THE BELL SYSTEM
Technical Journal

Volume 49 October 1970 Number 8

On the Distribution of Numbers R. W. Hamming 1609

A Mathematical Study of a Model of Magnetic Domain Interactions R. L. Graham 1627

Dielectric Guide with Curved Axis and Truncated Parabolic Index E. A. J. Marcatili 1645

Radiation Losses of the Dominant Mode in Round Dielectric Waveguides D. Marcuse 1665

Excitation of the Dominant Mode of a Round Fiber by a Gaussian Beam D. Marcuse 1695

The Capacity of the Gaussian Channel with Feedback P. M. Ebert 1705

New Theorems on the Equations of Nonlinear DC Transistor Networks A. N. Willson, Jr. 1713

Theorems on the Computation of the Transient Response of Nonlinear Networks Containing Transistors and Diodes I. W. Sandberg 1739

Characterization of Second-Harmonic Effects in IMPATT Diodes C. A. Brackett 1777

An Analysis of Adaptive Retransmission Arrays in a Fading Environment Y. S. Yeh 1811

Microwave Line-of-Sight Propagation With and Without Frequency Diversity W. T. Barnett 1827

Computed Transmission Through Rain at Microwave and Visible Frequencies D. E. Setzer 1873

(Continued inside back cover)

THE BELL SYSTEM TECHNICAL JOURNAL

THE BELL SYSTEM TECHNICAL JOURNAL is published ten times a year by the American Telephone and Telegraph Company, H. I. Romnes, Chairman and President, J. J. Scanlon, Vice President and Treasurer, R. W. Ehrlich, Secretary. 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 49 October 1970 Number 8

On the Distribution of Numbers

By R. W. HAMMING

(Manuscript received March 17, 1970)

This paper examines the distribution of the mantissas of floating point numbers and shows how the arithmetic operations of a computer transform various distributions toward the limiting distribution

$$r(x) = \frac{1}{x \ln b} \qquad (1/b \leqq x \leqq 1)$$

(where b *is the base of the number system). The paper also gives a number of applications to hardware, software, and general computing which show that this distribution is not merely an amusing curiosity. A brief examination of the distribution of exponents is included.*

I. INTRODUCTION

The main purpose of this paper is to examine, from the computing machine's point of view, the well-known (to comparatively few people) unequal distribution of the "mantissas" of "naturally occurring" sets of numbers. The observed probability density distributions are often close to the reciprocal density distribution

$$r(t) = \frac{1}{t \ln b} \qquad (1/b \leqq t \leqq 1), \tag{1}$$

where b is the number base (usually 2, 8, 10, or 16). The corresponding cumulative probability distribution is

$$R(t) = \int_{1/b}^{t} r(x)\,dx = \int_{1/b}^{t} \frac{dx}{x \ln b}$$
$$= \frac{\ln t + \ln b}{\ln b} \tag{2}$$

where, of course,

$$R(1/b) = 0 \quad \text{and} \quad R(1) = 1.$$

From the cumulative distribution, it follows that the probability of observing the leading digit N of a number that is drawn at random from $r(t)$ is

$$R(N+1) - R(N) = \frac{\ln (N+1) - \ln (N)}{\ln b}, \tag{3}$$

and this is usually what is measured in experiments.

A typical experiment is that of tabulating the number of physical constants in a table having a given leading digit (see Table I and Ref. 1, p. 7). The result looks reasonable. Many other examples of observing the reciprocal distribution have been reported. For references see Refs. 2 and 3.

The reciprocal distribution has been explained in many ways. One popular but not immediately obvious explanation for the distribution of physical constants is as follows. Consider the distribution of the leading

TABLE I—THE DISTRIBUTION OF THE LEADING DIGITS OF 50 PHYSICAL CONSTANTS

Leading digit N	Number of cases observed	Expected number eq. (3)	Difference
1	16	15	1
2	11	9	2
3	2	6	−4
4	5	5	0
5	6	4	2
6	4	3	1
7	2	3	−1
8	1	3	−2
9	3	2	1
	50	50	

digits of the set of *all* the physical constants that might occur. If the units of measurement were to be changed then the corresponding leading digit of any particular physical constant would probably change, but it is difficult to believe that the distribution itself would change significantly. To believe so seems to indicate a belief that either the present units of measurement or else the new set have some intimate connection with the real world. An alternative, and more elegant, explanation is given by Roger Pinkham in his classic paper (Ref. 2). The explanation given in the present paper is based on how the computer transforms distributions during arithmetic operations. In particular the paper shows how, from any reasonable distributions, repeated multiplications and/or divisions rapidly move the distributions toward the reciprocal distribution. The effect for addition and subtraction is somewhat different. The paper also shows the persistance of the reciprocal distribution once it is attained.

Since floating point numbers are the basis of most of numerical analysis one may well ask why this obvious and experimentally well-verified distribution is so often ignored. Is it because it appears to contradict the usually accepted model of the number system in which numbers correspond to points on a homogeneous straight line? Not only are the floating point numbers not uniformly spaced in a computer (the difference between the two largest possible numbers is very large, while the distance between the two smallest positive number is very small, and zero is relatively isolated), but the reciprocal distribution shows that even in intervals in which the numbers are equally spaced they are not equally likely to occur.

Thus in analogy with non-Euclidean geometry this paper proposes an alternative to the conventional identification of numbers with points on a homogeneous straight line. Instead of adopting a measure for sets that is invariant under translation

$$x' = x + k,$$

we often prefer a measure that is invariant under scaling, namely

$$x' = kx \qquad (k \neq 0).$$

The reciprocal distribution is of practical as well as theoretical interest as we shall show in Section VII. In view of these examples, it is hoped that by adopting the machine's point of view with respect to how numbers are transformed by arithmetical operations, the computer scientists will become more aware of the importance of this distribution in many situations including numerical analysis.

II. THE MODEL

The floating point numbers in a computing machine form a discrete, finite set. As is true in so many applications of mathematics to practical problems, we shall approximate a discrete distribution by a continuous one of sufficient smoothness. Anyone familiar with the upper and lower Riemann Integral sums can appreciate the degree of approximation being made (provided common sense is used in choosing the values of the curve between the given points). In the limit of the Riemann sum all the $| \Delta x_i |$ become less than any given $\epsilon > 0$; we of course need to stop at the granularity of the number system used, typically 10^{-8} or smaller.

In principle, it is possible to carry this error estimate throughout all the subsequent steps of the mathematics to see how much the mathematics errs from reality; but it is customary to recognize that a little intuition will suffice to convince the user that the error will be much less than the accuracy of the experiments that the theory is designed to account for. Thus we have no need to get excited about such things as the Banach measure of a set (Ref. 4); we do not intend in this paper to let the mathematics obscure what is going on. The fact that computers are finite and operate at a finite speed for a finite length of time spares us from taking seriously all the confusions that can arise in mathematics when dealing with the infinite.

III. THE BASIC FORMULAS

In this section we derive the basic formulas which describe how distributions are combined and transformed by the four arithmetic operations of a computer. Let $f(x)$ be the density distribution of the factor x, $g(y)$ be the density distribution of the factor y, and $h(z)$ be the density distribution of the result z of the arithmetic operation. Further, let $F(x)$, $G(y)$, and $H(z)$ be the corresponding cumulative distributions.

For both multiplication and division, the mantissas are directly combined and the exponents do not enter into the formation of the distribution of the result of the operation. Thus, it is sufficient in these cases to consider the distributions for $(1/b \leqq x, y \leqq 1)$.

For multiplication, an examination of Fig. 1 shows that when the product falls in the shaded regions then the mantissa of the product is in the interval $(1/b, z)$. Thus the cumulative distribution $H(z)$ is given by

$$H(z) = \int_{1/b}^{z} \int_{1/b}^{z/bx} f(x)g(y)\, dy\, dx + \int_{1/b}^{z} \int_{1/bx}^{1} f(x)g(y)\, dy\, dx + \int_{z}^{1} \int_{1/bx}^{z/x} f(x)g(y)\, dy\, dx$$

$$= \int_{1/b}^{z} f(x)[G(z/bx) - G(1/b) + G(1) - G(1/bx)]\,dx$$

$$+ \int_{z}^{1} f(x)[G(z/x) - G(1/bx)]\,dx.$$

Differentiating with respect to z to get the density distribution we have

$$h(z) = f(z)[G(1/b) - G(1/b) + G(1) - G(1/bz) - G(1) + G(1/bz)]$$

$$+ \int_{1/b}^{z} f(x)g(z/bx)(1/bx)\,dx + \int_{z}^{1} f(x)g(z/x)(1/x)\,dx$$

$$= \frac{1}{b}\int_{1/b}^{z} \frac{f(x)}{x}\, g(z/bx)\,dx + \int_{z}^{1} \frac{f(x)}{x}\, g(z/x)\,dx. \tag{4}$$

Similarly for division. The shaded region of Fig. 2 shows where the quotient x/y is less than z; thus the cumulative distribution for the quotient is

$$H(z) = \int_{1/b}^{z}\int_{1/b}^{z} f(x)g(y)\,dy\,dx + \int_{1/b}^{z}\int_{x/z}^{1} f(x)g(y)\,dy\,dx$$

$$+ \int_{z}^{1}\int_{x/bz}^{z} f(x)g(y)\,dy\,dx$$

$$= \int_{1/b}^{z} f(x)[G(x) - G(1/b) + G(1) - G(x/z)]\,dx$$

$$+ \int_{z}^{1} f(x)[G(x) - G(x/bz)]\,dx.$$

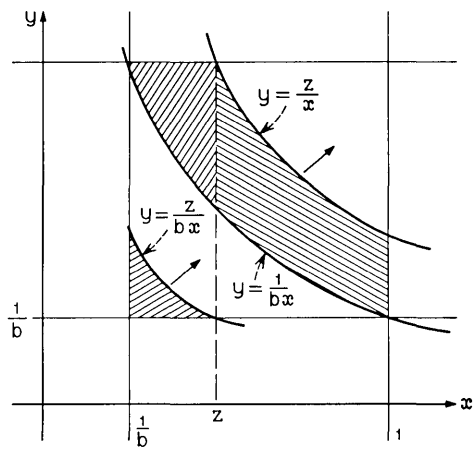


Fig. 1—The cumulative probability distribution for the product $z = xy$.

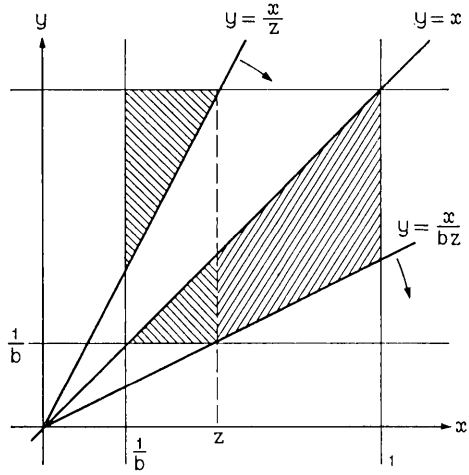


Fig. 2—The cumulative probability distribution for the quotient $z = x/y$.

Again differentiating with respect to z to get the density distribution we have

$$h(z) = f(z)[G(z) - G(1/b) + G(1) - G(1) - G(z) + G(1/b)] + \int_{1/b}^{z} f(x)[-g(x/z)(-x/z^2)]\,dx + \int_{z}^{1} f(x)[-g(x/bz)(-x/bz^2)]\,dx$$

$$= \frac{1}{z^2}\int_{1/b}^{z} xf(x)g(x/z)\,dx + \frac{1}{bz^2}\int_{z}^{1} xf(x)g(x/bz)\,dx. \tag{5}$$

For both addition and subtraction the difference in the exponents of the two numbers x and y is used to shift one mantissa with respect to the other *before* they are combined. For addition, we may suppose that one of the numbers, say x, lies in the range $z/2 \leqq x \leqq z$. The other term, y, therefore lies in the range $z/2 \geqq y \geqq z \cdot b^{-k}$, where k is the number of digits in the mantissa and we set $b^{-k} = \epsilon$. Thus the density distribution of the sum is

$$h(z) = \int_{z/2}^{z(1-\epsilon)} f(x)g(z - x)\,dx. \tag{6}$$

For subtraction we suppose, without loss of generality, that $x \geqq y > 0$, and

$$z = x - y$$

with $z \leqq x \leqq z/\epsilon$. Then the density distribution is given by

$$h(z) = \int_z^{z/\epsilon} x)g(z + x)\, dx. \tag{7}$$

We have now derived the basic relations for the density distributions that arise from combining two numbers from arbitrary distributions according to the four arithmetic operations of a computer.

IV. THE PERSISTENCE OF THE RECIPROCAL DISTRIBUTION

In this section, we first show for both multiplication and division that if one of the factors x or y comes from the reciprocal distribution, and regardless of the distribution of the other factor, then $h(z)$ is the reciprocal distribution. In particular, if a number is chosen from the reciprocal distribution, then its reciprocal is also from the reciprocal distribution. For addition and subtraction we show somewhat less.

For the product set

$$g(y) = \frac{1}{y \ln b} \tag{8}$$

in equation (4). We get for any distribution $f(x)$

$$\begin{aligned} h(z) &= \frac{1}{b} \int_{1/b}^{z} \frac{f(x)}{x} \cdot \frac{bx}{z \ln b}\, dx + \int_z^1 \frac{f(x)}{x} \frac{x}{z \ln b}\, dx \\ &= \frac{1}{z \ln b} \left[\int_{1/b}^{z} f(x)\, dx + \int_z^1 f(x)\, dx \right] = \frac{1}{z \ln b}. \end{aligned} \tag{9}$$

Obviously since $z = xy$, the same applies if we assume that $f(x)$ is the reciprocal distribution.

For the quotient, again assume equation (8) and put it in equation (5).

$$\begin{aligned} h(z) &= \frac{1}{z^2} \int_{1/b}^{z} xf(x) \frac{z}{x \ln b}\, dx + \frac{1}{bz^2} \int_z^1 xf(x) \frac{bz}{x \ln b}\, dx \\ &= \frac{1}{z \ln b} \left\{ \int_{1/b}^{z} f(x)\, dx + \int_z^1 f(x)\, dx \right\} = \frac{1}{z \ln b}. \end{aligned} \tag{10}$$

In the special case of $f(x)$ being the "spike distribution" with all of its probability at $x = 1$ we see that the reciprocal of a variable having the reciprocal distribution has the reciprocal distribution. The case of x having the reciprocal distribution and producing the reciprocal distribution, regardless of the distribution of the denominator, is covered by the product form, or can be worked out directly if desired.

