process $N(t)$ is assumed to consist of a broadband background component plus narrowband interference that is statistically independent of the background noise. The covariance function of the broadband background noise is assumed to be a continous covariance function that is given in the form†

$$\begin{aligned} R_1(\tau) &= R_1(\tau_1 \; ; \tau) \\ &= \frac{C_0}{\tau_1} \rho\left(\frac{|\tau|}{\tau_1}\right) \cos \omega_1 \tau, \end{aligned} \tag{3}$$

where $\rho(x)$ satisfies the conditions

$$\rho(0) = 1, \tag{4}$$

$$\int_0^\infty |\rho(x)| \, dx < \infty. \tag{5}$$

This specification of $R_1(\tau)$ has the properties:

† For example, consider the exponential covariance

$$R_{1E}(\tau) = \frac{C_0}{\tau_1} \exp\left(-\beta \frac{|\tau|}{\tau_1}\right) \cos \omega_1 \tau.$$

(*i*) The total average broadband noise power $R_1(0)$ increases linearly with τ_1^{-1} where $\tau_1 > 0$ is defined to be the broadband noise "correlation time."

(*ii*) The average broadband noise power observed in any fixed band of finite extent approaches a finite constant as the correlation time τ_1 approaches zero.

Finally, the covariance function of the narrowband interference is assumed to be given by $R_2(\tau) \cos \omega_2\tau$ where $S_2(f)$, the Fourier transform of $R_2(\tau)$, occupies a narrow band centered at zero frequency. Therefore, the covariance function of the noise process $N(t)$ is given by

$$R_N(\tau) = R_1(\tau) + R_2(\tau) \cos \omega_2\tau \qquad (6)$$

where $R_1(\tau)$ satisfies equation (3).

It was stated that the covariance functions just specified are suggested by the spectrum analysis application, and this is true in the following sense: it is often the case that one desires to analyze narrowband signals that lie at *a priori* unknown locations within a relatively wide band, and in fact it may be that the total bandwidth to be searched is a significant fraction of the band center frequency. Given such a spectrum analysis problem, it is proposed that the situation of greatest interest is that in which the average noise power in the narrow band actually occupied by the signal may or may not be comparable to the average signal power, but in which the total average noise power is much larger than the average signal power by virtue of the large noise bandwidth. Having such a situation in mind, it is seen that the model for the broadband covariance function $R_1(\tau)$ specified in equation (3) does in fact exhibit the desired behavior when the correlation time τ_1 is appropriately small.

However, in addition to this "weak-signal" situation in which the narrowband signal power $R_S(0)$ is much smaller than the broadband noise power $R_1(0)$, it is also of interest to allow the presence of "strong" narrowband signals whose average power is comparable to that of the broadband background noise. The presence of such strong narrowband signals is expected to be obvious at the limiter output, and in fact these signals are of interest since we expect that their presence will lead to the generation of intermodulation products that may interfere with the analysis of any weak signals that are present. In order to examine this situation, a narrowband interference has been included, and it is convenient to consider this interfering signal to be part of the additive noise $N(t)$.

Before proceeding with the analysis of the problem stated above, it is noted that the ideal limiter described in Fig. 1 has received a great deal of attention in the literature. The noiseless case has been considered and output amplitudes examined when the input consists of a collection of sinusoids.[1,2] The noise-alone case has been examined and results obtained for the autocorrelation function and power spectral density at the limiter output both for the case of broadband gaussian noise alone $[R_1(\tau)]$ and for the case of narrowband gaussian noise alone $[R_2(\tau) \cos \omega_2\tau]$.[3,4] The ratio of output signal-to-noise ratio (SNR) to input SNR has been evaluated for the case in which the input consists of one or two sinusoids plus narrowband gaussian noise.[5-7] These same workers have examined the strengths of intermodulation products, and the analysis of output signal and intermodulation product power has been extended to the case of an arbitrary number of sinusoids plus gaussian noise.[8,9] In addition, analysis of the limiter has played an important part in studies of the performance of angle-modulation systems, and these analyses have generally assumed that the limiter is driven by a narrowband process.

On the other hand, it does not appear that much has been reported for the situation in which the limiter is driven by a narrowband signal plus noise that includes a broadband component. Known results that have application to this situation include those of Manasse, and others, which apply when the limiter is driven by a "weak" narrowband signal plus narrowband gaussian noise whose bandwidth is much larger than that of the signal,[10] plus approximate results that apply when the input includes a narrowband component that is "much stronger" than the sum of the other inputs present.[11] We address this problem by examining the the output power spectral density when the limiter input is given by equation (1); namely, the input is made up of a narrowband gaussian signal plus a gaussian noise consisting of a broadband background component plus narrowband interference. In particular, this examination is carried out by calculating the output power spectral density in Section II, as the broadband noise correlation time τ_1 approaches zero. This calculated result is then used in Section III to evaluate three performance measures. An example of a system to which these performance measures apply is a spectrum analyzer preceded by the ideal limiter.

(*i*) The degradation in the ratio (SNR) of average signal power to average noise power in the spectral band occupied by the signal is calculated. This degradation is important because the signal-to-noise power ratio measured in the signal band is often one of the important parameters in determining system performance.

(*ii*) The ratio (SIR) of average output signal power to average image power is calculated where, if the narrowband signal is centered at a frequency f_0 and the narrowband interference is centered at a frequency f_2, then the signal image is defined to be that intermodulation product centered at the frequency $| 2f_2 - f_0 |$. This is the strongest of the intermodulation products of the signal with the additive noise, and thus it is reasonable to use the SIR as an indication of whether or not these intermodulation products will have a significant effect on system performance.

(*iii*) The ratio $\mathrm{S_2NR_0}$ of average output interference power to average output broadband noise power in the spectral band occupied by the interference is calculated. As discussed previously, the distinction in this work between signal and interference is made based upon average power at the limiter input. That is, it has been assumed that the presence of any narrowband signal having an average power comparable to that of the broadband background noise will be obvious at the limiter output, and that such an input may in fact interfere with the analysis of other narrowband inputs. $\mathrm{S_2NR_0}$ is calculated to check the assumption that in fact the presence and location of such an interfering signal will be obvious upon analyzing the power spectrum at the limiter output.

Since the performance measures listed above are calculated as the broadband noise correlation time τ_1 approaches zero, it follows that they will all apply in practice to situations in which the broadband component of the input noise has been shaped by a low-pass filter whose bandwidth is large compared with the center frequencies of the narrowband inputs that may be present. An example of a situation in which such a model is viable occurs in the spectrum analysis of underwater acoustical signals.

On the other hand, the SNR and $\mathrm{S_2NR_0}$ results obtained will not apply directly to communication situations in which the bandwidth of the additive broadband noise is much larger than that of the narrowband signal but much smaller than the system center frequency. This situation is discussed in Section IV, and it is pointed out there that the results can be modified to encompass this situation by letting the center frequencies of both the narrowband signal and additive noise increase linearly with τ_1^{-1}.

II. THE OUTPUT POWER SPECTRAL DENSITY

The output power spectral density can be calculated by using the expression for the output autocorrelation function $R_Y(\tau)$ given by Davenport and Root (Ref. 12, p. 308)

$$\begin{aligned} R_Y(\tau) &= \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \frac{2^{k+m}\Gamma^2[(k+m)/2]}{\pi^2 k!\, m!\, [R_S(0) + R_N(0)]^{k+m}} R_S^k(\tau) R_N^m(\tau), \quad k + m \text{ odd} \\ &= 0, \qquad \text{otherwise}, \end{aligned} \tag{7}$$

where $\Gamma(x)$ denotes the gamma function; in conjunction with the expression for $R_Y(\tau)$ given by Van Vleck (Ref. 3, p. 23)

$$R_Y(\tau) = \frac{2}{\pi} \arcsin \left[\frac{R_S(\tau) + R_N(\tau)}{R_S(0) + R_N(0)} \right]. \tag{8}$$

Defining α to be the fraction of the average noise power due to the broadband background noise,

$$\alpha \triangleq \frac{R_1(0)}{R_N(0)} = \frac{R_1(0)}{R_1(0) + R_2(0)}, \tag{9}$$

it is seen that the ratio η_S of average signal power to average noise power at the limiter input is given by

$$\eta_S \triangleq \frac{R_S(0)}{R_N(0)} = \alpha \frac{R_S(0)}{C_0} \tau_1 . \tag{10}$$

Now, it was pointed out in Section I that we are interested in the situation in which the signal-to-noise power ratio η_S is small, and in fact the case of interest is that in which η_S is small because τ_1 is small, that is, η_S is small due to the large bandwidth of the observed broadband background noise. Motivated by this, it is shown in Appendix A, using the expressions for $R_Y(\tau)$ given by equations (7) and (8), that when $\alpha > 0$ the output power spectral density $S_Y(f)$ is given by

$$\begin{aligned} S_Y(f) = \frac{4}{\pi} \Bigg\{ &\int_0^{\infty} \arcsin \rho_N(\tau) \cos \omega\tau \, d\tau \\ &+ \alpha \frac{R_S(0)}{C_0} \tau_1 \int_0^{\infty} [\rho_S(\tau) - (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau] \\ &\cdot [1 - (1 - \alpha)^2 \rho_2^2(\tau) \cos^2 \omega_2\tau]^{-\frac{1}{2}} \cos \omega\tau \, d\tau \Bigg\} + o(\tau_1) \end{aligned} \tag{11}$$

as $\tau_1 \to 0$, uniformly in f, where

$$\rho_\gamma(\tau) \triangleq \frac{R_\gamma(\tau)}{R_\gamma(0)}, \qquad \gamma = S, N, 0, 1, 2, \tag{12}$$

are assumed to be absolutely integrable.

Equation (11) exhibits the components that dominate the output power spectral density when the broadband noise correlation time τ_1

approaches zero. In particular, inspection of equation (11) shows that these dominant contributions include a component that is just the output power spectral density observed when the noise $N(t)$ alone is present at the limiter input, a component that has the spectral characteristics of the signal $S(t)$, and a component that is due to interaction of the signal with the interference component $[\rho_2(\tau)\cos\omega_2\tau]$ of the noise. In order to quantitatively analyze these components where, in particular, we desire to use $S_Y(f)$ to calculate the performance measures discussed in Section I it is convenient to make use of the fact that both the signal $S(t)$ and the interference component of the noise have been assumed to be narrowband processes, plus the fact that the broadband component of the noise becomes white across any fixed band of finite extent when $\tau_1 \to 0$. These properties can be exploited by expanding both $[1 - (1-\alpha)^2\rho_2^2(\tau)\cos\omega_2^2\tau]^{-\frac{1}{2}}$ and $\arcsin\rho_N(\tau)$ followed by an appropriate collection of terms. This is carried out in Appendix B and the result is

$$S_Y(f) = S_{Y_1}(f) + S_{Y_2}(f) + \frac{4}{\pi^2}\alpha\frac{R_S(0)}{C_0}\tau_1\sum_{m=0}^{\infty}\epsilon_m\frac{\Gamma^2(m+\frac{1}{2})}{\Gamma(2m+1)}(1-\alpha)^{2m}$$

$$\cdot\int_0^\infty {}_2F_1[m+\tfrac{1}{2}, m+\tfrac{1}{2}; 2m+1; (1-\alpha)^2\rho_2^2(\tau)]$$

$$\cdot\rho_S(\tau)\rho_2^{2m}(\tau)\cos 2m\omega_2\tau\cos\omega\tau\, d\tau$$

$$-\frac{8}{\pi^2}\alpha\frac{R_S(0)}{C_0}\tau_1\sum_{m=0}^{\infty}\frac{\Gamma(m+\frac{1}{2})\Gamma(m+\frac{3}{2})}{\Gamma(2m+2)}(1-\alpha)^{2m+1}$$

$$\cdot\int_0^\infty {}_2F_1[m+\tfrac{1}{2}, m+\tfrac{3}{2}; 2m+2; (1-\alpha)^2\rho_2^2(\tau)]$$

$$\cdot\rho_2^{2m+1}(\tau)\cos(2m+1)\omega_2\tau\cos\omega\tau\, d\tau + o(\tau_1) \tag{13}$$

as $\tau_1 \to 0$, uniformly in f, where ${}_2F_1(a, b; c; x)$ is Gauss's hypergeometric function (Ref. 13, p. 556), ϵ_m is the Neumann factor $\epsilon_0 = 1$, $\epsilon_m = 2(m = 1, 2, \cdots)$, and where $S_{Y_1}(f)$ and $S_{Y_2}(f)$ are given:

$$S_{Y_1}(f) = \frac{4}{\pi}\tau_1\int_0^\infty \{\arcsin[\alpha\rho(x) + 1 - \alpha]$$

$$-\arcsin(1-\alpha)\}\, dx + o(\tau_1) \tag{14}$$

as $\tau_1 \to 0$, for all $f \ll f_{\max} < \infty$ for arbitrary fixed $f_{\max}$† and

† Recall from equation (3) that

$$\rho_1(\tau) = \rho\left(\frac{|\tau|}{\tau_1}\right)\cos\omega_1\tau$$

where $\rho(x)$ satisfies equations (4) and (5).

$$S_{Y_2}(f) = \frac{4}{\pi^2} \sum_{m=0}^{\infty} \frac{\Gamma^2(m + \frac{1}{2})}{\Gamma(2m + 2)} (1 - \alpha)^{2m+1}$$
$$\cdot \int_0^{\infty} {}_2F_1[m + \tfrac{1}{2}, m + \tfrac{1}{2}; 2m + 2; (1 - \alpha)^2 \rho_2^2(\tau)]$$
$$\cdot \rho_2^{2m+1}(\tau) \cos (2m + 1)\omega_2 \tau \cos \omega\tau \, d\tau. \tag{15}$$

The expression given by equations (13), (14), and (15) exhibits in a useful fashion the components that dominate the output power spectral density when the broadband noise correlation time approaches zero. To see this more clearly, it is convenient to assume that the narrowband interference in fact has a line spectrum, that is,

$$\rho_2(\tau) \equiv 1. \tag{16}$$

This assumption is convenient since it simplifies the calculations without obscuring the most important effects that result from the presence of narrowband interference. This assumption is applied in Appendix B to equations (13), (14), and (15), and it is shown that, when $\rho_2(\tau) \equiv 1$, we can write

$$\lim_{\tau_1 \to 0} S_Y(f) = S_{Y_1}(f) + S_{Y_2}(f) + \frac{\alpha}{\pi^2 C_0} \tau_1 \sum_{m=0}^{\infty} \epsilon_m \frac{\Gamma^2(m + \frac{1}{2})}{\Gamma(2m + 1)} (1 - \alpha)^{2m}$$
$$\cdot {}_2F_1[m + \tfrac{1}{2}, m + \tfrac{1}{2}; 2m + 1; (1 - \alpha)^2]$$
$$\cdot [S_S(f - 2mf_2) + S_S(f + 2mf_2)] \tag{17}$$

where $S_{Y_1}(f)$ is given by equation (14),

$$S_{Y_2}(f) = \frac{1}{\pi^2} \sum_{m=0}^{\infty} \frac{\Gamma^2(m + \frac{1}{2})}{\Gamma(2m + 2)} (1 - \alpha)^{2m+1}$$
$$\cdot {}_2F_1[m + \tfrac{1}{2}, m + \tfrac{1}{2}; 2m + 2; (1 - \alpha)^2]$$
$$\cdot \{ \delta[f - (2m + 1)f_2] + \delta[f + (2m + 1)f_2] \} \tag{18}$$

where $\delta(x)$ denotes the Dirac-delta function, and where

$$S_S(f) = 2 \int_0^{\infty} R_S(\tau) \cos \omega\tau \, d\tau \tag{19}$$

is the power spectral density of the signal $S(t)$. Equations (17), (14), and (18) give the representation we desire, and they demonstrate that there are three contributions that dominate the output power spectral density when the broadband noise correlation time τ_1 approaches zero.

(*i*) There is a component $S_{Y_1}(f)$ that becomes white across any frequency band of finite extent as $\tau_1 \to 0$. When $\alpha = 1$, this component is just the output power spectral density that would be observed if the broadband component of the noise was present alone at the limiter input. Moreover it is necessary to specify the broadband covariance function $R_1(\tau)$ in order to calculate $S_{Y_1}(f)$. For example, if $R_1(\tau)$ is the "triangular" covariance function

$$\begin{aligned} R_{1\Delta}(\tau) &\triangleq \frac{C_0}{\tau_1}\left(1 - \frac{|\tau|}{\tau_1}\right), & |\tau| &\leqq \tau_1 \\ &\triangleq 0, & |\tau| &> \tau_1 , \end{aligned} \tag{20}$$

then equation (14) gives the result

$$S_{Y_{1\Delta}}(f) = \frac{4}{\pi}\frac{\tau_1}{\alpha}\left[\frac{\pi}{2} - \arcsin(1-\alpha) - (2\alpha - \alpha^2)^{\frac{1}{2}}\right] + o(\tau_1) \tag{21}$$

as $\tau_1 \to 0$, for all $f \leqq f_{\max} < \infty$ for arbitrary fixed $f_{\max}$.

(*ii*) There is a component $S_{Y_2}(f)$ consisting of line spectra located at $|f| = kf_2$, $k = 1, 3, \cdots$. When $\alpha = 0$, this component is just the power spectral density that would be observed if the narrowband interference was present alone at the limiter input.

(*iii*) There is a component consisting of a term that has the spectral characteristics of the signal plus terms that are intermodulation products of the signal with the narrowband interference component of the noise.

2.1 *Noise Consisting of Broadband Component Alone*

It is clear from inspection of equations (17), (14), and (18) that the output power spectral density is greatly simplified when the additive noise consists only of the broadband component ($\alpha = 1$), and in fact it is seen that in this case equation (13) reduces to the simple result

$$S_Y(f) = S_{Y_1}(f) + \frac{2}{\pi}\frac{\tau_1}{C_0} S_S(f) + o(\tau_1) \tag{22}$$

as $\tau_1 \to 0$, uniformly in f. Moreover, the calculation of $S_{Y_1}(f)$ is simplified when $\alpha = 1$. For example, if $R_1(\tau)$ is given by the triangular function in equation (20), then it is seen that, when $\alpha = 1$,

$$S_{Y_{1\Delta}}(f) = \frac{4}{\pi}\int_0^{\tau_1} \arcsin\left(1 - \frac{\tau}{\tau_1}\right) \cos\omega\tau \, d\tau. \tag{23}$$

This integral can be evaluated using Erdelyi [Ref. 14, item 4.8(1)], and we find

$$S_{Y_{1\Delta}}(f) = 2\tau_1\left[J_0(\omega\tau_1)\operatorname{sinc}(2f\tau_1) - \frac{1}{\omega\tau_1} H_0(\omega\tau_1)\cos\omega\tau_1\right] \tag{24}$$

where $J_\nu(x)$ denotes the Bessel function of the first kind of order ν and $H_\nu(x)$ is a Struve function of order ν (Ref. 14, p. 372).[†] Note that equation (24) holds for all f and for all τ_1 . $S_{Y1\Delta}(f)$ is plotted in Fig. 2 along with

$$S_{1\Delta}(f) = C_0 \operatorname{sinc}^2 (f\tau_1), \tag{25}$$

the power spectral density at the limiter input corresponding to $R_{1\Delta}(\tau)$. The plotted data are normalized so that both processes have the same average power. Thus the data plotted in Fig. 2 show explicitly how the ideal limiter redistributes the average broadband noise power across the band and demonstrate in particular the power-spreading effect that takes place due to the limiter nonlinearity.

III. EVALUATION OF PERFORMANCE MEASURES

It is now desired to use the output power spectral density results derived above to evaluate the performance measures discussed in Section I. These calculations use directly the results derived above except that the assumption that the narrowband interference has a line spectrum can be relaxed. That is, the results derived below continue to be useful as long as the interference is a narrowband gaussian process with the covariance function $R_2(\tau) \cos \omega_2\tau$ specified in Section I.

3.1 *Degradation in Signal-to-Noise Power Ratio*

The degradation in signal-to-noise power ratio in the spectral band occupied by the signal is obtained by calculating the ratio $\text{SNR}_0/\text{SNR}_\text{I}$ of output signal-to-noise power ratio to input signal-to-noise power ratio, where these SNR's are calculated in the spectral band B occupied by the signal. Moreover, we assume that:

(*i*) The band B contains significant contributions from only the narrowband signal and the broadband component of the noise, that is, the narrowband interference and intermodulation products of the narrowband signal with the narrowband interference have negligible power in the band B.

(*ii*) $R_1(\tau)$ is the triangular function in equation (20) since it is necessary to specify the covariance function of the broadband component of the noise.

Making these assumptions, the ratio $\text{SNR}_0/\text{SNR}_\text{I}$ measured in the

[†] Note that $\operatorname{sinc} x \triangleq \dfrac{\sin \pi x}{\pi x}$.

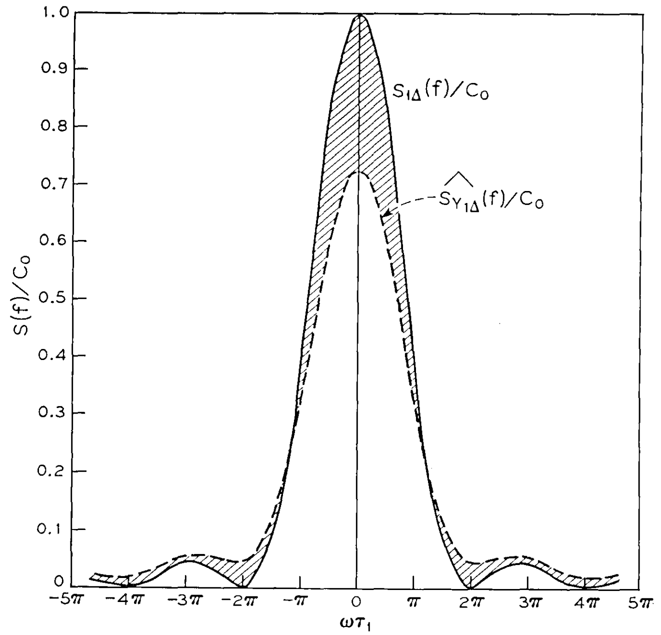


Fig. 2 — Normalized power spectral density.

$$\widehat{S_{Y_1\Delta}}(f) = \frac{R_{1\Delta}(0)}{R_{Y_1\Delta}(0)} S_{Y_1\Delta}(f); \; S_{1\Delta}(f) = C_0 \operatorname{sinc}^2 (f\tau_1).$$

band B of finite extent can be calculated using $S_Y(f)$ given by equation (17), and it is seen that

$$\lim_{\tau_1 \to 0} \frac{\mathrm{SNR}_0}{\mathrm{SNR}_\mathrm{I}} = \frac{\int_B S_{Y_S}(f)\, df \int_B S_1(f)\, df}{\int_B S_{Y_1}(f)\, df \int_B S_S(f)\, df} \tag{26}$$

where $S_{Y_1}(f)$ is given by equation (21), $S_S(f)$ is the power spectral density of the narrowband signal $S(t)$, and $S_{Y_S}(f)$ and $S_1(f)$ are given: $S_{Y_S}(f)$ is defined to be the contribution to $S_Y(f)$ that has the spectral characteristics of the signal $S(t)$ and thus is determined by setting $m = 0$ in the sum in equation (17). This gives

$$S_{Y_S}(f) = \frac{2}{\pi} \frac{\alpha}{C_0} \, {}_2F_1[\tfrac{1}{2}, \tfrac{1}{2}; 1; (1 - \alpha)^2]\tau_1 S_S(f) \tag{27}$$

which (using p. 387 of Ref. 14) can be written as

$$S_{Y_S}(f) = \frac{4}{\pi^2}\frac{\alpha}{C_0}K(1 - \alpha)\tau_1 S_S(f) \qquad (28)$$

where $K(k)$ denotes the complete elliptic integral of the first kind. $S_1(f)$ is defined to be the power spectral density at the limiter input due to the broadband component of the noise and thus, using equation (25), is given by

$$S_1(f) = C_0 + O(\tau_1^2) \qquad (29)$$

as $\tau_1 \to 0$, for all $f \leqq f_{\max} < \infty$ for arbitrary fixed $f_{\max}$. Thus, making the appropriate substitutions into equation (26) yields

$$\lim_{\tau_1 \to 0} \frac{\text{SNR}_0}{\text{SNR}_I} = \frac{\alpha^2 K(1 - \alpha)}{\pi\left[\frac{\pi}{2} - \arcsin(1 - \alpha) - (2\alpha - \alpha^2)^{\frac{1}{2}}\right]}. \qquad (30)$$

This relative signal-to-noise power ratio result is plotted in Fig. 3 and demonstrates the expected result that the degradation in the signal band increases when there is a strong narrowband interfering signal present at the limiter input. However, it is important to note that the

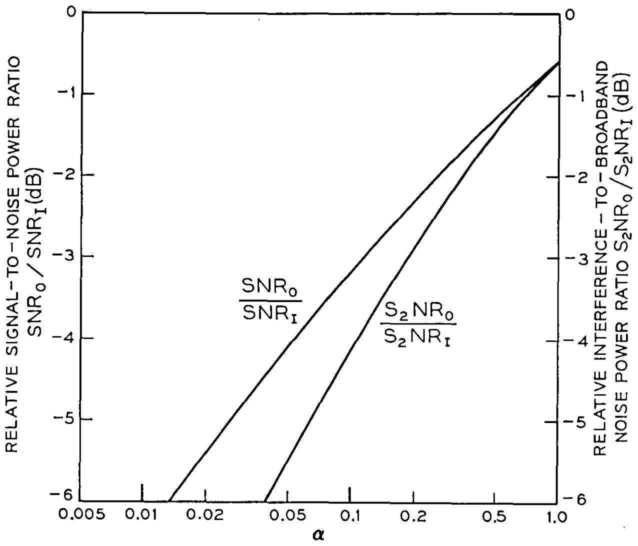


Fig. 3 — Relative signal-to-noise power ratios.

$$\alpha \triangleq \frac{R_1(0)}{R_1(0) + R_2(0)}.$$

narrowband interference must be very strong to cause a significant increase in the degradation. In particular, it is seen that the degradation is less than about 1.3 dB as long as α is greater than 0.5, that is, as long as the broadband noise power is greater than the narrowband interference power.

3.2 *Signal-to-Image Power Ratio*

The signal-to-image power ratio (SIR) is obtained by calculating the ratio of average output signal power to average image power where the image has been defined to be that narrowband component of $S_Y(f)$ centered at the frequency $| 2f_2 - f_0 |$. The SIR can be calculated using $S_Y(f)$ given by equation (17), but it should be noted that, when $\tau_1 \to 0$, the SIR does not depend on the particular choice of $R_1(\tau)$ within the class specified by equation (3). Using equation (17), it is seen that

$$\lim_{\tau_1 \to 0} \text{SIR} = \frac{\int_{-\infty}^{\infty} S_{Y_S}(f)\, df}{\frac{1}{2}\int_{-\infty}^{\infty} S_{Y_I}(f)\, df} \tag{31}$$

where $S_{Y_S}(f)$ is given by equation (28) and $S_{Y_I}(f)$ is found by setting $m = 1$ in the sum in equation (17). That is,

$$S_{Y_I}(f) = \frac{1}{4\pi} \frac{\alpha(1-\alpha)^2}{C_0} \, {}_2F_1[\tfrac{3}{2}, \tfrac{3}{2}; 3; (1-\alpha)^2]\tau_1 \cdot [S_S(f - 2f_2) + S_S(f + 2f_2)], \tag{32}$$

which, using Abramowitz and Stegun [Ref. 13, item 15.2.1] together with Price [Ref. 15, p. 10] and Dwight [Ref. 16, items 788.1, 788.2], can be written as

$$S_{Y_I}(f) = \frac{8}{\pi^2} \frac{\alpha}{(1-\alpha)^2 C_0} \left[\frac{1 + 2\alpha - \alpha^2}{2} K(1-\alpha) - E(1-\alpha) \right] \tau_1 \cdot [S_S(f - 2f_2) + S_S(f + 2f_2)] \tag{33}$$

where $E(k)$ denotes the complete elliptic integral of the second kind. Making the appropriate substitutions, there results

$$\lim_{\tau_1 \to 0} \text{SIR} = \frac{(1-\alpha)^2 K(1-\alpha)}{(1 + 2\alpha - \alpha^2)K(1-\alpha) - 2E(1-\alpha)}. \tag{34}$$

This SIR result is plotted in Fig. 4 and demonstrates that the signal-to-image power ratio decreases when there is a strong narrowband interfering signal present at the limiter input. In fact, equation (34)

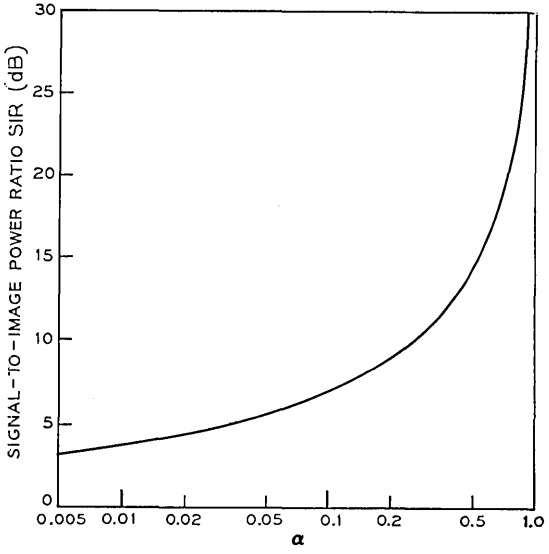


Fig. 4 — Signal-to-image power ratio.

$$\alpha \triangleq \frac{R_1(0)}{R_1(0) + R_2(0)}.$$

has the limiting behavior

$$\lim_{\alpha \to 0} \lim_{r_1 \to 0} \text{SIR} = 1, \tag{35}$$

which agrees with the approximate result obtained when one assumes that the input to the limiter includes a narrowband component that is much stronger than the sum of the other input components present.[11] However, the most interesting result demonstrated by Fig. 4 is that the narrowband interference must be very strong for the image power to be comparable to the signal power at the limiter output. In particular, it is seen that the SIR is greater than about 14.5 dB as long as the broadband noise power is greater than the narrowband interference power.

3.3 *Output Interference-to-Broadband Noise Power Ratio*

The output interference-to-broadband noise power ratio S_2NR_0 is obtained by calculating the ratio of average output interference power to average output broadband noise power, measured in the spectral band B_2† occupied by the interference. In order to perform this calculation it is necessary to specify the broadband covariance function, and is is assumed that $R_1(\tau)$ is the triangular function in equation (20). Having specified $R_1(\tau)$ in this manner, $\mathrm{S_2NR_0}$ can be calculated using $S_Y(f)$ given by equation (17), and it is seen that

$$\lim_{\tau_1 \to 0} \mathrm{S_2NR_0} = \frac{\int_{B_2} S_{Y_2}(f)\, df}{\int_{B_2} S_{Y_1}(f)\, df} \tag{36}$$

where $S_{Y_1}(f)$ is given by equation (21) and $S_{Y_2}(f)$ is given by equation (18). Proceeding with these substitutions and making the assumption that the components of $S_{Y_2}(f)$ concentrated at (odd) harmonics of the fundamental frequency f_2 contribute negligible power in the band B_2, there results

$$\lim_{\tau_1 \to 0} \mathrm{S_2NR_0} = \frac{\alpha(1-\alpha)\; {}_2F_1[\frac{1}{2}, \frac{1}{2}; 2; (1-\alpha)^2]}{2\tau_1\left[\frac{\pi}{2} - \arcsin(1-\alpha) - (2\alpha-\alpha^2)^{\frac{1}{2}}\right]\left(\int_{B_2} df\right)}, \tag{37}$$

which, making use of Price [Ref. 15, p. 10], can be written as

$$\lim_{\tau_1 \to 0} \mathrm{S_2NR_0} = \frac{2\alpha[E(1-\alpha) - (2\alpha-\alpha^2)K(1-\alpha)]}{\pi(1-\alpha)\left[\frac{\pi}{2} - \arcsin(1-\alpha) - (2\alpha-\alpha^2)^{\frac{1}{2}}\right]W\tau_1} \tag{38}$$

where

$$W \triangleq \int_{B_2} df. \tag{39}$$

The normalized power ratio $\lim_{\tau_1 \to 0} W\tau_1(\mathrm{S_2NR_0})$ is plotted in Fig. 5, and the plotted data are seen to support the intuitive assumption made in Section I that the presence and location of a narrowband input having an average power comparable to that of the broadband background noise will be obvious at the limiter output.