Thus, if in a long sequence of multiplications and divisions at least one

factor has the reciprocal distribution, then regardless of how the distributions of the other factors are chosen the result is still the reciprocal distribution; the reciprocal distribution persists under multiplication and division and cannot be broken by any choices for the other factors.

For addition let x come from the reciprocal distribution for some range with normalization factor N_1 , and y also come from a reciprocal distribution with its corresponding range and normalization factor N_2 . Then writing $\epsilon = b^{-k}$

$$\begin{aligned} h(z) &\quad \int_{z/2}^{z(1-\epsilon)} \frac{N_1}{x} \cdot \frac{N_2}{z-x}\, dx \\ &= N_1 N_2 \int_{z/2}^{z(1-\epsilon)} \frac{1}{z}\left[\frac{1}{x} + \frac{1}{z-x}\right] dx \\ &= \frac{N_1 N_2}{z} \ln\left[\frac{x}{z-x}\right]\Bigg|_{z/2}^{z(1-\epsilon)} \\ &= \frac{N_3}{z} \end{aligned} \tag{11}$$

where N_3 is some constant.

Similarly for subtraction (different N_i)

$$\begin{aligned} h(z) &= \int_{z}^{z/\epsilon} \frac{N_1}{x} \cdot \frac{N_2}{z+x}\, dx \\ &= N_1 N_2 \int_{z}^{z/\epsilon} \frac{1}{z}\left[\frac{1}{x} - \frac{1}{z+x}\right] dx \\ &= \frac{N_1 N_2}{z} \ln\left[\frac{x}{z+x}\right]\Bigg|_{z}^{z/\epsilon} \\ &= \frac{N_3}{z}. \end{aligned} \tag{12}$$

It should be noted, however, that in the last two cases the assumption of the reciprocal distribution for such great ranges is suspicious to say the least, since we know from experience that all exponents are not equally likely. That the reciprocal distribution over a large range implies the equally likely distribution of the relevant exponents can be seen by examining the base 16 number system in exponents, but where the mantissas are in binary. Thus the mantissas can have one of the forms:

$$0.1xxx\cdots$$
$$0.01xx\cdots$$
$$0.001x\cdots$$
$$0.0001\cdots .$$

If we assume

$$p(x) = \frac{1}{x \ln 16} \qquad (\tfrac{1}{16} \leqq x \leqq 1),$$

what are the probabilities of each of the four forms? For the first one

$$\int_{\frac{1}{2}}^{1} \frac{1}{x \ln 16}\, dx = \frac{1}{4 \ln 2} [\ln 1 - \ln \tfrac{1}{2}]$$
$$= \tfrac{1}{4}.$$

Similarly, each of the others is $\frac{1}{4}$. This result is quite different from that of the flat distribution (see Table II).

V. THE APPROACH TO THE RECIPROCAL DISTRIBUTION

Having shown that once it arises the reciprocal distribution persists for multiplication and division, we need to show how it can arise. For this we need a measure of how far a distribution $h(z)$ is from the reciprocal distribution $r(z)$. It is obvious that

$$\int_{1/b}^{1} [h(z) - r(z)]\, dz = 0 \tag{13}$$

for any $h(z)$ and this does not provide a useful measure of distance. We shall define the distance of $h(z)$ from the reciprocal distribution $r(z)$ by

$$\max_{1/b \leqq z \leqq 1} \left| \frac{h(z) - r(z)}{r(z)} \right| \equiv D\{h(z)\} = D\{h\}, \tag{14}$$

which measures the maximum of the difference *relative* to the reciprocal distribution (it is natural to use the relative error when dealing with floating point numbers).

TABLE II—PROBABILITY OF OBSERVING MANTISSAS WITH LEADING ZEROS IN BASE 16 NUMBERS WHEN WRITTEN IN BASE 2

Form	Range	Binary Exponent	Probabilities	
			Flat	Reciprocal
0.0001....	$1/16 \leqq x \leqq 1/8$	−3	1/15	1/4
0.001x....	$1/8 \leqq x \leqq 1/4$	−2	2/15	1/4
0.01xx....	$1/4 \leqq x \leqq 1/2$	−1	4/15	1/4
0.1xxx....	$1/2 \leqq x \leqq 1$	0	8/15	1/4

We showed in equation (9) that for a product,

$$r(z) = \frac{1}{b}\int_{1/b}^{z} \frac{f(x)}{x} r(z/bx)\, dx + \int_{z}^{1} \frac{f(x)}{x} r(z/x)\, dx.$$

Subtracting this from equation (4) and dividing by $r(z)$ we have

$$\frac{h(z) - r(z)}{r(z)} = \frac{1}{b}\int_{1/b}^{z} \frac{f(x)}{x}\left[\frac{g(z/bx) - r(z/bx)}{r(z)}\right] dx + \int_{z}^{1} \frac{f(x)}{x}\left[\frac{g(z/x) - r(z/x)}{r(z)}\right] dx.$$

But

$$bxr(z) = \frac{bx}{z \ln b} = r(z/bx)$$

$$xr(z) = \frac{x}{z \ln b} = r(z/x),$$

and we have

$$\frac{h(z) - r(z)}{r(z)} = \int_{1/b}^{z} f(x)\left[\frac{g(z/bx) - r(z/bx)}{r(z/bx)}\right] dx + \int_{z}^{1} f(x)\left[\frac{g(z/x) - r(z/x)}{r(z/x)}\right] dx. \tag{15}$$

Since $f(x) \geqq 0$ for $(1/b \leqq x \leqq 1)$,

$$\left|\frac{h(z) - r(z)}{r(z)}\right| \leqq \int_{1/b}^{z} f(x)\, D\{g\}\, dx + \int_{z}^{1} f(x)\, D\{g\}\, dx \leqq D\{g\}$$

for all z. From this it follows that

$$D\{h\} \leqq D\{g\} \tag{16}$$

regardless of the choice of $f(x)$.

We note that the equality would hold if $f(x)$ were a single spike at $x = 1$, say, but that in view of equation (13), we generally expect a great deal of cancellation in the square brackets of equation (15) as it is integrated over the range.

It is easy to examine the rapidity of the approach in the case of all the factors coming from the flat distribution

$$p(x) = \frac{1}{1 - 1/b} = \frac{b}{b - 1}.$$

Equation (14) gives for two factors

$$h(z) = \frac{1}{b}\left(\frac{b}{b-1}\right)^2 \int_{1/b}^{z} \frac{dx}{x} + \left(\frac{b}{b-1}\right)^2 \int_{z}^{1} \frac{dx}{x}$$

$$= \frac{b}{(b-1)^2}\{\ln b - (b-1)\ln z\}.$$

In the base $b = 10$, this is

$$h(z) = \tfrac{10}{81}\{\ln 10 - 9 \ln z\}, \tag{17}$$

which (for the proper range) is given by Ref. 5 (p. 37). The distance of the flat distribution is

$$\max_{1/10 \leqq x \leqq 1} \left| z \frac{10 \ln 10}{9} - 1 \right| = \frac{10 \ln 10}{9} - 1 = 1.558\ldots$$

while the distance of equation (17) is equal to $0.3454\cdots$. See Table III for further results.

Similarly for division using equations (10) and (5), we have

$$\frac{h(z) - r(z)}{r(z)} = \frac{1}{z^2}\int_{1/b}^{z} xf(x)\left[\frac{g(x/z) - r(x/z)}{r(z)}\right] dx$$

$$+ \frac{1}{bz^2}\int_{z}^{1} xf(x)\left[\frac{g(x/bz) - r(x/bz)}{r(z)}\right] dx.$$

But

$$z^2 \frac{r(z)}{x} = r(x/z)$$

$$bz^2 \frac{r(z)}{x} = r(x/bz),$$

and we have

$$\left| \frac{h(z) - r(z)}{r(z)} \right| \leqq D\{g\}\left\{\int_{1/b}^{z} f(x)\,dx + \int_{z}^{1} f(x)\,dx\right\}$$

TABLE III—THE DISTANCE OF A CONTINUED PRODUCT AS A FUNCTION OF THE NUMBER OF FACTORS SELECTED FROM A FLAT DISTRIBUTION

Number of Factors	Distance
1	1.558
2	0.3454
3	0.0980
4	0.0289

or

$$D\{h\} \leqq D\{g\}.$$

In the case of flat distributions

$$h(z) = \frac{1}{2(b-1)}\left[b + \frac{1}{z^2}\right]$$

which for the base 10 is (see Ref. 5, p. 37)

$$h(z) = \frac{1}{18}\left[10 + \frac{1}{z^2}\right]$$

and has a distance of $0.4071 \cdots$.

For addition we select $g(y)$ as a reciprocal distribution (with suitable normalization factor N), subtract the corresponding equations and divide by $r(z)$ to get

$$\frac{h(z) - r(z)}{r(z)} = \int_{z/2}^{z(1-\epsilon)} \left[f(x)\frac{N_2}{z-x} - \frac{N_1}{x}\frac{N_2}{z-x}\right]\frac{dx}{r(z)}$$

$$= \int_{z/2}^{z(1-\epsilon)} \left[\frac{f(x) - \frac{N_1}{x}}{r(x)}\right]\cdot\left[\frac{N_2}{z-x}\frac{r(x)}{r(z)}\right] dx.$$

But by the mean value theorem for integrals

$$\frac{h(z) - r(z)}{r(z)} = \left[\frac{f(\theta) - \frac{N_1}{\theta}}{r(\theta)}\right]\int_{z/2}^{z(1-\epsilon)} \frac{N_2}{z-x}\cdot\frac{r(x)}{r(z)}\,dx,$$

where $z/2 \leqq \theta \leqq z(1-\epsilon)$. The integral has been shown in equation (11) to be exactly 1. Hence

$$D\{h(z)\} \leqq D\{f(x)\}.$$

A similar derivation works for subtraction.

In view of the dubious assumption of having the reciprocal distribution over a very large range we need to examine more carefully the behavior of the mantissas of sums of numbers selected from some distribution. Let us imagine a Monte Carlo experiment. We select numbers from the range $(0 < a \leqq x \leqq b)$ having the probability density distribution $p(x)$ with mean μ and variance σ^2. Divide the range into n equal intervals

$$(a, a+h), \quad (a+h, a+2h), \cdots, [a+(n-1)h, b],$$

where $h = (b - a)/n$. By counting how many numbers fall in each interval we get estimates of $p(x)$.

Let us add 2^m numbers of this set of numbers. The range for the sum is

$$(2^m a, 2^m b),$$

the mean $\mu_1 = 2^m\mu$ and $\sigma_1^2 = 2^k\sigma^2$. But the central limit theorem says that the distribution of the sum approaches a normal distribution about the mean with half width σ_1 . Suppose, for convenience, that μ fell in the middle of an interval. Then as m increases and we count the number of cases of mantissas in each interval (note that the m in the term 2^m appears in the exponent only) we will find more and more of them will fall in the interval containing μ (which has the same mantissa as μ_1); the distribution approaches a spike! This does not contradict the central limit theorem; it merely says that if $\mu \neq 0$ ($\mu = 0$ is the exceptional case), the distribution contracts as seen from the point of view of floating point numbers. In loose words, standing at the origin and viewing the rapidly receding mean μ_1 , the width of the distribution σ_1 seems to get narrower as compared to the sum—the sum recedes as 2^m, the half width changes as $2^{m/2}$.

VI. RANGE OF EXPONENTS

It is now clear that in order to examine carefully the effect of addition (and subtraction) on the reciprocal distribution, it is necessary to know the distribution of the exponents of the numbers to be combined. Unfortunately at this time about the only model we have is as follows. Assume a distribution of exponents. Under multiplication and division the exponents are added and subtracted (with, due to carries an extra 1 occasionally added, or subtracted) and by the central limit theorem we can expect: (*i*) that the distribution of the exponents will approach a normal distribution (assuming that overflow and underflow do not happen first) and (*ii*) that this distribution will gradually spread out proportional to the square root of the number of operations. Thus, it appears that in practice the distribution of exponents is probably not stationary. Addition tends to eliminate the smaller exponents, while subtraction tends to increase them.

Experience in numerical analysis shows that the range of the output numbers is usually much greater than the range of the input numbers, enough so to make one suspect that the variance increases as indicated in the above model.

As one thinks carefully about the matter of addition and subtraction it seems reasonable to believe that they will not greatly perturb the

reciprocal distribution; and the experimental data from "naturally occurring numbers", which must have included some additions and subtractions, seem to bear out this belief.

The feeling that under repeated additions and subtractions the central limit theorem applies to numbers (which is true), and therefore contradicts the reciprocal distribution of the mantissas, is typical of the "fixed point arithmetic" viewpoint of numbers—we are representing the sums and differences as floating point numbers, and it is the distribution of these mantissas and their possible approach to the reciprocal distribution that is of relevance here.

VII. APPLICATIONS OF THE RECIPROCAL DISTRIBUTION

Besides accounting for the experimentally found distributions, the reciprocal distribution is relevant to many optimization situations.

As a first example,[6] consider the problem of placing the decimal (binary) point in the number representation system in order to minimize the number of normalization shifts after the computation of a product. (It was probably the minimization of normalizing shifts that caused IBM to adopt the base 16 in the system 360). If the point is placed before the first digit, then products of the form

$$\begin{array}{l} 0.xxx\ldots \\ \underline{0.xxx\ldots} \\ 0.0xx\ldots \end{array}$$

will require a shift to normalize the result; while if it is placed after the first digit, then products like

$$\begin{array}{l} x.x\ldots \\ \underline{x.x\ldots} \\ xx.x\ldots \end{array}$$

will require a shift. Clearly these two cases have complementary probabilities. For the reciprocal distributions the probability p of

$$xy \leqq 1/b$$

is

$$\begin{aligned} p &= \int_{1/b}^{1} \int_{1/b}^{1/bx} \frac{1}{x \ln b} \frac{1}{y \ln b}\, dy\, dx \\ &= \int_{1/b}^{1} \frac{1}{\ln^2 b} \left(\frac{\ln 1/bx - \ln 1/b}{x} \right) dx = \int_{1/b}^{1} \frac{1}{\ln^2 b} \left(-\frac{\ln x}{x} \right) dx \\ &= \frac{1}{\ln^2 b} \left\{ -\frac{\ln^2 x}{2} \right\} \Bigg|_{1/b}^{1} = \frac{1}{2}. \end{aligned}$$

But for a flat distribution,

$$p = \left(\frac{b}{b-1}\right)^2 \int_{1/b}^{1} \int_{1/b}^{1/bx} dy\, dx = \left(\frac{b}{b-1}\right)^2 \int_{1/b}^{1} \left(\frac{1}{bx} - \frac{1}{b}\right) dx$$

$$= \left(\frac{b}{b-1}\right)^2 \frac{1}{b}\left[\ln b - \left(1 - \frac{1}{b}\right)\right]$$

$$= \frac{b \ln b - (b-1)}{(b-1)^2}.$$

For $b = 2$ this is

$$p = 2 \ln 2 - 1 \cong 0.38.$$

As a second application, consider the estimation of the effect of the representation error of numbers in base 2 and base 16. In Ref. 7 McKeeman reports that the maximum relative representation error (MRRE) and the average relative representation error (ARRE) are as shown in Table IV, where the average is over the reciprocal distribution.