A result of perhaps more interest than $\mathrm{S_2NR_0}$ is the ratio $\mathrm{S_2NR_0}/\mathrm{S_2NR_I}$ of output interference-to-broadband noise power ratio to input interference-to-broadband noise power ratio. This calculation can be carried out in the same way that $\mathrm{SNR_0}/\mathrm{SNR_I}$ was calculated earlier, and we find

† This calculation is not of interest if the interference truly has a line spectrum (that we can resolve). However, it is of interest here since these results are useful as long as the interference is a narrowband gaussian process.

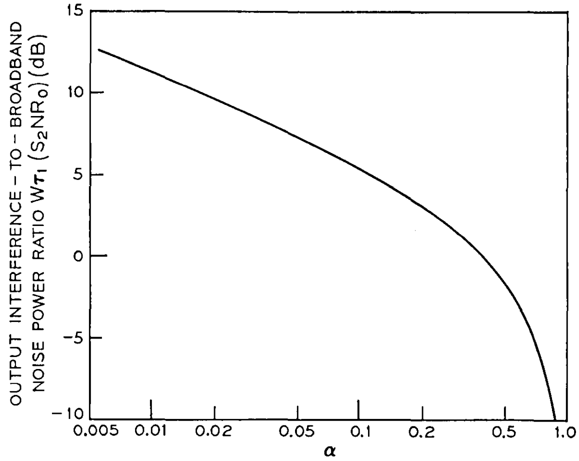


Fig. 5 — Normalized output interference-to-broadband noise power ratio.

$$\alpha = \frac{R_1(0)}{R_1(0) + R_2(0)}.$$

$$\lim_{\tau_1 \to 0} \frac{\mathrm{S_2NR_0}}{\mathrm{S_2NR_I}} = \frac{\int_{B_2} S_{r_2}(f)\, df \int_{B_2} S_1(f)\, df}{\int_{B_2} S_{r_1}(f)\, df \int_{B_2} S_2(f)\, df} \tag{40}$$

where S_{r_1} (f) is given by equation (21), S_{r_2} (f) by equation (18), S_1 (f) by equation (29), and

$$\int_{B_2} S_2(f)\, df = R_2(0). \tag{41}$$

Making these substitutions and using the definition of α in equation (9) yields

$$\lim_{\tau_1 \to 0} \frac{\mathrm{S_2NR_0}}{\mathrm{S_2NR_I}} = \frac{2\alpha^2[E(1-\alpha) - (2\alpha - \alpha^2)K(1-\alpha)]}{\pi(1-\alpha)^2\left[\frac{\pi}{2} - \arcsin(1-\alpha) - (2\alpha - \alpha^2)^{\frac{1}{2}}\right]}. \tag{42}$$

This relative (interfering) signal-to-noise power ratio result is plotted in Fig. 3 and is particularly interesting since the plotted data can be viewed as a plot of $\mathrm{S_2NR_0/S_2NR_I}$ versus the input interfering signal-to-total broadband noise power ratio $\mathrm{S_2N_TR_I}$. That is, it is seen that the ratio of average input interfering-signal power to total average input broadband noise power is given by

$$S_2N_TR_I \triangleq \frac{R_2(0)}{R_1(0)} = \frac{1-\alpha}{\alpha}. \tag{43}$$

With this interpretation in mind, the plotted data show that there is a degradation in signal-to-noise power ratio in the signal band at all levels of input signal-to-noise power ratio as $\tau_1 \to 0$, and that this degradation increases monotonically with increasing input signal-to-total noise power ratio. We note the contrast of this result to that found by Davenport for the case in which the limiter is driven by an unmodulated sinusoid plus narrowband Gaussian noise where he shows that there is an enhancement in signal-to-noise ratio (measured in the narrow noise band) at high input signal-to-noise ratios.[5] It is also noted that the data plotted in Fig. 5 together with that in Fig. 3 show, that although the degradation increases monotonically with $S_2N_TR_I$, it does not increase as rapidly as $W_{T1}(S_2NR_I)$ itself is increasing.

IV. CONCLUSIONS

This paper has concentrated on analyzing the power spectral density at the output of an ideal limiter when the input is driven by a narrowband gaussian signal plus an additive gaussian noise that consists of a broadband background component plus a narrowband interference. Conclusions that can be drawn from this work depend upon the system in which the limiter is used, and one is led to the following conclusions when this system consists of a spectrum analyzer preceded by the ideal limiter: Spectrum analyzer performance will be degraded by the presence of the limiter, and this degradation can be substantial when there is a strong narrowband interfering signal present at the limiter input. This intuitive conclusion follows from the fact that the signal-to-noise power ratio SNR measured in the signal band may be significantly degraded by the presence of the limiter when there is a strong narrowband interfering signal present at the limiter input, plus the fact that intermodulation products of the narrowband signal with the narrowband interference may be troublesome as indicated by a decreased signal-to-image power ratio SIR.

However, it is important to note that the results also indicate that the degradation in performance can be minimized by making the bandwidth observed by the limiter sufficiently wide that the average broadband noise power dominates both the signal and interference powers. This conclusion follows from the fact that such a procedure minimizes both the degradation in SNR and the decrease in SIR mentioned above since it ultimately requires that α approach unity. In particular, the data plotted in Fig. 3 show that the signal-to-noise power ratio SNR is degraded by less than about 1.3 dB as long as the total average broadband noise power is greater than the average narrowband interference power. In addition, the data plotted in Fig. 4 show that the signal-to-image power ratio SIR is greater than about 14.5 dB as long as the total average broadband noise power is greater than the average narrowband interference power. This SIR result is interesting since it is indicative of the fact that intermodulation products do not grow as rapidly with increasing interfering-signal power in the situation analyzed here as they do when the ideal limiter is driven by two sinusoids plus narrowband Gaussian noise. This conclusion follows from comparison of Fig. 4 with the results of Jones as presented in his Fig. 4.[7] The difference in behavior appears to be due primarily to the fact that the strong narrowband signal in this analysis is a gaussian process and not a sinusoid.

It is of course true that the conclusions reached above based on the data plotted in Fig. 3 are conclusions based on the assumption that the broadband covariance function $R_1(\tau)$ is the triangular function specified in equation (20). This example was chosen as a typical example that is computationally convenient for studying the degradation in signal-to-noise power ratio SNR as a function of interfering-signal strength. It is also of interest to study the dependence of the degradation in SNR on the choice of $R_1(\tau)$, and it is noted that this can be accomplished by using $S_{Y_1}(f)$ given by equation (14) instead of $S_{Y_{1\Delta}}(f)$ given by equation (21) in the calculation of $\text{SNR}_0/\text{SNR}_I$.

Finally, it is emphasized that the results leading to the above conclusions are asymptotic results that apply when the broadband noise correlation time τ_1 approaches zero. As discussed in Section I, our interest in small τ_1 stems from a desire to model the situation in which the average noise power in the spectral band occupied by the narrowband signal may be comparable to the average signal power but in which the total average noise power is much larger than the average signal power by virtue of the large noise bandwidth observed by the limiter. Thus we have a practical interest in the situation of small τ_1 , although it is of course true that the situation of engineering importance is that in which τ_1 although small is greater than zero; for example, $\alpha < 1$ makes physical sense only if $\tau_1 > 0$. With this in mind, it is of interest to determine the conditions that must be satisfied for the results of this work to be useful when $\tau_1 > 0$, and inspection of the analysis performed leads to the following conclusions (when the broadband noise covariance function $R_1(\tau)$ is written such that the bandwidth of the broadband noise is approximately τ_1^{-1}): In order for the power spectral density result given by equation (11) and the signal-to-image power ratio result plotted in Fig. 4 to remain useful, it is necessary that certain conditions be satisfied:

(*i*) The broadband noise correlation time must itself satisfy the condition $\tau_1 \ll 1$.

(*ii*) The input signal-to-noise power ratio

$$\eta_S \triangleq \frac{R_S(0)}{R_N(0)} = \alpha \frac{R_S(0)}{C_0} \tau_1 \tag{10}$$

must satisfy the condition $\eta_S \ll 1$.

In addition to these conditions, in order for the power spectral density results given by (13) and (17) and the signal-to-noise power ratio results plotted in Fig. 3 and 5 to remain useful, it is necessary that the condition

$$\omega_i \tau_1 \ll 1, \qquad i = 0, 1, 2, \tag{44}$$

be satisfied. This last condition requires that the bandwidth of the broadband background noise be much larger than the largest of the center frequencies ω_0, ω_1, and ω_2. The necessity of this condition was noted in Section I, and it was pointed out that this condition is not satisfied in communications situations in which the bandwidth of the broadband noise is much larger than that of the narrowband signals that may be present but much smaller than their center frequencies. However, inspection of the derivation of equations (13) and (17) shows that, if we set

$$\omega_0 = \omega_1 = \hat{\omega}_0/\tau_1 \qquad \text{and} \qquad \omega_2 = \hat{\omega}_0/\tau_1 + \omega_\Delta , \tag{45}$$

then we have constructed a model for these "narrowband" communications situations for which equation (13) and (17) hold except for the term $S_{Y_1}(f)$ which is now given by

$$S_{Y_1}(f) = \frac{4}{\pi} \int_0^\infty \{\arcsin [\alpha\rho_1(\tau) + (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau] - \arcsin [(1 - \alpha)\rho_2(\tau) \cos \omega_2\tau]\} \cos \omega\tau \, d\tau. \tag{46}$$

Signal-to-noise power ratio results corresponding to those plotted in Figs. 3 and 5 can be calculated (numerically) using equation (17) with $S_{Y_1}(f)$ given by equation (46) after making the simplifications that follow from the definitions of ω_0, ω_1, and ω_2 given in equation (45). When $\alpha = 1$ and $\hat{\omega}_0$ is large, the signal-to-noise ratio result corresponding to Fig. 3 will reduce to the result derived by Manasse, and others.[10]

APPENDIX A

Calculation of Output Power Spectral Density

Using the characteristic function method discussed by Rice [Ref. 17] it can be shown [Ref. 12, p. 308] that, if the input to the ideal limiter of Fig. 1 is given by equation (1), then the autocorrelation function at the limiter output

$$R_Y(\tau) \triangleq \langle Y(t)Y^*(t-\tau)\rangle_{\text{av}} \tag{47}$$

is given by equation (7). Defining the input signal-to-noise power ratio η_S according to equation (10), it follows that

$$\begin{aligned} R_Y(\tau) &= \sum_{k=0}^{\infty}\sum_{m=0}^{\infty} \frac{2^{k+m}\Gamma^2[(k+m)/2]}{\pi^2 k!\, m!}\,\eta_S^k \\ &\qquad \cdot\left[\frac{\rho_S(\tau)}{1+\eta_S}\right]^k\left[\frac{\rho_N(\tau)}{1+\eta_S}\right]^m, \qquad k+m \quad \text{odd} \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{48}$$

It was pointed out in the text that we are interested in the situation where η_S is small due to the large bandwidth of the broadband background noise. Motivated by this, it is noted that, upon summing on m, equation (48) can be written as

$$\begin{aligned} R_Y(\tau) &= \sum_{k=0(\text{even})}^{\infty} \frac{2^{k+1}}{\pi^2 k!}\Gamma^2\left(\frac{k+1}{2}\right) \\ &\quad \cdot {}_2F_1\left\{\frac{k+1}{2},\frac{k+1}{2};\frac{3}{2};\left[\frac{\rho_N(\tau)}{1+\eta_S}\right]^2\right\}\frac{\rho_N(\tau)}{1+\eta_S}\left[\frac{\rho_S(\tau)}{1+\eta_S}\right]^k\eta_S^k \\ &\quad + \sum_{k=1(\text{odd})}^{\infty}\frac{2^k}{\pi^2 k!}\Gamma^2\left(\frac{k}{2}\right){}_2F_1\left\{\frac{k}{2},\frac{k}{2};\frac{1}{2};\left[\frac{\rho_N(\tau)}{1+\eta_S}\right]^2\right\}\left[\frac{\rho_S(\tau)}{1+\eta_S}\right]^k\eta_S^k . \end{aligned} \tag{49}$$

Noting that ${}_2F_1(a, b; c; x)$ is finite for all $|x| < 1$ as long as $c \neq m \cdot (m = 0, -1, -2, \cdots)$† [Ref. 13, p. 556], it follows that

† Gauss's hypergeometric function is also absolutely convergent at $|x| = 1$ as long as Re $(c - a - b) > 0$. Thus in fact

$$ {}_2F_1(\tfrac{1}{2}, \tfrac{1}{2}; \tfrac{3}{2}; 1) = \frac{\pi}{2}$$

which implies that the series

$$\arcsin x = x + \frac{1}{2.3}x^3 + \frac{1.3}{2.4.5}x^5 + \cdots$$

converges for all $|x| \leqq 1$.

$$R_Y(\tau) = \frac{2}{\pi}\left\{ {}_2F_1\left[\frac{1}{2}, \frac{1}{2}; \frac{3}{2}; \left(\frac{\rho_N(\tau)}{1+\eta_S}\right)^2\right]\frac{\rho_N(\tau)}{1+\eta_S} \right.$$
$$\left. + {}_2F_1\left[\frac{1}{2}, \frac{1}{2}; \frac{1}{2}; \left(\frac{\rho_N(\tau)}{1+\eta_S}\right)^2\right]\frac{\rho_S(\tau)}{1+\eta_S}\,\eta_S \right\}$$
$$+ O[\eta_S^2\rho_S^2(\tau)\rho_N(\tau)] + O[\eta_S^3\rho_S^3(\tau)] \tag{50}$$

as $\eta_S \to 0$, for all τ such that $|\ \rho_N(\tau)\ | < 1$. Moreover, by expressing the hypergeometric function in the first term of equation (50) in its series form and then appropriately collecting terms, it can be shown that

$$\frac{\rho_N(\tau)}{1+\eta_S}\,{}_2F_1\left[\frac{1}{2}, \frac{1}{2}; \frac{3}{2}; \left(\frac{\rho_N(\tau)}{1+\eta_S}\right)^2\right]$$
$$= \arcsin \rho_N(\tau) - \rho_N(\tau)[1 - \rho_N^2(\tau)]^{-\frac{1}{2}}\eta_S + O[\eta_S^2\rho_N(\tau)] \tag{51}$$

as $\eta_S \to 0$, for all τ such that $|\ \rho_N(\tau)\ | < 1$. Also, it is immediately recognized that, in the second term in equation (50),

$$\frac{\rho_S(\tau)}{1+\eta_S}\,{}_2F_1\left[\frac{1}{2}, \frac{1}{2}; \frac{1}{2}; \left(\frac{\rho_N(\tau)}{1+\eta_S}\right)^2\right]$$
$$= \rho_S(\tau)[1 - \rho_N^2(\tau)]^{-\frac{1}{2}} + O[\eta_S\rho_S(\tau)] \tag{52}$$

as $\eta_S \to 0$, for all τ such that $|\ \rho_N(\tau)\ | < 1$. Therefore, recalling that the noise $N(t)$ contains a broadband component so that in fact

$$|\ \rho_N(\tau)\ | < 1 \tag{53}$$

for all $|\ \tau\ | > 0$,† it is concluded upon substitution of equations (51) and (52) into equation (50) that

$$R_Y(\tau) = \frac{2}{\pi}\{\arcsin \rho_N(\tau) + [\rho_S(\tau) - \rho_N(\tau)][1 - \rho_N^2(\tau)]^{-\frac{1}{2}}\eta_S\}$$
$$+ O[\eta_S^2\rho_S(\tau)] + O[\eta_S^2\rho_N(\tau)] \tag{54}$$

as $\eta_S \to 0$, for all $|\ \tau\ | > 0$.

In order to calculate the power spectral density $S_Y(f)$ at the limiter output it is necessary to evaluate

$$S_Y(f) \triangleq 2\int_0^\infty R_Y(\tau)\cos\omega\tau\, d\tau, \qquad \omega \triangleq 2\pi f. \tag{55}$$

† Note that this follows from the integrability condition placed on the broadband covariance function $R_1(\tau)$ by (5). This integrability condition implies that $|\ \rho(x)\ | < \rho(0)$ for all $|\ x\ | > 0$ and requires that the power spectrum of the broadband noise contain no line components.

As it stands, $R_Y(\tau)$ given by equation (54) is not enough because of the difficulty as $\tau \rightarrow 0$. It is not clear from the foregoing analysis whether or not the representation given by equation (54) is valid as $\tau \rightarrow 0$ when $\eta_S \rightarrow 0$, and in fact this representation may be valid for all $\rho_S(\tau)$ and $\rho_N(\tau)$ of interest [compare Ref. 18].* In any event, the difficulties involved in evaluating the remainder terms in order to examine this possibility can be circumvented by using the well-known result that $R_Y(\tau)$ is also given by equation (8).[3] Thus,

$$R_Y(\tau) = \frac{2}{\pi} \arcsin \frac{\rho_N(\tau) + \eta_S \rho_S(\tau)}{1 + \eta_S} \tag{56}$$

which implies that

$$R_Y(\tau) = \frac{2}{\pi} \arcsin \rho_N(\tau) + o(1) \tag{57}$$

as $\eta_S \rightarrow 0$, uniformly in τ. In fact, making use of the expressions for $R_Y(\tau)$ given by equations (54) and (57) in conjunction with the expression for η_S given by equation (10) and the integrability condition in equation (5), it is seen that, if $R_1(\tau)$ can be written in the form specified by equation (3) and the parameters α and $R_S(0)/C_0$ satisfy the conditions $\alpha > 0$, $R_S(0)/C_0 < \infty$, then $R_Y(\tau)$ can be expressed:†

$$\begin{aligned} R_Y(\tau) &= \frac{2}{\pi} \arcsin \rho_N(\tau) + o(1), \qquad 0 \leqq |\tau| \leqq \tau_1 \\ &= \frac{2}{\pi} \left\{ \arcsin \rho_N(\tau) + \alpha \frac{R_S(0)}{C_0} [\rho_S(\tau) - \rho_N(\tau)][1 - \rho_N^2(\tau)]^{-\frac{1}{2}} \tau_1 \right\} \\ &\quad + O[\tau_1^2 \rho_S(\tau)] + O[\tau_1^2 \rho_N(\tau)], \qquad |\tau| \geqq \tau_1 , \end{aligned} \tag{58}$$

sa $\tau_1 \rightarrow 0$. Substituting this result into equation (55) and assuming that the integrability conditions

$$\int_0^\infty |\rho_S(\tau)| \, d\tau < \infty \tag{59}$$

$$\int_0^\infty |\rho_N(\tau)| \, d\tau < \infty \tag{60}$$

are satisfied, there results

* McFadden derives a similar expression for the case of a weak sinusoid in additive gaussian noise and asserts that the expansion is valid at $\tau = 0$ as long as $\rho_N(\tau)$ satisfies certain differentiability conditions.

† Another method for obtaining equation (58) is to expand equation (56) in a Taylor series about $\rho_N(\tau)$.

$$\begin{aligned} S_Y(f) &= \frac{4}{\pi}\left\{\int_0^\infty \arcsin \rho_N(\tau) \cos \omega\tau \, d\tau \right. \\ &\quad \left. + \alpha \frac{R_S(0)}{C_0} \tau_1 \int_{\tau_1}^\infty [\rho_S(\tau) - \rho_N(\tau)][1 - \rho_N^2(\tau)]^{-\frac{1}{2}} \cos \omega\tau \, d\tau \right\} + o(\tau_1) \end{aligned} \tag{61}$$

as $\tau_1 \rightarrow 0$, uniformly in f. This result can immediately be simplified by observing that the predominant contributions to $S_Y(f)$ due to interaction of the signal and noise processes are due to interaction of the signal process with the narrowband interference component of the noise. In fact, noting that

$$\rho_N(\tau) = \alpha\rho_1(\tau) + (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau, \tag{62}$$

it can be seen that equation (61) reduces to equation (11).

APPENDIX B

Derivation of Output-Power Spectral Density Expansion

It is shown in Appendix A that the output power spectral density can be expressed according to equation (11); namely, that

$$\begin{aligned} S_Y(f) = S_{Y_N}(f) + \frac{4}{\pi} \alpha \frac{R_S(0)}{C_0} \tau_1 \int_0^\infty [\rho_S(\tau) - (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau] \\ \cdot [1 - (1 - \alpha)^2 \rho_2^2(\tau) \cos^2 \omega_2\tau]^{-\frac{1}{2}} \cos \omega\tau \, d\tau + o(\tau_1) \end{aligned} \tag{63}$$

as $\tau_1 \rightarrow 0$, uniformly in f, where

$$S_{Y_N}(f) \triangleq \frac{4}{\pi} \int_0^\infty \arcsin [\alpha\rho_1(\tau) + (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau] \cos \omega\tau \, d\tau \tag{64}$$

is the output power spectral density when the noise $N(t)$ alone is present at the limiter input. $S_Y(f)$ can be put in a more useful form by expanding both $[1 - (1 - \alpha)^2\rho_2^2(\tau) \cos^2 \omega_2\tau]^{-\frac{1}{2}}$ and $\arcsin [\alpha\rho_1(\tau) + (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau]$. Proceeding with expansion of the latter it is seen that [Ref. 13, item 15.1.6]

$$\begin{aligned} &\arcsin [\alpha\rho_1(\tau) + (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau] \\ &\quad = \frac{1}{2\pi^{\frac{1}{2}}} \sum_{m=0}^{\infty} \frac{\Gamma^2(m + \frac{1}{2})}{\Gamma(m + \frac{3}{2})m!} [\alpha\rho_1(\tau) + (1 - \alpha)\rho_2(\tau) \cos \omega_2\tau]^{2m+1} \\ &\quad = \frac{1}{2\pi^{\frac{1}{2}}} \sum_{m=0}^{\infty} \frac{\Gamma^2(m + \frac{1}{2})}{\Gamma(m + \frac{3}{2})m!} \sum_{j=0}^{2m+1} \frac{(2m + 1)!}{(2m + 1 - j)!\, j!} \\ &\quad\quad \cdot [(1 - \alpha)\rho_2(\tau) \cos \omega_2\tau]^j [\alpha\rho_1(\tau)]^{2m+1-j} \end{aligned}$$

$$= \arcsin\,[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau] + \frac{1}{2\pi^{\frac{1}{2}}}\sum_{m=0}^{\infty}\frac{\Gamma^2(m+\frac{1}{2})}{\Gamma(m+\frac{3}{2})m!}$$

$$\cdot\sum_{j=0}^{2m}\frac{(2m+1)!}{(2m+1-j)!\,j!}\,[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau]^j[\alpha\rho_1(\tau)]^{2m+1-j}. \qquad (65)$$

Thus, substituting equation (65) into equation (64), we have

$$S_{Y_N}(f) = S_{Y_1}(f) + S_{Y_2}(f), \qquad (66)$$

where

$$S_{Y_1}(f) = \frac{2}{\pi^{\frac{3}{2}}}\int_0^{\infty}\sum_{m=0}^{\infty}\frac{\Gamma^2(m+\frac{1}{2})}{\Gamma(m+\frac{3}{2})m!}\sum_{j=0}^{2m}\frac{(2m+1)!}{(2m+1-j)!\,j!}$$

$$\cdot[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau]^j[\alpha\rho_1(\tau)]^{2m+1-j}\,\cos\omega\tau\,d\tau \qquad (67)$$

and

$$S_{Y_2}(f) = \frac{4}{\pi}\int_0^{\infty}\arcsin\,[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau]\,\cos\omega\tau\,d\tau. \qquad (68)$$

We have succeeded in breaking $S_Y(f)$ into a broadband component $S_{Y_1}(f)$ plus a component $S_{Y_2}(f)$ consisting of narrowband contributions. In fact, letting $x \triangleq \tau/\tau_1$, it can be seen, using the integrability condition of equation (5), that

$$\int_0^{\infty}[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau]^j[\alpha\rho_1(\tau)]^{2m+1-j}\,\cos\omega\tau\,d\tau$$

$$= \tau_1\int_0^{\infty}[(1-\alpha)\rho_2(\tau_1x)\,\cos\omega_2\tau_1x]^j[\alpha\rho(x)\,\cos\omega_1\tau_1x]^{2m+1-j}\,\cos\omega\tau_1x\,dx$$

$$= \tau_1\int_0^{\infty}(1-\alpha)^j\alpha^{2m+1-j}\rho^{2m+1-j}(x)\,dx + o(\tau_1) \qquad (69)$$

as $\tau_1 \to 0$, for all $f \leqq f_{\max} < \infty$ for arbitrary fixed $f_{\max}$, as long as $j < 2m+1$. Moreover, using this integrability condition plus the fact that the series in the integrand is absolutely convergent, it can be shown that

$$S_{Y_1}(f) = \frac{2}{\pi^{\frac{3}{2}}}\tau_1\int_0^{\infty}\sum_{m=0}^{\infty}\frac{\Gamma^2(m+\frac{1}{2})}{\Gamma(m+\frac{3}{2})m!}\sum_{j=0}^{2m}\frac{(2m+1)!}{(2m+1-j)!\,j!}$$

$$\cdot(1-\alpha)^j[\alpha\rho(x)]^{2m+1-j}\,dx + o(\tau_1) \qquad (70)$$

as $\tau_1 \to 0$, for all $f \leqq f_{\max} < \infty$, which can be written as

$$S_{Y_1}(f) = \frac{4}{\pi}\tau_1\int_0^{\infty}\{\arcsin\,[\alpha\rho(x) + 1 - \alpha]$$

$$- \arcsin\,(1-\alpha)\}\,dx + o(\tau_1) \qquad (71)$$

as $\tau_1 \rightarrow 0$, for all $f \leqq f_{\max} < \infty$ for arbitrary fixed $f_{\max}$. Thus it is seen that the broadband component $S_{Y_1}(f)$ becomes white across any frequency band of finite extent as $\tau_1 \rightarrow 0$ and moreover that, if $\alpha = 1$, then $S_{Y_1}(f)$ is just the output power spectral density that would be observed if the broadband component of the noise was present alone at the limiter input.

Turning now to $S_{Y_2}(f)$ given by equation (68), it is seen that

$$\begin{aligned}
&\arcsin\,[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau] \\
&\quad = \frac{1}{2\pi^{\frac12}} \sum_{k=0}^{\infty} \frac{\Gamma^2(k+\frac12)}{\Gamma(k+\frac32)k!} [(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau]^{2k+1} \\
&\quad = \frac{1}{2\pi^{\frac12}} \sum_{k=0}^{\infty} \frac{\Gamma^2(k+\frac12)}{\Gamma(k+\frac32)k!} [(1-\alpha)\rho_2(\tau)]^{2k+1} \\
&\qquad \cdot \sum_{r=0}^{k} \frac{(2k+1)!}{(2k+1-r)!\,r!\,2^{2k}} \cos\,(2k+1-2r)\omega_2\tau. \qquad (72)
\end{aligned}$$

Now, letting $k - r \triangleq m$ and then interchanging the order of summation on k and m, there results

$$\begin{aligned}
&\arcsin\,[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau] \\
&\quad = \frac{1}{2\pi^{\frac12}} \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} \frac{\Gamma^2(k+\frac12)(2k+1)!}{\Gamma(k+\frac32)k!\,(k+m+1)!\,(k-m)!\,2^{2k}} \\
&\qquad \cdot [(1-\alpha)\rho_2(\tau)]^{2k+1} \cos\,(2m+1)\omega_2\tau. \qquad (73)
\end{aligned}$$

However [Ref. 13, item 6.1.18],

$$(2k+1)! = (2\pi)^{-\frac12} 2^{2k+\frac32} \Gamma(k+1)\Gamma(k+\tfrac32) \qquad (74)$$

so that equation (73) can be rewritten as

$$\begin{aligned}
&\arcsin\,[(1-\alpha)\rho_2(\tau)\,\cos\omega_2\tau] \\
&\quad = \frac{1}{\pi} \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} \frac{\Gamma^2(k+\frac12)}{(k+m+1)!\,(k-m)!} [(1-\alpha)\rho_2(\tau)]^{2k+1} \cos\,(2m+1)\omega_2\tau \\
&\quad = \frac{1}{\pi} \sum_{m=0}^{\infty} \sum_{j=0}^{\infty} \frac{\Gamma^2(j+m+\frac12)}{\Gamma(j+2m+2)j!} [(1-\alpha)\rho_2(\tau)]^{2j+2m+1} \cos\,(2m+1)\omega_2\tau \\
&\quad = \frac{1}{\pi} \sum_{m=0}^{\infty} \frac{\Gamma^2(m+\frac12)}{\Gamma(2m+2)} \, {}_2F_1\{m+\tfrac12,\, m+\tfrac12;\, 2m+2;\, [(1-\alpha)\rho_2(\tau)]^2\} \\
&\qquad \cdot [(1-\alpha)\rho_2(\tau)]^{2m+1} \cos\,(2m+1)\omega_2\tau. \qquad (75)
\end{aligned}$$

Substituting this result into equation (68), we obtain the result stated in equation (15).

The expansion of $[1 - (1 - \alpha)^2 \rho_2^2(\tau) \cos^2 {}_2\tau]^{-\frac{1}{2}}$ in the second term in equation (63) can be pursued in a manner identical to that used above for the expansion of arcsin $[(1 - \alpha)\rho_2(\tau) \cos \omega_2\tau]$, and the result obtained is that given in (13).

It is pointed out in the text that the assumption $\rho_2(\tau) \equiv 1$ greatly simplifies the expression for $S_Y(f)$ without obscuring the most important effects that result from the presence of narrowband interference. In particular, it is seen that the assumption $\rho_2(\tau) \equiv 1$ violates the integrability condition in equation (60). As a result, equation (13) does not hold uniformly in f under this assumption since the points $f = \pm kf_2$, $k = 1, 3, \cdots$, must be excluded. However, it is observed that equation (13) can be made to hold at these points as $\tau_1 \rightarrow 0$ by addition of the remainder term

$$O\left(\tau_1^2 \int_0^\infty \mid \rho_2(\tau) \mid d\tau\right). \tag{76}$$

Moreover, it is seen from equation (15) that, when $\rho_2(\tau) \equiv 1$, $S_{Y_2}(f)$ is nonzero only at $f = \pm kf_2$, $k = 1, 3, \cdots$, and its value at these points is

$$O\left(\int_0^\infty \mid \rho_2(\tau) \mid d\tau\right). \tag{77}$$

Thus in fact it can be seen that, when $\rho_2(\tau) \equiv 1$, it is meaningful to write $S_Y(f)$ as given by equation (17).