A third example is the application to roundoff propagation. If x_1 has an error ϵ_1 and x_2 has error ϵ_2, then in the product

$$\begin{array}{l} x_1 + \epsilon_1 \\ \underline{x_2 + \epsilon_2\qquad\qquad} \\ x_1x_2 + x_1\epsilon_2 + x_2\epsilon_1 + \epsilon_1\epsilon_2 \end{array}$$

it is the leading digits that control the estimate of the propagated error. For the reciprocal distribution the mean is

$$\bar{x} = \int_{1/b}^{1} \frac{x}{x \ln b}\, dx = \frac{1 - 1/b}{\ln b} = \frac{b-1}{b \ln b}.$$

For base 2, this is

$$\bar{x} = \frac{1}{2 \ln 2} \cong 0.72134.$$

TABLE IV—MAXIMUM RELATIVE REPRESENTATION ERROR AND AVERAGE RELATIVE REPRESENTATION ERROR

	MRRE	ARRE
binary	$1/2 \times 2^{-37}$	0.18×2^{-37}
octal	2^{-37}	0.21×10^{-37}
hexadecimal	2^{-37}	0.17×2^{-37}

The second moment about the mean is

$$M_2 = \frac{1}{\ln b}\int_{1/b}^{1}\frac{(x-\bar{x})^2}{x}\,dx = \frac{b-1}{b^2\ln b}\left\{\frac{b+1}{2} - \frac{b-1}{\ln b}\right\}$$

which for $b = 2$ is

$$M_2 = \frac{1}{4\ln 2}\left(\frac{3}{2} - \frac{1}{\ln 2}\right) \cong 0.020674.$$

For the flat distribution, $\bar{x} = 0.75$ and $M_2 = 0.020833$.

Thus we see that the effect of the reciprocal distribution on the average roundoff propagation is surprisingly small.

Another example in which the reciprocal distribution must be considered is that of producing "random" floating point mantissas. To generate these mantissas we use the earlier result that a long sequence of multiplications of numbers from a flat distribution will approximate a reciprocal distribution. Thus random mantissas can be generated by

$$Y_n = Y_{n-1}\cdot r_n \qquad \text{(shifted)}$$

where r_n is from the usual (flat) random number generator and "shifted" means after each product the leading zeros are shifted off. How well does this work? Experimental verification* is given by 8192 trials. Counting the number of mantissas falling in each of N catagories (see Table V).

The last two columns of Table V give the sign changes observed in the difference between the observed and theoretical reciprocal distribution. The expected number of sign changes might be expected to be $(N - 1)/2$, but since for $N = 2$ it is clear that one sign change will occur (because the mean of the residuals is zero) we have used $N/2$ as the expected number. The chi-square test shows that the two distributions are close; the sign change test shows that the residuals are not systematically distributed. From these tests, we see that the generator "works." It is interesting to note that the period of this generator may well be much longer than that of the underlying flat random number generator.

It is easy to see as a general rule that when we try to optimize a library routine for minimum mean running time (as against the Chebyshev minimax run time) we need to consider the distribution of the input data. Hence floating point numerical routines need to consider the reciprocal distribution; the square root, log, exponential, and sine

* Thanks to Brian Kernighan.

TABLE V—DISTRIBUTION OF 8192 RANDOM MANTISSAS

N	χ^2	Degrees of Freedom	Residuals	
			Sign Changes	Expected
64	61.392	63	30	32
32	22.804	31	14	16
16	11.150	15	8	8
8	7.724	7	5	4
4	3.261	3	2	2
2	1.467	1	1	1

are all examples. In the case of the exponential and sine, some study of the exponents is also necessary.

REFERENCES

1. *Handbook of Mathematical Functions*, AMS 55, Nat. Bureau of Standards, 1964.
2. Pinkham, R. S., "On the Distribution of First Significant Digits," Annal. Math. Statistics, *32* (1961), pp. 1223–1230.
3. Adhikari, A. K., and Sarkar, B. P., "Distribution of most Significant Digit in Certain Functions Whose Arguments are Random Variables," Indian J. of Statistics, Series B, *30*, Parts 1 and 2 (1968), pp. 47–58.
4. Raimi, R. A., "On the Distribution of First Significant Digits," Amer. Math. Monthly, *74*, No.2 (February 1969), pp. 342–348.
5. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, New York: McGraw-Hill, 1962.
6. Hamming, R. W., and Mammel, W. L., "A Note on the Location of the Binary Point in a Computing Machine," IEEE Trans. Electronic Computers, *EC-14*, No. 2 (February 1965), pp. 260–1.
7. McKeeman, W. M., "Representation Error for Real Number in Binary Computer Arithmetic," IEEE Trans. Electronic Computers, *EC-16*, No. 6 (June 1967), pp. 682.

A Mathematical Study of a Model of Magnetic Domain Interactions

By R. L. GRAHAM

(Manuscript received March 18, 1970)

In this paper, we initiate a study into the combinatorial aspects of a model of the interactions between discrete magnetic domains and their potential use in information processing devices. Starting with a simple model suggested by W. Shockley, we demonstrate certain (surprising) capabilities as well as inherent limitations upon the possible applications of the interactions described by this model. It should be noted that this simple model does not take into account all of the possible interactions between magnetic domains.

I. INTRODUCTION

The subject of discrete magnetic domains in certain orthoferrite materials has been under active investigation during the past several years, both from a theoretical physical viewpoint as well as that of the device-oriented physicist (for example, see Refs. 1–6). Considerable progress has resulted from these efforts, although needless to say, the end is certainly not in sight. Particular attention has been directed toward the problem of applying this new technology to the very important area of information processing devices, an area in which it seems to have natural and significant applications.[1,7] It is our intention in this paper to examine certain mathematical aspects of these applications for a simple model of magnetic domain interactions suggested by W. Shockley.

II. DESCRIPTION OF THE MODEL

We shall begin by giving a very brief description of the physical situation and its translation into the mathematical model under consideration. The reader whose interests motivate him to seek a more technical explanation is referred to Refs. 6 or 8.

Roughly speaking, thin platelets of certain orthoferrite materials

possess the property that under suitable (magnetic) conditions, small ($\sim$ 3 mils) discrete cylindrical magnetic domains, hereafter called "bubbles", may be stably supported. Moreover, these bubbles may be manipulated by the application of external magnetic fields as well as by their own mutual interaction (which in general causes two bubbles to repel one another). In a suitable physical environment, the location of a bubble in a piece of orthoferrite can be restricted to a finite set of possible positions within the material; these are ordinarily arranged in a rectangular array. It is possible to apply a local magnetic field to specific locations within the array with the following results:†

(*i*) If a bubble already occupies the position at which the field was applied, then nothing happens.

(*ii*) If no bubble occupies the position at which the field was applied *and* no bubble occupies any "nearby" position as well, then (still) nothing happens.

(*iii*) If no bubble occupies the position at which the field was applied but at least one bubble occupies some "nearby" position, then some bubble at a nearby position will leave its original position and now occupy the position selected by the field.

To eliminate the annoying indeterminancy in item (*iii*) it is possible to apply "holding" fields to all but one of the "nearby" sites which has the effect that only a bubble at the unheld position can move.

The mathematical model which will correspond to the preceding description will be phrased in the terminology of *graph theory*. The discrete positions at which bubbles may lie correspond to the set V of *vertices* of a graph G. Two sites which are "nearby" or "adjacent" to one another (this is assumed to be a symmetric relation) correspond to two vertices of G which are joined by an *edge* of G. Suppose bubbles are located at (the sites corresponding to) the subset X of vertices V. We define a *command* to be a *directed edge* $e = (v_1, v_2)$ with v_1 and v_2 adjacent vertices of G. The command e transforms the locations of the bubbles from X to X^e where

$$X^e = \begin{cases} X - \{v_1\} \cup \{v_2\} & \text{if} \quad v_1 \in X, \quad v_2 \notin X; \\ X, & \text{otherwise.} \end{cases}$$

In other words, if there is a bubble at v_1 but no bubble at v_2 and the

† Of course, "careless" application of a magnetic field to an orthoferrite with bubbles can annihilate bubbles, create bubbles, split bubbles in two, deform bubbles into strips, and so on; but these pathological (though certainly useful) operations will not be considered in our model.

command $e = (v_1, v_2)$ is applied to X, then the bubble at v_1 is moved to v_2. Otherwise, the command e has no effect on X. A *program* is defined to be sequence $P = (e_1, e_2, \cdots, e_r)$ of commands e_i. In general, a program P maps the set 2^V of all subsets of V into itself by $X^P = (\cdots(X^{e_1})^{e_2})\cdots)^{e_r}$. It is the purpose of this paper to investigate the mathematical properties of these maps.

III. SOME BASIC PROPERTIES OF PROGRAMS

We begin by making the assumption that G is the *complete* graph on n vertices, that is, *all* pairs of vertices of G are joined by an edge.† As mentioned in the previous section, a program P is a sequence of directed edges $(e_1, e_2, \cdots, e_r)$ and P acts on a subset X of the vertices V of G by

$$X^P = (\cdots((X^{e_1})^{e_2})\cdots)^{e_r}$$

where for $e = (v, v')$,

$$X^e = \begin{cases} X - \{v\} \cup \{v'\} & \text{if} \quad v \in X, \quad v' \notin X; \\ X, & \text{otherwise.} \end{cases}$$

If $X \subseteq V$ then $|X|$ denotes the cardinality of X. We note

Fact 1: For all $X \subseteq V$, and all programs P, $|X^P| = |X|$.

This follows immediately from the definition of X^P.

The first interesting result we state is due to W. Shockley who called it the

Non-decreasing Overlap Theorem: (Shockley) For all X, $Y \subseteq V$ and all programs P,

$$|X^P \cap Y^P| \geqq |X \cap Y|.$$

Proof: Assume for some $P = (e_1, \cdots, e_r)$ and subsets $X, Y \subseteq V$ we have $|X^P \cap Y^P| < |X \cap Y|$. Since $X^P = (\cdots((X^{e_1})^{e_2})\cdots)^{e_r}$, there must exist a *least* j such that

$$|X^{P_{j+1}} \cap Y^{P_{j+1}}| < |X^{P_j} \cap Y^{P_j}|$$

where P_k denotes the program $(e_1, \cdots, e_k)$. Thus, for $\hat{X} = X^{P_j}$, $\hat{Y} = Y^{P_j}$ and $e = e_{j+1} = (a, b)$ we have

$$|\hat{X}^e \cap \hat{Y}^e| < |\hat{X} \cap \hat{Y}|.$$

† Nothing essential is lost by this simplifying assumption. The vertices and edges of the present model should not be confused with any incidental physical vertices or edges in a particular device. An edge of the model may be generated for example by transferring bubbles from a storage zone to an interaction zone and then returning the resultant to the storage zone.

If $c \neq a$, $c \neq b$ then $c \in \hat{X} \cap \hat{Y}$ implies $c \in \hat{X}^e \cap \hat{Y}^e$. If either $a \in \hat{X} \cap \hat{Y}$ or $b \in \hat{X} \cap \hat{Y}$ but not both then $b \in \hat{X}^e \cap \hat{Y}^e$. If both $a \in \hat{X} \cap \hat{Y}$ and $b \in \hat{X} \cap \hat{Y}$ then $a \in \hat{X}^e \cap \hat{Y}^e$ and $b \in \hat{X}^e \cap \hat{Y}^e$. Hence, in any case

$$| \hat{X}^e \cap \hat{Y}^e | \geqq | \hat{X} \cap \hat{Y} |$$

which is a *contradiction*. This proves the theorem.

Shockley noted that this result shows that there is no *replicating program* P^*. By a replicating program, we mean the following: Starting with two fixed sets of vertices V' and V'' with $V' \cap V'' = \varnothing$ and 1-to-1 map θ: $V'' \to V'$, we require that for each $X \subseteq V$,

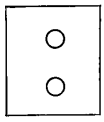
$$X^{P^*} \cap V' = X \cap V' \quad \text{and} \quad \theta(X^{P^*} \cap V'') = X \cap V'.$$

In other words, P^* does not disturb $X \cap V'$ and in V'', P^* creates a "copy" of $X \cap V'$.

To show this, suppose there were such a program P^*. By choosing two subsets X and X' differing in a single element of V, their images X^{P^*} and X'^{P^*} must differ in *two* points, namely, one in V' and the corresponding point (under θ) in V''. This, however, contradicts the non-decreasing overlap (NDO) theorem and therefore P^* cannot exist.

Another consequence of the NDO theorem is the nonexistence of a program P^+ which performs binary addition in the following way.

Suppose V' denotes a set of $m \geqq 1$ pairs of vertices of G, V'' denotes another set of m pairs of vertices disjoint from V', and V''' denotes a set of $m + 1$ pairs of vertices, disjoint from V' and V''. We can imagine these sets arranged as shown in Fig. 1.

We can represent an integer M, $0 \leqq M < 2^m$, in the m pairs of V' by letting the jth pair of V' denote the jth binary digit in the binary expansion of M. This can be done, for example, by assuming that for each pair 

, either $U_0 \in X$, $U_1 \notin X$, which will correspond to a 0, or $U_0 \notin X$, $U_1 \in X$, which will correspond to a 1. Thus, for $m = 5$ the configuration $\{V_1, U_2, U_3, V_4, V_5\}$ (Fig. 2) would denote the integer $10011_{(2)} = 19$.

The addition program P^+ would operate by starting with V''' in some fixed configuration (for example, all zeros) and with arbitrary integers M', M'' loaded into V', V'', respectively, to form the initial state X; after applying P^+ to X we should get the sum $M' + M''$ in V'''.

The reason that P^+ cannot exist as described is precisely that the NDO theorem would be violated. For consider the two additions:

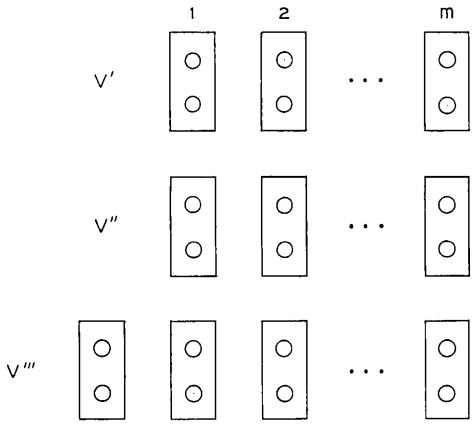


Fig. 1—Symbolic arrangement of vertex locations for addition.

$0 + (2^m - 1) = 2^m - 1$ and $1 + (2^m - 1) = 2^m$. The initial configurations differ in only *two* positions. The final configurations differ in at least $m + 1$ however, since $2^m - 1 = \overbrace{11 \cdots 1}^{m}{}_{(2)}$ and $2^m = \overbrace{100 \cdots 0}^{m}{}_{(2)}$. Thus, by the NDO theorem we get a contradiction and our assertion is proved.

We give another example of a program which does not exist. If $e = (a, b)$ is a command and $a, b \in X$ then $X^e = X$. In the case that a and b are both in X, we say that there is *interference* as e acts on X. (We can think of the bubble at b as interfering with the attempted movement of the bubble at a to vertex b.) Similarly, if $P = (e_1, \cdots, e_n)$ we say that there is interference as P acts on X if for some i there is interference as e_i acts on $X^{e_1, \cdots, e_i}$. We note

Fact 2: If P acts on X with no interference then

$$X^P = \bigcup_{x \in X} \{x\}^P.$$

Proof: It is sufficient to establish this for the case $P = e = (a, b)$. In this case

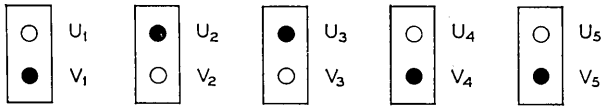


Fig. 2—A typical configuration representing an integer.

$$\{x\}^P = \begin{cases} b, & \text{if } x = a; \\ x, & \text{otherwise.} \end{cases}$$

Thus

$$\bigcup_{x \in X} \{x\}^P = \begin{cases} X - \{a\} \cup \{b\}, & \text{if } a \in X; \\ X, & \text{otherwise.} \end{cases}$$

But by the hypothesis of no interference, we cannot have both a and $b \in X$. Thus

$$X^P = \begin{cases} X - \{a\} \cup \{b\}, & \text{if } a \in X \\ X, & \text{otherwise} \end{cases} = \bigcup_{x \in X} \{x\}^P.$$

and the fact is established.

Fact 3: For $X = \{a, b, c, z\}$, there does not exist a program P such that

$$\{a, b\}^P = \{c, z\},$$
$$\{b, c\}^P = \{a, z\},$$
$$\{c, a\}^P = \{b, z\}.$$

Proof: Suppose such a P exists. If P acts on these sets with no interference then we would have by Fact 2,

$$\{c, z\} = \{\{a\}^P, \{b\}^P\},$$
$$\{a, z\} = \{\{b\}^P, \{c\}^P\},$$
$$\{b, z\} = \{\{c\}^P, \{a\}^P\},$$

which is impossible since the union of the left-hand sides of the equations cannot equal the union of the right-hand sides. Thus, if $P = (e_1, \cdots, e_n)$ we may assume that there is a *least* i, $1 \leqq i \leqq n$, with $P_{i-1} = (e_1, \cdots, e_{i-1})$ such that e_i acts on at least one of the sets $\{a, b\}^{P_{i-1}}$, $\{b, c\}^{P_{i-1}}$, $\{c, a\}^{P_{i-1}}$ with interference. To be specific, assume that it is the set $\{a, b\}^{P_{i-1}}$, that is, $e_i = (\{a\}^{P_{i-1}}, \{b\}^{P_{i-1}})$ (the other two cases are similar). By Fact 2 we have

$$\{a, b\}^{P_{i-1}} = \{\{a\}^{P_{i-1}}, \{b\}^{P_{i-1}}\},$$
$$\{b, c\}^{P_{i-1}} = \{\{b\}^{P_{i-1}}, \{c\}^{P_{i-1}}\},$$
$$\{c, a\}^{P_{i-1}} = \{\{c\}^{P_{i-1}}, \{a\}^{P_{i-1}}\}.$$

Therefore

$$\{b, c\}^{P_i} = \{\{b\}^{P_{i-1}}, \{c\}^{P_{i-1}}\}^{e_i} = \{\{b\}^{P_{i-1}}, \{c\}^{P_{i-1}}\}$$

and

$$\{c, a\}^{P_i} = \{\{c\}^{P_{i-1}}, \{a\}^{P_{i-1}}\}^{e_i} = \{\{c\}^{P_{i-1}}, \{b\}^{P_{i-1}}\}.$$

Hence,

$$\{a, z\} = \{b, c\}^P = \{c, a\}^P = \{b, z\}$$

which is a *contradiction.* This proves the Fact 3.

Note that the nonexistence of the program of Fact 3 does not follow directly from Fact 1 or the NDO theorem. A similar argument can be given to show that for $X = (a, b, c, d, A, B, C, D, z)$ there is no program P such that

$$\{a, c\}^P = \{A, z\},$$

$$\{a, d\}^P = \{B, z\},$$

$$\{b, c\}^P = \{C, z\},$$

$$\{b, d\}^P = \{D, z\}.$$

IV. THE 2-VALUED BOOLEAN FUNCTIONS

Our attention will now be focussed on the positive aspects of the model. In particular we shall be concerned with the problem of representing the Boolean functions of m variables with appropriate programs. The way in which a function is to be represented is as follows. Suppose $m = 2$ and consider the function f: $\{0, 1\} \times \{0, 1\} \to \{0, 1\}$ by

x	y	$f(x, y)$
0	0	0
0	1	1
1	0	1
1	1	1

If the values 1 and 0 are interpreted as "true" and "false", respectively, then f is just the truth function of the familiar operation of alternation. V will be the set of six vertices (x_0 , x_1 , y_0 , y_1 , f_0 , f_1) which we indicate in Fig. 3. It is not difficult to show that no generality is lost by assuming there are no additional vertices. In fact, by using the pair of positions x_0 , x_1 in which to observe the result of the program, instead of providing the separate positions f_0 , f_1 , it is true that if a Boolean function of $m \geqq 2$ variables can be represented by a program in this general way, then it can be represented using just $2m$ vertices. The program $P(f)$

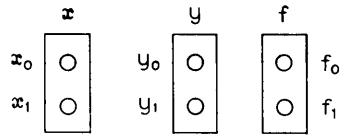


Fig. 3—Symbolic arrangement of vertex locations for computing Boolean functions of two variables.

which represents f is required to have the property that

$$f_0 \in \{x_0 , y_0\}^{P(f)}, \qquad f_1 \notin \{x_0 , y_0\}^{P(f)},$$
$$f_0 \in \{x_0 , y_1\}^{P(f)}, \qquad f_1 \notin \{x_0 , y_1\}^{P(f)},$$
$$f_0 \in \{x_1 , y_0\}^{P(f)}, \qquad f_1 \notin \{x_1 , y_0\}^{P(f)},$$
$$f_0 \notin \{x_1 , y_1\}^{P(f)}, \qquad f_1 \in \{x_1 , y_1\}^{P(f)}.$$

The correspondence between the indices of the vertices of V and the values of the variables of f is immediate. In terms of bubbles, one may think of the configurations shown in Fig. 4 as representing a 0 and 1 respectively (compare Fig. 2); $P(f)$ is required to map each of the four possible initial states of the x_i-pair and y_i-pair into the correct value in the f_i-pair.

It is not difficult in this case to find an appropriate $P(f)$, for example, we can take

$$P(f) = (x_0 , y_0)(x_0 , f_0)(x_1 , y_1)(y_1 , f_1).$$

This is easily checked, as shown in Table I. We can write the preceding result in the shorthand form

f	$P(f)$
(0, 0, 0, 1)	$(x_0 , y_0)(x_0 , f_0)(x_1 , y_1)(y_1 , f_1)$.

Note that if $\bar{f}$ is defined by $\bar{f}(x, y) = 1 - f(x, y)$, that is, $\bar{f}$ is the complement of f, then we can take

$$P(\bar{f}) = P(f)(x_0 , x_1)(x_0 , y_1)(x_0 , y_2)(f_1 , x_0)(f_0 , f_1)(x_0 , f_0)$$

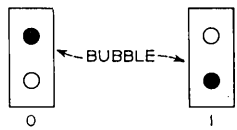


Fig. 4—Configurations which represent 0 and 1.

TABLE I—CUMULATIVE EFFECT OF $P(f)$

(x, y)	(x_0, y_0)	(x_0, f_0)	(x_1, y_1)	(y_1, f_1)	$f(x, y)$
$(0, 0) \leftrightarrow \{x_0, y_0\}$	$\{x_0, y_0\}$	$\{f_0, y_0\}$	$\{f_0, y_0\}$	$\{f_0, y_0\}$	$\leftrightarrow$ 0
$(0, 1) \leftrightarrow \{x_0, y_1\}$	$\{y_1, y_0\}$	$\{y_1, y_0\}$	$\{y_1, y_0\}$	$\{f_1, y_0\}$	$\leftrightarrow$ 1
$(1, 0) \leftrightarrow \{x_1, y_0\}$	$\{x_0, y_0\}$	$\{x_1, y_0\}$	$\{y_1, y_0\}$	$\{f_1, y_0\}$	$\leftrightarrow$ 1
$(1, 1) \leftrightarrow \{x_1, y_1\}$	$\{x_1, y_1\}$	$\{x_1, y_1\}$	$\{x_1, y_1\}$	$\{x_1, f_1\}$	$\leftrightarrow$ 1

as a program which represents $\bar{f}$ (we leave this to the reader to verify). Table II, together with this remark about $\bar{f}$, show that *all* of the 16 possible 2-valued Boolean functions of two variables can be represented by programs.

A question which naturally arises at this point is whether *all* Boolean functions of m variables can be represented by programs in this manner. For $m = 1$, the answer is in the affirmative (the specific programs are left to the reader to discover); for $m = 2$, we have given the required 16 programs; for $m = 3$, the answer is in the affirmative but the number ($2^{2^3} = 256$) of programs prohibits their listing here; for $m = 4$, the answer is once again in the affirmative but the calculations necessary to establish this are much too long to be exhibited (there are, after all, $2^{2^4} = 65536$ functions to consider). The cases $m = 3$ and $m = 4$ were established by J. H. Spencer.[9]

One may note that since all Boolean functions of two variables can be represented, then in particular the Sheffer stroke function given by

x	y	$f(x, y)$
0	0	1
0	1	0
1	0	0
1	1	0

TABLE II—PROGRAMS FOR BOOLEAN FUNCTIONS OF 2 VARIABLES

f	$P(f)$
(0, 0, 0, 0)	(x_0, f_0) (x_1, f_0)
(0, 0, 0, 1)	(x_1, y_1) (x_1, f_1) (x_0, f_0) (y_0, f_0)
(0, 0, 1, 0)	(x_1, y_0) (x_1, f_1) (x_0, f_0) (y_1, f_0)
(0, 0, 1, 1)	(x_1, f_1) (x_0, f_0)
(0, 1, 0, 0)	(x_0, y_1) (x_0, f_1) (x_1, f_0) (y_0, f_0)
(0, 1, 0, 1)	(y_0, f_0) (y_1, f_1)
(0, 1, 1, 0)	(x_0, y_0) (x_0, f_0) (x_1, y_1) (y_1, y_0) (x_1, f_0) (y_1, f_1)
(0, 1, 1, 1)	(x_0, y_0) (x_0, f_0) (x_1, f_1) (y_1, f_1)

can also be represented. It is well known that any Boolean function of m variables can be generated by expressions containing just the variables and the stroke function.[10] Hence, one is tempted to conclude that any Boolean function is representable by a program. The flaw in this line of reasoning is that in order to express a particular Boolean function in terms of the stroke function, many occurrences of the stroke function and the variables are usually required. This in turn requires many "copies" of the variables to be available to the program in order to represent f. But we initially have only one pair of positions which indicates the value of any particular variable and by the NDO theorem we have seen that there cannot exist a "replication" program which would form extra copies of the values of the variables. Hence, within this model, we cannot use this technique to generate all the Boolean functions. It is certainly true however that if the model were extended to include bubble interactions which would allow replication of configurations (and such are known to exist physically), then all Boolean functions of m variables could be represented exactly in the manner described.

These initial results create considerable optimism concerning the possibility of representing all the Boolean functions of m variables. Such hopes are shattered however by the result (which we later prove) that *there exists a Boolean function of 11 variables which cannot be represented by any program of this type.* In fact, even though the fraction of the total number of Boolean functions of 11 variables which *can* be represented by programs can be shown to be $< 10^{-163}$, the author is currently unable to exhibit any specific function which cannot be represented. Clearly, our understanding of this is less than complete. It is not unreasonable to hope that the representable functions could eventually be effectively characterized.

We now restrict ourselves (without loss of generality) to representing the Boolean functions of m variables in the following way. We shall take $V = \{x_1 , x_1' , x_2 , x_2' , \cdots x_m , x_m'\}$ to be a set of $2m$ vertices which we imagine to be arranged in pairs as illustrated in Fig. 5. As before, a bubble in the $x_i(x_i')$ location of the pair (x_i , x_i') will denote that the ith variable of the function f has the value 0(1). The way in which a program $P(f)$ represents f is as follows. Choose a distinguished vertex $\alpha \in V$. There is an obvious 1-1 correspondence between $\{0, 1\}^m$ and the class $\bar{C}$ of all subsets $X \subseteq V$ such that X intersects each $\{x_i , x_i'\}$ in exactly one element given by

$$a = (a_1 , \cdots , a_m) \leftrightarrow \{y_i \in V : y_i = x_i \text{ if } a_i = 0,$$
$$y_i = x_i' \text{ if } a_i = 1,\ 1 \leqq 1 \leqq m\} = X.$$

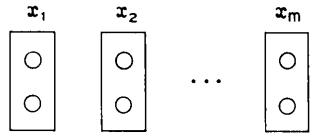


Fig. 5—Symbolic arrangement of vertex locations for computing Boolean functions of m variables.