REFERENCES

1. Granlund, J., "Interference in Frequency-Modulation Reception," M.I.T. Res. Laboratory of Elec. Technical Rep. No. 252, Cambridge, Massachusetts, January 1949.
2. Sollfrey, W., "Hard Limiting of Three and Four Sinusoidal Signals," Rand Memorandum RM-4653-NASA, Santa Monica, California, July 1965.
3. Van Vleck, J. H., "Spectrum of Clipped Noise," Harvard University Radio Res. Laboratory Rep. No. 51, Cambridge, Massachusetts, 1943.
4. Middleton, D., "The Response of Biased, Saturated Linear, and Quadratic Rectifiers to Random Noise," J. Appl. Phys., *17*, No. 10 (October 1946), pp. 788–801.
5. Davenport, W. B., Jr., "Signal-to-Noise Ratios in Bandpass Limiters," J. Appl. Phys., *24*, No. 6 (June 1953), pp. 720–727.
6. Blachman, N. M., "The Output Signal-to-Noise Ratio of a Power-Law Device," J. Appl. Phys., *24*, No. 6 (June 1953), pp. 783–785.
7. Jones, J. J., "Hard-Limiting of Two Signals in Random Noise," IEEE Trans. Inform. Theory, *IT-9*, No. 1 (January 1963), pp. 34–42.
8. Shaft, P. D., "Limiting of Several Signals and its Effect on Communication System Performance," IEEE Trans. Commun. Technology, *COM-13*, No. 4 (December 1965), pp. 504–512.
9. Kirlin, R. L., "Hard Limiter Intermodulation With Low Input Signal-to-Noise Ratio," IEEE Trans. Commun. Technology, *COM-15*, No. 4 (August 1967), pp. 653–654.

10. Manasse, R., Price, R., and Lerner, R. M., "Loss of Signal Detectability in Band-Pass Limiters," IEEE Trans. Inform. Theory, *IT-4*, No. 1 (March 1958), pp. 34–38.
11. Darlington, S., unpublished work.
12. Davenport, W. B., Jr., and Root, W. L., *An Introduction to the Theory of Random Signals and Noise*, New York: McGraw-Hill, 1958.
13. Abramowitz, M., and Stegun, I. A., *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, New York: Dover Publications, 1965.
14. Erdelyi, A., ed., *Tables of Integral Transforms, I*, New York: McGraw-Hill, 1954.
15. Price, R., "A Note on the Envelope and Phase-Modulated Components of Narrow-Band Gaussian Noise," IRE Trans. Inform. Theory, *IT-1*, No. 2 (September 1955), pp. 9–13.
16. Dwight, H. B., *Tables of Integrals and Other Mathematical Data*, New York: Macmillan, 1961.
17. Rice, S. O., "Mathematical Analysis of Random Noise," B.S.T.J., *23*, No. 3 July 1944), pp. 282–332, *24*, No. 1 (January 1945), pp. 46–156.
18. McFadden, J. A., "The Correlation Function of a Sine Wave Plus Noise After Extreme Clippings," IRE Trans. Inform. Theory, *IT-2*, No. 2 (June 1956), pp. 82–83.

Rate-Distortion Functions for Gaussian Markov Processes*

By BARRY J. BUNIN

(Manuscript received June 5, 1969)

The rate-distortion function with a mean square error distortion criterion is investigated for a class of Gaussian Markov sources. It is found that for rates greater than a certain minimum, the rate-distortion function is equivalent to that of an independent letter source. This minimum rate was found to be less than n *bits per symbol, where* n *is the order of the Markov sequence. Comparisons between the rate-distortion function, and two quantizing systems are made.*

I. INTRODUCTION

Suppose in the communication system of Fig. 1, the source emits a sequence of continuous-valued random variables. The exact specification of such variates requires an infinite number of binary digits. Hence exact transmission would require a channel of infinite capacity. Since no physical channels possess infinite capacity, we see that exact transmission is impossible through this system.

However, if we are willing to accept some error in our specification of the source output, then finitely many binary digits are necessary. In the study of digital encoding systems, a useful quantity to know is the fewest number of binary digits necessary to represent an analog signal within a certain error. Such a quantity would give us a performance criterion with which to compare existing systems, and also tell us how much improvement is possible.

The quantity we seek is given by Shannon's rate-distortion function.[1,2] The rate-distortion function gives, for any bit rate, the minimum possible error achievable.

In this paper we study the rate-distortion functions for the important

* This research was partially supported by the Air Force Office of Scientific Research under Contract AF 49(638)-1600. This paper is part of a dissertation submitted in 1969 to the Faculty of the Polytechnic Institute of Brooklyn, in ρartial fulfillment of the requirements for the Ph.D. degree in systems science.

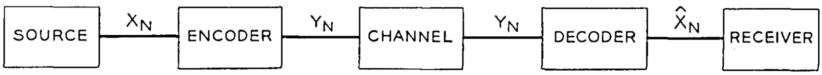


Fig. 1 — General communication system.

class of gaussian Markov sources. We measure our error by the mean square error criterion. Also, the performance of two quantizing systems, differential PCM and block quantizing, is compared to the rate-distortion bound.

II. DISCUSSION OF RESULTS

We have studied the rate-distortion functions of gaussian Markov sources with a mean square error criterion. We express our results in Fig. 2 by plotting signal-to-noise ratio in dB, versus bit rate R. The signal-to-noise ratio is given by

$$\mathrm{S/N} = 10 \log_{10} \frac{\sigma^2}{D} \tag{1}$$

where σ^2 is the variance of the source output, and D is the mean square error.

It was found that for rates R greater than a certain $R_{\min}$, the rate distortion function is given by

$$R = \tfrac{1}{2} \log_2 \frac{\sigma_m^2}{D} \qquad 0 \leqq D \leqq \sigma_m^2 \tag{2}$$

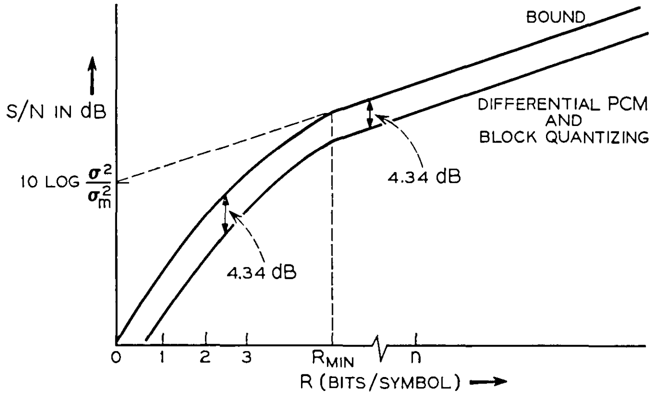


Fig. 2 — Rate-distortion bound of a Markov-n source compared with block quantizing system and differential PCM.

or

$$S/N = 6.02R + 10 \log_{10} \frac{\sigma^2}{\sigma_m^2} \tag{3}$$

where σ_m^2 is the minimum mean square prediction error one step ahead. The point $R_{\min}$ occurs in the interval $(0, n)$ where n is the order of the Markov process that the source emits. The exact location of $R_{\min}$ depends on the exact shape of the power spectral density of the process, as we shall see. At $R = R_{\min}$, the rate-distortion function has a discontinuity in the third derivative.

If the source were followed by the optimum prediction system of Fig. 3 then the output sequence produced would be uncorrelated with variance σ_m^2. Such a sequence has the rate-distortion function given by (2). Hence for rates greater than $R_{\min}$ the sequences at the input and output of the prediction system have equal rate-distortion functions. For rates less than $R_{\min}$ they do not.

A lower bound on the performance achievable by the block quantizing system of Fig. 4 was found. The result is also shown in Fig. 2, where it is seen that this system can be made to perform within 4.34 dB of the bound.

Also shown in Fig. 2 is the performance bound for a differential PCM system (see Fig. 5) as derived by O'Neal. This bound however, holds only for high bit rates.

III. RATE DISTORTION FUNCTIONS FOR MARKOV-N SOURCES

3.1 *Introduction*

Consider again the communication system of Fig. 1. The source emits the discrete time, stationary random process x_t, $t = 0, \pm 1, \pm 2, \cdots$. After N seconds, a column N vector X is obtained, and after encoding, transmission and decoding, the receiver obtains a replica $\hat{X}$ of X. The mean square error between the transmitted and received vectors is

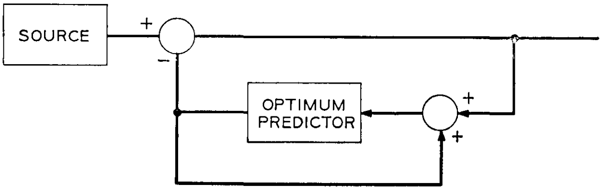


Fig. 3 — Predictive communication system.

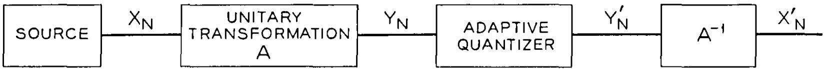


Fig. 4 — Block quantizer for correlated source.

defined by

$$D = \frac{1}{N} E(X - \hat{X})^T(X - \hat{X}) \tag{4}$$

where E denotes expectation and X^T is the transpose of X. It is reasonable to ask what the minimum bit rate is, at which we must transmit, so as to be able to achieve a mean square error less than some prescribed amount. The answer is given by Shannon's rate-distortion function which is defined as follows:[1,2]

$$R(D) = \lim_{N\to\infty} \min \frac{1}{N} \iint p(X_N)p(\hat{X}_N \mid X_N) \cdot \log_2 \frac{p(\hat{X}_N \mid X_N)}{p(\hat{X}_N)}\, dX_N\, d\hat{X}_N \tag{5}$$

where the minimization is taken over all $p(\hat{X}_N \mid X_N)$ satisfying

$$\langle D\rangle = \frac{1}{N} \iint (X_N - \hat{X}_N)^T(X_N - \hat{X}_N) \cdot p(X_N)p(\hat{X}_N \mid X_N)\, dX_N\, d\hat{X}_N \leqq D \tag{6}$$

and where

$p(X_N)$ = probability measure of the source vector X_N
$p(\hat{X}_N \mid X_N)$ = conditional probability measure of $\hat{X}_N$ given X_N
$p(\hat{X}_N)$ = probability measure induced on $\hat{X}_N$ by $p(X_N)$ and $p(\hat{X}_N \mid X_N)$.

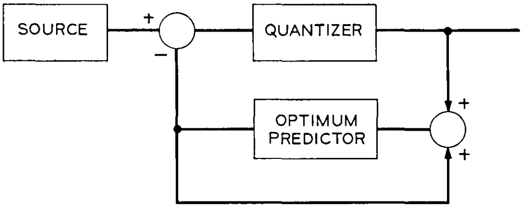


Fig. 5 — Differential pulse code modulation system.

(The subscript N is included to emphasize that we are dealing with an N-vector.)

Suppose the source emits a stationary gaussian time series with correlations $E(x_j x_k) = r_{j-k} = r_\tau$. Then the discrete time power spectral density is given by

$$f(\lambda) = \sum_{\tau=-\infty}^{\infty} r_\tau e^{i\tau\lambda} \qquad -\pi \leqq \lambda \leqq \pi \tag{7}$$

and the rate distortion function is given parametrically by[3] (see Fig. 6 for interpretation)

$$R(\phi) = \frac{1}{2} \int_A \log \frac{f(\lambda)}{\phi} \frac{d\lambda}{2\pi} \qquad 8(a)$$

$$D(\phi) = \int_A \phi \frac{d\lambda}{2\pi} + \int_{A'} f(\lambda) \frac{d\lambda}{2\pi} \qquad 8(b)$$

$$A = \{\lambda : f(\lambda) \geqq \phi\}$$

$$A' = \{\lambda : f(\lambda) < \phi\}$$

and

$$A \cup A' = (-\pi, \pi).$$

Hence, if we are given a distortion D, from (8b) we can find ϕ, and then from (8a) we can find the theoretically minimum rate R necessary to achieve a mean square error less than or equal to D. If $\{x_t\}$ consists

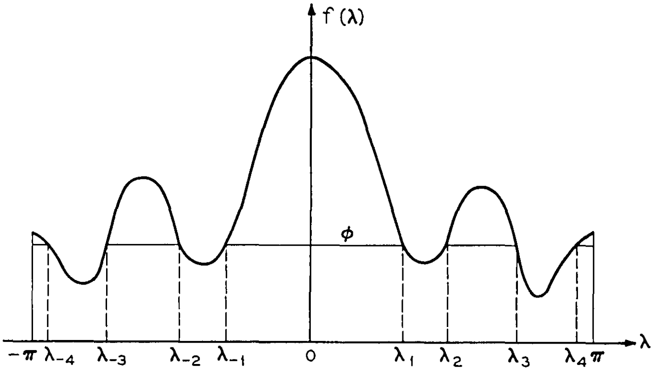


Fig. 6 — Graphical interpretation of equations 8a and b. The set $A = (-\pi, \lambda_{-4}) \cup (\lambda_{-3}, \lambda_{-2}) \cup (\lambda_{-1}, \lambda_1) \cup (\lambda_2, \lambda_3) \cup (\lambda_4, \pi)$. $A' = (\lambda_{-4}, \lambda_{-3}) \cup (\lambda_{-2}, \lambda_{-1}) \cup (\lambda_1, \lambda_2) \cup (\lambda_3, \lambda_4)$.

of independent Gaussian variates, with variance σ^2, then $f(\lambda) = \sigma^2$ and (8a) becomes

$$R(D) = \tfrac{1}{2} \log_2 \frac{\sigma^2}{D} \text{ bits/symbol.} \tag{9}$$

If we restrict the class of sources to be wide sense Markov of order n, then $f(\lambda)$ assumes the following form:

$$f(\lambda) = \frac{K}{\prod_{j=1}^{n} | e^{i\lambda} - a_j |^2} \tag{10}$$

with $0 < a_j < 1$, $a_j \neq a_k$ if $j \neq k$, and K is chosen to satisfy

$$\sigma^2 \equiv E\{x_r^2\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda)\, d\lambda. \tag{11}$$

In the remainder of this paper we consider some properties of the rate distortion function as given by (8a) and (8b) for processes with power spectral density (10).*

3.2 *The Markov-n Sequence*

In this section we present some results from prediction theory. For details and proofs see Refs. 6 and 7.

A process with power spectral density given in (10) is known as a Markov-n process.[7] Performing the indicated multiplication in (10) results in

$$f(\lambda) = \frac{K}{\prod_{j=1}^{n} | e^{i\lambda} - a_j |^2} = \frac{K}{| e^{in\lambda} + b_1 e^{i(n-1)\lambda} + \cdots + b_n |^2}. \tag{12}$$

A sequence with the spectrum (12) can be shown to satisfy the autoregressive relation

$$x_n + \sum_{i=1}^{n} b_i x_{n-i} = \epsilon_n \tag{13}$$

where $\{\epsilon_n\}$ is a sequence of uncorrelated random variables with variance K.

Writing (13) in the form

$$x_n = -\sum_{i=1}^{n} b_i x_{n-i} + \epsilon_n \tag{14}$$

* T. Berger, in a recent paper considers similar properties for the Weiner process[4].

it can be shown by the orthogonality principle (Ref. 8, Section VII-C) that the best linear predictor in the mean square sense, of x_n given the infinite past is just

$$\hat{x}_n = -\sum_{i=1}^{n} b_i x_{n-i} . \tag{15}$$

Hence for a Markov-n process the best prediction involves only the n previous samples.

The error is

$$e \equiv x_n - \hat{x}_n = \epsilon_n . \tag{16}$$

The minimum mean square error is thus

$$\sigma_m^2 \equiv E(\epsilon_n)^2 = K. \tag{17}$$

From (10) and (17)

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log_2 f(\lambda)\, d\lambda = \log_2 \sigma_m^2 - \frac{1}{2\pi}\sum_{j=1}^{n}\int_{-\pi}^{\pi} \log_2 | e^{i\lambda} - a_j |^2. \tag{18}$$

From Peirce's tables,[9] number 540, it can be shown that the integral is zero (recalling that $0 < a_j < 1$). We state our conclusion as a theorem.

Theorem 1: *For a sequence with spectrum given in* (10) *the minimum mean square error resulting from an optimal prediction one step ahead is* σ_m^2 , *where*

$$log_2\, \sigma_m^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} log_2\, f(\lambda)\, d\lambda. \tag{19}$$

Theorem 1 is a special case of the theorem proved in Ref. 6, page 183.

3.3 *Evaluation of R(D) for D* $\leqq f(\pi)$

We next consider the particular form that equations (8a) and (8b) assume when $f(\lambda)$ is as given in (10).

Theorem 2: *Given a process with*

$$f(\lambda) = \frac{K}{\prod_{j=1}^{n} | e^{i\lambda} - a_j |^2}$$

for some integer n. *For mean square errors satisfying* $0 \leqq D \leqq f(\pi)$, R(D) *is given by*

$$R(D) = \tfrac{1}{2}\, log_2 \frac{\sigma_m^2}{D}\ bits/symbol. \tag{20}$$

Proof: From (8a) and (8b)

$$R(\phi) = \frac{1}{2}\int_A \log_2 \frac{f(\lambda)}{\phi}\frac{d\lambda}{2\pi}$$

$$D = \frac{1}{2\pi}\int_A \phi\, d\lambda + \int_{A'} f(\lambda)\, d\lambda.$$

The power spectral density $f(\lambda)$ is monotonically decreasing with a minimum at $\lambda = \pi$. Hence for ϕ in the range $0 \leqq \phi \leqq f(\pi)$, $A = (-\pi, \pi)$, $A' = \emptyset$, and

$$D = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi\, d\lambda = \phi. \tag{21}$$

It follows that

$$R(\phi) = R(D) = \frac{1}{2}\int_{-\pi}^{\pi} \log_2 f(\lambda)\frac{d\lambda}{2\pi} - \tfrac{1}{2}\log_2 D. \tag{22}$$

From Theorem 1 the first term is $\frac{1}{2}\log \sigma_m^2$ so $R(D) = \frac{1}{2}\log_2 \sigma_m^2/D$ which holds for $0 < D \leqq f(\pi)$. This is (20).

The rate-distortion function (20) is precisely the rate-distortion function of a process consisting of independent gaussian random variables with mean 0 and variance σ_m^2 [see (9)].

Figure 7 illustrates why the rate-distortion function depends on $f(\pi)$ in this way. The shape of the spectrum of D in (8b) is that which would be assumed by water if it were poured into a container shaped as $f(\lambda)$. As we pour in water, it distributes itself uniformly so long as its level is below $f(\pi)$. Hence D is independent of $f(\lambda)$ so long as $D < f(\pi)$. Once $D = f(\pi)$ the exact shape of $f(\lambda)$ comes into play.

Consider next the predictive communication system of Fig. 4. The source emits the gaussian process with power spectral density (10). The

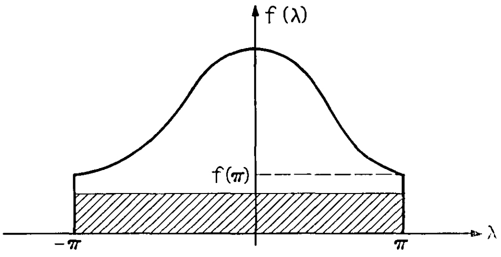


Fig. 7 — Typical Markov spectrum, illustrating water filling interpretation of the rate-distortion function.

optimum predictor makes a prediction of x_n based on $\{x_k\}_{k=0}^{n-1}$. This prediction is then subtracted from x_n and the error is transmitted. The transmitted sequence is thus the sequence $\{\epsilon_n\}$ [see (14)] which is a sequence of uncorrelated gaussian random variables with variance σ_m^2 . Its rate-distortion function is thus also given by (20), for D in the interval $0 < D \leqq \sigma_m^2$.

From (1)

$$\begin{aligned} \text{S/N} &= 10 \log_{10} \frac{\sigma^2}{D} \\ &= 10 \log_{10} \frac{\sigma^2 \sigma_m^2}{\sigma_m^2 D} \\ &= 3.01 \log_2 \frac{\sigma_m^2}{D} + 10 \log_{10} \frac{\sigma^2}{\sigma_m^2} \\ &= 6.02R + 10 \log_{10} \frac{\sigma^2}{\sigma_m^2} \end{aligned} \tag{23}$$

since R is given by (20). Hence S/N is a linear function of R over the range of R for which $0 \leqq D \leqq f(\pi)$. This range depends on n, the order of the Markov process, as given in theorem 3.

Theorem 3: For an n*th order gaussian Markov process, the rate-distortion function is given by*

$$\mathrm{R(D)} = \tfrac{1}{2}\, log_2 \frac{\sigma_{\mathrm{m}}^2}{\mathrm{D}} \ bits/symbol$$

for rates $\mathrm{R} \geqq \mathrm{R}_{min}$ *. The value of* R_{min} *depends on the exact shape of the power spectral density* $\mathrm{f}(\lambda)$ *and assumes a value satisfying*

$$0 < \mathrm{R}_{min} < \mathrm{n} \quad bits/symbol \tag{24}$$

depending on the a_i*'s of* $\mathrm{f}(\lambda)$ *[see* (10)].

Proof: From (10)

$$f(\lambda) = \frac{K}{\prod_{j=1}^{n} | e^{i\lambda} - a_j |^2}.$$

From this

$$f(\pi) = \frac{K}{\prod_{j=1}^{n} | 1 + a_j |^2}. \tag{25}$$

At $D = f(\pi)$

$$R_{\min} = R(f(\pi)) = \tfrac{1}{2} \log_2 \frac{\sigma_m^2}{f(\pi)} \text{ bits/symbol} \tag{26}$$

which from Theorem 1 is

$$= \frac{1}{2}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log_2 f(\lambda)\, d\lambda - \log_2 f(\pi)\right]$$

$$= \frac{1}{2}\left[\log_2 K - \frac{1}{2\pi}\sum_{j=1}^{n}\int_{-\pi}^{\pi} \log_2 |\, e^{i\lambda} - a_j\,|^2\, d\lambda \right.$$
$$\left. - \log_2 K + \sum_{j=1}^{n} \log_2 (1 + a_j)^2\right]\cdot \tag{27}$$

As in (18) the integral is zero and

$$R_{\min} = \sum_{j=1}^{n} \log_2 (1 + a_j) \text{ bits/symbol.} \tag{28}$$

Since $|\, a_j\,| < 1$, $R_{\min} < n$ bits/symbol. Hence, $0 < R_{\min} < n$ bits/symbol, which is the desired result.

3.4 *Behavior of $R(D)$ at $D = f(\pi)$*

With $f(\lambda)$ as given in (10), the rate-distortion function is, from (20)

$$R(D) = \tfrac{1}{2} \log_2 \frac{\sigma_m^2}{D}$$

for $0 < D \leqq f(\pi)$, and from (8a) and (8b)

$$R(\lambda) = \frac{1}{2\pi}\int_0^{\lambda} \log_2 \frac{f(\gamma)}{f(\lambda)}\, d\gamma \tag{29a}$$

$$D(\lambda) = \frac{1}{\pi}\left[\int_0^{\lambda} f(\lambda)\, d\gamma + \int_{\lambda}^{\pi} f(\gamma)\, d\gamma\right] \tag{29b}$$

for $f(\pi) \leqq D \leqq \sigma^2$. Writing (8a) and (8b) in this form follows from the observation that for a monotonically decreasing power spectral density the set A equals the simply connected interval $(0, \lambda)$ and $\phi = f(\lambda)$, for the appropriate λ.

From (20)

$$\frac{d^nR}{dD^n} = (-1)^n \frac{(n-1)!}{2} D^{-n} \ln 2 \qquad 0 < D < f(\pi) \tag{30}$$

and from (29)

$$\frac{dR}{dD} = \frac{1}{2}\frac{1}{f(\lambda)} \ln 2 \tag{31}$$

$$\frac{d^2R}{dD^2} = \frac{\pi}{2}\frac{1}{\lambda f^2(\lambda)} \ln 2 \tag{32}$$

$$\frac{d^3R}{dD^3} = -\frac{\pi^2}{2}\frac{f^2(\lambda) + 2\lambda f(\lambda)f'(\lambda)}{\lambda^3 f^4(\lambda) f'(\lambda)} \ln 2 \tag{33}$$

for $f(\pi) < D < \sigma^2$, where $f'(\lambda) = df(\lambda)/d\lambda$

From (30), (31), and (32) we see that dR/dD and d^2R/dD^2 are continuous at $D = f(\pi)$. But from (33) we see that $d^3R/dD^3 \rightarrow -\infty$ as $D \rightarrow f(\pi)$ from above (since $f'(\pi) \rightarrow 0$), whereas d^3R/dD^3 is bounded as $D \rightarrow f(\pi)$ from below. Hence d^3R/dD^3 is discontinuous at $D = f(\pi)$.

IV. QUANTIZING CORRELATED SOURCES

4.1 *Introduction*

Consider a source that emits a sequence of independent gaussian random variables of mean 0, variance σ^2. It is desired to optimally quantize the source by using an M level quantizer. Max[10] has shown that by optimally choosing the quantizer input ranges and output levels, a mean square quantization error of

$$D_q = K(M)\frac{\sigma^2}{M^2} \tag{34}$$

can be achieved where $K(M)$ is a function of M. Further, it is shown numerically that $K(M) \leqq 2.72$, and that the inequality becomes an equality as $M \rightarrow \infty$. Hence for any M

$$D_q \leqq 2.72\frac{\sigma^2}{M^2}. \tag{35}$$

For an M level quantizer the number of bits/symbol is $R = \log_2 M$, so that (35) can be written

$$D_q \leqq 2.72\frac{\sigma^2}{2^{2R}}. \tag{36}$$

The rate-distortion function of the process is from (9)

$$R = \tfrac{1}{2}\log_2\frac{\sigma^2}{D}$$

so that the minimum possible mean square error achievable with a fixed bit rate R is

$$D_{\min} \equiv \frac{\sigma^2}{2^{2R}}. \tag{37}$$

Hence Max's scheme can be made to achieve a mean square error satisfying

$$D_q \leqq 2.72\, D_{\min} \tag{38}$$

where $D_{\min}$ is the minimum mean square error as given by rate-distortion theory.

In this section we find a bound on a quantizing system studied by Huang and Schultheiss.[11] Our result is that (38) holds also for correlated sources, when $D_{\min}$ is as given by the appropriate rate-distortion funtion. For the case of Markov sources we plot this result in Fig. 2.

4.2 *Description of the System*

Referring to Fig. 4, the source emits correlated gaussian variates (not necessarily Markov), of mean 0 and with correlation matrix $\mathcal{R} = E(XX^T)$. The operator A accumulates source N-vectors X, and rotates them in such a way that

$$Y = AX \tag{39}$$

and

$$E(YY^T) = E(AXX^TA^T) = AE(XX^T)A^T = A\mathcal{R}A^T = J \tag{40}$$

where J is a diagonal matrix whose ith entry is λ_i, the ith eigenvalue of $\mathcal{R}$. Hence Y is an N-vector whose components are independent random variables with mean 0 and variance λ_i, and A is a unitary transformation.

The sequence of independent variates $\{y_i\}$ (the components of Y_N) are then quantized step by step.[10,11] The jth quantization can be optimized to produce a mean square error of

$$D_j = K(M_j)\lambda_j M_j^{-2} < 2.72\lambda_j M_j^{-2} \tag{41}$$

where M_j is the number of quantization levels used to quantize y_j. Denoting the output of the quantizer by the vector Y', the average mean square error is

$$D = \frac{1}{N} E(Y - Y')^T(Y - Y') = \frac{1}{N} E(Y - Y')^T A^T A(Y - Y')$$

$$= \frac{1}{N} E(X - X')^T(X - X') \tag{42}$$

where we have used the fact that for a unitary transformation $AA^T = AA^{-1} = I$, the identity matrix. Hence the system mean square error equals the quantizer mean square error.

From (41) and (42)

$$D = \frac{1}{N} E(Y - Y')^T(Y - Y') = \frac{1}{N} E \sum_{i=1}^{N} (y_i - y_i')^2$$

$$\leqq \frac{1}{N} 2.72 \sum_{i=1}^{N} \lambda_i M_i^{-2} \equiv D_u \,. \tag{43}$$

4.3 *Optimization over the* M_j

We next tighten the upper bound by optimally choosing the M_j's subject to the following constraints.

(*i*) $M_j \geqq 1$ for every j. The quantizer must have at least one output level.

(*ii*) The bit rate is limited by the channel capacity, C bits per symbol. We can thus use $M = 2^C$ levels per symbol or M^N levels per vector. This implies the constraint

$$M^N = \prod_{i=1}^{N} M_i \,. \tag{44}$$

Hence we wish to minimize the right side of (43) subject to (44), while keeping in mind constraint (*i*).

With ν a Lagrange multiplier, we form

$$F = D_u + \nu M^N. \tag{45}$$

A differentiation with respect to M_k yields

$$\frac{\lambda_k}{M_k^2} = \mu \tag{46}$$

where μ is a constant. Using (44) to solve for the constant gives

$$M_k = M \left[\frac{\lambda_k}{\left(\prod_{i=1}^{N} \lambda_i \right)^{1/N}} \right]^{1/2} \tag{47}$$

and

$$D_u = \frac{2.72}{M^2} \left(\prod_{i=1}^{N} \lambda_i \right)^{1/N}. \tag{48}$$

However constraint (*i*) will only hold if in (47)

$$\lambda_k \geqq \frac{\left(\prod_{i=1}^{N} \lambda_i\right)^{1/N}}{M^2}. \tag{49}$$

for every k.

The right side of (49) can be written

$$\frac{\left(\prod_{i=1}^{N} \lambda_i\right)^{1/N}}{M^2} = \frac{2^{\frac{1}{N}\sum_{i=1}^{N} \log_2 \lambda_i}}{M^2} \tag{50}$$

$$\sim = \frac{2^{\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \log_2 f(\lambda)\, d\lambda\right\}}}{M^2} \tag{51}$$

$$= \frac{\sigma_m^2}{M^2} \tag{52}$$

where we have used the fact that the eigenvalues of $\mathfrak{R}$ approach the ordinates of $f(\lambda)$ equally spaced in $(-\pi, \pi)$ as $N \to \infty$ (see Ref. 6), and then applied the definition of a Riemann integral. Finally, we used (19). Hence the constraint (i) is met if

$$\lambda_k \geqq \frac{\sigma_m^2}{M^2} \tag{53}$$

for all k. Using (50), (51), and (52), (48) becomes

$$D_u = 2.72 \frac{\sigma_m^2}{M^2}. \tag{54}$$

In terms of signal to noise ratio we get

$$\begin{aligned} \text{S/N} &= 10 \log_{10} \frac{\sigma^2}{D} \geqq 10 \log_{10} \frac{\sigma^2}{D_u} \\ &= 10 \log_{10} \frac{\sigma^2}{\sigma_m^2} + 20 \log_{10} 2 \log_2 M - 4.34 \\ &= 10 \log_{10} \frac{\sigma^2}{\sigma_m^2} + 6.02R - 4.34 \qquad (55) \end{aligned}$$

for

$$R > \tfrac{1}{2} \log_2 \frac{\sigma_m^2}{f(\pi)} \tag{56}$$

and where we used the relation

$$R = \log_2 M. \tag{57}$$

Suppose, however, that for some λ_k's (53) is not met. Specifically, arrange the eigenvalues such that $\lambda_1 \geqq \lambda_2 \geqq \lambda_3 \cdots \geqq \lambda_N$ and suppose that (47) yields

$$M_k \geqq 1 \qquad k = 1, 2, \cdots J \tag{58a}$$

$$M_k < 1 \qquad k = J + 1 \cdots N. \tag{58b}$$

Set those M_k in (58b) equal to one, and reoptimize over the M_k of (58a), the expression

$$D_J = 2.72 \sum_{k=1}^{J} \frac{\lambda_k}{M_k^2} \tag{59}$$

subject to the constraint

$$\prod_{k=1}^{J} M_k = M^N. \tag{60}$$

We would find that optimally

$$\frac{\lambda_k}{M_k^2} = \frac{\left(\prod_{i=1}^{J} \lambda_i\right)^{1/J}}{M^{2N/J}} = \gamma \qquad k = 1 \cdots J \tag{61}$$

where the right side of (61) is a constant. Without loss of generality, we can assume that all M_k obtained from (61) are greater than or equal to one. Otherwise we would set the infeasible M_k equal to one, and reoptimize. The procedure would return us to an equation similar to (61). As $N \rightarrow \infty$

$$\begin{aligned} D_q &\leqq 2.72 \frac{1}{N} \left(\sum_{i=1}^{J} \frac{\lambda_i}{M_i^2} + \sum_{i=J+1}^{N} \lambda_i \right) \\ &= 2.72 \frac{1}{N} \left(\sum_{i=1}^{J} \gamma + \sum_{i=J+1}^{N} \lambda_i \right) \\ \sim &= 2.72 \left[\frac{1}{2\pi} \int_A \gamma \, d\lambda + \frac{1}{2\pi} \int_{A'} f(\lambda) \, d\lambda \right] \end{aligned} \tag{62}$$

where A and A' are as given in (8) with ϕ replaced by γ.