Let $A_i \subseteq \{0, 1\}^m$, $i = 0, 1$, be the set of all $a \in \{0, 1\}^m$ such that $f(a) = i$, and let $\bar{C}_i$ be the corresponding subsets of $\bar{C}$. Our object is to find a program $P(f)$ which distinguishes between the sets $\bar{C}_0$ and $\bar{C}_1$. (Note that $\bar{C}_0 \cup \bar{C}_1 = \bar{C}$.) Specifically we shall say that $P(f)$ represents f if

$$\alpha \in X^{P(f)} \quad \text{for all} \quad X \in \bar{C}_0 ,$$

$$\alpha \notin X^{P(f)} \quad \text{for all} \quad X \in \bar{C}_1 .$$

Let C denote the subset of all subsets $x \subseteq V$ with $| x | = m$ and for x and y distinct elements of V, let $C(x)$ be the set of elements of C which contain x with $C(y)$ defined similarly.† Consider the effect of the command (x, y) on the members of $C(x)$ and $C(y)$. There are four cases:

(*i*) $X \in C(x)$, $X \in C(y)$.
Then $X^{(x,y)} = X$ and $X^{(x,y)} \in C(x)$, $X^{(x,y)} \in C(y)$.
(*ii*) $X \in C(x)$, $X \notin C(y)$.
Then $X^{(x,y)} = X - \{x\} \cup \{y\}$ and $X^{(x,y)} \notin C(x)$, $X^{(x,y)} \in C(y)$.
(*iii*) $X \notin C(x)$, $X \in C(y)$.
Then $X^{(x,y)} = X$ and $X^{(x,y)} \notin C(x)$, $X^{(x,y)} \in C(y)$.
(*iv*) $X \notin C(x)$, $X \notin C(y)$.
Then $X^{(x,y)} = X$ and $X^{(x,y)} \notin C(x)$, $X^{(x,y)} \notin C(y)$.

Hence, after the application of (x, y) to all the sets in C, the new sets $C'(x)$, $C'(y)$ (which now consist of all the subsets in C which contain x and y respectively) are related to $C(x)$ and $C(y)$ by

$$C'(x) = C(x) \cap C(y),$$

$$C'(y) = C(x) \cup C(y).$$

Stated in these terms, the object of the program $P(f)$ is finally to have $C'^{\cdots'}(\alpha) \cap \bar{C} = \bar{C}_0$ after it has been applied to all the sets in C.

We give an example which illustrates these concepts. Let f be the Boolean function of three variables defined by:

† This approach was first suggested by J. H. Spencer.[9]

x	y	z	$f(x, y, z)$
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

$$V = \{x_1 , x_1' , x_2 , x_2' , x_3 , x_3'\} \quad \text{and we take} \quad \alpha = x_1' .$$

$$\bar{C}_0 = \{\{x_1 , x_2 , x_3\}, \{x_1 , x_2' , x_3\}, \{x_1 , x_2' , x_3'\}, \{x_1' , x_2 , x_3\}\},$$

$$\bar{C}_1 = \{\{x_1 , x_2 , x_3'\}, \{x_1' , x_2 , x_3'\}, \{x_1' , x_2' , x_3\}, \{x_1' , x_2' , x_3'\}\}.$$

A program $P(f)$ which achieves the separation is

$$P(f) = (x_1' , x_2)(x_1' , x_3)(x_2' , x_3)(x_1 , x_3)(x_1 , x_1').$$

That is,

$$X \in \bar{C}_0 \Rightarrow \alpha = x_1' \in X^{P(f)},$$

$$X \notin \bar{C}_0 \Rightarrow x_1' \notin X^{P(f)}.$$

If $C(x)$ denotes the initial subset of C consisting of all the sets in C which contain x then we may conveniently record the sequential changes which occur in each current $C(x)$ in terms of the original $C(y)$'s as the successive commands of $P(f)$ are applied as shown in Table III. A little computation shows that the final set in the x_1'-row, the final $C(x_1')$, when intersected with $\bar{C}$ gives exactly

$$\{\{x_1 , x_2 , x_3\}, \{x_1 , x_2' , x_3\}, \{x_1 , x_2' , x_3'\}, \{x_1' , x_2 , x_3\}\}$$

which equals $\bar{C}_0$ as required.

In general the problem of representing Boolean functions reduces to the following problem. We start with the $2m$ classes $C^{(0)}(y) = C(y) \cap \bar{C}$, $y \in V$. We are then allowed to replace two of the classes $C^{(0)}(y)$ and $C^{(0)}(y')$ by two (possibly) new classes $C^{(0)}(y) \cap C^{(0)}(y')$ and $C^{(0)}(y) \cup C^{(0)}(y')$. We can repeat this operation as many times as desired with any pair of classes currently in the list. Our objective is to eventually generate a specified subset C^* of $\bar{C}$.

We have already mentioned that for $m = 1$, 2, 3 and 4 it is possible

TABLE III—CUMULATIVE EFFECT OF $P(f)$

		(x_1', x_2)	(x_1', x_3)
x_1:	$C(x_1)$	$C(x_1)$	$C(x_1)$
x_1':	$C(x_1')$	$C(x_1') \cap C(x_2)$	$C(x_1') \cap C(x_2) \cap C(x_3)$
x_2:	$C(x_2)$	$C(x_1') \cup C(x_2)$	$C(x_1') \cup C(x_2)$
x_2':	$C(x_2')$	$C(x_2')$	$C(x_2')$
x_3:	$C(x_3)$	$C(x_3)$	$(C(x_1') \cap C(x_2)) \cup C(x_3)$
x_3':	$C(x_3')$	$C(x_3')$	$C(x_3')$

		(x_2', x_3)	(x_1, x_3)
x_1:	$C(x_1)$		$C(x_1) \cap (C(x_2') \cup (C(x_1') \cap C(x_2)) \cup C(x_3))$
x_1':	$C(x_1') \cap C(x_2) \cap C(x_3)$		$C(x_1') \cap C(x_2) \cap C(x_3)$
x_2:	$C(x_1') \cup C(x_2)$		$C(x_1') \cup C(x_2)$
x_2':	$C(x_2') \cap ((C(x_1') \cap C(x_2)) \cup C(x_3))$		$C(x_2') \cap ((C(x_1') \cap C(x_2)) \cup C(x_3))$
x_3:	$C(x_2') \cup (C(x_1') \cap C(x_2)) \cup C(x_3)$		$C(x_1) \cup C(x_2') \cup (C(x_1') \cap C(x_2)) \cup C(x_3)$
x_3':	$C(x_3')$		$C(x_3')$

	(x_1, x_1')
x_1:	$C(x_1) \cap (C(x_2') \cup (C(x_1') \cap C(x_2)) \cup C(x_3)) \cap C(x_1') \cap C(x_2) \cap C(x_3)$
x_1':	$(C(x_1) \cap (C(x_2') \cup (C(x_1') \cap C(x_2))) \cup C(x_3)) \cup (C(x_1') \cap C(x_2) \cap C(x_3))$
x_2:	$C(x_1') \cup C(x_2)$
x_2':	$C(x_2') \cap ((C(x_1') \cap C(x_2)) \cup C(x_3))$
x_3:	$C(x_1) \cup C(x_2') \cup (C(x_1') \cap C(x_2)) \cup C(x_3)$
x_3':	$C(x_3')$

to generate *any* subset of C in this manner. We proceed to show that for $m = 11$, there is a subset of C which cannot be generated. We first need several preliminary observations.

To begin with, for $a, b \in V$, let A and B denote the current sets $C^{(i)}(a)$ and $C^{(i)}(b)$, respectively, after the ith command of the program P has been executed. In other words, at this point in time $C^{(i)}(a)$ is the class of all the original subsets of C which now contain a. For example, if $a = x_2'$ in the preceding example, then after the fifth (and final) command of $P(f)$, $C^{(5)}(x_2')$ is $C^{(0)}(x_2') \bigcap (C^{(0)}(x_2') \bigcap C^{(0)}(x_2) \bigcup C^{(0)}(x_3))$. It is immediate that if $C^{(i)}(a) \subseteq C^{(i)}(b)$ then the application of the command (a, b) as the $(i + 1)$-st command of the program changes nothing. Hence *we can assume that we only use commands* (a, b) *for which at the time of their application* $C^{(i)}(a) \not\subseteq C^{(i)}(b) \not\subseteq C^{(i)}(a)$ (we say that $C^{(i)}(a)$ and $C^{(i)}(b)$ are incomparable).

Initially all the starting classes $C^{(0)}(x)$, $x \in V$, are mutually incomparable. In general suppose we have a family of classes $D = \{A_i ; 1 \leqq i \leqq t\}$, $A_i \subseteq \bar{C}$, with exactly r of the $\binom{t}{2}$ pairs of A_i being comparable and assume A_1 and A_2 are incomparable. Consider the family $D' =$

$D - \{A_1\} - \{A_2\} \cup \{A_1 \cap A_1\} \cup \{A_1 \cup A_2\}$. We wish to determine how many pairs of the classes of D' are comparable. By definition $D' = \{A_1 \cap A_2 , A_1 \cup A_2 , A_3 , A_4 , \cdots, A_t\}$. Of course for $i, j \geqq 3$, the comparability between the classes A_i and A_j remains unchanged. There are several cases:

(*i*) $A_i \supseteq A_1 , A_i \supseteq A_2 .$
Then $A_i \supseteq A_1 \cup A_2 , A_i \supseteq A_1 \cap A_2 .$
(*ii*) $A_i \supseteq A_1 , A_i \not\supseteq A_2 .$
Then $A_i \supseteq A_1 \cap A_2 .$
(*iii*) $A_i \not\supseteq A_1 , A_i \supseteq A_2 .$
Then $A_i \supseteq A_1 \cap A_2 .$
(*iv*) $A_i \subseteq A_1 , A_i \subseteq A_2 .$
Then $A_i \subseteq A_1 \cap A_2 , A_i \subseteq A_1 \cup A_2 .$
(*v*) $A_i \subseteq A_1 , A_i \not\subseteq A_2 .$
Then $A_i \subseteq A_1 \cup A_2 .$
(*vi*) $A_i \not\subseteq A_1 , A_i \subseteq A_2 .$
Then $A_i \subseteq A_1 \cup A_2 .$

Finally, we have a most important *new* comparability in D', namely $A_1 \cap A_2 \subseteq A_1 \cup A_2$. Thus, at least $r + 1$ pairs of classes of D' are comparable. An immediate consequence of this observation is

Fact 4: We can assume that no program $P(f)$ consists of more than $\binom{2m}{2}$ commands.

Proof: Since after i (nontrivial) commands of a program $P(f)$ have been applied, we must have (by induction) at least i pairs of the classes $C^{(i)}(x)$, $x \in V$, being comparable and since there are just $2m$ classes and therefore $\binom{2m}{2}$ pairs of classes, then $P(f)$ must have $\leqq \binom{2m}{2}$ commands.

Theorem. There exists a Boolean function of 11 variables which cannot be represented by a program.

Proof: It is sufficient to show that for $m = 11$, there is a subset C^* of $\bar{C}$ which cannot be generated by starting with the $2m$ classes $C^{(0)}(x)$, $x \in V$, and recursively applying the transformation $A, B \to A \cap B$, $A \cup B$. Consider a typical program $P = (e_1 , e_2 , \cdots , e_t)$ and the corresponding expressions $C^{(i)}(t)$, presented in Table IV.

In choosing the ith command e_i of P there are at most $\binom{2m}{2} - i + 1$ possibilities for e_i since after $(e_1 , \cdots , e_{i-1})$ has been applied, at least $i - 1$ of the pairs $C^{(i-1)}(x)$, $C^{(i-1)}(y)$ are comparable and thus neither (x, y) nor (y, x) can be the next command e_i . Therefore there are at most

$$\prod_{i=1}^{\binom{2m}{2}} \left[\binom{2m}{2} - i + 1\right] = [m(2m - 1)]!$$

TABLE IV—CUMULATIVE EFFECT OF P

P:		e_1	e_2		e_i		e_t
x_1:	$C^{(0)}(x_1)$	$C^{(1)}(x_1)$	$C^{(2)}(x_1)$	$\cdots$	$C^{(i)}(x_1)$	$\cdots \cdots$	$C^{(t)}(x_1)$
x_1':	$C^{(0)}(x_1')$	$C^{(1)}(x_1')$	$C^{(2)}(x_1')$	$\cdots$	$C^{(i)}(x_1')$	$\cdots \cdots$	$C^{(t)}(x_1')$
.	.	.	.	.	.		.
.	.	.	.	.	.		.
.	.	.	.	.	.		.
x_m:	$C^{(0)}(x_m)$	$C^{(1)}(x_m)$	$C^{(2)}(x_m)$	$\cdots$	$C^{(i)}(x_m)$	$\cdots \cdots$	$C^{(t)}(x_m)$
x_m':	$C^{(0)}(x_{m1}')$	$C^{(1)}(x_m')$	$C^{(2)}(x_m')$	$\cdots$	$C^{(i)}(x_m')$	$\cdots \cdots$	$C^{(t)}(x_m')$

choices for the sequence of e_i, since $t \leqq \binom{2m}{2} = m(2m-1)$ by Fact 4. Also, for $i \geqq 1$, each column $C^{(i)}(x)$, $x \in V$, contains at most two new classes which did not occur in the preceding column since only two classes are changed at each step. Hence there are at most

$$[m(2m-1)]!\binom{2m}{2} + 2m$$

classes which can be generated by these rules where the additional term $+2m$ comes from the $2m$ initial sets $C^{(0)}(x)$, $x \in V$. On the other hand, since $\bar{C}$ contains 2^m sets $X \subseteq V$, then there are 2^{2^m} subsets of $\bar{C}$ which we must try to generate. We are doomed to failure however since

$$\left\{[m(2m-1)]!\binom{2m}{2} + 2m\right\} \Big/ 2^{2^m} \to 0$$

as $m \to \infty$. We list these expressions for several small values of m in Table V. Thus, not only are we guaranteed a single Boolean function of 11 variables which cannot be represented by a program, but in fact we have at least 10^{615} of them. It seems quite likely that there exist Boolean functions of five variables which cannot be represented. However, at present, no specific example of a Boolean function is known which cannot be represented by a program.

TABLE V—BOUNDS ON THE NUMBER OF BOOLEAN FUNCTIONS WHICH CAN BE GENERATED

m	$[m(2m-1)]!\binom{2m}{2} + 2m$	2^{2^m}
2	4324	16
3	19615115520006	256
10	$>10^{355}$	$<10^{309}$
11	$<10^{453}$	$>10^{616}$

V. SOME REMARKS

A number of partial results are known concerning the preceding problems which we shall only mention briefly here.

The generation of Boolean functions as described has the following very natural geometrical interpretation. For a fixed integer n, consider the set of the 2^n vertices of an n dimensional cube C^n and let $A_1, \cdots, A_{2n}$ represent the $2n$ sets of 2^{n-1} vertices which each lie on one $(n-1)$-dimensional "face". In other words, if the vertices of C^n are labelled by binary n-tuples in the usual way, then each A_i corresponds to a set of 2^{n-1} n-tuples in which some component is constant. As before, we are allowed to replace any two sets A and B in the class of $2n$ sets of $A \cap B$ and $A \cup B$. We can repeat this transformation as often as desired. The question is: which subsets $X \subseteq C^n$ can be generated in this manner. We have shown that there exists a set $X \subseteq C^{11}$ which cannot be so generated.