Similarly

$$\gamma = \frac{\left(\prod_{i=1}^{J} \lambda_i\right)^{1/J}}{M^{2N/J}} \tag{63}$$

which, upon rearrangement, becomes

$$\begin{aligned} R \equiv \log_2 M &= \frac{1}{2N} \sum_{i=1}^{J} \log_2 \frac{\lambda_i}{\gamma} \\ \sim &= \frac{1}{4\pi} \int_A \log_2 \frac{f(\lambda)}{\gamma} \, d\lambda. \end{aligned} \tag{64}$$

By comparing (8a) and (8b) with (62) and (64) we see that (62) has the optimal spectrum for a rate given by (64). This implies that our procedure of setting infeasible M_k's equal to one does indeed lead to an optimum result.

Further, the terms in brackets in (62) is the minimum mean square error for a rate given by (64). Hence the quantization procedure has yielded

$$D_q \leqq 2.72 \, D_{\min}$$

which is (38).

This result is plotted in dB in Fig. 2, for the case of a Markov-n process.

There is an approximation involved in obtaining this result. The M_i obtained may not be integers. However, the large M_i will be little affected by rounding, and the looseness of the bound of (38) for small M_i counteracts the effects of rounding the small M_i. In fact, for very small M_i the bound is conservative, as we can see from Fig. 2. Clearly S/N should approach zero as R goes to zero. Hence our lower bound on S/N is loose in this range.

V ACKNOWLEDGMENT

I would like to thank Professor J. K. Wolf of the Polytechnic Institute of Brooklyn for his able guidance during the course of this work.

REFERENCES

1. Shannon, C. E., "Coding Theorems for a Discrete Source with a Fidelity Criterion," IRE Nat. Conv. Record, *7*, part 4, (March 1959), pp. 142–163.
2. Gallager, R. G., *Information Theory and Reliable Communication*, New York: Wiley, 1968.
3. Kolmogorov, A. N., "On the Shannon Theory of Information Transmission in the Case of Continuous Signals," IRE Trans. Inform. Theory, *IT-2*, No. 4 (December 1956), pp. 102–108.
4. Berger, T., "Information Rates of Weiner Processes," IEEE Symp. Inform. Theory, Ellenville, New York, January 28, 1969.
5. O'Neal, J. B., Jr., "A Bound on Signal to Quantizing Noise Ratios for Digital Encoding Systems," Proc. IEEE, *55*, No. 3 (March 1967), pp. 287–292.
6. Grenander, U., and Szego, G., *Toeplitz Forms and Their Applications*, Berkeley: University of California Press, 1958.
7. Yaglom, A. M., *Introduction to the Theory of Stationary Random Functions*, Englewood Cliffs, New Jersey: Prentice-Hall, 1962.
8. Rosenblatt, M., *Random Process*, New York: Oxford University Press, 1962.
9. Pierce, B. O., *A Short Table of Integrals*, New York: Ginn and Company, 1956.
10. Max, J., "Quantizing for Minimum Distortion," IRE Trans. Inform. Theory, *IT-6*, No. 1 (March 1960), pp. 7–12.
11. Huang, J. J. Y., and Schultheiss, P. M., "Block Quantization of Correlated Gaussian Random Variables," IEEE Trans. Commun. Theory, *CS-11*, No. 3 (September 1963), pp. 289–296.

The Optimum Linear Modulator for a Gaussian Source Used with a Gaussian Channel

By RANDOLPH J. PILC

(Manuscript received June 12, 1969)

The optimum linear modulator and demodulator which provide transmission of a gaussian vector source through an additive gaussian vector channel are derived in this paper. The measure of performance that is used is the transmission distortion, which is defined here as the mean square error between the source output and the decoder output. It is assumed that the source and channel are mutually independent but that correlations can exist among the components of each. The performance of the best linear system is then compared with the distortion shown by Shannon to be theoretically obtainable when no functional constraint is imposed at the modulator other than an energy constraint. Although the precise form of this optimum modulator is not known for general gaussian vector sources and channels, it is known to be nonlinear and to require arbitrarily long coding block lengths. However, it is a commonly held notion that when the source and channel dimensionalities are equal the optimum modulator is linear and requires a block length of only one. It is shown here that this belief is incorrect except in very particular situations which are described. Some relations between the optimum linear modulator-demodulator pair and Shannon's test channel are discussed, and an example is included which shows that the nonoptimality of linear devices can be quite small.

I. INTRODUCTION

We are concerned here with the transmission of a gaussian vector source over an additive gaussian vector channel. The mean square difference between the source and decoder outputs is used to measure the transmission distortion in the system and is, therefore, attempted to be minimized in the design of the encoder and decoder. In this design the encoder is constrained to present only a limited energy to the channel, thus constraining the transmission capacity of the system.[1] It is because the transmission capacity of the system is limited in this way that the given gaussian vector source cannot be transmitted with arbitrarily small error.

The distortion which necessarily must exist in the system is prescribed by Shannon's rate-distortion theory.[2] This theory states that when the transmission rate in a system is limited to R, the transmission of the source must include an average distortion of at least d_R, which in general is a function of the source statistics and the distortion measure. The theory further states that the distortion level d_R is attainable with some modulator-demodulator pair. Unfortunately, the precise form of this modulator and demodulator is not known in general, except that it is nonlinear[3,4] and that it requires the use of arbitrarily long coding block lengths.[2]

Since the nonlinearity of the optimum encoder is probably a very complex twisting of the source space locus within the channel input space, the implementation of the optimum encoder, even if it were known, would be extraordinarily complex. Of course, the long coding block length requirement does nothing to help the situation. For these reasons we study in this paper the optimum linear transmission system, restricting both the encoder and decoder to be linear operators. Such a system uses a block length of only one and is very simple to implement. (It is later shown that increasing the block length does not improve the performance.)

The degradation in performance with the use of the optimum linear system is found by comparing the resulting distortion to that of the optimum nonlinear system as found by Shannon. Contrary to popular belief, the best linear system does not provide the minimum attainable distortion, *even* when source and channel dimensionalities are equal, except in very particular situations that are described. However, in many cases the difference is small. At the end of the paper we discuss some relations between the optimum linear modulator-demodulator pair and Shannon's test channel.[2]

II. THE LINEAR TRANSMISSION SYSTEM

The system considered is shown in Fig. 1. The N_s dimensional zero-mean source vector w is linearly modulated by A to form the input to the N_c dimensional additive gaussian noise channel. We assume the noise vector n to be independent of w. The linear demodulator B extracts from the received vector y an estimate $\hat{w}$ of the source

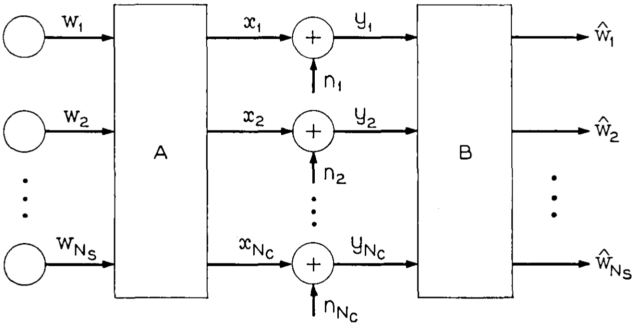


Fig. 1 — The linear system.

which is presented to the user. In summary

$$\hat{w} = By = B(x + n) = B(Aw + n). \tag{1}$$

The measure of distortion in the system is taken to be the sum of mean-square errors between the components of w and $\hat{w}$, that is

$$d = E[| w - \hat{w} |^2] = E\left[\sum_{i=1}^{N_s} (w_i - \hat{w}_i)^2 \right]. \tag{2}$$

The modulation matrices, A and B, are sought which minimize this distortion, their choice subject only to an average channel input energy constraint,

$$S_T = E\left[\sum_{i=1}^{N_c} x_i^2 \right] = \sum_{i=1}^{N_c} \operatorname{Var} x_i , \tag{3}$$

$$\leqq S_0 , \tag{4}$$

which obviously will be met with equality in the optimum system.

It is well known that the minimum mean square error estimate of any quantity (here the source vector w) based on the observation of a second quantity (here the channel output vector y) is the conditional expected value of the first given the second.[4] Further, the average error made with such an estimate is the conditional variance of the first given the second. Therefore, we have

$$\begin{aligned} \hat{w}_i &= E(w_i \mid y); \qquad i = 1, 2, \cdots , N_s \\ d &= \sum_{i=1}^{N_s} \operatorname{Var} (w_i \mid y). \end{aligned} \tag{5}$$

The required conditional density $p(w|y)$ can be found from

$$p(w) = k_1 \exp [-\tfrac{1}{2}w^t\Phi_w^{-1}w]$$

and

$$p(y \mid w) = k_2 \exp [-\tfrac{1}{2}(y - Aw)^t\Phi_n^{-1}(y - Aw)]$$

by application of Bayes rule. The result is

$$p(w \mid y) = k_3 \exp [-\tfrac{1}{2}(w - \hat{w})^t\Phi_{w \mid y}^{-1}(w - \hat{w})]$$

with

$$\Phi_{w \mid y}^{-1} = A^t\Phi_n^{-1}A + \Phi_w^{-1} \tag{6}$$

and

$$\hat{w}^t = y^t\Phi_n^{-1}A\Phi_{w \mid y} . \tag{7}$$

From these equations we have one immediate result, that is, the optimum demodulator matrix is given in terms of A by

$$B = \Phi_{w \mid y}A^t\Phi_n^{-1}. \tag{8}$$

If we now rewrite equations (5) and (3) as

$$d = \text{trace } \Phi_{w \mid y} \tag{9}$$

$$S_T = \text{trace } \Phi_x \tag{10}$$

we can restate our problem as that of finding the matrix A which minimizes the trace of $\Phi_{w \mid y}$ subject to a constrained maximum trace of Φ_x .

III. THE SOLUTION UNDER CERTAIN ASSUMPTIONS

We first restrict our attention to systems in which the source and channel dimensionalities are equal, $N_s = N_c = N$, and in which the correlation matrices Φ_w and Φ_n are diagonal. From equation (6) we have

$$\Phi_w\Phi_{w \mid y}^{-1} = \Phi_wA^t\Phi_n^{-1}A + I \tag{11}$$

and from equation (1) that $\Phi_x = A\Phi_wA^t$ and $\Phi_y = \Phi_x + \Phi_n$, which provides

$$\Phi_y\Phi_n^{-1} = A\Phi_wA^t\Phi_n^{-1} + I. \tag{12}$$

Noting that Φ_y enters these equations in a more symmetric way than dose Φ_x , we recast the energy constraint in equation (10) to be in terms of the received energy at the channel output. This energy equals

$$\begin{aligned} S_R &= E\left[\sum_{i=1}^{N_c} y_i^2\right] = \sum_{i=1}^{N_c} \operatorname{Var} y_i \\ &= \operatorname{trace} \Phi_y \\ &= \operatorname{trace} \Phi_x + \operatorname{trace} \Phi_n \end{aligned}$$

which, if trace $\Phi_n \equiv N_0$, is constrained to satisfy

$$S_R \leqq S_0 + N_0 . \tag{13}$$

3.1 *The Proof that the Optimum Modulator Matrix is Diagonal*

If we denote the characteristic polynomial of a matrix M in the variable λ by

$$\text{c.p. } [M, \lambda] = \det (M - \lambda I)$$

and state that M_i is square, we can use the following two matrix properties:[5]

$$(i) \qquad \text{c.p. } [M_1 M_2 , \lambda] = \text{c.p. } [M_2 M_1 , \lambda] \tag{14}$$

$$(ii) \qquad \text{c.p. } [M_1 , \lambda] = \text{c.p. } [M_1 + I, \lambda - 1] \tag{15}$$

to conclude from equations (11) and (12) that

$$\text{c.p. } [\Phi_w \Phi_{w|y}^{-1} , \lambda] = \text{c.p. } [\Phi_y \Phi_n^{-1}, \lambda]. \tag{16}$$

It is this equation which provides the important relations among the correlation matrices in the system.

We note that the set of matrix pairs $\Phi_{w|y}$, Φ_y which are consistent with equation (16) include many pairs which do not satisfy both equations (11) and (12) for any given A. The latter equations of course specify the relations among $\Phi_{w|y}$ and Φ_y which must exist in the communication problem under consideration. Nevertheless, we will work with equation (16) to perform the optimization and then show that the solutions for $\Phi_{w|y}$ and Φ_y can be realized with some modulator matrix A and, therefore, are consistent with the more restrictive equations (11) and (12).

Equation (14) and the assumed diagonal form of Φ_w and Φ_n allows us to rewrite equation (16) as

$$\text{c.p. } [\Phi_w^{\frac{1}{2}} \Phi_{w|y}^{-1} \Phi_w^{\frac{1}{2}} , \lambda] = \text{c.p. } [\Phi_n^{-\frac{1}{2}} \Phi_y \Phi_n^{-\frac{1}{2}}, \lambda].$$

As Φ_w and Φ_n are system constants not under the control of the user, any specification of Φ_y completely determines the roots of $\Phi_n^{-\frac{1}{2}} \Phi_y \Phi_n^{-\frac{1}{2}}$, which we denote by $\{\alpha_i\}$, $i = 1, 2, \cdots , N$. The roots of $\Phi_w^{-\frac{1}{2}} \Phi_{w|y} \Phi_w^{-\frac{1}{2}}$

are also determined and are equal to $\{\alpha_i^{-1}\}$. Our claim now is that among all matrices Φ with roots $\{\alpha_i^{-1}\}$, the one which produces the minimum trace of $\Phi_{w|v} = \Phi_w^{\frac{1}{2}}\Phi\Phi_w^{\frac{1}{2}}$ is diagonal.

If φ_{ij} are used to denote the elements of Φ, the trace of $\Phi_{w|v}$ equals

$$\text{trace } \Phi_{w|v} = \sum_{i=1}^{N} \sigma_i^2 \varphi_{ii} \, .$$

At this point we impose, without loss of generality, that the variances σ_i^2 be ordered such that $\sigma_1^2 \geqq \sigma_2^2 \geqq \cdots \geqq \sigma_N^2$. Since the minimum trace of $\Phi_{w|v}$ is sought, clearly the diagonal elements φ_{ii} should correspondingly satisfy $\varphi_{11} \leqq \varphi_{22} \leqq \cdots \leqq \varphi_{NN}$. This presents no restriction on Φ as a simultaneous interchange of rows and columns produces no change in the characteristic equation of Φ.

Now consider any nondiagonal candidate for the desired Φ. In particular, let $\varphi_{mk} = \varphi_{km}$, $m > k$, be nonzero. Because the submatrix

$$\Phi(km) = \begin{bmatrix} \varphi_{kk} & \varphi_{km} \\ \varphi_{mk} & \varphi_{mm} \end{bmatrix}$$

is itself a correlation matrix, it can be diagonalized by some orthogonal matrix T such that

$$\Phi'(km) = T\Phi(km)T^t = \begin{bmatrix} \varphi'_{kk} & 0 \\ 0 & \varphi'_{mm} \end{bmatrix}.$$

From (14) it is known that the characteristic polynomials of $\Phi(km)$ and $\Phi'(km)$ are equal. The trace and determinant of each are therefore equal. It follows that $\varphi'_{kk} = \varphi_{kk} - c$ and $\varphi'_{mm} = \varphi_{mm} + c$; $c > 0$, or that the larger diagonal element is increased and that the smaller one is decreased.

The diagonalization of the submatrix $\Phi(km)$ within Φ can be effected by an orthogonal matrix Q which contains T in the appropriate submatrix position and identity matrix elements in the other positions:

$$\begin{aligned} q_{ij} &= t_{ij} \; ; \quad (i, j) = (k, k), (k, m), (m, k), (m, m) \\ q_{ij} &= \delta_{ij} \; ; \quad \text{other } (i, j). \end{aligned}$$

We then have $\Phi' = Q\Phi Q^t$ with only the elements in Φ' in rows and columns k and m changed from those in Φ. If Φ' is used to generate a new correlation matrix $\Phi'_{w|v} = \Phi_w^{\frac{1}{2}}\Phi'\Phi_w^{\frac{1}{2}}$, we have

$$\text{tr } \Phi'_{w|v} = \sum_{i=1}^{N} \sigma_i^2 \varphi'_{ii} = \sum_{i=1}^{N} \sigma_i^2 \varphi_{ii} - c(\sigma_k^2 - \sigma_m^2)$$

$$
\begin{aligned}
&= \operatorname{tr} \Phi_{w|} \; - c(\sigma_k^2 - \sigma_m^2) \\
&\leqq \operatorname{tr} \Phi_{w|y} \, , \qquad (17)
\end{aligned}
$$

which establishes the claim of this section. That is, any nondiagonal correlation matrix Φ with roots $\{\alpha_i^{-1}\}$ conjectured as providing a minimum trace correlation matrix $\Phi_w^{\frac{1}{2}}\Phi\Phi_w^{\frac{1}{2}} = \Phi_{w|y}$ can be improved upon by Φ'. The desired matrix for Φ is therefore diagonal and equal to

$$\Phi = [\alpha_i^{-1}\delta_{ij}] \qquad (18)$$

with the corresponding form of $\Phi_{w|y}$ equal to

$$\Phi_{w|y} = [\sigma_i^2\alpha_i^{-1}\delta_{ij}]. \qquad (19)$$

It follows that among all matrices $\Phi_{w|y}$ consistent with equation (16) with any given Φ_y , the one with minimum trace is diagonal.

An identical argument yields the symmetric conclusion. That is, for any specified $\Phi_{w|y}$ the matrix Φ_y with minimum trace among those consistent with equation (16) is also diagonal and equal to

$$\Phi_y = [\sigma_{ni}^2\alpha_i\delta_{ij}]. \qquad (20)$$

The argument assumes only that the noise variances are ordered $\sigma_{n1}^2 \leqq \sigma_{n2}^2 \leqq \cdots \leqq \sigma_{nN}^2$.

We can now state that the minimization of the trace of $\Phi_{w|y}$ over all pairs $\Phi_{w|y}$, Φ_y which satisfy equation (16) and the constraint equation (13) is obtained with a pair of diagonal matrices parametrically related as in equations (19) and (20). Any pair not so related can be altered, one matrix at a time, to decrease either the error (trace $\Phi_{w|y}$) or the received energy (trace Φ_y). Although we have worked with pairs $\Phi_{w|y}$, Φ_y consistent with equation (16) rather than the smaller set satisfying equations (11) and (12), the solution forms for $\Phi_{w|y}$ and Φ_y are still valid as they do satisfy these equations.

The modulator matrix which produces the correlation matrices $\Phi_{w|y}$ and Φ_y in the optimum form can be found from either equation (11) or (12) to be

$$A = \left[\frac{\sigma_{ni}}{\sigma_i}(\alpha_i - 1)^{\frac{1}{2}}\delta_{ij}\right]. \qquad (21)$$

Equations (12), (14), and (15) and the fact that $\Phi_n^{-\frac{1}{2}}A\Phi_wA^t\Phi_n^{-\frac{1}{2}}$ has nonnegative roots (it is a correlation matrix) can be used to show that $\alpha_i \geqq 1$, $i = 1, 2, \cdots , N$ which guarantees that the elements of A are real. It remains to solve for the set of roots $\{\alpha_i\}$ which provides the desired optimization.

3.2 *The Optimum Diagonal Modulator Matrix*

In terms of the set $\{\alpha_i\}$, the distortion which is to be minimized is given by

$$d = \operatorname{trace} \Phi_{w|y} = \sum_{i=1}^{N} \sigma_i^2 \alpha_i^{-1}$$

and the received energy constraint by

$$S_R = \operatorname{trace} \Phi_y = \sum_{i=1}^{N} \sigma_{ni}^2 \alpha_i \leqq S_0 + N_0 .$$

A further constraint is that $\alpha_i \geqq 1$, $i = 1, 2, \cdots, N$. As the set of permissible α_i's is a convex set and the functions $d(\alpha_i)$ and $S_R(\alpha_i)$ are convex functions, the Kuhn–Tucker theorem is applicable.[6] This states that at the point of minimization:

$$\begin{aligned} \frac{\partial}{\partial \alpha_i}\left[d + \frac{1}{\lambda^2} S_R\right] &= 0 \quad \text{if} \quad \alpha_i > 1 \\ &< 0 \quad \text{if} \quad \alpha_i = 1. \end{aligned}$$

Therefore we have

$$\begin{aligned} -\sigma_i^2 \alpha_i^{-2} + \frac{1}{\lambda^2} \sigma_{ni}^2 &= 0 \quad \text{if} \quad \alpha_i > 1 \\ &< 0 \quad \text{if} \quad \alpha_i = 1 \end{aligned}$$

or

$$\alpha_i = \max \left[\left(\frac{\sigma_i}{\lambda \sigma_{ni}}\right), 1\right] \cdot \tag{22}$$

It has already been observed that $\alpha_1 \geqq \alpha_2 \geqq \cdots \geqq \alpha_N$ and that $\alpha_i = 1$ corresponds to $a_{ii} = 0$ or no transmission of the ith source component. If we let N' denote the last α_i strictly greater than one we have the following solution for the optimum modulator matrix

$$A = \begin{bmatrix} \dfrac{\sigma_{ni}}{\sigma_i}\left(\dfrac{\sigma_i}{\lambda \sigma_{ni}} - 1\right)^{\frac{1}{2}} \delta_{ij} & 0 \\ 0 & 0 \end{bmatrix}; \qquad 1 \leqq i, j \leqq N'. \tag{23}$$

The solution for the distortion in the optimum linear system follows directly from equation (19):

$$d = \sum_{i=1}^{N'} \lambda \sigma_i \sigma_{ni} + \sum_{i=N'+1}^{N} \sigma_i^2 , \tag{24}$$

as does the solution for the total received energy from equation (20):

$$S_R = \sum_{i=1}^{N'} \frac{1}{\lambda} \sigma_i \sigma_{ni} + \sum_{i=N'+1}^{N} \sigma_{ni}^2 \, . \tag{25}$$

In these equations, the parameter λ is chosen to satisfy the constraint in equation (13) with equality. It should be remembered in the solution for λ that N' is a function of λ, being equal to the largest value of i for which $\sigma_i/\sigma_{ni} \geqq \lambda$. For completeness, we give the optimum demodulator matrix:

$$B = \begin{bmatrix} \lambda\left(\dfrac{\sigma_i}{\lambda\sigma_{ni}} - 1\right)^{\frac{1}{2}} \delta_{ij} & 0 \\ 0 & 0 \end{bmatrix}; \qquad 1 \leqq i, j \leqq N'. \tag{26}$$

IV. ELIMINATION OF THE ASSUMPTIONS

4.1 *A Source and Channel with Nonindependent Components*

We now consider systems in which Φ_w and Φ_n are not diagonal. Let P and R be the orthogonal matrices which respectively diagonalize these two correlation matrices, that is, $\Phi_{w'} = P\Phi_w P^t$ and $\Phi_{n'} = R\Phi_n R^t$ with $\Phi_{w'}$ and $\Phi_{n'}$ diagonal. Using the previous results, we can find the optimum modulator matrix A' in the primed system containing the correlation matrices $\Phi_{w'}$ and $\Phi_{n'}$. Now consider the use of the modulator matrix $A = R^t A' P$ in the system with Φ_w and Φ_n. From equation (6) and $\Phi_y = A\Phi_w A^t + \Phi_n$, it can be easily shown that using A' in the primed system and A in the unprimed system each produces the same distortion and uses the same energy. Consequently, A must be the optimum matrix in the unprimed system. If it is not, and A_0 is better, $A_0' = RA_0P^t$ would be a better choice than A' for modulator in the primed system contrary to A' being optimum.

4.2 *Nonequal Source and Channel Dimensionality*

When $N_s \neq N_c$, we can appropriately modify either the source or channel to restore the equality. For example, when $N_s < N_c$, $N_c - N_s$ source components of arbitrarily small variance, say ϵ, are added to the original source vector. The optimum modulator is then found as a function of ϵ by the previous method, and finally the limit taken as ϵ goes to zero. Similarly, when $N_c < N_s$, $N_s - N_c$, channel components of arbitrarily large noise variance, say $1/\epsilon$, are added to the original channel, the optimum modulator found, and the limit taken as ϵ goes to zero. We have seen that whenever either the source has a component with small variance or the channel has a component with large noise variance, the number of source components actually transmitted, N', is smaller than N. Since the optimum modulator matrix is diagonal, N' is also the number of channel components actually used. Therefore, the limiting modulator form in both of the above situations is attained for a nonzero value of ϵ, say ϵ_1. This modulator form is then optimum for all $\epsilon < \epsilon_1 \neq 0$.

V. COMPARISON OF OPTIMUM LINEAR AND NONLINEAR MODULATORS

In 1959 C. E. Shannon introduced a relation between d_R, the minimum attainable transmission distortion of a source, and R, the total information rate used in transmission.[2] This relation involves only the source statistics and the distortion measure in use. From it one is able to conclude that any channel with capacity R can be used to transmit the source with a transmission distortion arbitrarily close to d_R. One need only use a "sufficiently complex" encoder and decoder.

Another part of rate-distortion theory is the idea of a "test channel." Associated with each point on the rate-distortion curve, (d_R, R), is such a test channel which has the significance that among all channels that transmit the source at a rate equal to R, it provides the minimum transmission distortion d_R. Therefore, if there exist pre- and post-operators which can transform a given capacity R channel into the test channel for the source at (d_R, R), these operators must be optimum. An obvious necessary condition for this transformation, which is not always met, is that the capacity of the test channel at (d_R, R) be equal to R.

For a gaussian source with variance σ^2 and squared difference distortion, Shannon has found[2] both the rate distortion expression, $d_R = \sigma^2 e^{-2R}$ and the test channel:

$$\begin{tikzcd} w & \oplus \arrow[l] & \hat{w}. \arrow[l] \\ & n \arrow[u] & \end{tikzcd} \tag{27}$$

In this reverse channel, $\hat{w}$ and n are independent gauss variables with respective variances $\sigma^2 - d_R$ and d_R. It can be shown that this channel is identical to the forward channel:

$$\begin{tikzcd} w \arrow[r] & \otimes \arrow[r] & \oplus \arrow[r] & \hat{w} \\ & A_1 \arrow[u] & n \arrow[u] & \end{tikzcd} \tag{28}$$

with $A_1 = (\sigma^2 - d_R)/\sigma^2$, $\sigma_n^2 = A_1 d_R$, and the independence between w and n. A similar form is given by Gallager in Ref. 7. Still another form of the test channel is:

$$\begin{tikzcd} w \arrow[r] & \otimes \arrow[r] & \oplus \arrow[r] & \otimes \arrow[r] & \hat{w} \\ & A \arrow[u] & n \arrow[u] & B \arrow[u] & \end{tikzcd} \tag{29}$$

with $A^2 = (\sigma^2 - d_R)\sigma_n^2/\sigma^2 d_R$, $B^2 = (\sigma^2 - d_R)d_R/\sigma^2\sigma_n^2$, and n *any* given additive gaussian noise.

Now consider a single dimensional gaussian channel of capacity R. Since the received energy S_R is accordingly restricted to $\sigma_n^2 \exp(2R)$, we have from equations (23) through (26) that the optimum linear operators are

$$a_{11}^2 = \frac{\sigma_n^2}{\sigma^2}\left(\frac{\sigma}{\lambda\sigma_n} - 1\right) = \frac{\sigma_n^2}{\sigma^2}\left(\frac{\sigma^2 - d}{d}\right)$$

$$b_{11}^2 = \lambda^2\left(\frac{\sigma}{\lambda\sigma_n} - 1\right) = \frac{d(\sigma^2 - d)}{\sigma^2\sigma_n^2}$$

$$\lambda = \frac{d}{\sigma\sigma_n} = \frac{\sigma\sigma_n}{S_R}.$$

Note that the distortion d equals $\sigma^2 \exp(2R)$, and that a_{11} and b_{11} agree precisely with the test channel parameters in (29). Therefore, we can conclude that in this case the operators in equations (23) and (26) are optimum, even outside the linear class.

The rate-distortion curve and the test channel for gaussian vector sources can also be found from Shannon's results. The results for the N-dimensional source with variances $\sigma_1^2, \sigma_2^2, \cdots, \sigma_N^2$ are (we continue to assume that $\sigma_1^2 \geqq \sigma_2^2 \geqq \cdots \geqq \sigma_N^2$):

$$d_R = N\left\{e^{-2R}\prod_{i=1}^{N}\sigma_i^2\right\}^{1/N}; \qquad 0 \leqq d_R \leqq N\sigma_N^2$$

$$= \sigma_N^2 + (N-1)\left\{e^{-2R}\prod_{i=1}^{N-1}\sigma_i^2\right\}^{1/N-1};$$

$$N\sigma_N^2 \leqq d_R \leqq \sigma_N^2 + (N-1)\sigma_{N-1}^2$$

$$= \sigma_N^2 + \sigma_{N-1}^2 + (N-2)\left\{e^{-2R}\prod_{i=1}^{N-2}\sigma_i^2\right\}^{1/N-2};$$

$$\sigma_N^2 + (N-1)\sigma_{N-1}^2 \leqq d_R \leqq \sigma_N^2 + \sigma_{N-1}^2 + (N-2)\sigma_{N-2}^2$$

$$\vdots$$

$$= \sigma_N^2 + \cdots + \sigma_2^2 + \sigma_1^2 e^{-2R};$$

$$\sigma_N^2 + \cdots + 2\sigma_2^2 \leqq d_R \leqq \sigma_N^2 + \cdots + \sigma_1^2$$

$$= \sigma_N^2 + \cdots + \sigma_1^2 \, ; \qquad R = 0. \tag{30}$$

This expression can also be applied to a gaussian vector source with correlated components if the variances σ_i^2 are interpreted as those in the diagonalized correlation matrix $\Phi_{w'} = P\Phi_w P^t$. The test channel for $N > 1$ is the product of elementary test channels given in (29) with

$$A = A_1 , A_2 , \cdots , A_N ,$$

$$A_i^2 = \frac{(\sigma_i^2 - d_i)}{\sigma_i^2 d_i} \sigma_{ni}^2 ,$$

$$d_i = \min (\sigma_i^2 , d_R),$$

$$\sigma_n^2 = \sigma_{n1}^2 , \sigma_{n2}^2 , \cdots , \sigma_{nN}^2 = \text{any noise vector.}$$

Let us now presume that the vector channel provided for use has the additive noise variances given by the vector σ_n^2 and is constrained to have an output energy level equal to S_R . This equivalently specifies the channel capacity as

$$R = \max_{S_{Ri}} \sum_{i=1}^{N} \frac{1}{2} \log \frac{S_{Ri}}{\sigma_{ni}^2} \tag{31}$$

with

$$S_{Ri} = \max (S, \sigma_{ni}^2)$$

and S adjusted to have $\sum S_{Ri} = S_R$. The comparison between the minimum attainable transmission distortion using linear transmitter and receiver operators (equation 24) and using unrestricted transmitter and receiver operators (equation 30) now reveals that contrary to the single dimension case, when $N > 1$ the linear operators are *not,* in general, optimum. The only exception is when both the vectors σ^2 and σ_n^2 are uniform. Some intuition as to why the single and multidimensional cases are different might be provided by the following.

The test channel at (d_R , R), for example, the one including the noise vector σ_n^2 in its form, is a result of a minimization of mutual information under a distortion constraint. It does not, therefore, necessarily divide the total energy presented to the gaussian vector channel in a way which uses this channel to capacity. Since this channel, by definition, transmits information at a rate equal to R, its total capacity is (except for the special case noted previously) strictly greater than R. Consequently, when the same additive noise channel, σ_n^2, is to be used for transmission but is stated to have a capacity of only R, it cannot be transformed into the test channel by any pre- and postoperators.

The impossiblity of such a transformation can also be observed by noting that the allowed total input energy on the given capacity R channel is restricted to a lower level than present on the test channel The uniqueness of the test channel, which is formed with linear operators, and the continuity of both the mutual information and distortion with the modulator matrix then precludes the possibility of attaining the test channel's performance with the given capacity R channel and linear operators.

One could argue that the comparison to this point is not fair in that Shannon allows modulators and demodulators that operate on blocks of letters, whereas the results in equations (23), (24), and (25) were derived using a coding block length of one. However, the previous results show that the optimum linear modulator does not mix independent source components before presentation to the channel, assuming the channel has already been rotated in N-space so as to have independent noise components. Neither does it cross-couple sets of source components having no cross dependence when presentation is to a channel with sets of noise components of equal respective dimensionalities also having no cross dependence. Therefore, if successive source and channel (vector) events are independent, and their dimensionalities filled out to be equal by adding either zero variance source components or infinite variance noise components, there is no memory introduced by the optimum linear modulator among elements of the encoded block. The consequence is that the distortion and the energy are only scaled by the block length in use.

VI. AN EXAMPLE

We cite here just one example which shows that at least in many cases the performance of the optimum linear modulator- demodulator pair compares favorably with that theoretically obtainable with more complex operators. We take $\sigma_1 = \sigma_2 = 1$, $\sigma_{n1} = a$, $\sigma_{n2} = ae^{2\varphi}$ and use a and φ as parameters that generate a set of different channels. To better compare the two performances, we fix the channel capacity at C which in turn fixes Shannon's minimum attainable distortion at $d_C = 2e^{-C}$. The total allowed received energy is thus specified according to equation (31).

Upon solution for λ and d in equations (24) and (25) we have the following expression for the ratio between the distortion obtainable with linear operators and that theoretically attainable:

$$\frac{d(\varphi)}{d_C} = \begin{cases} \cosh^2 \varphi & ; \quad 0 \leqq \varphi \leqq \tfrac{1}{2}C \\ \dfrac{\cosh^2 \varphi}{\cosh(2\varphi - C)} & ; \quad \tfrac{1}{2}C \leqq \varphi \leqq C \\ \cosh C & ; \quad C \leqq \varphi . \end{cases}$$

We illustrate this function for several different values of capacity in Fig. 2. At $\varphi = 0$ (where both the vectors σ^2 and σ_n^2 are uniform) we see that $d(0) = d_C$ indicating the optimality of the linear modulator and demodulator for this case. Using a term introduced in Ref. 8, we can therefore say that when $\varphi = 0$ the source and channel are "matched." As φ increases, the source-channel mismatch increases and the nonoptimality of linear operators also increases. As the figure illustrates, the nonoptimality ratio, $d(\varphi)/d_C$, can be quite large when both the channel capacity is high and the additive noise vector is highly skewed in variance. However, over a significant region of interest,

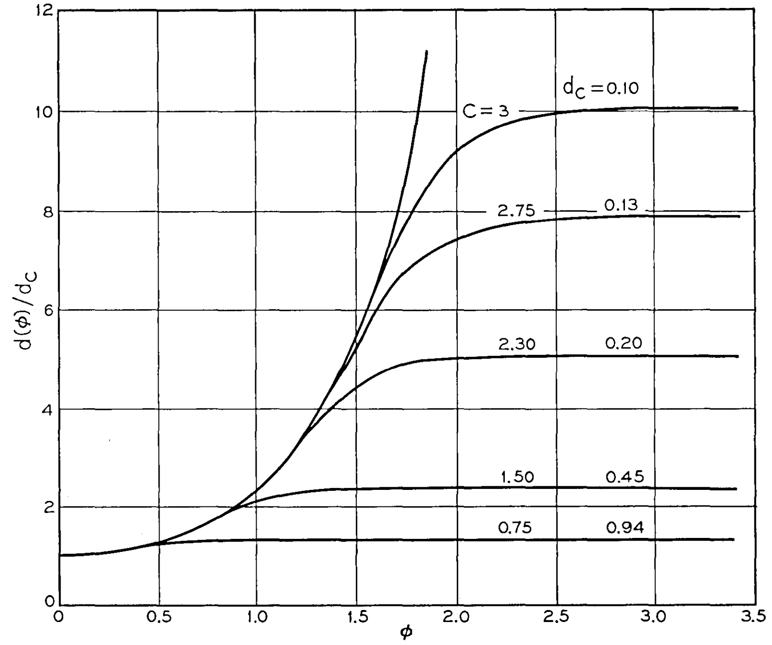


Fig. 2 — The linear system nonoptimality for $N = 2, \sigma_1 = \sigma_2 = 1, \sigma_{n1} = 1, \sigma_{n2} = \exp 2\phi$.

$\varphi \leqq 1$ (reflecting a noise component variance ratio of about 50), the nonoptimality ratio is small.

VII. SUMMARY

In this paper we have derived the optimum linear modulator and demodulator for the transmission of a gaussian vector source through an additive gaussian vector channel. It was found that when both the source and channel components are independent, both the modulator and demodulator matrices are diagonal. This specifies the separate amplification, transmission, and decoding of each source component. When both the source and channel components are correlated, the optimum modulator matrix was found to be the cascade of three matrices: (i) the orthogonal matrix which diagonalizes the source correlation matrix, (*ii*) the optimum modulator matrix which transmits this newly formed independent component source over the independent component additive noise channel which is formed by (*iii*) the orthogonal transformation matrix that diagonalizes the noise correlation matrix. We have found that in general the best linear system does not provide a distortion as small as that stated by Shannon to be attainable with a channel of the same capacity. The only exception is when both the source and channel noise variance vectors are uniform. The nonoptimality of linear modulators and demodulators can be quite large in some cases but, in many other situations, can be small enough to justify the use of these very simple operators.

REFERENCES

1. Fano, R. M., *Transmission of Information*, New York: John Wiley, 1961.
2. Shannon, C. E., "Coding Theorems for a Discrete Source with a Fidelity Criterion," IRE Nat. Conv. Record, Part 4 (1959), pp. 142–163.
3. Shannon, C. E., "Communication in the Presence of Noise," Proc. IRE, *37*, No. 1 (January 1949), pp. 10–21.
4. Wozencraft, J. M., and Jacobs, I. M., *Principles of Communication*, New York: John Wiley, (1965).
5. Hoffman, K., and Kunze, R., *Linear Algebra*, Englewood Cliffs, New Jersey: Prentice-Hall, 1961.
6. Kuhn, H. W., and Tucker, A. W., "Nonlinear Programming," Proc. 2nd Berkeley Symp. Math. Stat. and Prob., 1951, pp. 481–492.
7. Gallager, R. G., *Information Theory and Reliable Communication*, New York: John Wiley, 1968.
8. Pilc, R. J., "Coding Theorems for Discrete Source-Channel Pairs," Ph.D. thesis, Department of Electrical Engineering, M.I.T., Cambridge, Massachusetts, (February 1967).

Communication Systems Which Minimize Coding Noise

By BROCKWAY McMILLAN

(Manuscript received May 28, 1969)

The problem of minimizing coding or quantizing noise in a communication system is posed in a general setting. It is shown that if the messages to be transmitted are sample sequences drawn from a discrete-time random process meeting a certain simply stated criterion of "randomness" and if there exists a quantized communication system which is optimal in that it introduces a minimum amount of coding noise, then this optimal system can be realized using a transmitter of special form. Specifically, the optimum transmitter is one which quantizes each message sample according to a scheme that depends only upon the quantized material already transmitted, rather than upon the (unquantized) material that has been previously offered for transmission. It follows that only digital storage is required at the transmitter or receiver. If the receiver is limited, a priori, *to have only a given finite amount of storage, and if the system is optimum within this constraint, the transmitter need have only the same amount of storage.*

I. INTRODUCTION: THE MODEL

Shannon's theory of communication, shows how to defeat noise introduced in a communication medium by restricting the repertoire of transmitted signals to a discrete set.[1] If the messages to be transmitted are not already in an appropriately discrete form, noise in the medium is then eliminated only at the expense of noise, here called coding noise, caused by the failure of the restricted family of available signals to represent faithfully the full family of possible messages. The amount of coding noise introduced is of course subject to control by design.

This paper considers one aspect of the problem of minimizing coding noise. Noise in the medium is not considered. The paper limits attention to systems in which the random process representing the message is a discrete-time or sampled-data process. The sampling noise caused by creating such a process out of a continuous-time process is not considered.

The problem of selecting a coding scheme that maximizes the rate of communication over a noisy channel is not considered. Rather, the paper starts at the point that a coding scheme has been found, that is optimum according a fairly general criterion of fidelity. What is then shown is that the transmitter and receiver—encoder and decoder—of the system are of a special form.

A Q-coded communication system is defined by a discrete set Q and by three jointly distributed random processes, $\{x_n , q_n , y_n \mid n = 0, \pm 1, \pm 2, \cdots\}$. For purposes of this paper, the set Q will be either

(*i*) the set $\{1, 2, \cdots , M\}$, where M is a given positive integer > 1, or
(*ii*) the set $\{1, 2, 3, \cdots\}$ of all positive integers.

The process $\{x_n\}$ represents periodic samples derived from the message offered for transmission, each x_n is a real random variable. $\{q_n\}$ represents the transmitted signals; for each n, q_n is a random variable, taking values from the set Q and measurable on the sample space of $\{x_n , x_{n-1} , x_{n-2} , \cdots\}$. That is, for each n, the value of the integer variable q_n depends only upon, and is determined (apart perhaps from events of probability zero) by the present and past of the message. $\{y_n\}$ represents the version of the message reconstructed at the receiver; for each n, y_n is a real random variable measurable on the sample space of $\{q_n , q_{n-1} , q_{n-2} , \cdots\}$. Therefore for each n, y_n depends only upon, and is determined (apart perhaps from events of probability zero) by the present and past of the transmitted signal.

The model at this point is very general. It provides that at each time, n a discrete valued random variable q_n be generated in some way out of the material $\{x_n , x_{n-1} , x_{n-2} , \cdots\}$ then available from the message process, and that subsequently at the receiver a y_n be generated out of the material $\{q_n , q_{n-1} , \cdots\}$ there currently available. If all three processes $\{x_n , q_n , y_n\}$ are stationary we can call the system stationary. The question of stationarity does not enter in what follows.

What remains to be specified in this model is that in some sense the process $\{y_n\}$ is to represent the process $\{x_n\}$. At the start it appears natural to consider three cases; it develops that two are simply special cases of the third, one of them not interesting in the framework of this paper.

We start with a given sequence $\{\psi_n \mid n = 0, \pm 1, \pm 2, \cdots\}$ of functions, in which each ψ_n is a real valued Borel measurable function $\psi_n(x, y)$ of the real variables x, y. The use of a sequence $\{\psi_n\}$ here is a largely decorative generality that costs nothing. The conventional case is that in which all ψ_n are the same function ψ. These functions define a fidelity criterion as follows:

Case (*i*), the *delay-free case*:

Here we choose to regard y_n as a replica of x_n, and evaluate our communication system at each time n by the quantity

$$E\{\psi_n(x_n , y_n)\}, \tag{1}$$

where E denotes expectation over the message ensemble.

Case (*ii*), the case of *fixed delay*:

Here we are given a fixed integer $d \geqq 0$ and we choose to regard y_n as a replica of x_{n-d}, thus allowing q_n to take advantage not only of $\{x_{n-d}, x_{n-d-1}, \cdots\}$ (the present and past of x_{n-d}) but also of $\{x_n, x_{n-1}, \cdots, x_{n-d+1}\}$ (a limited span of the "future" of x_{n-d}) in representing x_{n-d}. Here the criterion relative to x_{n-d} is (by a convention we will use with respect to indices)

$$E\{\psi_n(x_{n-d} , y_n)\}. \tag{2}$$

If $d = 0$, this case reduces to case *i*.

Case (*iii*), *block encoding* with cycle time c:

This is the situation that arises naturally in Shannon's theory. We are given a fixed integer $c \geqq 1$, and the transmission process is repetitive with a cycle of length c. By a choice of time origin, we can describe it as follows. Let Q_1 be a discrete set with $M_1 < \infty$ members. At time 0 the transmitter examines $\{x_0, x_{-1}, \cdots\}$ and generates a Q_1-discrete variable which we shall call $\hat{q}_0$. At time c, the transmitter then examines $\{x_c, x_{c-1}, \cdots\}$ and produces $\hat{q}_1$; the process repeats with period c. For transmission, the random variable $\hat{q}_0$ is encoded into the string $\{q_c, q_{c-1}, \cdots, q_1\}$ of random variables each being Q-discrete, where $M^c \geqq M_1$. At time c, all of $\hat{q}_0, \hat{q}_{-1}, \cdots$ are available at the receiver, being represented by the sequence $\{q_c, q_{c-1}, q_{c-2}, \cdots\}$. From these, the sequence $\{y_{2c-1}, y_{2c-2}, \cdots, y_c\}$ is generated, representing $x_0, x_{-1}, \cdots, x_{-c+1}$, respectively. We think of these y's as being presented to the output of the receiver in the order of their indices, y_c at time c, and so on.

If one follows through the functional dependencies here, he sees that indeed the processes $\{x_n, q_n, y_n\}$ are so related that each q_n depends at most upon $\{x_n, x_{n-1}, \cdots\}$, and each y_n at most upon $\{q_n, q_{n-1}, \cdots\}$. Indeed, except at times which recur with period c, q_n is not "up to date," depending in fact only on x's strictly prior to x_n. Similarly, y_n is only periodically up to date; at other times it depends only upon q's that are actually earlier than q_n.

In the situation as just described, the criterion of fidelity becomes $E\{\psi_n(x_{n-2c+1}, y_n)\}$. Case *iii* is then also a special case of case *ii*, in which

$d = 2c - 1 \geqq 1$. What makes it special is that in case *ii*, q_n and y_n are permitted to be up to date at each value of n, however in case *iii* the block coding process restricts the currency of the data upon which most of the q's and y's depend.

Actually, case *iii* as just described will turn out not to be covered, in general, by the theorems to be proved. This happens because, as is later be stated more precisely, we are interested only in communication systems that minimize (2) for each n, in comparison with all possible competing systems. Clearly, to impose the restrictions immanent in case *iii* upon one's reportoire of coding schemes limits the domain within which a minimum is to be sought. The system that brings about an absolute minimum is simply not, in general, to be found in this restricted domain.

The previous observation is not to be entered as a criticism of Shannon's theory. Typically, in a noisy medium, it is necessary to use a highly redundant encoding $\{q_c, q_{c-1}, \cdots, q_1\}$ to represent $\hat{q}_0$, so that the inefficiencies (as measured by expression 2) that are imposed by the block-coding process are needed in order to ensure that the y_n in (2) is an approximately error free replica of x_{n-d}. We must remember that (2) measures the noise introduced by the coding process, not by the noisy medium. It is interesting to a designer only if the latter noise has been eliminated. The price of this elimination is that one may not be able to minimize (2) in competition with systems that are not restricted to be of block coding form.

A true engineering solution to the problems reflected in the remarks immediately above would consider (2) in which the expectation is taken over the joint ensemble of message and noise. The solution should balance coding noise against channel noise at, say, a fixed delay, to minimize (2). This paper is very far from solving such a problem.

It does not follow that the results of this paper are without interest in the search for coding schemes to eliminate noise. Given a Q-coded communication system which does minimize (2), the $\{q_n\}$ process is in digital form. This $\{q_n\}$ process can then be redundantly encoded according to Shannon's theory, and recovered with few errors (and typically much delay) at the receiver. The $\{y_n\}$ process then results (perhaps delayed) and has few errors. Then (2) does measure the total amount of noise introduced in this operation.

II. STATEMENT OF RESULTS

Given the message process $\{x_n\}$, the sequence $\{\psi_n\}$, and the delay $d \geqq 0$, a Q-coded communication system $\{x_n, q_n, y_n\}$ will be called

$\{\psi_n, d\}$-optimal if

(*i*) For each $n = 0, \pm 1, \pm 2, \cdots$

$$E\{|\psi_n(x_{n-d}, y_n)|\} < \infty, \tag{3}$$

and

(*ii*) For any other Q-coded communication system $\{x_n, q_n', y_n'\}$,

$$E\{\psi_n(x_{n-d}, y_n)\} \leqq E\{\psi_n(x_{n-d}', y_n')\}, \tag{4}$$

for each $n = 0, \pm 1, \pm 2, \cdots$.

The simplest result of this paper is of such a form as to illustrate the nature of all of the results. We define a class K of functions ψ, and a class, here called *CCD,* of message processes $\{x_n\}$, such that the following theorem is true.

Theorem 1: Let $\{\mathrm{x_n}, \mathrm{q_n}, \mathrm{y_n}\}$ *be a given* Q*-coded communication system that is* $\{\psi_\mathrm{n}, 0\}$*-optimal. If each* $\psi_\mathrm{n} \varepsilon \mathrm{K}$, $\mathrm{n} = 0, \pm 1, \pm 2, \cdots$, *and if* $\{\mathrm{x_n}\} \varepsilon$ CCD, *then each* $\mathrm{q_n}$ *is equal with probability one to a random variable measurable on the sample space of* $\{\mathrm{x_n}, \mathrm{q_{n-1}}, \mathrm{q_{n-2}}, \cdots\}$.

The force of this theorem is that it simplifies, in principle at least, the requirements for memory at the transmitter. Only the digital sequence $\{q_{n-1}, q_{n-2}, \cdots\}$ need be in storage at time n. The proof of the theorem will also develop a standard structure for the optimum transmitter difficult to summarize easily in a theorem.

The definition of the class K is long and is deferred to Section III. Suffice it here to say that K is a large class that includes the conventional

$$\psi^1(x, y) = |x - y|, \qquad \psi^2(x, y) = (x - y)^2,$$

and any other continuous strictly increasing function of ψ^1.

We define *CCD,* and a related class *CCDf,* thus:

CCD consists of those processes $\{x_n\}$ such that: for each $n = 0, \pm 1, \pm 2, \cdots$, if z is a random variable measurable on the sample space of $\{x_{n-1}, x_{n-2}, \cdots\}$, then the probability that $z = x_n$ is zero:

$$P\{z = x_n\} = 0. \tag{5}$$

CCDf consists of those processes $\{x_n\}$ such that: for each $n = 0, \pm 1, \pm 2, \cdots$, if **A** is a finite Borel field or the completion of a finite Borel field, and if z is a random variable measurable on the smallest Borel field containing **A** and the sample space of $\{x_{n-1}, x_{n-2}, \cdots\}$, then (5) holds.

Read *CCD* as "continuous conditional distribution." If $\{x_n\} \varepsilon$ *CCD* and if x_n has a conditional distribution given $\{x_{n-1}, x_{n-2}, \cdots\}$, that distribution must be continuous.

We now define a more restricted class of Q-coded communication systems and a corresponding notion of optimality.

Given an integer $m \geqq 0$, a Q-coded communication system $\{x_n, q_n, y_n\}$ will be said to have *decoder memory span m* if for each $n = 0, \pm 1, \pm 2, \cdots$ y_n is measurable on the sample space of $\{q_n, q_{n-1}, \cdots, q_{n-m}\}$.

A Q-coded communication system $\{x_n, q_n, y_n\}$ will be called $\{\psi_n, d, m\}$-optimal if it has decoder memory span m, if (3) holds for every n, and if (4) holds for every n and for every $\{x_n, q_n', y_n'\}$ which has decoder memory span m.

In the case of $\{\psi_n, d, m\}$ optimality, then, the competition is restricted to systems with decoder memory span m. We can put $m = \infty$ to refer to the case of $\{\psi_n, d\}$ optimality defined earlier.

Perhaps our most surprising result is that case *ii* of our model, which includes case *i* as a special case, is also included in case *i*. This is shown by Theorem 2.

Theorem 2: Let $\{x_n, q_n, y_n\}$ *be a given* Q*-coded communication system that is* $\{\psi_n, d\}$*-optimal. If each* $\psi_n \varepsilon K$, $n = 0, \pm 1, \pm 2, \cdots$, *if* M, *the number of elements of Q, is finite, and if* $\{x_n\} \varepsilon$ *CCDf, then each* q_n *is equal with probability one to a random variable measurable on the sample space of* $\{x_{n-d}, q_{n-1}, q_{n-2}, \cdots\}$. *Furthermore, the system* $\{x_n, q_n', y_n'\}$, *where*

$$\begin{aligned} q_n' &= q_{n+d}, \\ y_n' &= y_{n+d}, \end{aligned} \qquad n = 0, \pm 1, \pm 2, \cdots, \tag{6}$$

is a Q-coded communication system that is $\{\psi_n', 0\}$ *optimal, where*

$$\psi_n' = \psi_{n+d}, \qquad n = 0, \pm 1, \pm 2, \cdots. \tag{7}$$

Finally, we state a theorem that includes the two preceding ones.

Theorem 3: Let $\{x_n, q_n, y_n\}$ *be a given* Q*-coded communication system that is* $\{\psi_n, d, m\}$*-optimal. If each* $\psi_n \varepsilon K$, $n = 0, \pm 1, \pm 2, \cdots$, *if* $M < \infty$, *and if* $\{x_n\} \varepsilon$ *CCDf, then each* q_n *is equal with probability one to a random variable measurable on the sample space of* $\{x_{n-d}, q_{n-1}, \cdots, q_{n-m}\}$ ($\{x_{n-d}\}$ *if* $m = 0$). *The system as defined by* (6) *is a* Q*-coded communication system with decoder memory span m that is* $\{\psi_n', 0, m\}$*-optimal, where* ψ_n' *is given by* (7). *If, in the initial hypotheses,* $d = 0$, *then it suffices that* $\{x_n\} \varepsilon$ *CCD and the restriction* $M < \infty$ *may be removed. If* $m < \infty$, *the hypothesis* $\{x_r\} \varepsilon$ *CCDf may be replaced by:*

For each n $= 0, \pm 1, \pm 2, \cdots$, *if* z *is a random variable that takes only finitely many values, then* $\mathrm{P}\{\mathrm{x_n} = \mathrm{z}\} = 0$.

Theorem 1 shows the basic facts about measurability in the present context. Theorem 2 adds the fact that delay $d > 0$ gains no advantage (since the "future" of x_{n-d} is not known at the receiver, even if it is at the transmitter). Finally, Theorem 3 includes these facts and shows that a limitation on the memory span of the receiver allows a corresponding simplification of the transmitter.

In the proofs of these theorems it is seen that they are true for classes of process slightly larger than *CCD* or *CCDf*. In particular, the final conclusion of Theorem 3 opens the case of finite memory span to any process $\{x_n\}$ that has a little additive nonsingular Gaussian noise in each sample.

III. THE CLASS K

The class K of cost functions allowed by these theorems can be very general. The definition below seems more inclusive than is called for by the applications I can think of; at the cost of elaboration, it can be enlarged further.

We let K be the class of all functions $\psi(x, y)$ of two real variables x, y with the following properties.

(*i*) $\psi(x, y)$ is continuous;
(*ii*) for all x, y, $\psi(x, y) \geqq 0$;
(*iii*) for all x, $\psi(x, x) = 0$;
(*iv*) for each y, there are at most countably many solutions x to the equation

$$\psi(x, y) = 0, \tag{8}$$

in the sense that: there exist Borel measurable functions $g_k(y)$, $k = 1, 2, 3, \cdots$, such that if (8) holds, then for some k, $x = g_k(y)$.

v) If $y_1 \neq y_2$, there are at most countably many solutions to the equation

$$\psi(x, y_1) = \psi(x, y_2), \tag{9}$$

in the sense that: there exist Borel measurable functions $f_k(y, z)$, $k = 1, 2, 3, \cdots$ such that if (9) holds and if $y_1 \neq y_2$, then for some k, $x = f_k(y_1, y_2)$.

It follows from this definition that $\psi^1 \varepsilon K$, where $\psi^1(x, y) = | x - y |$. Then also $\psi^2 \varepsilon K$, where $\psi^2(x, y) = (x - y)^2$. Similarly any other continuous strictly monotone function of ψ^1 is also in K. In all of these instances, (8) has the unique solution $y = x$, and (9) has a unique solution given by $2x = y_1 + y_2$.

IV. PROOFS

Let $\{\Omega, \mathbf{B}, P\}$ be a probability space: A set Ω of points ω, a Borel field $\mathbf{B}$ of subsets of Ω, and a probability measure P on $\mathbf{B}$ with respect to which $\mathbf{B}$ is complete. This probability space is assumed given and fixed.

A random variable x is a real-valued function $x(\omega)$ defined on Ω and measurable $\mathbf{B}$.

If $\mathbf{F} \subseteq \mathbf{B}$ is a Borel field, a random variable x is said to be *essentially measurable* $\mathbf{F}$ if x is equal with probability one to a random variable x' which is measurable $\mathbf{F}$. If $\mathbf{F}$ is complete, such an x is then itself measurable $\mathbf{F}$.

If $\mathbf{F} \subseteq \mathbf{B}$ is a Borel field and x a random variable, $\{x\} \vee \mathbf{F}$ denotes the smallest Borel field such that: x is measurable $\{x\} \vee \mathbf{F}$ and $\mathbf{F} \subseteq \{x\} \vee \mathbf{F}$.

A random variable taking its values in the set Q will be called Q-discrete.

Denote by $[x \mid q, \mathbf{F} \mid y, \mathbf{G}]$ a mathematical object of the following kind:

x is a random variable,

q is a Q-discrete random variable,

$\mathbf{F}$ is a Borel field, $\mathbf{F} \subseteq \mathbf{B}$, and q is essentially measurable on the field determined by $\mathbf{F}$ and the sample space of x,

y is a random variable,

$\mathbf{G}$ is a Borel field, $\mathbf{G} \subseteq \{x\} \vee \mathbf{F}$, and y is essentially measurable on the field determined by $\mathbf{G}$ and the sample space of q.

For convenience let $CQAx$ ("conditionally quantized approximation to x") denote the class of all objects of the kind described, based on the given probability space $\{\Omega, \mathbf{B}, P\}$, the given x, and the given set Q.

Given a Q-coded communication system $\{x_n, q_n, y_n\}$, given a delay d and a memory span m, let $\mathbf{X}_{n,d}$ be the sample space of the selection $\{x_n, x_{n-1}, \cdots\}$ of random variables from which the specific variable x_{n-d} has been deleted. Let $\mathbf{Q}_{n,m}$ be the sample space of the random variables $\{q_{n-1}, q_{n-2}, \cdots, q_{n-m}\}$. Then it is easy to see that $\{x_n, q_n, y_n\}$ is a Q-coded communication system with decoder memory span m if and only if for each $n = 0, \pm 1, \pm 2, \cdots$

$$[x_{n-d} \mid q_n, \mathbf{X}_{n,d} \mid y_n, \mathbf{Q}_{n,m}] \; \varepsilon \; CQAx_{n-d} .$$

Given ψ, a $[x \mid q, \mathbf{F} \mid y, \mathbf{G}] \in CQAx$ will be called weakly ψ-optimal if:

(*i*) $E\{\mid \psi(x, y) \mid\} < \infty$,

(*ii*) If random variables q' and y' are such that $[x \mid q', \mathbf{F} \mid y', \mathbf{G}] \in CQAx$, then $E\{\psi(x, y)\} \leqq E\{\psi(x, y')\}$.

The qualifier "weakly" in this definition signals the fact that the fields **F** and **G** are not allowed to vary in the competition for optimality.

Lemma 1: If $\{\mathrm{x_n}, \mathrm{q_n}, \mathrm{y_n}\}$ *is a* Q*-coded communication system with decoder memory span* m, *and if* $\{\mathrm{x_n}, \mathrm{q_n}, \mathrm{y_n}\}$ *is* $\{\psi_{\mathrm{n}}, \mathrm{d}, \mathrm{m}\}$*-optimal, then for each* n $[\mathrm{x_{n-d}} \mid \mathrm{q_n}, \mathbf{X}_{\mathrm{n,d}} \mid \mathrm{y_n}, \mathbf{Q}_{\mathrm{n,m}}]$ *is weakly* ψ_{n}*-optimal.*

Proof: Fix an n; for convenience identify it as $n = 0$. Suppose that we are given random variables q' and y', which we shall here call q_0' and y_0', such that

$$[x_{-d} \mid q_0', \mathbf{X}_{0,d} \mid y_0', \mathbf{Q}_{0,m}] \in CQAx_d .$$

Define a new Q-coded communication system $\{x_n, q_n', y_n'\}$ thus:

For $n < 0$, $q_n' = q_n$, $y_n' = y_n$;

For $n = 0$, q_0' and y_0' are those above;

For $n > 0$, $q_n' = 1$ and $y_n' = 0$.

That this is a Q-coded communication system with decoder memory span m follows at once from the definitions. Furthermore, the sample space of $\{q_{-1}', q_{-2}', \cdots q_m'\}$ is $\mathbf{Q}_{0,m}$. Because $\{x_n, q_n, y_n\}$ is $\{\psi_n, d, m\}$-optimal, we conclude that $E\{\mid \psi_0(x_{-d}, y_0) \mid\} < \infty$ and that $E\{\psi_0(x_{-d}, y_0)\} \leqq E\{\psi_0(x_{-d}, y_0')\}$.

These, however, prove that $[x_{-d} \mid q_0, \mathbf{X}_{0,d} \mid y_0, \mathbf{Q}_{0,m}]$ is weakly ψ_0-optimal. Clearly this proof can be repeated for any other value of n.