More generally, suppose we start with a class of n formal sets $X_1, \cdots, X_n$ and ask which formal expressions in the X_i can be generated using the transformation $X, Y \to X \cap Y, X \cup Y$ iteratively. It can be shown[11] for example, that all the elementary symmetric functions (using $\cap$ and $\cup$ in place of the usual $\cdot$ and $+$) can be generated. Let us call a well-formed expression E in the X_k's *symmetric in X_i and X_j* if the substitution $X_i \to X_j$, $X_j \to X_i$, leaves E unchanged. Thus we can write E in the form

$$E = (X_i \cap X_j \cap W_1) \cup ((X_i \cup X_j) \cap W_2) \cup W_3$$

where the W_i are well-formed (possibly empty) expressions in the X_k's *not* involving X_i or X_j. We say that we *collapse X_i and X_j in E* if we apply the transformation $X_i \cap X_j \to X_i$, $X_i \cup X_j \to X_j$, to form

$$E' = (X_i \cap W_1) \cup (X_j \cap W_2) \cup W_3 .$$

Certainly, if E can be generated using the transformations $X, Y \to X \cap Y, X \cup Y$ starting from $X_1, \cdots, X_n$, then there is a sequence of collapses starting with E and ending with some single variable X_i. A basic theorem can be proved which asserts that if it is possible to generate E, then no matter how we collapse symmetric variables starting with the expression E we must reach some single variable X_j. In other words in attempting to collapse E to a single variable, we can never make a "bad" move. Once the structure of the expressions E which can be generated is sufficiently well understood, perhaps the representable subsets of C^n can then be determined.

Another line of research suggested by this bubble model is in the

following direction. For binary sequences x and y, define $d(x, y)$, the (Hamming) distance between x and y, to be the number of positions in which the sequences x and y differ. The fact which prevented the existence of a program which could add two integers expressed to the base 2 was the fact that there are pairs of additions in which the binary expansions of the corresponding summands are close together (in the metric d) but whose sums are not close, thus conflicting with the NDO theorem. What we would like is a mapping $m \rightarrow \tau(m)$ of integers into binary sequences for which we have

$$d(\tau(m), \tau(n)) + d(\tau(m'), \tau(n')) \geqq d(\tau(m + n), \tau(m' + n')).$$

With only this constraint there are trivial solutions, for example,

$$m \rightarrow \underbrace{111 \cdots 1}_{m}.$$

With this mapping we are essentially expressing m to the base 1 (well-known by many cultures to be inefficient for representing large numbers, say, those exceeding 10). Hence, we might require in addition that the number of binary sequences of length t which are in the range of the mapping τ to be at least α^t for some fixed $\alpha > 1$. Is it possible to find a suitable τ for which an addition program *is* possible in this model of bubble interactions?

Finally, we have just considered just one rather simple model in this paper. Physically, many other bubble interactions are possible (although some presently operate with significantly smaller margins than others) and this of course would lead to other models. It would be very interesting to understand the corresponding questions in some of these other models.

VI. ACKNOWLEDGMENTS

The author wishes to express his indebtedness to W. Shockley, A. H. Bobeck, D. E. Knuth, A. A. Thiele and especially to J. H. Spencer for many interesting ideas and discussions on this subject.

REFERENCES

1. Bobeck, A. H., "Properties and Device Applications of Magnetic Domains in Orthoferrites," B.S.T.J., *46*, No. 8 (October 1967), pp. 1901–1925.
2. Galt, J. K., "Motion of Individual Domain Walls in a Nickel-Iron Ferrite," B.S.T.J., *33*, No. 5 (September 1954), pp. 1023–1054.
3. Gyorgy, E. M., and Hagedorn, F. B., "Analysis of Domain Wall Motion in Canted Antiferromagnets," J. Appl. Phys., *39*, No. 1 (January 1968), pp. 88–90.

4. Kooy, C., and Enz, U., "Experimental and Theoretical Study of the Domain Configuration in Thin Layers of $BaFe_{12}O_{19}$," Philips Res. Rep., *15*, No. 1 (February 1960), pp. 7–29.
5. Michaelis, P. C., "A New Method of Propagating Domains in Thin Ferromagnetic Films," J. Appl. Phys., *39*, No. 2, Part 2 (February 1968), pp. 1224–1226.
6. Thiele, A. A., "The Theory of Cylindrical Magnetic Domains," B.S.T.J., *48*, No. 10 (December 1969), pp. 3287–3336.
7. Smith, D. O., "Proposal for Magnetic Domain-Wall Storage and Logic," I.R.E. Trans. on Elec. Computers, *10*, No. 4 (December 1961), pp. 709–711.
8. Dillon, J. F., Jr., "Domains and Domain Walls," *Magnetism*, Vol. III, New York: Academic Press, (1963), pp. 450–453.
9. Spencer, J. H., unpublished work.
10. Graham, R. L., "On n-Valued Functionally Complete Truth Functions," Jour. Sym. Logic, *32*, No. 2 (June 1967), pp. 190–195.
11. Kurshan, R. P., "All Terminal Bubbles Programs Yield the Elementary Symmetric Polynomials," B.S.T.J., this issue, pp. 1995–1998.

Dielectric Guide with Curved Axis and Truncated Parabolic Index

By E. A. J. MARCATILI

(Manuscript received May 5, 1970)

We find the field configurations and the propagation constants of the guided modes in a dielectric waveguide with curved axis and rectangular cross-section. Outside the guide, the refractive index is uniform. Inside, the index profile in the radial direction (intersection of the meridional plane and the plane of curvature) follows a parabolic law with the maximum at the center of the guide; in the direction perpendicular to the plane of curvature the index is either uniform or parabolic, again with the maximum at the center of the guide. The guide with mixed profiles has been proposed as an easy-to-support, low-loss, ribbon-like guide for millimeter and optical waves while the other, with parabolic profile in both directions, is similar to the "SELFOC®" or "GRIN" image transmitting guides.

The axial field components are small compared to the transverse components and consequently the modes are almost of the TEM kind. Within the guide the field distribution along a quadratic profile is a parabolic cylinder function of order close to an integer, and is sinusoidal along the uniform profile. The field components outside of the guide decay almost with exponential law.

Inside the SELFOC-like guide, the field distribution of the fundamental mode is gaussian and except for the attenuation the characteristics of the beam are similar to those obtained for a guide in which the parabolic index profile is not truncated.

The attenuation constant α of any mode is very sensitive to the radius of curvature R. Doubling R reduces α by several orders of magnitude.

Fixing R and the difference of refractive index between the center of the guide and the edge of it, the attenuation constant α passes through a minimum for a guide width measured in the plane of curvature which is only a few beam-widths.

Radiation loss for the fundamental gaussian mode is negligibly small if the distance between the center of the beam and the edge of the guide is two or more half beam-widths.

Guides with rectangular index profile in the plane of curvature have less radiation loss than similar guides with truncated parabolic profile.

I. INTRODUCTION

A dielectric guide in which the refractive index decreases with parabolic law away from its axis acts as a lens-like medium.[1,2] The transmission through it is known even if the axis is not straight[3] and if the parabolic decrease is different in two orthogonal directions[4] (astigmatic guide).

Though extremely useful in many respects the parabolic medium is not realizable since it has ever-decreasing refractive index away from the axis and this in turn produces an untenable physical result. Thus though we know that in any realizable dielectric guide with curved axis, radiation losses are inevitable,[5] the modes in the parabolic medium with curved axis can have no radiation loss since the refractive index tending towards infinity far away from the axis prevents it.

A more realistic model is achieved by truncating the parabolic index distribution. We begin, in Section II, studying the two dimensional guide, Fig. 1a, in which the index profile, Fig. 1b, varies as a truncated parabolic function along the x axis and is independent of y while outside of the guide the index is uniform.

Later, this guide is modified in such a way that along y, the index profile is either rectangular, Fig. 2a, or another truncated parabolic function, Fig. 2b.

The first of these guides has the index distribution of the dielectric thin-film guide proposed in Ref. 6 as a low-loss, easy-to-support ribbon-like guide for millimeter and optical waves. It has also the configuration of a possible guide for integrated optics.[7] This guide, with curved axis has been analyzed in Ref. 8 ignoring radiation due to curvature. In Section II, both the phase and attenuation coefficients of the guided modes are evaluated and compared to those in a similar guide with rectangular index profiles along both x and y.

The results obtained for the guide with truncated parabolic profiles along x and y, Fig. 2b, are applicable, at least in order of magnitude, to "SELFOC"[9] or "GRIN"[10] fibers, and tubular gas lenses[11] with curved axes.

Finally conclusions are drawn in Section III, while all the mathematics are given in the Appendix.

II. MODES IN THE CURVED GUIDE

Consider the two-dimensional curved guide in Fig. 1a. The parabolic refractive index within the guide is independent of y and equal to

$$n_i = n\left[1 - \Delta\left(1 + \frac{2x}{a}\right)^2\right], \tag{1}$$

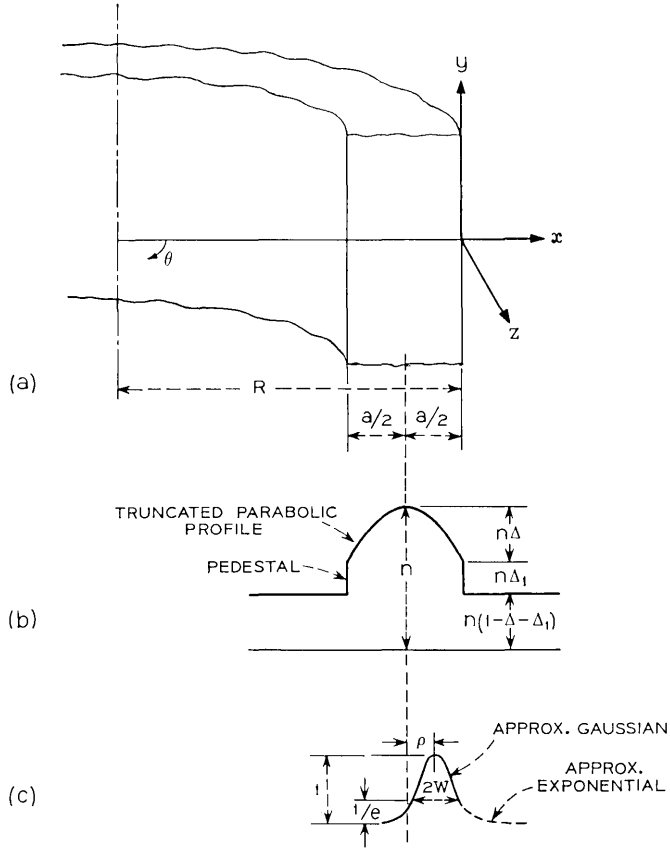


Fig. 1—(a) Two-dimensional truncated parabolic guide; (b) Refractive index profile; (c) Electric field distribution of the fundamental mode.

where a is the width of the guide, n, the refractive index in the center of it and $n(1 - \Delta)$, the refractive index at the edges. Outside the guide, the index is

$$n_o = n(1 - \Delta - \Delta_1). \tag{2}$$

We make the following assumptions:

$$\begin{aligned} \Delta &\ll 1 \\ \Delta_1 &\ll 1 \end{aligned} \tag{3}$$

and

$$\frac{\lambda}{a\sqrt{\Delta}} \ll 1 - \frac{a}{4\Delta R} \tag{4}$$

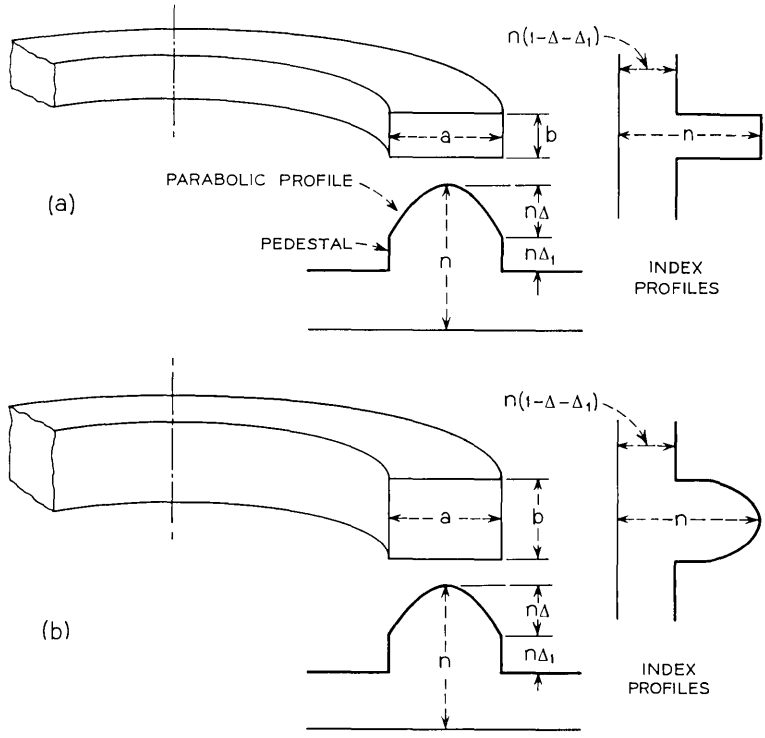


Fig. 2—(a) Inhomogeneous dielectric thin film guide; (b) "SELFOC®" or "GRIN" guides with rectangular cross-section.

where λ is the free-space wavelength and R the radius of curvature of the guide. The physical significance of inequality (3) is that the guided modes will have phase velocities quite comparable to that of a plane wave in a uniform medium of refractive index n. The inequality (4) insures that the amplitude of the field components at the edge of the guide are small compared to their maxima within the guide. In other words, most of the electromagnetic field is well confined within the guide, Fig. 1c, and consequently the loss per wavelength is small compared to unity. Considering only guided modes with field configurations independent of y, we can group them in two families: TE and TM. The field components of any mode of the first family are E_y, H_x and H_z while those of the second are H_y, E_x and E_z. In each family the transverse components are far larger than the axial components and consequently both families are essentially of the TEM kind.

The transverse components E_y , H_x , H_y and E_z of both families have the same functional dependence within and without the guide. Therefore we will talk from now on of the E field meaning either one of those four components.