The proof of this lemma indicates, deliberately, the force of the notion of $\{\psi_n, d, m\}$-optimality for $\{x_n, q_n, y_n\}$. The competing communication system $\{x_n, q_n', y_n'\}$ used in the proof sacrificed all reasonable behavior for $n > 0$, yet was still allowed to compete at $n = 0$. In particular, notice that even if $\{x_n, q_n, y_n\}$ is stationary, it must compete with nonstationary systems designed to excel at only one value of n. The theorems of Section II are not proved for stationary systems which are known only to minimize each $E\{\psi_n(x_{n-d}, y_n)\}$ against competing systems drawn from the class of stationary systems.

Given a Borel field $\mathbf{G} \subset \mathbf{B}$, we define $CCD(\mathbf{G})$ analogously to CCD: $CCD(\mathbf{G})$ is the class of all random variables x such that:

If z is a random variable measurable $\mathbf{G}$, then $P\{x = z\} = 0$.

The results of this paper all derive from Theorem 4.

Theorem 4: Let $[\mathrm{x} \mid \mathrm{q}, \mathbf{F} \mid \mathrm{y}, \mathbf{G}] \varepsilon \mathrm{CQA}x$ *and suppose that it is weakly* ψ*-optimal. If* Q *is a finite set, or if* ψ *is Borel measurable and for each* x *is bounded from below, then there exists a* Q*-discrete random variable* q' *and a random variable* y' *such that*

(*i*) $[\mathrm{x} \mid \mathrm{q}', \mathbf{G} \mid \mathrm{y}', \mathbf{G}] \varepsilon \mathrm{CQAx}$,

(*ii*) $\psi(\mathrm{x}, \mathrm{y}') = \psi(\mathrm{x}, \mathrm{y})$ *with probability one. In particular, also, the object i is weakly* ψ*-optimal.*

If $\psi \varepsilon \mathrm{K}$ *and* $\mathrm{x} \varepsilon \mathrm{CCD}(\mathrm{G})$ *then also*

(*iii*) $\mathrm{q}' = \mathrm{q}$ *with probability one, and*

(*iv*) $\mathrm{y}' = \mathrm{y}$ *with probability one.*

It then follows that the given q *is essentially measurable on the Borel field* $\{\mathrm{x}\} \mathrm{v}\mathbf{G}$, *determined by* **G** *and the sample space of x.*

We wish to use the given $[x \mid q, \mathbf{F} \mid y, \mathbf{G}]$ as a model for some

$$[x_{n-d} \mid q_n , \mathbf{X}_{n,d} \mid y_n , \mathbf{Q}_{n,m}]$$

in a *Q*-coded communication system. Conclusions *i* and *ii* show that for any given *n* we can find a q_n' essentially measurable $\{x_{n-d}\} \vee \mathbf{Q}_{n,m}$ and a y_n' such that, according to the criterion defined by ψ, y_n' represents x_{n-d} as well as y_n did. Without conclusion *iii*, however, the substitution of q_n' for q_n can alter the subsequent Borel fields $\mathbf{Q}_{n+k,m}'$, $k \geqq 0$, to the point that we are no longer sure that $[x_{n+k-d} \mid q_{n+k}' , \mathbf{X}_{n+k,d} \mid y_{n+k}' , \mathbf{Q}_{n+k,m}']$, $k > 0$ is weakly ψ_{n+k}-optimal. Without *iii*, therefore, one cannot apply Theorem 4 to prove the other theorems.

It is convenient now to invoke a lemma which is a simple theorem from measure theory. The lemma provides a standard form for the variables *q* and *y* of an object $[x \mid q, \mathbf{F} \mid y, \mathbf{G}] \varepsilon CQAx$.

Theorem 2: Given a Q*-discrete random variable* q *and a Borel field* **G**, *if y is a random variable measurable on the Borel field determined by* **G** *and the sample space of* q, *then there exist random variables* $\{\mathrm{z}_{\mathrm{p}} , \mathrm{p} \varepsilon \mathrm{Q}\}$ *such that*

(*i*) *each* z_{p} *is measurable* **G** *and*

(*ii*) *for each* $\omega \varepsilon \Omega$, *if* $\mathrm{q}(\omega) = \mathrm{p}$ *then* $y(\omega) = \mathrm{z}_{\mathrm{p}}(\omega)$.

Conversely, of course, given $\{\mathrm{Z}_{\mathrm{p}} , \mathrm{p} \varepsilon \mathrm{Q}\}$, *each measurable* **G**, *any* y *defined by ii is measurable on the field determined by* **G** *and the sample space of* q.

The proof of this lemma consists in showing that the class of random variables of the type of *y* above, as the $\{z_p , p \varepsilon Q\}$ are selected arbitrarily from the class of variables measurable **G**, exhausts the class of all random variables measurable on the Borel field determined by **G** and the sample space of q. The proof is a straightforward exercise in measure theory and is omitted.

To begin the main argument, given $[x \mid q, \mathbf{F} \mid y, \mathbf{G}] \varepsilon\ CDAx$ and a Borel measurable function $\psi(x, y)$, if for each x $\psi(x, y)$ is bounded from below, or if Q is a finite set, we can define the random variable

$$\xi(\omega) = \inf_{r \varepsilon Q} \psi(x(\omega), z_r(\omega)).$$

Then ξ is measurable $\{x\} \vee \mathbf{G}$.

Given $p\ \varepsilon\ Q$ and $r\ \varepsilon\ Q$, we define sets T_p^* , T_{pr} , T_p by

$$\begin{aligned} T_p^* &= \{\omega \mid \psi(x(\omega), z_p(\omega)) = \xi(\omega)\}, \\ T_{pr} &= \{\omega \mid \psi(x(\omega), z_p(\omega)) = \psi(x(\omega), z_r(\omega))\}, \\ T_p &= T_p^* - \bigcup_{\substack{r \neq p \\ r \varepsilon Q}} T_{pr} . \end{aligned}$$

Clearly each of these sets is measurable $\{x\} \vee \mathbf{G}$. T_p^* is the set where the index p minimizes $\psi(x, z_p)$, and T_p is that subset of T_p^* where this minimizing index is unique. It follows that if $r \neq p$ then

$$T_p \wedge T_r^* = \emptyset, \tag{10}$$

and as a consequence, $T_p \wedge T_r = \emptyset, r \neq p$.
Clearly

$$T_{pr} = T_{rp} .$$

Also

$$T_r^* \wedge T_{pr} = T_p^* \wedge T_{pr} , \tag{11}$$

since either side is the set where an index minimizing $\psi(x, z_s)$ can be equal either to p or to r.

In terms of these sets, the argument to be used can be outlined briefly. First, one shows that the T_p^* essentially cover Ω, in the sense that there is a null set N such that

$$\Omega - N = \bigcup_{p \varepsilon Q} T_p^* . \tag{12}$$

This follows without argument, and with $N = \emptyset$, if Q is finite; it results from ψ-optimality in general.

Second, by definition

$$T_p^* - T_p \leq \bigcup_{\substack{r \varepsilon Q \\ r \neq p}} T_{pr} . \tag{13}$$

Third, one observes that for p, $r \varepsilon Q$ and $p \neq r$, T_{pr} consists of the set S_{pr}

$$S_{pr} = \{\omega \mid z_p(\omega) = z_r(\omega)\}$$

plus a disjoint remainder $T_{pr} - S_{pr}$. The hypothesis $x \varepsilon CCD(\mathbf{G})$ allows one to show that this remainder is a null set. Over the set S_{pr}, on the other hand, the information about x conveyed by the family $\{z_p , p \varepsilon Q\}$ is redundant. The hypothesis of ψ-optimality can then be violated, unless S_{pr} is also a null set. It follows then that each T_{pr} is a null set, and from (12) and (13) then that the T_p partition Ω apart from a null set. From this the full theorem follows quickly.

To proceed with (12), given $p \varepsilon Q$, let N_p be the set

$$N_p = \{\omega \mid q(\omega) = p\} \wedge \{\Omega - \bigcup_{r \varepsilon Q} T_r^*\}.$$

Fix an $\omega \varepsilon N_p$; then $y(\omega) = z_p(\omega)$ but $\omega \not\varepsilon T_p^*$, so that $\xi(\omega) < \psi(x(\omega), z_p(\omega))$. It follows that there is some $r \varepsilon Q$, $r \neq p$, such that

$$\psi(x(\omega), z_r(\omega)) < \psi(x(\omega), z_p(\omega)), \tag{14}$$

and indeed, since Q is bounded from below, that there is a least such r, call it $r_p^*(\omega)$. Notice that N_p is measurable on the Borel field determined by the sample space of $\{x\}$, by $\mathbf{F}$, and by $\mathbf{G}$. Since $\mathbf{G} \subseteq \{x\} \vee \mathbf{F}$, it follows that N_p is measurable $\{x\} \vee \mathbf{F}$. That subset R_{pk} of N_p where $r_p^*(\omega) = k$ is empty if $k = p$; otherwise

$$R_{pk} = N_p \wedge \{\omega \mid \psi(x(\omega), z_1(\omega)) < \psi(x(\omega), z_p(\omega))\} \quad \text{if} \quad k = 1 \neq p,$$

$$R_{pk} = N_p \wedge \{\omega \mid \psi(x(\omega), z_k(\omega)) < \psi(x(\omega), z_p(\omega))\} \wedge$$
$$\cdot \bigcap_{j=1}^{k-1} \{\omega \mid \psi(x(\omega), z_j(\omega)) \geqq \psi(x(\omega), z_p(\omega))\} \quad \text{if} \quad k > 1, \quad k \neq p.$$

It follows from these equalities that R_{pk} and r_p^* are measurable $\{x\} \vee \mathbf{F}$.

We now define the Q-discrete random variable q' by

If $p \varepsilon Q$ and $\omega \varepsilon N_p$, $q'(\omega) = r_p^*(\omega)$;

If $\omega \varepsilon \Omega - \bigcup_{p \varepsilon Q} N_p$, then $q'(\omega)$ is the least value of $r \varepsilon Q$ such that $\omega \varepsilon T_r^*$.

Since the N_p cover the complement of $\bigcup_r T_r^*$, and since Q is bounded from below, this defines $q'(\omega)$ for each $\omega \varepsilon \Omega$; clearly q' is Q-discrete. Given $k \varepsilon Q$, the set where $q' \leqq k$ consists of the union of

$$\bigcup_{p \varepsilon Q} R_{pk}$$

with the set V_k, where

$$V_1 = T_1^*$$

$$V_k = (\Omega - T_1^*) \wedge \cdots \wedge (\Omega - T_{k-1}^*) \wedge T_k^*, \qquad k > 1.$$

Since each V_k is measurable $\{x\} \vee \mathbf{G} \subseteq \{x\} \vee \mathbf{F}$, it follows that q' is measurable $\{x\} \vee \mathbf{F}$. Furthermore, over $\Omega - \bigcup_{p \varepsilon Q} N_p$, q' is equal to a random variable that is measurable $\{x\} \vee \mathbf{G}$, since each V_k is measurable on this latter field.

We now define the random variable y' by

$$y'(\omega) = z_{q'(\omega)}(\omega), \qquad \omega \varepsilon \Omega.$$

Then y' is measurable on $\mathbf{G}$ and the sample space of q'. It follows that $[x \mid q', \mathbf{F} \mid y', \mathbf{G}] \varepsilon CQAx$, and from the hypothesis of weak ψ-optimality then that

$$E\{\psi(x, y)\} \leqq E\{\psi(x, y')\}. \tag{15}$$

But now we claim that for all $\omega \varepsilon \Omega$

$$\psi(x(\omega), y'(\omega)) \leqq \psi(x(\omega), y(\omega)). \tag{16}$$

First, if $\omega \varepsilon N_p$, we have

$$\begin{aligned} \psi(x(\omega), y'(\omega)) &= \psi(x(\omega), z_{r_p^*(\omega)}(\omega)) < \psi(x(\omega), z_p(\omega)) \\ &= \psi(x(\omega), y(\omega)), \end{aligned} \tag{17}$$

the inequality being by definition of r_p^*. Therefore strict inequality prevails in (16) for $\omega \varepsilon \bigcup_{p \varepsilon Q} N_p$. Consider now an $\omega \varepsilon (\Omega - \bigcup_{r \varepsilon Q} N_r) \wedge \{\omega' \mid q'(\omega') = p\}$. For this ω we have $\omega \varepsilon T_p^*$ and $\psi(x(\omega), y'(\omega)) = \psi(x(\omega), z_p(\omega)) \leqq \psi(x(\omega), z_r(\omega))$ for any $r \varepsilon Q$, by definition of T_p^*. But then (16) follows for this ω because $y(\omega) = z_r(\omega)$ for some $r \varepsilon Q$.

Now from (16), by taking expectations, we conclude the inequality opposite in sense to (15), hence (15) is an equality, and (16) is then an equality with probability one. Therefore *ii* of Theorem 4 is proved. Now by (17), (16) is a strict inequality over $N = \bigcup_{p \varepsilon Q} N_p$. Hence this latter set is a null set. Therefore *i* of Theorem 4 is proved, since q' is equal, over the complement of N, to a variable that is measurable $\{x\} \vee \mathbf{G}$, as we noted earlier. Finally, since

$$\Omega - \bigcup_{p \varepsilon Q} T_p^* = \bigcup_{r \varepsilon Q} N_r = N$$

the T_p^* essentially cover Ω. This is (12), as was to be proved.

It would be possible at this point to invoke the hypotheses $\psi \varepsilon K$

and $x \varepsilon CCD(\mathbf{G})$ to conclude *iv* of the Theorem. It will be more efficient to prove *iii* and *iv* together. To do so requires, as our earlier outline suggests, that we examine the sets $T_p^* \wedge T_{pr}$ over which redundancy prevails (because on $T_p^* \wedge T_{pr}$ either of z_p or z_r, where $r \neq p$, could be used to define the same value of y minimizing $\psi(x, y)$.

We have concluded (12), that except for $\omega \varepsilon N$, a null set, for each ω there is at least one $p \varepsilon Q$ such that $\xi(\omega) = \psi(x(\omega), z_p(\omega))$, that is, the minimizing index is uniquely p for $\omega \varepsilon T_p - N$.

Now define, as earlier, for $r \neq p$,

$$S_{pr} = \{\omega \mid z_p(\omega) = z_r(\omega)\}.$$

Then if $\omega \varepsilon T_{pr} - S_{pr}$, we have

$$\psi(x(\omega), z_p(\omega)) = \psi(x(\omega), z_r(\omega)), \qquad z_p(\omega) \neq z_r(\omega).$$

Since $\psi \varepsilon K$, it follows that for some $k = 1, 2, \cdots$ we have

$$x(\omega) = f_k(z_p(\omega), z_r(\omega)). \tag{18}$$

Now let A_{kpr} be the set of all ω such that (18) holds. We have just showed that

$$T_{pr} - S_{pr} \subseteq \bigcup_{k=1}^{\infty} A_{kpr} . \tag{19}$$

But now, since f_k is Borel measurable and each z_p is measureable $\mathbf{G}$, (18) constrains x on A_{kpr} to be equal to a random variable measurable $\mathbf{G}$. Since $x \varepsilon CCD(\mathbf{G})$, then A_{kpr} is a subset of some null set,

$$P\{A_{kpr}\} = 0, \qquad k = 1, 2, \cdots ,$$

and

$$\sum_{k=1}^{\infty} P\{A_{kpr}\} = 0.$$

This last with (19) makes $P\{T_{pr} - S_{pr}\} = 0$. Indeed, finally, since Q is countable,

$$P\{\bigcup_{p \varepsilon Q} \bigcup_{\substack{r \varepsilon Q \\ r \neq p}} (T_{pr} - S_{pr})\} = 0.$$

It is important later that by definition, S_{pr} is measurable $\mathbf{G}$ and therefore that, by (19), T_{pr} is essentially measurable $\mathbf{G}$.

We now define a new Q-discrete random variable q'' and a corresponding y''. The construction depends upon an arbitrarily chosen

$p_0 \,\varepsilon\, Q$ and an arbitrarily chosen real number a, although the notation will not emphasize this dependence. Later it will be shown that $q'' = q'$ and $y'' = y'$ each with probability one, so that the dependence upon p_0 and a is not essential.

Fix a $p_0 \,\varepsilon\, Q$ and select a real number a. Define the random variable $z''_{p_0}(\omega)$ by:

$$\text{if} \quad \omega \,\varepsilon \bigcup_{\substack{r \,\varepsilon\, Q\\ r \neq p_0}} T_{p_0 r}\,, \qquad z''_{p_0}(\omega) = a,$$

$$\text{otherwise,} \quad z''_{p_0}(\omega) = z_{p_0}(\omega).$$

Then $z''_{p_0}(\omega)$ is measurable **G**. Define

$$z''_p = z_p\,, \qquad p \,\varepsilon\, Q, \qquad p \neq p_0\,.$$

Then certainly each z''_p, $p \,\varepsilon\, Q$, is measurable **G**. Define the Q-discrete random variable $q''(\omega)$ by

$$\text{If } \omega \,\varepsilon\, T_{p_0} \vee [(T^*_{p_0} - T_{p_0}) \wedge \{\omega' \mid \psi(x(\omega'), a) < \psi(x(\omega'), z_{p_0}(\omega'))\}]$$

then $q''(\omega) = p_0$;

$$\text{if} \quad \omega \,\varepsilon\, (T^*_{p_0} - T_{p_0}) \wedge \{\omega' \mid \psi(x(\omega'), a) \geqq \psi(x(\omega'), z_{p_0}(\omega'))\}$$

then $q''(\omega)$ is the least value of $p \,\varepsilon\, Q$ such that $p \neq p_0$ and $\omega \,\varepsilon\, T^*_p$;

$$\text{if} \quad \omega \,\varepsilon\, \Omega - T^*_{p_0}\,, \text{ then } q''(\omega) = q'(\omega).$$

It is easily seen that this defines q'' for all $\omega \,\varepsilon\, \Omega$.

We now define the random variable y'' by $y''(\omega) = z''_{q''(\omega)}(\omega)$. Then y'' is measurable on **G** and the sample space of q'', so that by construction $[x \mid q'', \mathbf{F} \mid y'', \mathbf{G}] \,\varepsilon\, CQAx$. Applying the hypothesis of weak ψ-optimality, we conclude that

$$\int_\Omega [\psi(x, y'') - \psi(x, y)]\, dP = E\{\psi(x, y'')\} - E\{\psi(x, y)\} \geqq 0. \tag{20}$$

We now partition the domain Ω of integration into the four sets

$$\begin{aligned} A_1 &= T_{p_0} \\ A_2 &= (T^*_{p_0} - T_{p_0}) \wedge \{\omega \mid \psi(x(\omega), a) < \psi(x(\omega), z_{p_0}(\omega))\}, \\ A_3 &= (T^*_{p_0} - T_{p_0}) \wedge \{\omega \mid \psi(x(\omega), a) \geqq \psi(x(\omega), z_{p_0}(\omega))\}, \\ A_4 &= \Omega - T^*_{p_0}\,. \end{aligned}$$

That this is a partition follows from the definition and the fact, already proved, that $T_{p_0} \subseteq T^*_{p_0}$. We consider the four resulting integrals individually, in the order of the listing.

If $\omega \varepsilon T_{p_0}$ then either $\omega \varepsilon T_{p_0} \wedge N$, or $\omega \varepsilon T_{p_0} - N$. We may ignore the first case. For the second, by definition of T_{p_0}, if $r \neq p_0$

$$\psi(x(\omega), z_{p_0}(\omega)) < \psi(x(\omega), z_r(\omega)). \tag{21}$$

Also, by definition

$$\omega \neq \bigcup_{\substack{r \varepsilon Q \\ r \neq p_0}} T_{p_0 r} ,$$

and therefore by definition $z''_{p_0}(\omega) = z_{p_0}(\omega)$, and $q''(\omega) = p_0$. Then

$$\psi(x(\omega), y''(\omega)) = \psi(x(\omega), z''_{p_0}(\omega)) = \psi(x(\omega), z_{p_0}(\omega))$$

and from the inequality (21) we conclude that the integrand

$$\psi(x(\omega), y''(\omega)) - \psi(x(\omega), y(\omega)) < 0,$$

since $y(\omega)$ is equal to some $z_r(\omega)$, $r \varepsilon Q$. Hence the integral over A_1 is not positive.

If $\omega \varepsilon A_2$, then by definition $q''(\omega) = p_0$ and

$$y''(\omega) = z''_{p_0}(\omega).$$

Again, we ignore the contribution of $A_2 \wedge N$. If $\omega \varepsilon A_2 - N$ then by (13),

$$\omega \varepsilon \bigcup_{\substack{r \varepsilon Q \\ r \neq p_0}} T_{p_0 r} .$$

Then by definition $z''_{p_0}(\omega) = a$. Hence, the integrand

$$\begin{aligned} &\psi(x(\omega), y''(\omega)) - \psi(x(\omega), y(\omega)) \\ &\quad = [\psi(x(\omega), a) - \psi(x(\omega), z_{p_0}(\omega))] + [\psi(x(\omega), z_{p_0}(\omega)) - \psi(x(\omega), y(\omega))]. \end{aligned}$$

The first bracket on the right is <0 by definition of A_2, and the second is $\leqq 0$ because $\omega \varepsilon T^*_{p_0}$ and by definition of $T^*_{p_0}$ we have $\psi(x(\omega), z_{p_0}(\omega)) \leqq \psi(x(\omega), z_r(\omega))$ for all $r \varepsilon Q$; among the latter is $\psi(x(\omega), y(\omega))$. Hence the second integral is not positive, and its integrand is strictly negative.

Now consider $\omega \varepsilon A_3$. We ignore the integral over $A_3 \wedge N_1$. If $\omega \varepsilon A_3 - N_1$, then $q''(\omega) = p \neq p_0$ and $\omega \varepsilon T^*_p$ for some $p \varepsilon Q$. For this ω we have

$$\psi(x(\omega), y''(\omega)) = \psi(x(\omega), z''_p(\omega)) = \psi(x(\omega), z_p(\omega)) \leqq \psi(x(\omega), z_r(\omega))$$

for all $r \varepsilon Q$; here the first equality is by definition of y'', the second by definition of z_p'' since $p \neq p_0$, and the inequality is by definition of T_p^*. But the inequality makes the integrand in (20) $\leqq 0$, since $y(\omega) = z_r(\omega)$ for some $r \, \varepsilon \, Q$. Therefore, the integral over A_3 is not positive.

Over A_4, the integrand of (20) is

$$[\psi(x, y'') - \psi(x, y')] + [\psi(x, y') - \psi(x, y)].$$

The second bracket vanishes with probability one by *ii* of Theorem 4, already proved. The first bracket is

$$\psi(x(\omega), z_{q'(\omega)}''(\omega)) - \psi(z(\omega), z_{q'(\omega)}(\omega))$$

and this vanishes for all $\omega \, \varepsilon \, A_4$ by the definitions because over A_4, $\xi(\omega) < \psi(x(\omega), z_{p_0}(\omega))$ so that $q'(\omega) \neq p_0$; therefore by definition $z_{q'(\omega)}''(\omega) = z_{q'(\omega)}(\omega)$.

We conclude from these calculations that the integral (20) cannot be positive. By (20), therefore, the integral vanishes. But the argument showed that the integrand was $\leqq 0$ with probability one, hence indeed, the integrand vanishes with probability one:

$$\psi(x, y'') = \psi(x, y) \quad \text{with probability one.}$$

In particular, over A_2, the integrand was strictly < 0. Therefore A_2 has probability zero. We shall now exploit this fact.

In the argument above, a was any real number. Let $\{a_n\}$ be a countable dense set of real numbers and let

$$W_n = \{\omega \mid \psi(x(\omega), a_n) < \psi(x(\omega), z_{p_0}(\omega))\}.$$

We have just proved that $P\{A_2\} = 0$, which is to say that we could have proved, for each n, that

$$P\{(T_{p_0}^* - T_{p_0}) \wedge W_n\} = 0.$$

Then also

$$N_2 = \bigcup_n (T_{p_0}^* - T_{p_0}) \wedge W_n$$

is a null set. Now if $\omega \, \varepsilon \, N_2$, then $\omega \, \varepsilon \, T_{p_0}^* - T_{p_0}$ and also there is some number a_n such that

$$\psi(x(\omega), a_n) < \psi(x(\omega), z_{p_0}(\omega)). \tag{22}$$

Conversely, if $\omega \, \varepsilon \, T_{p_0}^* - T_{p_0}$ and there is a number a_n such that (22) is true, then $\omega \, \varepsilon \, N_2$. Therefore if $\omega \, \varepsilon \, (T_{p_0}^* - T_{p_0}) - N_2$, then for every number a_n we have

$$\psi(x(\omega), a_n) \geqq \psi(x(\omega), z_{p_0}(\omega)). \tag{23}$$

Given an $\omega \varepsilon (T_{p_0}^* - T_{p_0}) - N_2$, choose a sequence $a_n \rightarrow x(\omega)$. Assume that $\psi \varepsilon K$. Then ψ is continuous and from (23) we have

$$0 = \psi(x(\omega), x(\omega)) = \lim \psi(x(\omega), a_n) \geqq \psi(x(\omega), z_{p_0}(\omega)) \geqq 0.$$

Notice, incidentally, that it suffices here for each x that $\psi(x, y)$ be continuous for y in some neighborhood of x. This is an example of one way in which K can be enlarged.

From this and item *iv* in the definition of K, there is some integer K such that

$$x(\omega) = g_k(z_{p_0}(\omega)). \tag{24}$$

Let C_k be the set of all ω such that (24) holds. Since g_k is Borel measurable, over C_k, (24) constrains x to be equal to a function measurable **G**. If $x \varepsilon CCD(\mathbf{G})$, then C_k is a null set. But we have just showed above that

$$(T_{p_0}^* - T_{p_0}) - N_2 \subseteq \bigcup_{k=1}^{\infty} C_k .$$

Therefore

$$P\{T_{p_0}^* - T_{p_0}\} = 0.$$

Since p_0 was arbitrary, this can be proved for each $p_0 \varepsilon Q$; therefore from (12) the T_p, $p \varepsilon Q$ essentially cover Ω. We proved along with definitions that the T_p are pairwise disjoint, hence they partition $\Omega - N_3$, where N_3 is some null set.

We continue the argument using the selected p_0. For $\omega \varepsilon \Omega - N_3$, either $\omega \varepsilon T_{p_0}$ or $\omega \varepsilon T_r$ where $r \varepsilon Q$ but $r \neq p_0$. In this latter case, however, as we proved with the definitions, $\omega \varepsilon \Omega - T_{p_0}^*$; then by definition $q''(\omega) = q'(\omega)$. If $\omega \varepsilon T_{p_0}$, by the definitions $q''(\omega) = q'(\omega) = p_0$. Therefore

$$q'' = q' \quad \text{with probability one.} \tag{25}$$

Furthermore we know that if $\omega \varepsilon T_p$, then $q'(\omega) = p$. From (25)

$$y''(\omega) = z''_{q'(\omega)}(\omega). \tag{26}$$

If $\omega \varepsilon \Omega - T_{p_0}$, except at most on a null set we have $q''(\omega) \neq p_0$ and from (26) and the definition of z''_p

$$y''(\omega) = z''_{q'(\omega)}(\omega) = z_{q'(\omega)}(\omega) = y'(\omega), \qquad \omega \varepsilon (\Omega - T_{p_0}) \wedge N_5 \tag{27}$$

where N_5 is a null set. Now if $\omega \varepsilon T_{p_0} - N$, we showed earlier that $z''_{p_0}(\omega) = z_{p_0}(\omega)$. Hence the equalities in (27) hold for $\omega \varepsilon T_{p_0} - N$ as well, so that

$$y'' = y' \quad \text{with probability one.} \tag{28}$$

Equalities (25) and (28) free the constructions from any dependence, except on a null set, upon the initially selected p_0 and a. We need the Theorem to make identification with q and y.

Let S_p be that subset of T_p where $q(\omega) \neq p$. Then if $\omega \varepsilon S_p$, by definition of T_p,
$\psi(x(\omega), y'(\omega)) = \psi(x(\omega), z_p(\omega)) < \psi(x(\omega), z_{q(\omega)}(\omega)) = \psi(x(\omega), y(\omega))$. From *ii* of Theorem 4, then, $P\{S_p\} = 0$, and $P\{\bigcup_{r \varepsilon Q} S_r\} = 0$. Since the T_p, $p \varepsilon Q$, essentially partition Ω, it follows that $q' = q$ with probability one, and at once that $y(\omega) = z_{q(\omega)}(\omega) = z_{q'(\omega)}(\omega) = y'(\omega)$ with probability one. These conclusions are *iii* and *iv* of the Theorem, the proof of which is now complete.

To prove Theorem 1, let $\{x_n, q_n, y_n\}$ be a given Q-coded communication system that is $\{\psi_n, 0\}$ optimal. Given n, by Lemma 1,

$$[x_n \mid q_n, \mathbf{X}_{n,0} \mid y_n, \mathbf{Q}_{n,\infty}] \varepsilon CQAx_n$$

and is weakly ψ_n-optimal. If $\psi_n \varepsilon K$ and $x_n \varepsilon CCD(\mathbf{Q}_{n,\infty})$, Theorem 4 proves that q_n is measurable on $\{x_n\} \vee \mathbf{Q}_{n,\infty}$. But $\mathbf{Q}_{n,\infty}$ is the sample space of $\{q_{n-1}, q_{n-2}, \cdots\}$, and is therefore contained in the sample space of $\{x_{n-1}, x_{n-2}, \cdots\}$, since by hypothesis $\{x_n, q_n, y_n\}$ is a Q-coded communication system. The hypothesis $\{x_n\} \varepsilon CCD$ of Theorem 1 then implies that for the given n, $x_n \varepsilon CCD(\mathbf{Q}_{n,\infty})$, and Theorem 4 establishes Theorem 1.

Turning to Theorem 3, let $\{x_n, q_n, y_n\}$ be a given Q-coded communication system with decoder memory span m, and suppose that it is $\{\psi_n, d, m\}$-optimal. By Lemma 1, then, given n, $[x_{n-d} \mid q_n, \mathbf{X}_{n,d} \mid y_n, \mathbf{Q}_{n,m}] \varepsilon CQAx_{n-d}$ and is weakly ψ_n-optimal. By the hypotheses of Theorem 3, $\psi_n \varepsilon K$, and $\{x_n\} \varepsilon CCDf$. Consider $\mathbf{Q}_{n,m}$, the sample space of $\{q_{n-1}, q_{n-2}, \cdots, q_{n-m}\}$. Suppose first that $m > d$; then this sample space is the smallest Borel field which contains both the sample space of $\{q_{n-1}, \cdots, q_{n-d}\}$ and that of $\{q_{n-d-1}, \cdots, q_{n-m}\}$. Since $M < \infty$, the first of these is a finite field, and the second is a subfield of $\{x_{n-d-1}, x_{n-d-2}, \cdots\}$ (since $\{x_n, q_n, y_n\}$ is indeed a Q-coded communication system). The hypothesis $\{x_n\} \varepsilon CCDf$ then implies that $x_n \varepsilon CCD(\mathbf{Q}_{n,m})$. If $m \leqq d$, the subfield of $\{x_{n-d-1}, \cdots\}$ is empty, but the reasoning and conclusion are still valid. Then Theorem 4 applies and we conclude that q_n is measurable on the sample space of $\{x_{n-d}, q_{n-1}, \cdots, q_{n-m}\}$. This is the first conclusion of Theorem 3. We note now that a weaker hypothesis than $\{x_n\} \varepsilon CCDf$ could suffice here. Indeed, if $m < \infty$, it is sufficient that: if A is a finite field then $x_n \varepsilon CCD(\mathrm{A})$. This is the final conclusion of Theorem 3.