Within the guide, and subject to the conditions (3) and (4), the E field distribution for the pth mode is essentially

$$E = \exp\left[-\left(\frac{x + \frac{a}{2} - \rho}{w}\right)^2\right] \mathrm{He}_p\left(2\,\frac{x + \frac{a}{2} - \rho}{w}\right) \exp\left[i(k_z z - \omega t)\right] \tag{5}$$

in which the first two factors describe the field distribution along x, and the last gives the propagating wave dependence along the curvilinear z axis. Similarly to the field distribution in the lens-like medium ($a = \infty$), the first factor is a gaussian with its maximum located at a distance

$$\rho = \frac{a^2}{8\Delta R} \tag{6}$$

from the center of the guide. The normalizing $1/e$ half-width is

$$w = \sqrt{\frac{a\lambda}{\pi n \sqrt{8\Delta}}}. \tag{7}$$

The second factor in equation (5) is a Hermite polynomial of order p which is also centered at $x = -(a/2) + \rho$ and the argument is normalized to $w/2$. Strictly speaking the expression (5) should have, instead of the Hermite polynomial, a Hermite function of order close to p. Interested readers can find the details in the Appendix.

For the fundamental mode $p = 0$ the Hermite polynomial is unity and the transverse field distribution is the well-known gaussian.

The propagation constant $k_z = \beta + i\alpha$ in equation (5) is complex and the phase and attenuation constants calculated in equations (36) and (37) are

$$\beta = \beta_\infty\left[1 + \frac{a^2}{16\Delta R^2} - \frac{2}{wkn}\,\frac{1 - M}{K(1 + M)}\right] \tag{8}$$

and

$$\alpha = \frac{\left[\frac{a}{w}(1 - d)\right]^{2p+1}}{2\sqrt{\pi\Delta}\, dRp!} \exp\left\{-\frac{\Re}{3}\left[(1 - d)^2 + \frac{\Delta_1}{\Delta} - \left(p + \frac{1}{2}\right)\left(\frac{2w}{a}\right)^2\right]^{\frac{3}{2}} - \frac{a^2}{2w^2}(1 - d)^2\right\} \tag{9}$$

in which

$$\beta_\infty = kn\sqrt{1 - \left(\frac{2}{wkn}\right)^2\left(p + \frac{1}{2}\right)}, \tag{10}$$

$$d = \frac{2\rho}{a} = \frac{1}{\mathfrak{R}}\left(\frac{a}{w}\right)^2, \tag{11}$$

$$\mathfrak{R} = \frac{4\pi n}{\lambda}(2\Delta)^{\frac{3}{2}}R, \tag{12}$$

and the values of M and K can be found in equations (38) and (39). Let us discuss the physical meaning of some of these formulas.

The phase constant β given in equation (8) is the product of the phase constant β_∞(10) of the lens-like medium with straight axis ($R = a = \infty$), multiplied by a bracket essentially equal to one; the two small terms contained therein take into account the curvature of the axis and the truncation of the parabolic profile.

More interesting is the attenuation constant (9). The value $\sqrt{2\Delta}\ R\alpha$ which is the normalized attenuation per radian has been plotted in Fig. 3 for the fundamental mode $p = 0$ and $\Delta_1 = 0$. The abscissa is the square of the guide width a normalized to the beam-width $2w$ or its equivalent $(\pi na/\lambda)\sqrt{\Delta/2}$ which is the guide width normalized to the free wavelength. The parameter used for the solid curves is the normalized radius of curvature $\mathfrak{R}$(12). For a given radius of curvature the loss per radian is highly sensitive to the width of the guide and passes through a minimum at width

$$\frac{a}{2w} = \left(\frac{\mathfrak{R}}{8}\right)^{\frac{1}{3}}.$$

For a wide range of values of $\mathfrak{R}$, say 10 to 1000, that minimum loss occurs when the guide width is only a few beam-widths.

The dotted lines are curves of constant d, that is constant ratio $2\rho/a$ between the beam displacement from the guide axis ρ and the guide half-width $a/2$. It is easy to understand the downward trend of these curves for large abscissas. Consider a guide with fixed geometry and decrease the wavelength λ of operation. The beam remains at the same distance ρ from the guide axis but it becomes narrower and consequently the field at the edge of the guide and the radiation loss decrease. It is surprising that the minimum radiation loss of the solid curves occurs when the beam displacement is a small part of the gidue width (d of the order of 0.1).

Why do the solid lines have a minimum? For very narrow guides

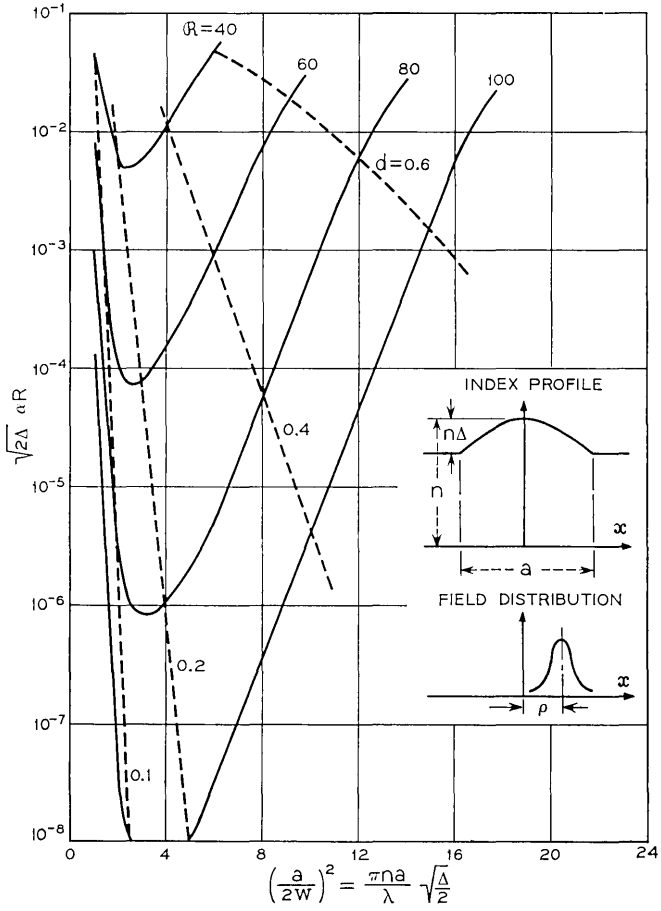


Fig. 3—Radiation loss in curved guides with truncated parabolic index profile.

$$w = \sqrt{\frac{a\lambda}{\pi n\sqrt{8\Delta}}}\ ; \qquad d = \frac{2\rho}{a}\ ; \qquad \rho = \frac{a^2}{8\Delta R}\ .$$

($a/2w \ll 1$), most of the electromagnetic field travels outside of the guide and any curvature of the axis introduces substantial radiation losses to this loosely guided beam. On the other hand, for very wide guides ($a/2w \gg 1$), any curvature of the axis displaces the beam close to one edge of the guide (d close to unity) and once again substantial losses occur. There must be a minimum in between.

It is interesting to compare the losses in these guides of truncated

parabolic index profile with guides of identical width but with rectangular index profile of height $n\Delta$. In Fig. 4, the solid curves are a repetition of some of those in Fig. 3, while the dotted ones have been reproduced from Ref. 12. The abscissa is again $(a/2w)^2$ which is identical to $(\pi/4)a/A$ in which

$$A = \frac{\lambda}{n\sqrt{8\Delta}}$$

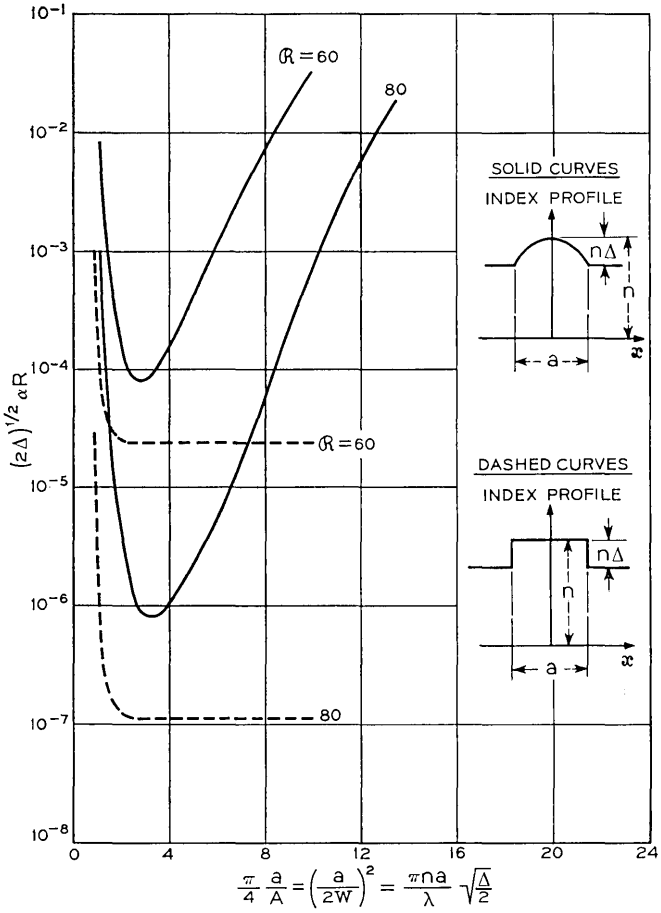


Fig. 4—Radiation loss in curved guides with truncated parabolic index profile (solid curves) and with rectangular index profile (dashed curves).

$$A = \frac{\lambda}{n\sqrt{8\Delta}}; \qquad w = \sqrt{\frac{aA}{\pi}}; \qquad \mathcal{R} = \frac{4\pi nR}{\lambda}(2\Delta)^{\frac{1}{2}}.$$

is a dimension such that for $a < A$, the guide with rectangular index profile supports a single mode and for $a > A$, the guide is multimode.

For the same radius of curvature, guide width, and same Δ on axis, the guide with truncated parabolic profile has more loss than the guide with rectangular profile. The difference is very marked for large abscissas, but this result should not be surprising because in the case of curved guides with truncated parabolic profile the beam travels close to one edge of the guide where there is little difference of refractive index between the inside and outside, while in the case of rectangular profile, though most of the power travels also close to one edge of the guide the full difference of refractive index $n\Delta$ is there to help in the guidance.

In Fig. 5 we have plotted again the attenuation per radian as a function of $(a/2w)^2$, but this time we use as parameter, the value of

$$h = \frac{\frac{a}{2} - \rho}{w}$$

which is the number of beam half-widths between the center of the beam and the external edge of the guide. The curves have asymptotes (dashed lines) parallel to both coordinates.

For $h \geqq 2$, $\Delta = 0.01$, the attenuation per radian αR turns out to be smaller than 0.003, which is very small for most purposes.

If the truncated parabolic profile is on a pedestal ($\Delta_1 \neq 0$), the losses are even smaller than those depicted in Fig. 4. The influence of Δ_1 in the attenuation constant (9), appears in the bracket of the exponent. The other two terms are in general small compared to unity. Therefore even a modest value of Δ_1, say $\Delta_1 = \Delta$, is enough to reduce the losses depicted in Figs. 3 and 5 by several orders of magnitude.

What happens when $p \neq 0$. From equation (9) we find as expected that for a given guide the radiation loss increases fast with the order p of the mode. The highest order mode that travels only slightly influenced by the guide width is characterized by

$$p_{\max} = \left[\frac{\frac{a}{2} - \rho}{w}\right]^2 - \tfrac{1}{2} = h^2 - \tfrac{1}{2}.$$

Naturally $p_{\max}$ is independent of Δ_1, and when the beam center is close to a beam half-width from the edge, $p_{\max} = 0$.

It is shown in the Appendix that if the refractive index profile along y, Fig. 1a, is not uniform but has either rectangular or truncated parabolic shape, Figs. 2a and 2b, the guides have different phase constants

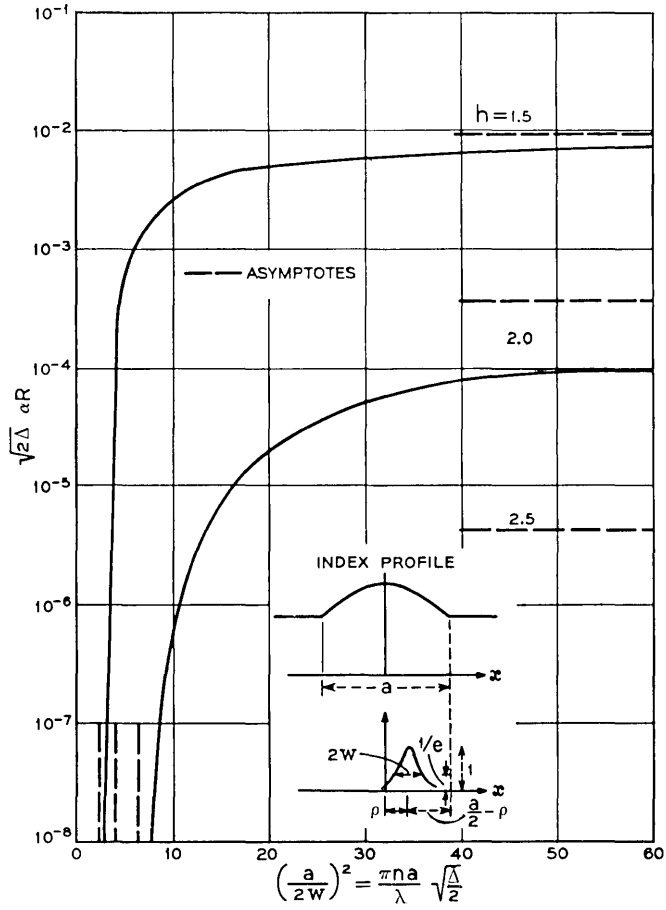


Fig. 5—Radiation loss in curved guides with truncated parabolic index profile.

$$h = \frac{(a/2) - \rho}{w}\,; \qquad w = \sqrt{\frac{a\lambda}{\pi n\sqrt{8\Delta}}}\,; \qquad \rho = \frac{a^2}{8\Delta R}\,.$$

than equation (8) but practically the same attenuation constant (9) provided that most of the electromagnetic field travels within the guide. Therefore everything said about attenuation in this section applies to the three guides.

For the following examples we will only use Figs. 3, 4 and 5 since all the important results and formulas are there.

2.1 *Example A*

For a guide such that

$$n = 1.5,$$
$$\Delta = 0.01,$$
$$\Delta_1 = 0,$$
$$a = 0.1 \text{ mm},$$
$$\lambda = 1\mu,$$

what is the radius of curvature R for which the loss per radian is of the order of 10^{-3}?