Given that q_n is essentially measurable on $\{x_{n-d}, q_{n-1}, \cdots, q_{n-m}\}$, for each n, we conclude by induction that q_n is essentially measurable $\{x_{n-d}, x_{n-d-1}, q_{n-2}, \cdots, q_{n-m-1}\}$, $\cdots$ and finally then that q_n is essentially measurable $\{x_{n-d}, x_{n-d-1}, \cdots\}$. Define

$$q'_n = q_{n+d},$$

$$y'_n = y_{n+d}, \qquad n = 0, \pm 1, \cdots .$$

Then it is a simple translation of notation to verify that $\{x_n, q'_n, y'_n\}$ is a Q-coded communication system with decoder memory span m that is $\{\psi'_n, 0, m\}$-optimal, where $\psi'_n = \psi_{n+d}$, $n = 0, \pm 1, \cdots$. This is the second conclusion of Theorem 3.

Finally, if $d = 0$, then "$\{x_n\} \varepsilon CCDf$" may be replaced by: "$\{x_n\} \varepsilon CCD$." Then M is unrestricted, since no "future" is involved that must be restricted to a finite field. This completes the proof.

Theorem 2 is a limiting case of Theorem 3, proved by putting $m = \infty$ everywhere in the proof of Theorem 3.

V. A COROLLARY

It is a consequence of Lemma 2 and of the proof of Theorem 4 that, given ω, in a set of probability one, $q(\omega)$ is that unique value of p which minimizes $\psi(x(\omega), z_p(\omega))$. (This was remarked in connection with equation 25.) Applying this to the situation of Theorem 1, one sees that the transmitter of a delay-free Q-coded communication system $\{x_n, q_n, y_n\}$ satisfying Theorem 1 has the block diagram form shown in Fig. 1. (If $d > 0$, one simply puts an analog delay line in the input lead, ahead of the rest of the system.)

This block diagram can be described thus: at time subsequent to $t = n - 1$ and prior to $t = n$, the transmitter has in its digital store the values $q_{n-1}, q_{n-2}, \cdots$ of the previously transmitted signals. From these, quantities $z_{1,n}, z_{2,n}, z_{3,n}, \cdots$ are constructed. These are the z_p of Lemma 2, for the particular random variable y_n. When x_n becomes available, quantities $\psi_n(x_n, z_{1,n})$, $\psi_n(x_n, z_{2,n})$, $\cdots$ are constructed and the comparator identifies the least of these (unique with probability one). The transmitted q_n is that value of the index which identifies the least $\psi_n(x_n, z_{p,n})$. This index is transmitted to the receiver as q_n and is also stored in the transmitter's memory for the next cycle. The receiver can be realized using a portion of the transmitter, as suggested in Fig. 2. Each function generator in these diagrams can

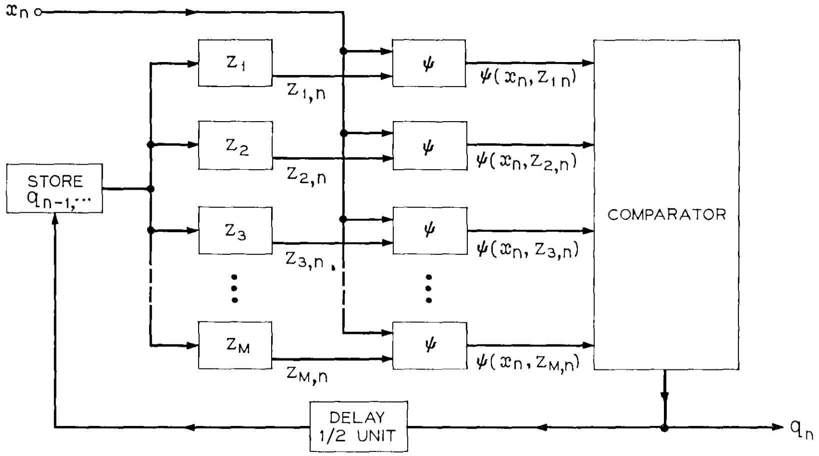


Fig. 1 — Generalized form of optimum transmitter.

of course be nonstationary. Connections to a master "clock" are not shown.

VI. REMARKS ON K AND CCD

One might ask to what degree are the central hypotheses of Theorem 4 necessary to the conclusions. The theorem itself provides a partial answer: conclusions *i* and *ii* do not use $x \varepsilon CCD(\mathrm{G})$ at all, and use only a measurability and a boundedness property of ψ. The critical conclusions are the uniqueness conclusions *iii* and *iv*. Clearly, something

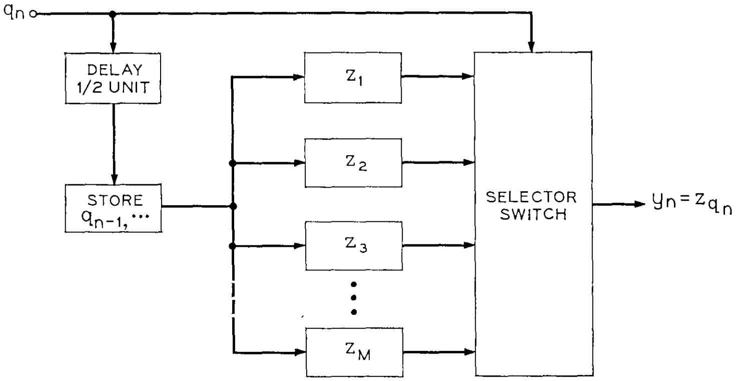


Fig. 2 — Form of receiver.

is required of $\psi(x, y)$ that makes it, in some sense, smaller when $y = x$ than elsewhere, and not too indifferent to the value of y when $y \neq x$, if uniqueness is to be expected from the hypothesis of ψ-optimality. As we have already noted, the hypothesis $\psi \in K$ is fairly weak in this regard, and could, in the presence of *CCD*, be made weaker at the expense of further elaboration of the proof.

The interesting hypothesis is $x \in CCD(\mathbf{G})$. This implies that if x has a conditional probability distribution relative to the field $\mathbf{G}$, then that distribution is continuous. It is easy to see that the ψ-optimum quantizing of a random variable x need not be unique if the distribution of x is not continuous, even when one uses $\psi(x, y) = (x - y)^2$. Since y in Theorem 2 ψ-optimally quantizes x for each event measurable on the conditioning field $\mathbf{G}$, something like $x \in CCD(\mathbf{G})$ is necessary if conclusion *iv* is to follow. Thus we conclude a loose kind of necessity for this hypothesis.

We notice finally that *iii* and *iv* were proved by confining the redundancy among the $\{z_p , p \in Q\}$ to a null set. In the application of this idea to the situation of Theorem 1, it seems likely that redundancy in the $\{z_{pn} , p \in Q\}$ for some fixed n might indeed be exploited to improve some

$$E\{\psi_{n+k}(x_{n+k} , y_{n+k})\}, \qquad k > 0, \tag{29}$$

by selection, among the minimizing z_{pn} to which $E\{\psi_n(x_n , y_n)\}$ is indifferent, one which actually contributes information about x_{n+k} and therefore allows a reduction in (29). I have no example to show this phenomenon, so its existence remains a conjecture. We have proved, of course, that its possible existence is ruled out by $x \in CCD(\mathbf{G})$.

REFERENCES

1. Shannon, C. E., "A Mathematical Theory of Communication," B.S.T.J., *27*, Nos. 3 and 4 (July and October 1948), pp. 379–423, 623–656.

The Equivalence of Certain Harper Codes

By MORGAN M. BUCHNER, Jr.

(Manuscript received May 12, 1969)

A class of binary encoding algorithms called Harper codes has been studied previously as a means of encoding numbers for transmission over an idealized binary channel. This paper considers a more general and practical transmission system model. For any Harper code, it presents a technique for obtaining the expression for the average absolute numerical error that occurs during transmission. It shows that all Harper codes do not exhibit the same average absolute numerical error for all transmission systems that satisfy the model. However, there is a subset of Harper codes such that all codes in the subset give identical performance. The paper defines the subset and presents an expression for the average absolute numerical error for any Harper code in the subset. The subset is important because it includes the natural binary representation, the Gray code, and the folded binary code.

I. INTRODUCTION

In order to send numerical data over a binary channel, each input number must be encoded into a suitable binary sequence for transmission. For example, when a sampler and quantizer are used, a binary sequence is assigned to each quantization level. For each sample, the number of the appropriate quantization level is transmitted by sending the binary sequence assigned to the level. But how should the binary sequences be assigned? One approach is to use the natural binary representation of each number. Alternatively, a Gray code might be used with the idea that its unit-distance properties are in some sense desirable.

If the transmission system is error-free and if the binary sequences are unique, it does not matter how the sequences are assigned. However, if transmission errors can occur, some assignment algorithms may be preferable to others. In this paper, the performance of certain binary encoding algorithms is considered. The average magnitude by which the number delivered to the destination differs from the transmitted number is used as the criterion of performance.

Previously, Harper presented a class of binary codes that we call Harper codes.[1] The class includes the natural binary representation, the Gray code, and the folded binary code. Reference 2 showed that for any set of 2^k input numbers all Harper codes exhibit the same mean magnitude error when used with a specific binary transmission system model (see Section II) and that, when the probability of transmission error is sufficiently small, Harper codes are optimum.

In this paper, a more general transmission system model is considered. For 2^k equally spaced input numbers, a means of obtaining the expression for the average absolute numerical error (hereafter called average numerical error) for any Harper code is presented. All Harper codes do not exhibit the same average numerical error except in the special case when the transmission system model reduces to the model used in Ref. 2. However, there does exist a subset of Harper codes such that all codes in the subset are equivalent in performance. The subset is defined and an expression is given for the average numerical error for any Harper code in the subset. The subset is important because it includes the natural binary representation, the Gray code, and the folded binary code.

II. SYSTEM MODEL AND PREVIOUS RESULTS

A system model is shown in Fig. 1. In general, we wish to send over a binary transmission system[†] any one of the 2^k equally likely numbers of the form $A + Bs$ where s is an integer, $0 \leqq s \leqq 2^k - 1$. At the transmitter, the binary encoder receives $A + Bs$ and, based upon s, sends a k-bit binary sequence assigned by a Harper code and denoted by $H_k(s)$. At the receiver, a binary decoder receives a k-bit binary sequence $H_k(r)$, $0 \leqq r \leqq 2^k - 1$, and generates $A + Br$. Let $\Pr[H_k(r) \mid H_k(s)]$ denote the probability of receiving $H_k(r)$ when $H_k(s)$ is sent. If all s are equally likely, the average numerical error (as in Ref. 3) that occurs is

$$ANE = \frac{B}{2^k} \sum_{r=0}^{2^k-1} \sum_{s=0}^{2^k-1} | r - s | \Pr[H_k(r) \mid H_k(s)]. \tag{1}$$

The average numerical error is dependent upon the binary encoding algorithm and the transmission system through $\Pr[H_k(r) \mid H_k(s)]$.

[†] It is important to distinguish between the binary transmission system and the channel. The transmission system includes the channel and the encoder and decoder for error control.

Harper codes are defined in terms of the vertices of the k-cube[1]. Assign 0 to an arbitrary vertex; that is, $H_k(0)$ is arbitrary. Having assigned $0, 1, 2, \cdots, l - 1$, assign l to an unnumbered vertex (not necessarily unique) that has the most numbered one-distant neighbors.[†] In the remainder of this paper, certain properties of Harper codes presented in Refs. 1 and 2 are used without specific reference.

We can now summarize the results in Ref. 2. In a binary transmission system as shown in Fig. 1, it was assumed that the errors between locations 1 and 2 are independent of the transmitted bits and occur independently of one another with probability p_1. For such a transmission system and for any set of 2^k input numbers, it was shown that all Harper codes yield the same mean magnitude error and, thus, are equivalent. Also, it was shown that when p_1 is sufficiently small, Harper codes are optimum for any set of 2^k input numbers because they minimize the mean magnitude error. Of course, the results in Ref. 2 are applicable to our set of 2^k equally spaced numbers and indicate that all Harper codes yield the same average numerical error for a transmission system that satisfies the model in Ref. 2.

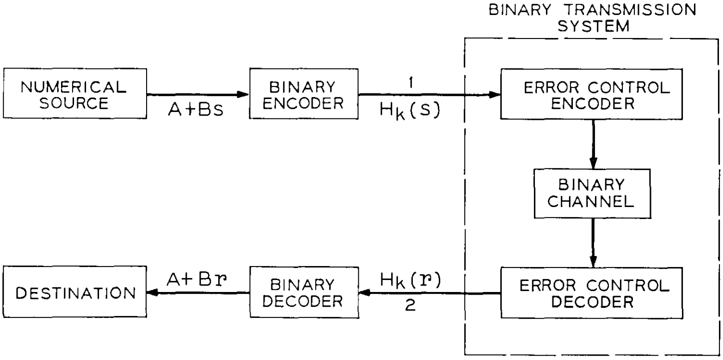


Fig. 1 — System model

However, the transmission system model in Ref. 2 is extremely restrictive. Channels with correlated errors are excluded. The model also excludes transmission systems using many types of error-correcting codes even if the actual channel is a memoryless binary symmetric channel with probability of bit error p. In fact, even the Hamming

[†] The weight of an n-tuple v is the number of nonzero components in v and is denoted by $w[v]$. The distance between two binary n-tuples u and v is $w[u \oplus v]$ where $\oplus$ denotes component by component modulo 2 addition of n-tuples.

perfect single error-correcting codes when used with a memoryless binary symmetric channel do not comply with the model in Ref. 2. The reason is that, in a Hamming code, all $H_k(s)$ of a particular weight are not encoded as code vectors of equal weight. Thus, all error patterns of equal weight in the Harper code sequences do not occur with equal probability. However, in order for a transmission system to satisfy the model in Ref. 2, all error patterns of equal weight must occur with equal probability. It follows that the Hamming code violates the model in Ref. 2.

An interesting approach to coding for numerical data transmission is found in unequal error-protection codes[4]. The idea behind unequal error-protection codes is to match the protection provided by the code to the numerical significance of the transmitted bits. Significant-bit codes (a subclass of unequal error-protection codes) have been shown to be effective in reducing the average numerical error and in many cases are easy to implement.[3,5] However, the transmission system model in Ref. 2 excludes unequal error-protection codes which is unfortunate because these codes are directly applicable to the basic problem considered in Ref. 2, that is, reducing the average numerical error.

Accordingly, it is important to examine the performance of Harper codes when a less restrictive and more practical transmission system model is used. For our model, we assume simply that the transmission system is binary and that the errors are independent of the transmitted bits. A binary transmission system satisfies this model if, for every integer r, $0 \leqq r \leqq 2^k - 1$, and integer s, $0 \leqq s \leqq 2^k - 1$, there exists an integer t, $0 \leqq t \leqq 2^k - 1$, such that

$$\Pr[H_k(r) \mid H_k(s)] = \Pr[H_k(t) \mid B_k(0)] \tag{2}$$

where

$$H_k(t) = H_k(r) \oplus H_k(s) \tag{3}$$

and $B_i(j)$ denotes the i-bit natural binary representation of the integer j, $0 \leqq j \leqq 2^i - 1$. Observe that equation (2) implies that the probability of a particular error pattern $H_k(t)$ in a Harper code sequence is independent of the transmitted sequence.

Because the transmission system model is not very restrictive, the results to be presented are applicable to a wide range of practical systems. For example, the model is satisfied by the important class of binary transmission systems composed of (*i*) a linear block code with a decoding scheme equivalent to Slepian's standard array[6], and

(*ii*) a binary symmetric channel in which the errors are independent of the transmitted bits.

III. THE AVERAGE NUMERICAL ERROR FOR A HARPER CODE

Let H' be a Harper code in which s is encoded as $H_k'(s)$. From the definition of a Harper code, it is possible that $H_k'(0) \neq B_k(0)$. We first show that if $H_k'(0) \neq B_k(0)$, then a Harper code H [in which s is encoded as $H_k(s)$] can be constructed such that (*i*) $H_k(0) = B_k(0)$ and, (*ii*) the performance of H' is identical to the performance of H. The average numerical error for H' is

$$ANE' = \frac{B}{2^k} \sum_{r=0}^{2^k-1} \sum_{s=0}^{2^k-1} | r - s | \Pr[H_k'(r) \mid H_k'(s)]. \tag{4}$$

Let H be a code whose elements are obtained from the elements of H' by the relation

$$H_k(s) = H_k'(s) \oplus H_k'(0). \tag{5}$$

From (5), $H_k(0) = B_k(0)$.

We now show that H is a Harper code. Clearly $H_k(0)$ satisfies the requirements for a Harper code. Suppose that $H_k(0)$ through $H_k(l - 1)$ have been determined by (5). Now, if $H_k'(s)$ is distance d from $H_k'(l)$, then $H_k(s)$ is distance d from $H_k(l)$. Thus, if $H_k'(l)$ is assigned to have the most numbered one-distant neighbors, $H_k(l)$ will have the most numbered one-distant neighbors. It follows that H is a Harper code.

The average numerical error for H is given by equation (1). We must show that the expression for ANE is identical to the expression for ANE'. From (2),

$$\Pr[H_k'(r) \mid H_k'(s)] = \Pr[H_k'(r) \oplus H_k'(s) \mid B_k(0)].$$

Also, from (2),

$$\Pr[H_k(r) \mid H_k(s)] = \Pr[H_k(r) \oplus H_k(s) \mid B_k(0)].$$

From (5),

$$H_k(r) \oplus H_k(s) = H_k'(r) \oplus H_k'(s).$$

Therefore,

$$\Pr[H_k(r) \mid H_k(s)] = \Pr[H_k'(r) \mid H_k'(s)]$$

and, by (1) and (4),

$$ANE = ANE'.$$

Thus, every Harper code is equivalent in performance to a Harper code in which

$$H_k(0) = B_k(0). \tag{6}$$

For convenience and without loss of generality, we shall consider the performance of Harper codes that satisfy (6). At the end of Section IV, we remove this restriction and give, in general terms, the structure of all Harper codes that are equivalent to the natural binary representation.

Now, let us consider the expression for the average numerical error for H. By substituting (2) into (1) and rewriting,

$$ANE = \frac{B}{2^k} \sum_{t=1}^{2^k-1} \sum_{s=0}^{2^k-1} | r_t - s | \Pr[H_k(t) \mid B_k(0)] \tag{7}$$

where r_t is the value of r in (3), that is,

$$H_k(r_t) = H_k(s) \oplus H_k(t). \tag{8}$$

Now, (7) can be written as

$$ANE = \frac{B}{2^k} \sum_{t=1}^{2^k-1} C_t \Pr[H_k(t) \mid B_k(0)] \tag{9}$$

where

$$C_t = \sum_{s=0}^{2^k-1} | r_t - s |. \tag{10}$$

The expression for the average numerical error is determined by specifying each C_t ($1 \leqq t \leqq 2^k - 1$).

In order to evaluate C_t, we proceed as follows. Divide the 2^k elements of H into $k + 1$ sets called levels. The 0-level contains $H_k(0)$ exclusively. For $1 \leqq j \leqq k$, the j-level is the set of $H_k(s)$ for which $2^{j-1} \leqq s \leqq 2^j - 1$. Because H is a Harper code, the elements of level j are in the shadow of the $(j - 1)$-subcube[†] formed by the elements of levels 0 through $j - 1$. From equation (6) and the definition of a Harper code, it follows that each element of the j-level has a one in a particular position which we call the j-position. Thus, the j-level consists of the k-tuples that

[†] A $(j - 1)$-subcube of the k-cube is a set of all k-tuples that are the same in $k - j + 1$ positions. The shadow of a $(j - 1)$-subcube is obtained by changing one of the $k - j + 1$ fixed positions.

have zeros in positions $j + 1$ through k, a one in position j, and all possible $(j - 1)$-tuples in positions 1 through $j - 1$.

Notice that the position numbers are determined by the structure of the Harper code and not by the order in which the bits are arranged for transmission. For example, in the Harper code shown in Table I, $\Pr[H_4(2) \mid B_4(0)]$ is the probability that no transmission errors occur in positions 1, 3, and 4 and that a transmission error occurs in position 2. If transmitted in the order shown in Table I, $\Pr[H_4(2) \mid B_4(0)]$ is the probability that the error sequence 0001 occurs.

We must determine C_t for each of the $2^k - 1$ nonzero values of t. Thus, we regard t as known and seek to determine C_t. Let σ be an integer such that

$$2^{\sigma-1} \leqq t \leqq 2^{\sigma} - 1. \tag{11}$$

Because H satisfies equation (6), $H_k(t)$ has a one in position σ. To evaluate C_t, we rewrite (10) to exhibit the levels of s as

$$C_t = \left(r_t + \sum_{j=1}^{\sigma} \sum_{s=2^{j-1}}^{2^j-1} | r_t - s |\right) + \sum_{j=\sigma+1}^{k} \sum_{s=2^{j-1}}^{2^j-1} | r_t - s | \tag{12}$$

TABLE I—A $k = 4$ HARPER CODE

s	$H_4(s)$	Level number
0	0 0 0 0	0
1	0 0 1 0	1
2	0 0 0 1	2
3	0 0 1 1	2
4	0 1 1 1	3
5	0 1 1 0	3
6	0 1 0 1	3
7	0 1 0 0	3
8	1 0 0 0	4
9	1 0 0 1	4
10	1 0 1 1	4
11	1 0 1 0	4
12	1 1 0 0	4
13	1 1 1 0	4
14	1 1 0 1	4
15	1 1 1 1	4

position 4 — position 2

position 3 — position 1

where the 0-level is shown individually as r_t and j indexes the levels from 1 to k. The parentheses enclose the contribution of levels 0 through σ. From Appendix A,

$$\left(r_t + \sum_{j=1}^{\sigma} \sum_{s=2^{j-1}}^{2^j-1} | r_t - s |\right) = 2^{2\sigma-1}. \tag{13}$$

Now, consider the set of $H_k(s)$ in the j-level where $\sigma + 1 \leqq j \leqq k$ and $2^{j-1} \leqq s \leqq 2^j - 1$. First, we define a run as follows.[†] In the j-level, there is a run in position m, $1 \leqq m \leqq j - 1$, that starts at s_0 and is of length $R(m, s_0)$ if and only if

(*i*) $R(m, s_0) = 2^l$ for some integer $l \geqq 0$,

(*ii*) the set of $H_k(s)$ for $s_0 \leqq s \leqq s_0 + 2^l - 1$ forms an l-subcube of the k-cube where m is one of the $k - l$ fixed positions,

(*iii*) the set of $H_k(s)$ for $s_0 + 2^l \leqq s \leqq s_0 + 2^{l+1} - 1$ forms an l-subcube that is in the shadow of the subcube in (*ii*),

(*iv*) the subcube in (*iii*) is distinguished from the subcube in (*ii*) by position m, and

(*v*) the $H_k(s)$ for $2^{j-1} \leqq s \leqq s_0 - 1$ can be divided into runs of length 2^l although perhaps not in position m.

An example from Table I will illustrate the definition of a run. Consider the 4-level. Then $H_4(8)$ starts a run of length 1 in position 2, a run of length 2 in position 1, and a run of length 4 in position 3. Thus,

$$R(1, 8) = 2 \qquad R(2, 8) = 1 \qquad R(3, 8) = 4.$$

Let $w[H_k(t)] = \omega$ and let $t_1, t_2, \cdots, t_\omega$ denote the ω nonzero positions in $H_k(t)$. Then $R(t_m, 2^{j-1})$ is the length of the run in position t_m that starts at 2^{j-1} (that is, the length of the first run in position t_m in the j-level). Let

$$\gamma_{j,1}(t) = \operatorname*{Max}_{m} R(t_m, 2^{j-1}).$$

From Appendix C,

$$\sum_{s=2^{j-1}}^{2^{j-1}+2\gamma_{j,1}(t)-1} | r_t - s |$$

$$= \sum_{s=2^{j-1}}^{2^{j-1}+\gamma_{j,1}(t)-1} (r_t - s) + \sum_{s=2^{j-1}+\gamma_{j,1}(t)}^{2^{j-1}+2\gamma_{j,1}(t)-1} (s - r_t) = 2\gamma_{j,1}^2(t).$$

[†] Appendix B contains a more complete discussion of the structure of the j-level of a Harper code and the relationship between the structure and the concept of a run. It is shown that runs are basic to the structure of Harper codes and that the definition of a run is meaningful and consistent.

The above process can be extended to obtain $\gamma_{j,i}(t)$ after $\gamma_{j,1}(t)$, $\gamma_{j,2}(t), \cdots, \gamma_{j,i-1}(t)$ are known. Specifically,

$$\gamma_{j,i}(t) = \operatorname*{Max}_{m} R\left(t_m, 2^{j-1} + 2\sum_{l=1}^{i-1} \gamma_{j,l}(t)\right).$$

Then

$$\sum_{s=2^{j-1}+2\sum_{l=1}^{i-1}\gamma_{j,l}(t)}^{2^{j-1}-1+2\sum_{l=1}^{i}\gamma_{j,l}(t)} | r_t - s | = 2\gamma_{j,i}^2(t).$$

By continuing the process, we eventually exhaust the 2^{j-1} values of s in the j-level. Let g_j denote the number of $\gamma_{j,i}(t)$ needed to cover the j-level, that is,

$$2\sum_{i=1}^{g_j} \gamma_{j,i}(t) = 2^{j-1}.$$

It follows that

$$\sum_{s=2^{j-1}}^{2^j-1} | r_t - s | = 2\sum_{i=1}^{g_j} \gamma_{j,i}^2(t). \tag{14}$$

From (12), (13), and (14),

$$C_t = 2^{2\sigma-1} + 2\sum_{j=\sigma+1}^{k}\sum_{i=1}^{g_j} \gamma_{j,i}^2(t). \tag{15}$$

By substituting (15) into (9),

$$ANE = \frac{B}{2^k}\sum_{t=1}^{2^k-1}\left(2^{2\sigma-1} + 2\sum_{j=\sigma+1}^{k}\sum_{i=1}^{g_j}\gamma_{j,i}^2(t)\right)\Pr[H_k(t) \mid B_k(0)] \tag{16}$$

where σ is given by (11). The expression in (16) is particularly useful because it consists exclusively of error probabilities conditional upon $B_k(0)$ being transmitted and the $\gamma_{j,i}(t)$ can be obtained directly from the Harper code. A numerical example that illustrates the use of (15) and (16) is given in Appendix D.

We now consider the condition under which two Harper codes give identical performance. Let H' be a Harper code that is not H (that is, H' cannot be obtained from H by a relationship of the form $H_k'(s) = H_k(s) \oplus B_k(s_1)$ where s_1 is an arbitrary integer, $0 \leqq s_1 \leqq 2^k - 1$).

From (9), for H',

$$ANE' = \frac{B}{2^k} \sum_{t'=1}^{2^k-1} C'_{t'} \Pr[H'_k(t') \mid B_k(0)].$$

Then H and H' exhibit identical performance for any transmission system that satisfies our model only if, for every t, $C'_{t'} = C_t$ where t' is determined by $H'_k(t') = H_k(t)$. Conversely, if $C'_{t'} \neq C_t$ for at least one value of t, the two codes may or may not give the same performance, depending upon the error statistics of the transmission system.

IV. CODES EQUIVALENT TO THE NATURAL BINARY REPRESENTATION

Because of the considerable structure in the natural binary representation, it is easy to use (15) to compute each C_t, $1 \leqq t \leqq 2^k - 1$. For a given t, we first find σ by (11), that is, $\sigma - 1$ is the largest power of 2 in t. Then, for each j, $\sigma + 1 \leqq j \leqq k$, we determine g_j and the $\gamma_{j,i}(t)$. For the natural binary representation,

$$\gamma_{j,i}(t) = 2^{\sigma-1} \tag{17}$$

for $1 \leqq i \leqq g_j$ so $g_j = 2^{j-\sigma-1}$. Therefore, by (15) and (17),

$$C_t = 2^{2\sigma-1} + 2 \sum_{j=\sigma+1}^{k} \sum_{i=1}^{2^{j-\sigma-1}} 2^{2\sigma-2} = 2^{k+\sigma-1}. \tag{18}$$

Notice that each C_t, $2^{\sigma-1} \leqq t \leqq 2^\sigma - 1$, is equal to $2^{k+\sigma-1}$. Thus, C_t is determined simply by the largest power of 2 in t. Substituting (18) into (9) and rewriting, we obtain

$$ANE_B = B \sum_{\sigma=1}^{k} 2^{\sigma-1} \sum_{t=2^{\sigma-1}}^{2^\sigma-1} \Pr[B_k(t) \mid B_k(0)] \tag{19}$$

where ANE_B denotes the average numerical error for the natural binary representation.

Is it possible to find a Harper code H that is not the natural binary representation but that exhibits performance that is identical to the natural binary representation for all transmission systems that satisfy our model? The answer is yes. We now show that a necessary and sufficient condition is that

$$\gamma_{j,i}(t) = 2^{\sigma-1} \tag{20a}$$

for $1 \leqq i \leqq g_j$ and

$$g_j = 2^{j-\sigma-1} \tag{20b}$$

for each t, $1 \leqq t \leqq 2^k - 1$, and for each j, $\sigma + 1 \leqq j \leqq k$ (where σ is chosen so that $2^{\sigma-1} \leqq t \leqq 2^\sigma - 1$).

If (20) is satisfied, then by (15), $C_t = 2^{k+\sigma-1}$. The average numerical error for H (denoted by ANE_H) is

$$ANE_H = B \sum_{\sigma=1}^{k} 2^{\sigma-1} \sum_{t=2^{\sigma-1}}^{2^\sigma-1} \Pr[H_k(t) \mid B_k(0)]. \tag{21}$$

By the definition of a Harper code and the definition of a level,

$$\sum_{t=2^{\sigma-1}}^{2^\sigma-1} \Pr[H_k(t) \mid B_k(0)] = \sum_{t=2^{\sigma-1}}^{2^\sigma-1} \Pr[B_k(t) \mid B_k(0)]. \tag{22}$$

Therefore, by (19), (21), and (22), $ANE_B = ANE_H$. It follows that (20) is a sufficient condition.

We now show by contradiction that (20) is a necessary condition. Consider the set of coefficients $C_{2^{\sigma-1}}$ for $1 \leqq \sigma \leqq k$. From (15),

$$C_{2^{\sigma-1}} = 2^{2\sigma-1} + 2 \sum_{j=\sigma+1}^{k} \sum_{i=1}^{g_j} \gamma_{j,i}^2(2^{\sigma-1}).$$

The term $2^{2\sigma-1}$ is independent of the particular Harper code used. Thus, we need only consider the summation part. Suppose that it is possible to arrange the $\gamma_{j,i}(2^{\sigma-1})$ so that they are not all equal to $2^{\sigma-1}$ but keep $C_{2^{\sigma-1}} = 2^{k+\sigma-1}$. If this is done, at least one $\gamma_{j,i}(2^{\sigma-1})$ will be less than $2^{\sigma-1}$ and at least one $\gamma_{j,i}(2^{\sigma-1})$ will be greater than $2^{\sigma-1}$. However, in order for one $\gamma_{j,i}(2^{\sigma-1})$ to be less than $2^{\sigma-1}$, there must exist a $\sigma' < \sigma$ such that $\gamma_{j',i'}(2^{\sigma'-1}) > 2^{\sigma'-1}$. But in order for $C_{2^{\sigma'-1}} = 2^{k+\sigma'-1}$, there must be at least one $\gamma_{j',i'}(2^{\sigma'-1}) < 2^{\sigma'-1}$. The argument continues until we reach $\gamma_{j'',i''}(2^0)$ where there must be at least one

$$\gamma_{j'',i''}(2^0) > 2^0. \tag{23}$$

However, in order for $C_{2^0} = 2^k$, (23) implies that there must be at least one $\gamma_{j'',i''}(2^0) < 2^0$, which is impossible. It follows that (20) must hold in order for a Harper code to be equivalent to the natural binary representation.

We can show the existence of a great many Harper codes other than the natural binary representation that satisfy (20) by presenting explicitly the structure implied by (20). At this point, we no longer assume that $H_k(0) = B_k(0)$ but state the structure for any Harper code. List the $H_k(s)$ sequentially as s runs from 0 to $2^k - 1$. For position i, $1 \leqq i \leqq k$, divide the s into 2^{k-i+1} consecutive intervals each of length 2^{i-1}. Let j index the intervals where $0 \leqq j \leqq 2^{k-i+1} - 1$.

A Harper code is equivalent to the natural binary representation if and only if, for every odd numbered interval (j odd), the binary digit in position i is the complement of the binary digit in position i in the immediately preceding even numbered interval (j even). The digit in position i in the even numbered intervals is arbitrary.

The structure is presented graphically in Table II for $k = 5$. The symbol $b_{i,j}$ denotes the binary digit in position i in the jth interval. For odd j, $b_{i,j} = b^*_{i,j-1}$ (where $b^*_{i,j-1} = 1 \oplus b_{i,j-1}$) and, thus, $b^*_{i,j-1}$ is shown in Table II for odd j. For all even j, $b_{i,j}$ can be assigned arbitrarily for each i.

The expression for the average numerical error of the Harper codes that are equivalent to the natural binary representation is interesting. From (21), the set of error probabilities $\Pr[H_k(t) \mid B_k(0)]$ for $2^{\sigma-1} \leqq t \leqq 2^{\sigma} - 1$ (that is, for t in the σ-level) all have the weighting coefficient $2^{\sigma-1}$. Thus, the cost of a particular error pattern is the numerical significance of the most significant bit in error. When one considers unequal error-protection codes, the structure in (21) is very convenient because the protection against transmission errors can be matched to the significance of the bit positions. However, for a Harper code that is not equivalent to the natural binary representation, the average numerical error does not exhibit the above structure. Therefore, unequal error-protection codes appear to be less applicable.

V. THE GRAY CODE AND THE FOLDED BINARY CODE

The Gray code and the folded binary code are of interest because of their possible applicability to numerical data transmission.[7,8] This section shows that both of these codes exhibit performance that is identical to the performance of the natural binary representation for all binary transmission systems that satisfy our model.

Let the k-bit binary representation of s be $B_k(s) = (b_k, b_{k-1}, \cdots, b_1)$ where b_i, $1 \leqq i \leqq k$, is the binary digit in position i and

$$s = \sum_{i=1}^{k} b_i 2^{i-1}.$$

As in Section III, the position numbers are defined in terms of the structure of the code, not the order in which the bits are transmitted. From Ref. 7, the Gray code representation of s, denoted by $G_k(s)$, is $G_k(s) = (b_k, b_k \oplus b_{k-1}, \cdots, b_2 \oplus b_1)$. We show that the Gray code is equivalent to the natural binary representation by showing that the structure of the Gray code conforms with the structure in Table II. Consider position i. As in the construction of Table II, divide the range for s into consecutive intervals each of length 2^{i-1} and number the intervals sequentially from 0 to $2^{k-i+1} - 1$. The binary digit in position i of $G_k(s)$ in an even numbered interval is $b_{i+1} \oplus b_i$ and the

Table II—Structure for a Harper Code Equivalent to the Natural Binary Representation; $k = 5$

	$H_5(s)$				
			Position Number		
s	5	4	3	2	1
0	$b_{5,0}$	$b_{4,0}$	$b_{3,0}$	$b_{2,0}$	$b_{1,0}$
1				↓	$b^*_{1,0}$
2				$b^*_{2,0}$	$b_{1,2}$
3			↓	↓	$b^*_{1,2}$
4			$b^*_{3,0}$	$b_{2,2}$	$b_{1,4}$
5				↓	$b^*_{1,4}$
6				$b^*_{2,2}$	$b_{1,6}$
7		↓	↓	↓	$b^*_{1,6}$
8		$b^*_{4,0}$	$b_{3,2}$	$b_{2,4}$	$b_{1,8}$
9				↓	$b^*_{1,8}$
10				$b^*_{2,4}$	$b_{1,10}$
11			↓	↓	$b^*_{1,10}$
12			$b^*_{3,2}$	$b_{2,6}$	$b_{1,12}$
13				↓	$b^*_{1,12}$
14				$b^*_{2,6}$	$b_{1,14}$
15	↓	↓	↓	↓	$b^*_{1,14}$
16	$b^*_{5,0}$	$b_{4,2}$	$b_{3,4}$	$b_{2,8}$	$b_{1,16}$
17				↓	$b^*_{1,16}$
18				$b^*_{2,8}$	$b_{1,18}$
19			↓	↓	$b^*_{1,18}$
20			$b^*_{3,4}$	$b_{2,10}$	$b_{1,20}$
21				↓	$b^*_{1,20}$
22				$b^*_{2,10}$	$b_{1,22}$
23		↓	↓	↓	$b^*_{1,22}$
24		$b^*_{4,2}$	$b_{3,6}$	$b_{2,12}$	$b_{1,24}$
25				↓	$b^*_{1,24}$
26				$b^*_{2,12}$	$b_{1,26}$
27			↓	↓	$b^*_{1,26}$
28			$b^*_{3,6}$	$b_{2,14}$	$b_{1,28}$
29				↓	$b^*_{1,28}$
30				$b^*_{2,14}$	$b_{1,30}$
31	↓	↓	↓	↓	$b^*_{1,30}$

binary digit in position i in the immediately following odd numbered interval is $b_{i+1} \oplus b_i^* = (b_{i+1} \oplus b_i)^*$. Therefore, from the structure in Table II, the Gray code is equivalent to the natural binary representation.

It is also interesting to consider the folded binary code[8]. Let $F_k(s)$ denote the representation of s. Then $F_k(s) = (b_k, b_k^* \oplus b_{k-1}, \cdots, b_k^* \oplus b_1)$ where $b_k^* = b_k \oplus 1$. As in the case of the Gray code, consider position i and divide the range for s into intervals of length 2^{i-1}. The binary digit in position i of $F_k(s)$ in an even numbered interval is $b_k^* \oplus b_i$. The binary digit in position i in the immediately following odd numbered interval is $b_k^* \oplus b_i^* = (b_k^* \oplus b_i)^*$. Therefore, from the structure in Table II, the folded binary code is equivalent to the natural binary representation.

VI. CONCLUSIONS

The model used in this paper for the binary transmission system is quite general and is satisfied by a wide range of practical systems including many that utilize error-correcting codes. A technique is presented for determining the average numerical error for any Harper code. All Harper codes do not exhibit equal performance for all transmission systems that satisfy the model. Because the performance of a given Harper code is closely related to the error statistics of the transmission system, it does not appear possible to specify a Harper code that is best for all applications. However, a subset of Harper codes is defined such that all codes in the subset give identical performance for all transmission systems covered by the model. The subset is important because it includes the natural binary representation, the Gray code, and the folded binary code. Unequal error-protection codes appear to be particularly applicable to Harper codes in the subset.

APPENDIX A

Contribution of Levels 0 through σ to C_t

To determine the contribution of levels 0 through σ to C_t, we must evaluate

$$r_t + \sum_{j=1}^{\sigma} \sum_{s=2^{j-1}}^{2^j-1} |r_t - s| = \sum_{s=0}^{2^\sigma-1} |r_t - s|.$$

From equation (8), for every s in the range $0 \leqq s \leq 2^{\sigma-1} - 1$, there exists a unique r_t in the range $2^{\sigma-1} \leqq r_t \leqq 2^\sigma - 1$. As s runs from 0 through $2^{\sigma-1} - 1$, every r_t in the range $2^{\sigma-1} \leqq r_t \leqq 2^\sigma - 1$ occurs once and only once. Similarly, as s runs from $2^{\sigma-1}$ through $2^\sigma - 1$, every r_t in the range $0 \leqq r_t \leqq 2^{\sigma-1} - 1$ occurs once and only once. Accordingly,

$$\sum_{s=0}^{2^\sigma-1} | r_t - s | = \sum_{s=0}^{2^{\sigma-1}-1} (r_t - s) + \sum_{s=2^{\sigma-1}}^{2^\sigma-1} (s - r_t) = 2^{2\sigma-1}.$$

APPENDIX B

The Structure of the j-Level of a Harper Code

Consider the set of $H_k(s)$ in the j-level of a Harper code where $2^{j-1} \leqq s \leqq 2^j - 1$. For clarity, Table III illustrates the ideas presented here by applying the ideas to the 4-level of the Harper code in Table I.

Let ρ be an integer, $1 \leqq \rho \leqq j - 1$. For each value of ρ, the j-level can be divided into $2^{j-\rho}$ sets of consecutive values of s each set of length $2^{\rho-1}$. The sets are numbered consecutively from 0 through $2^{j-\rho} - 1$ as follows. Let ξ be an integer, $0 \leqq \xi \leqq 2^{j-\rho-1} - 1$. For each value of ξ, there will be two sets; an even numbered set whose number is of the form 2ξ and an odd numbered set whose number is of the form $2\xi + 1$.

An even numbered set contains the $H_k(s)$ for $2^{j-1} + 2\xi 2^{\rho-1} \leqq s \leqq 2^{j-1} + (2\xi + 1)2^{\rho-1} - 1$ and forms a $(\rho - 1)$-subcube because H is a Harper code. Similarly, an odd numbered set contains the $H_k(s)$ for $2^{j-1} + (2\xi + 1)2^{\rho-1} \leqq s \leqq 2^{j-1} + (2\xi + 2)2^{\rho-1} - 1$ and forms a $(\rho - 1)$-subcube. The important point is that for each value of ξ, a useful relationship exists between set 2ξ and set $2\xi + 1$. Specifically,

TABLE III—DETAILS OF 4–LEVEL OF HARPER CODE IN TABLE I

		$\rho = 1$		$\rho = 2$		$\rho = 3$	
s	$H_4(s)$	Set	ξ	Set	ξ	Set	ξ
8	1 0 0 0	0	0	0	0	0	0
9	1 0 0 1	1	0	0	0	0	0
10	1 0 1 1	2	1	1	0	0	0
11	1 0 1 0	3	1	1	0	0	0
12	1 1 0 0	4	2	2	1	1	0
13	1 1 1 0	5	2	2	1	1	0
14	1 1 0 1	6	3	3	1	1	0
15	1 1 1 1	7	3	3	1	1	0

position 4 position 2

position 3 position 1

the $(\rho - 1)$-subcube formed by set $2\xi + 1$ is in the shadow of the $(\rho - 1)$-subcube formed by set 2ξ. Accordingly, all $H_k(s)$ in set $2\xi + 1$ differ in exactly one position from all $H_k(s)$ in set 2ξ. Denote the position that distinguishes the subcubes by m. Therefore, the 2ξ set consists of $2^{\rho-1}$ elements each of which has the same binary digit in position m. Similarly, the $2\xi + 1$ set consists of $2^{\rho-1}$ elements each of which has in position m the complement of the binary digit in position m in the elements of set 2ξ.

The above sets form what we call a run in position m of length $2^{\rho-1}$ that starts at $2^{j-1} + 2\xi 2^{\rho-1}$ (the first $H_k(s)$ in set 2ξ). The definition in Section III follows from the preceding sentence.

APPENDIX C

Contribution of First $2\gamma_{j,1}(t)$ Values of s in Level j to C_t

From equation (8), as s runs from 2^{j-1} through $2^{j-1} + \gamma_{j,1}(t) - 1$, every r_t in the range $2^{j-1} + \gamma_{j,1}(t) \leqq r_t \leqq 2^{j-1} + 2\gamma_{j,1}(t) - 1$ occurs once and only once. Similarly, as s runs from $2^{j-1} + \gamma_{j,1}(t)$ through $2^{j-1} + 2\gamma_{j,1}(t) - 1$, every r_t in the range $2^{j-1} \leqq r_t \leqq 2^{j-1} + \gamma_{j,1}(t) - 1$ occurs once and only once. Therefore,

$$\sum_{s=2^{j-1}}^{2^{j-1}+2\gamma_{j,1}(t)-1} | r_t - s |$$

$$= \sum_{s=2^{j-1}}^{2^{j-1}+\gamma_{j,1}(t)-1} (r_t - s) + \sum_{s=2^{j-1}+\gamma_{j,1}(t)}^{2^{j-1}+2\gamma_{j,1}(t)-1} (s - r_t) = 2\gamma_{j,1}^2(t).$$

APPENDIX D

Numerical Example to Illustrate Equations (15) *and* (16)

Consider the Harper code given in Table I. We show how to use equation (15) when $t = 2$ and $t = 3$ to find C_2 and C_3, respectively. For $t = 2$, $\sigma = 2$ so, from (15)

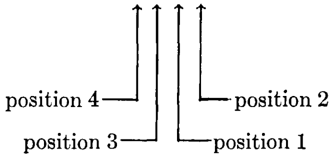
$$C_2 = 8 + 2 \sum_{j=3}^{4} \sum_{i=1}^{g_j} \gamma_{j,i}^2(2).$$

In the 3-level, $\gamma_{3,1}(2)$ and $\gamma_{3,2}(2)$ are shown in Table IV. Therefore, $g_3 = 2$. Also, in the 4-level, $\gamma_{4,1}(2)$, $\gamma_{4,2}(2)$ and $\gamma_{4,3}(2)$ are given in Table IV. Thus, $g_4 = 3$. It follows that

$$C_2 = 8 + 2\,(1^2 + 1^2 + 1^2 + 1^2 + 2^2) = 24.$$

TABLE IV—ILLUSTRATION OF EQUATION (15) APPLIED TO THE HARPER CODE IN TABLE I

s		$H_4(s)$	$\gamma_{j,i}(2)$	$\gamma_{i,j}(3)$
0	0-level	0 0 0 0		
1	1-level	0 0 1 0		
2	2-level	0 0 0 1		
3		0 0 1 1		
4	3-level	0 1 1 1	$\gamma_{3,1}(2) = 1$	$\gamma_{3,1}(3) = 2$
5		0 1 1 0		
6		0 1 0 1	$\gamma_{3,2}(2) = 1$	
7		0 1 0 0		
8	4-level	1 0 0 0	$\gamma_{4,1}(2) = 1$	$\gamma_{4,1}(3) = 2$
9		1 0 0 1		
10		1 0 1 1	$\gamma_{4,2}(2) = 1$	
11		1 0 1 0		
12		1 1 0 0	$\gamma_{4,3}(2) = 2$	$\gamma_{4,2}(3) = 2$
13		1 1 1 0		
14		1 1 0 1		
15		1 1 1 1		



Similarly, for $t = 3$, $\sigma = 2$ so, from (15),

$$C_3 = 8 + 2 \sum_{j=3}^{4} \sum_{i=1}^{g_j} \gamma_{j,i}^2(3).$$

In Table IV, $\gamma_{3,1}(3)$, $\gamma_{4,1}(3)$ and $\gamma_{4,2}(3)$ are given. Thus,

$$C_3 = 8 + 2\,(2^2 + 2^2 + 2^2) = 32.$$

By similar reasoning, the remaining C_t can be found. The expression for the average numerical error of the Harper code in Table I is

$$ANE = \frac{B}{16}\,(24\Pr[1 \mid 0] + 24\Pr[2 \mid 0] + 32\Pr[3 \mid 0] + 64\Pr[4 \mid 0]$$
$$+ 64\Pr[5 \mid 0] + 64\Pr[6 \mid 0] + 64\Pr[7 \mid 0] + 128\Pr[8 \mid 0]$$
$$+ 128\Pr[9 \mid 0] + 128\Pr[10 \mid 0] + 128\Pr[11 \mid 0] + 128\Pr[12 \mid 0]$$
$$+ 128\Pr[13 \mid 0] + 128\Pr[14 \mid 0] + 128\Pr[15 \mid 0]).$$

REFERENCES

1. Harper, L. H., "Optimal Assignments of Numbers to Vertices," J. Soc. Ind. Appl. Math., *12*, No. 1 (March 1964), pp. 131–135.
2. Bernstein, A. J., Steiglitz, K., and Hopcroft, J. E., "Encoding of Analog Signals for Binary Symmetric Channels," IEEE Trans. Inform. Theory, *IT-12*, No. 4 (October 1966), pp. 425–430.
3. Buchner, M. M., Jr., "Coding for Numerical Data Transmission," B.S.T.J., *46*, No. 5 (May-June 1967), pp. 1025–1041.
4. Masnick, B., and Wolf, J. K., "On Linear Unequal Error-Protection Codes," IEEE Trans. Inform. Theory, *IT-13*, No. 4 (October 1967), pp. 600–607.
5. Buchner, M. M., Jr., "A System Approach to Quantization and Transmission Error," B.S.T.J., *48*, No. 5 (May 1969), pp. 1219–1247.
6. Slepian, D., "A Class of Binary Signaling Alphabets," B.S.T.J., *35*, No. 1 (January 1956), pp. 203–234.
7. Gray, F., "Pulse Code Communication," U. S. Patent 2,632,058, March 17, 1953.
8. Dostis, I., "The Effects of Digital Errors on PCM Transmission of Compandored Speech," B.S.T.J., *44*, No. 10 (December 1965), pp. 2227–2243.

Contributors to This Issue

MORGAN M. BUCHNER, JR., B.E.S., 1961, Ph.D., 1965, Johns Hopkins University; U.S. Army active duty, 1966–1968; Bell Telephone Laboratories, 1965–66 and 1968—. Mr. Buchner has been interested in the design and performance of data transmission systems. Member, IEEE, Tau Beta Pi, Sigma Xi, Eta Kappa Nu.

BARRY J. BUNIN, B.E.E., 1963, Cooper Union; M.S., 1964, University of Pennsylvania; Ph.D. work completed, Polytechnic Institute of Brooklyn, 1969; Bell Telephone Laboratories 1963–1966, 1968—. Mr. Bunin has been concerned with the effects of impulse noise and cross-talk on repeater spacing for *Picturephone*® visual telephone applications. He is studying digital encoding of video signals. Member I.E.E.E., Eta Kappa Nu.

P. M. EBERT, B.S., 1958, University of Wisconsin; S.M., 1962, Sc.D., 1965, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1965—. Mr. Ebert has worked on problems in communications and information theory. Member, IEEE.

HARRY M. HALL, M. S. (E. E.), 1960, and Ph.D. (E. E.), 1966, Stanford University; U. S. Navy, 1960–1963; Bell Telephone Laboratories, 1966—. Mr. Hall has been concerned with problems in communication theory and sonar, with emphasis on nonlinear signal processing techniques. Member, IEEE, Tau Beta Pi, Sigma Xi.

D. C. HOGG, B. Sc., 1949, University of Western Ontario; M. Sc., 1950, Ph.D., 1953, McGill University; Bell Telephone Laboratories, 1953—. His work has included studies of artificial dielectrics for microwaves, diffraction of microwaves, and over-the-horizon, millimeter wave and optical propagation, sky noise and low-noise antennas. Fellow, IEEE, Chairman U. S. Commission 2 of Union de Radio Scientifique Internationale, Sigma Xi, American Association for the Advancement of Science.

HERWIG KOGELNIK, Dipl. -Ing. (electronic engineering), 1955, Dr. Techn., 1958, Technische Hochschule Wien, Austria; Ph.D., 1960, Oxford University, England. From 1955 to 1958 he was Assistant Professor at the Institut fur Hochfrequenztechnik in Vienna, engaged in microwave research and teaching. He won a British Council Scholarship to Oxford, 1958 to 1960, where he did research on electromagnetic radiation in magnetoplasmas and anisotropic media.

He joined Bell Telephone Laboratories in 1961, where he has been concerned with laser and holography research. He is head of the Coherent Optics Research Department. Member, American Physical Society, Elektrotechnischer Verein Osterreichs (Austria), IEEE.

Joseph B. Kruskal, Ph.B., 1948, B.S., 1948, M.S., 1949, University of Chicago; Ph.D., 1954, Princeton University; Logistics Research Project, George Washington University, 1950–53; Analytical Research Group, Princeton University, 1954–56; Mathematics Department, University of Wisconsin, 1956–58; Mathematics Department, University of Michigan, 1958–59; Visiting Professor, Statistics Department, Yale University, 1966–67; Bell Telephone Laboratories, 1959—. Mr. Kruskal has done research in several areas of mathematics, including combinatorics, statistics, and computer applications. Currently he is working in statistics, both theoretical and applied. Member, American Mathematical Society, The Classification Society, Mathematical Association of America, Society for Industrial and Applied Mathematics, Psychometric Society, American Statistical Association, Sigma Xi, Pi Mu Epsilon.

James E. Mazo, B.S., 1958, Massachusetts Institute of Technology; M.S., 1960, and Ph.D., 1963, Syracuse University; Research Associate, University of Indiana, 1963–64; Bell Telephone Laboratories, 1964—. Mr. Mazo was engaged in work on quantum scattering theory at Indiana University. Now he is doing theoretical analysis of data systems. Member, American Physical Society, IEEE, Sigma Xi.

Brockway McMillan, B. S. (mathematics), 1936, and Ph.D. (mathematics), 1939, Massachusetts Institute of Technology; instructor at M. I. T. and Princeton University, 1939–1943; Bell Telephone Laboratories, 1946—. Mr. McMillan was Assistant Secretary of the United States Air Force for Research and Development, 1961–63, and Under Secretary of the Air Force, 1963–65.

Mr. McMillan has served with a number of government agencies in advisory and consulting capacities. From September 1958 to March 1959 he was consultant to the White House office reporting to the President's Special Assistant for Science and Technology. Before that he had been associated with the Office of the Assistant Secretary of Defense for Research and Development, the Office of Defense Mobilization, and the Weapons System System Evaluation Group of the joint Chiefs of Staff.

He is now Vice President of Military Development at Bell Telephone Laboratories, and is in charge of the North Carolina Laboratories, Defense Systems, the Field Operations and Support Division, and the Safeguard Design Division.

Fellow, IEEE. Member, Society for Industrial and Applied Mathematics, American Mathematical Society, Mathematical Association of America, Institute of Mathematical Statistics, American Association for the Advancement of Science, National Academy of Engineering.

RANDOLPH J. PILC, B.E.E., 1960, City College of New York; M.E.E., 1962, New York University; Ph.D., 1967, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1960—. Mr. Pilc has been concerned with problems in data transmission and communication theory and is engaged in analysis of data systems. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

HARRISON E. ROWE, B.S., 1948, M.S., 1950, and Sc.D., 1952, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1952—. His fields of interest have included parametric amplifier theory, noise and communication theory, propagation in random media, and related problems in waveguide, radio, and optical systems. Senior member, IEEE; member, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

J. SALZ, B.S.E.E., 1955, M.S.E., 1956, Ph.D., 1961, University of Florida; Bell Telephone Laboratories, 1961—. He first worked on the remote line concentrators for the electronic switching system. He has since engaged in theoretical studies of data transmission systems. During the academic year 1967–68 he was on leave as Professor of Electrical Engineering at the University of Florida. Member, IEEE, Sigma Xi.

IRWIN W. SANDBERG, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of military systems, synthesis and analysis of active and time-varying networks, studies of properties of nonlinear systems, and some problems in communication theory and numerical analysis. He is head of the Systems Theory Research Department. Member, IEEE, Eta Kappa Nu, Sigma Xi, Tau Beta Pi.

Michael G. Taylor, B.A.Sc., 1961, and M.A.Sc., 1962, University of British Columbia, Vancouver, Canada; Ph.D., 1966, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1967—. Mr. Taylor has worked in the areas of digital data processing and communication theory. He is concerned with designing adaptive digital filters for use in digital data transmission systems. Member, IEEE, Tau Beta Pi, Sigma Xi.

B.S.T.J. BRIEF

Solving Nonlinear Network Equations Using Optimization Techniques

By ALLEN GERSHO

(Manuscript received September 10, 1969)

A class of nonlinear equations arising in transistor network analysis, as well as in other areas, has the form

$$f_i(x_i) + \sum_{j=1}^{n} a_{ij}x_j - b_i = 0 \qquad i = 1, 2, \cdots, n \tag{1}$$

or in matrix notation

$$\mathbf{F}(\mathbf{x}) + A\mathbf{x} - \mathbf{b} = \mathbf{0}, \tag{2}$$

where the nonlinearities $f_i(\cdot)$ are continuously differentiable, strictly monotone increasing functions. Results by Willson[1] and Sandberg and Willson[2,3] on nonlinear networks have included broad conditions for the existence and uniqueness of a solution to equation (2). However, convergent computational algorithms for finding the solution have been given only for restricted subclasses of the class of equations that have unique solutions.[1,2,4,5] These subclasses are characterized by a variety of restrictions on the matrix A and on the type of nonlinearities. In this brief we show that a single convergent algorithm exists for solving these equations under conditions virtually as broad as the known existence and uniqueness conditions. Peripherally, we obtain under these conditions a conceptually simple proof of the existence of a solution.

The approach is to use the old technique (probably due to Cauchy) of converting a root-finding problem to a minimization problem. Let

$$\mathbf{r}(\mathbf{x}) \triangleq \mathbf{F}(\mathbf{x}) + A\mathbf{x} - \mathbf{b}, \tag{3}$$

and define the scalar valued "potential" function

$$Q(\mathbf{x}) \triangleq \mathbf{r}^T B\mathbf{r} \tag{4}$$

where B is an arbitrarily chosen symmetric positive definite matrix and T denotes the transpose. Then $Q(\mathbf{x})$ is positive unless x is a solution of equation (2). Consequently, minimizing $Q(\mathbf{x})$ is equivalent to solving equation (2) if in fact the nonlinear equation (2) has a solution.

Since $Q(\mathbf{x})$ is continuous, we may regard it as a continuous surface and observe that if

$$Q(\mathbf{x}) \rightarrow \infty \quad \text{as} \quad ||\mathbf{x}|| \rightarrow \infty \tag{5}$$

the so-called "level sets",

$$\{\mathbf{x} : Q(\mathbf{x}) < c\},$$

are bounded for each number $c > 0$ and there must exist a point $\mathbf{x}^*$ where $Q(\mathbf{x})$ attains a global minimum. Under what conditions will this minimum satisfy $Q(\mathbf{x}^*) = 0$ so that $\mathbf{x}^*$ is a solution of equation (2)? From equations (3) and (4) the gradient of Q is easily found to be

$$\nabla Q(\mathbf{x}) = 2(D_\mathbf{x} + A^T)B\mathbf{r} \tag{6}$$

where $D_\mathbf{x}$ is the positive diagonal matrix whose ith diagonal element is $f_i'(x_i)$ where the prime denotes differentiation. Since the gradient must be zero at a minimum, either (i)

$$\mathbf{r}(\mathbf{x}^*) = \mathbf{0},$$

or (ii)

$$\det \{D_\mathbf{x} + A\} = 0 \quad \text{at} \quad \mathbf{x} = \mathbf{x}^*.$$

If A is in the class of matrices P_0 characterized by the property[3]

$$\det \{D + A\} \neq 0 \text{ for all diagonal matrices } D > 0, \tag{7}$$

it follows that condition (i) holds so that $\mathbf{x}^*$ is a solution of equation (3) for A in P_0 if condition (5) is satisfied. But Theorem 5 of Ref. 2 implies that condition (5) is satisfied if A is in P_0 and the range of the nonlinearities $f_i(\cdot)$ is the entire real line.* Uniqueness of the solution of equation (2) is very simply shown in Ref. 2. Reference 3 shows that the basic condition, A in P_0, is satisfied for large classes of transistor networks.

The minimum of a continuously differentiable function with bounded level sets can always be found by a gradient descent algorithm when the gradient has a unique root.[6] No assumption regarding convexity or the behavior of the Hessian matrix is necessary. Clearly, a sufficiently small change in $\mathbf{x}$ in the negative gradient direction will always decrease the potential $Q(\mathbf{x})$ unless $\mathbf{x}$ is already at a minimum. A sequence of iterations of this type, that is,

* Recently Sandberg[5] has shown that condition (5) holds without any requirements on the range of the nonlinearities if A is nonsingular as well as in P_0.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \nabla Q(\mathbf{x}_k), \tag{8}$$

monotonically reduces the potential $Q(\mathbf{x})$ and yields a bounded sequence of points $\mathbf{x}_k$ because the level sets are bounded. Convergence of the algorithm (8) is assured if the step sizes can be made large enough so that the potential $Q(\mathbf{x}_k)$ approaches zero rather than a positive limit. This can be achieved by making γ_k depend on the size of the gradient in such a way that γ_k cannot approach zero unless the gradient is approaching zero. Goldstein[6] gives the following procedure for selecting γ_k. Define the normalized potential drop:

$$g(\mathbf{x}, \gamma) = \frac{Q(\mathbf{x}) - Q[\mathbf{x} - \gamma \nabla Q(\mathbf{x})]}{\gamma \| \nabla Q(\mathbf{x}) \|^2}, \qquad \gamma > 0, \tag{9}$$

a continuous function of γ which assumes all values between 1 and 0 as γ ranges between zero and some positive value. Then for any δ with

$$0 < \delta < \tfrac{1}{2}$$

choose γ_k so that

$$\delta \leqq g(\mathbf{x}_k, \gamma_k) \leqq 1 - \delta \tag{10}$$

if $g(\mathbf{x}_k, \gamma_k) < \delta$; otherwise let $\gamma_k = 1$. Note that γ_k can be chosen by trial and error computation in each iteration. For small δ few trials are necessary; but the resulting drop in potential in each iteration is smaller so that more iterations are needed. With this method of choosing γ_k, convergence of the algorithm (8) is assured for any starting point $\mathbf{x}_0$.

In summary, using the optimization approach and a result of Ref. 2 we have shown the existence of a solution to equation (2) and the availability of a convergent algorithm to find the solution under the following conditions.

(I) the nonlinearities $f_i(\cdot)$ are continuously differentiable, strictly monotone increasing, and map the whole real line onto itself, and

(II) the matrix A is in the class P_0.

The original existence conditions given in Ref. 2 do not include the "continuously differentiable" assumption but are otherwise identical to conditions I and II above.

REFERENCES

1. Willson, A. N., Jr., "On the Solutions of Equations for Nonlinear Resistive Networks", B.S.T.J., *47,* No. 8 (October 1968), pp. 1755–1773.
2. Sandberg, I. W., and Willson, A. N., Jr., "Some Theorems on Properties of

dc Equations of Nonlinear Networks", B.S.T.J., *48,* No. 1 (January 1969), pp. 1–34.
3. Sandberg, I. W., and Willson, A. N., Jr., "Some Network-Theoretic Properties of Nonlinear dc Transistor Networks", B.S.T.J., *48,* No. 5 (May-June 1969), pp. 1293–1311.
4. Katzenelson, J., and Seitelman, L. H., "On Iterative Method for Solution of Networks of Nonlinear Monotone Resistors", IEEE Trans. Circuit Theory, *CT-13,* No. 3 (September 1966), pp. 317–323.
5. Sandberg, I. W., unpublished work.
6. Goldstein, A. A., *Constructive Real Analysis,* New York: Harper & Row, 1967, pp. 30–32.