We calculate the abscissa and ordinate of Fig. 5 to be

$$\left(\frac{a}{2w}\right)^2 = \frac{\pi n a}{\lambda}\sqrt{\frac{\Delta}{2}} = 33$$

and

$$\sqrt{2\Delta}\alpha R = 1.4 \cdot 10^{-4}.$$

The parameter h obtained from Fig. 5 is approximately 2 and we derive

$$R = \frac{a^2}{8\Delta\rho} = 3.9 \text{ mm}.$$

A very small radius indeed.

2.2 *Example B*

For integrated optics a guide with truncated-parabolic profile may have the following characteristics

$$n = 1.5,$$
$$\Delta = 0.01,$$
$$\Delta_1 = 0,$$
$$a = 10\mu,$$
$$\lambda = 0.5\mu,$$
$$R = 0.6 \text{ mm}.$$

What is the loss per radian?

From Fig. 3 or 4 we get the abscissa and parameter

$$\left(\frac{a}{2w}\right)^2 = 6.7,$$

$$\mathfrak{R} = \frac{4\pi nR}{\lambda}(2\Delta)^{\frac{3}{2}} \cong 60.$$

Consequently the loss per radian results

$$\alpha R = 0.018.$$

If instead of parabolic the index had been rectangular, from Fig. 4 we deduce that the loss per radian would have been 0.00018, two orders of magnitude smaller.

III. CONCLUSIONS

For losses small enough, the field configurations and phase constants of the modes in dielectric guides, Figs. 2a and 2b, with curved axis and parabolic index profile on a pedestal, are quite comparable to those in a similar guide in which the parabolic profile is extended to infinity.

The attenuation constant of a mode is very sensitive (exponential dependence) to the radius of curvature, size of the pedestal and order of the mode. The higher the order of the mode and the smaller the size of the pedestal the larger the loss.

Quantitative results about the attenuation constant for the fundamental gaussian mode in a guide without pedestal are given in Figs. 3, 4 and 5 and in typical examples at the end at the preceding section. We find in these figures the loss per radian αR as a function of the guide width a, using as parameter the radius of curvature R, or the ratio between beam displacement ρ and guide width or the ratio between the beam distance from the edge of the guide, $a/2 - \rho$ and the beam width w. The main conclusions are:

(*i*) Doubling R reduces the attenuation constant α several orders of magnitude.

(*ii*) For any R, there is a guide width that minimizes the loss per radian. That dimension is only a few beam-widths.

(*iii*) For comparable characteristics, guides with rectangular profiles have lower attenuation than those with truncated-parabolic profile. Therefore if the transmission of images is not important, such as in the case of the ribbon-like guide of Ref. 6

and guides for integrated optics, rectangular index profiles are more attractive than parabolic profiles.

(*iv*) The attenuation per 90° bend is smaller than 10^{-3} in a guide such that the distance between beam center and the external edge of the guide is larger than a couple of half beam-widths, that is, if

$$\frac{\frac{a}{2} - \rho}{2w} > 1.$$

APPENDIX

Modes in Curved Guides With Truncated-Parabolic Index Profile

We start studying the two-dimensional curved guide depicted in Fig. 1a in cylindrical coordinates. Later we will introduce a variation of the index profile along y.

The parabolic refractive index distribution within the guide is

$$n_i = n\left[1 - \Delta\left(1 + 2\frac{r - R}{a}\right)^2\right] \tag{13}$$

where a is the width of the guide, n the refractive index in the center and $n(1 - \Delta)$ the refractive index at the edges. The refractive index outside the guide is

$$n_o = n(1 - \Delta - \Delta_1). \tag{14}$$

Assuming that the electromagnetic field does not vary along y and that the only component along that direction is H_y, all the field components either inside or outside the guide are[13]

$$\left.\begin{aligned} H_y &= H \\ E_r &= \frac{H}{\omega\epsilon_0 n_{\substack{i\\o}} r} \\ E_\theta &= \frac{i}{\omega\epsilon_0 n_{\substack{i\\o}}^2}\frac{\partial H}{\partial r} \end{aligned}\right\} \exp\left[i(\nu\theta - \omega t)\right] \tag{15}$$

where ω is the angular frequency, ϵ_0 the refractive index of free space, and the indices i and o refer to the inside and outside of the guide.

The resulting wave equation for both media is

$$\frac{d^2H}{dr^2} + \frac{1}{r}\frac{dH}{dr} + \left(k^2 n_{\substack{i\\o}}^2 - \frac{\nu^2}{r^2}\right)H = 0 \tag{16}$$

in which $k = 2\pi/\lambda$ and λ is the free space wavelength. Within the guide n_i is given by equation (13) and the wave equation can be reduced to

$$\frac{d^2H}{d\xi^2} + [\eta + \tfrac{1}{2} - \tfrac{1}{4}(\xi + \xi_0)^2]H = 0 \tag{17}$$

by making the following substitutions

$$\xi = \frac{2(r - R)}{w}, \tag{18}$$

$$\xi_0 = \frac{a}{w}(1 - d),$$

$$\nu = k_z R, \tag{19}$$

$$\eta = \frac{k^2 n_1^2 - k_z^2}{4} w^2 - \frac{a^2 d}{2w^2}\left(1 - \frac{d}{2}\right) - \frac{1}{2}, \tag{20}$$

in which

$$w = \sqrt{\frac{aA}{\pi}} = \sqrt{\frac{a\lambda}{\pi n\sqrt{8\Delta}}}, \tag{21}$$

$$d = \frac{a^2}{w^2 \mathcal{R}} = \frac{a}{4\Delta R} = \frac{2\rho}{a}, \tag{22}$$

$$\mathcal{R} = \frac{4\pi n}{\lambda}(2\Delta)^{\frac{3}{2}} R, \tag{23}$$

and

$$A = \frac{\lambda}{n\sqrt{8\Delta}}. \tag{24}$$

Furthermore, equation (17) has been derived making the following simplifying assumptions

$$\begin{aligned} \Delta &\ll 1, \\ \Delta_1 &\ll 1, \\ \frac{\lambda}{a\sqrt{\Delta}} &\ll 1 - \frac{a}{4\Delta R}. \end{aligned} \tag{25}$$

The physical significance of w, d, A and the inequalities are given in the text.

The solution of equation (17) is[14]

$$H_i = D_\eta(\xi + \xi_0) = \exp\left[-\left(\frac{\xi + \xi_0}{2}\right)^2\right] \mathrm{He}_\eta(\xi + \xi_0) \tag{26}$$

where $D_\eta(\xi + \xi_0)$ is the parabolic cylinder function of order η and $\mathrm{He}_\eta\ (\xi + \xi_0)$ is the Hermite function of order η. Only if $a \to \infty$, η becomes an integer, the Hermite function is reduced to a polynomial and H_i becomes the well-known solution of the parabolic lens-like medium extending to infinity.[3]

Outside of the guide, that is for $r > R$, the refractive index n_o is uniform, equation (14), and the solution of the wave equation (16) is[13] the Hankel function of order ν and argument $kn_o r$. That is

$$H_o = H_\nu^{(1)}(kn_o r). \tag{27}$$

To match fields at the boundary $r = R$, the radial admittance H_ν/E_θ inside and outside the guide must be identical. With the help of equations (15), (26) and (27), we obtain the characteristic equation

$$k\frac{w}{2}\frac{nD_\eta(\xi_0)}{D'_\eta(\xi_0)} = \frac{H_\nu^{(1)}(kn_o R)}{H_\nu^{(1)\prime}(kn_0 R)} \tag{28}$$

in which the derivatives are taken with respect to the arguments of the functions.

We should have another boundary equation for the other side of the guide, $r = R - a$, but we are interested in guides with radius of curvature R small enough to push the field away from the center of the guide, and consequently the field at the interface $r = R - a$ is negligibly small.

To solve explicitly the boundary or characteristic equation (28) for k_z, we need asymptotic expansions of the functions involved. From the inequalities in equation (25), it can be deduced that

$$|\ \xi_0\ | \gg 1 \quad \text{and} \quad |\ \xi_0\ | \gg |\ \eta\ |. \tag{29}$$

The asymptotic expansion for $D_\eta(\xi_0)$ is then[14]

$$D_\eta(\xi_0) \cong \xi_0^\eta \exp\left(-\frac{\xi_0^2}{4} - i\pi\eta\right) + \frac{\sqrt{2\pi}}{\Gamma(-\eta)}\,\xi_0^{-\eta-1} \exp\left(\frac{\xi_0^2}{4}\right) \tag{30}$$

where $\Gamma(-\eta)$ is the gaussian function of argument $(-\eta)$.

The asymptotic expansion for the Hankel function results from observing that as a consequence of equation (25)

$$\begin{aligned} kn_o R &\gg 1, \\ k_z R &\gg 1, \\ \frac{kn_0}{k_z} &\simeq 1 \end{aligned} \tag{31}$$

and

$$(k_z^2 - k^2 n_o^2)^{\frac{3}{2}} \frac{R}{k_z^2} \gg 1.$$

Therefore we can replace the Hankel function by Watson's approximation.[14] This approximation involves Bessel functions of order one-third and large arguments. Keeping the first term of their asymptotic expansions, the Hankel function results

$$H_\nu^{(1)}(kn_oR) = \sqrt{\frac{2}{\pi R(k_z^2 - k^2 n_o^2)^{\frac{1}{2}}}} \left\{ -i \exp\left[\frac{R}{3k_z^2}(k_z^2 - k^2 n_o^2)^{\frac{3}{2}}\right] + \frac{1}{2}\exp\left[-\frac{R}{3k_z^2}(k_z^2 - k^2 n_o^2)^{\frac{3}{2}}\right]\right\}. \quad (32)$$

Substituting equations (30) and (32) in equation (28) we obtain a simplified version of the characteristic equation

$$\frac{1 + \frac{\sqrt{2\pi}}{\Gamma(-\eta)} \xi_0^{-2\eta-1} \exp\left(\frac{\xi_0^2}{2} + i\pi\eta\right)}{1 - \frac{\sqrt{2\pi}}{\Gamma(-\eta)} \xi_0^{-2\eta-1} \exp\left(\frac{\xi_0^2}{2} + i\pi\eta\right)} = \xi_0 \frac{1 + i \exp\left[-\frac{2}{3}\frac{R}{k_z^2}(k_z^2 - k^2 n_o^2)^{\frac{3}{2}}\right]}{(k_z^2 - k^2 n_o^2)^{\frac{1}{2}} w}. \quad (33)$$

To solve this equation for k_z we rewrite it as

$$\Gamma(-\eta) = F(\eta) \quad (34)$$

and notice that $F(\eta)$ is a large quantity. Therefore the gamma function is also large and hence η must be near a pole, which makes η close to an integer ρ. Then we can replace the gamma function by the first term of the Laurent series $(-1)^p/p\,!(p - \eta)$, and equation (34) becomes

$$\eta = p - \frac{(-1)^p}{p!\,F(p)}. \quad (35)$$

Substituting η by the value given in equation (20) we derive the explicit value of k_z . This propagation constant is complex, $k_z = \beta + i\alpha$, and the real and imaginary parts are the phase and attenuation constants of the pth mode:

$$\begin{aligned} \beta &= \text{Re}\, k_z \\ &= kn\left\{1 - \frac{2}{(wkn)^2}\left[p + \frac{1}{2} + \frac{1}{2}\Re d^2\left(1 - \frac{d}{2}\right) + \frac{1 - M}{K(1 + M)}\right]\right\} \end{aligned} \quad (36)$$

$$\alpha = \operatorname{Im} k_z = \frac{\exp\left[-\frac{\Re}{3}\left(\frac{1-d}{M}\right)^3\right]}{dKR\sqrt{2\Delta}} \frac{1+2M-M^2}{(1+M)^2} \tag{37}$$

where

$$M = \left[1 + \frac{\frac{\Delta_1}{\Delta} - \left(p+\frac{1}{2}\right)\frac{4}{\Re d}}{(1-d)^2}\right]^{-\frac{1}{2}} \tag{38}$$

$$K = \sqrt{2\pi}\, p! \frac{\exp\left[\frac{\Re d}{2}(1-d)^2\right]}{[\sqrt{\Re d}\,(1-d)]^{2p+1}}. \tag{39}$$

In equation (37), M affects the value of α mostly via the exponential and not via the fraction

$$\frac{1+2M-M^2}{(1+M)^2}$$

which for all practical purposes can be replaced by 1. Consequently the normalized loss per radian $\sqrt{2\Delta}R\alpha$ results

$$L = \sqrt{2\Delta}\, R\alpha = \frac{\exp\left[-\frac{\Re}{3}\left(\frac{1-d}{M}\right)^3\right]}{dK}. \tag{40}$$

Now we turn to guides in which the refractive index is a function of y, Figs. 2a and 2b.

Let us start with the ribbon-like structure of Fig. 2a and assume as in Ref. 6 that

$$\Delta_1 \gg \Delta. \tag{41}$$

Provided that most of the electromagnetic field travels within the ribbon, the attenuation per radian is still given by equation (40), but the phase constant is a slight modification of equation (36). From Ref. 12 is deduced

$$\beta_1 = \beta - \frac{kn}{2}\left[\frac{\pi(q+1)}{b}\right]^2 \begin{cases} \left(1+\dfrac{2(1-\Delta_1)^2 A_1}{\pi b}\right)^{-2} & \text{for field polarized along } y, \\ \left(1+\dfrac{2}{\pi}\dfrac{A_1}{b}\right)^{-2} & \text{for field polarized along } x, \end{cases} \tag{42}$$

where $q+1$ indicates the number of maxima of electric field within

the guide along y and

$$A_1 = \frac{\lambda}{n\sqrt{8\Delta_1}}. \tag{43}$$

Consider another guide, Fig. 2b, with rectangular cross-section and truncated parabolic index profile along both the x and y directions

$$n_i = n\left[1 - \Delta\left(1 + 2\frac{r - R}{a}\right)^2 - \Delta\left(\frac{2y}{b}\right)^2\right]. \tag{44}$$

Provided that most of the electromagnetic field is within the guide cross-section, the loss per radian is still given by equation (40), but the phase constant becomes[4]

$$\beta_2 = \beta - \frac{2}{w_2^2 kn}\left\{q + \frac{1}{2} + 2\frac{1 - \left(1 + \frac{\Delta_1}{\Delta}\right)^{-\frac{1}{2}}}{\sqrt{2\pi}\, q!\left(\frac{b}{w_2}\right)^{2q+1}\left[1 + \left(1 + \frac{\Delta_1}{\Delta}\right)^{-\frac{1}{2}}\right]}\right\} \tag{45}$$

where $q + 1$ is the number of maxima of the electric field along y and

$$w_2 = \sqrt{\frac{b\lambda}{\pi n\sqrt{8\Delta}}}. \tag{46}$$

If

$$p = q = 0$$

and

$$a = b$$

the guide has square cross-section and equations (40) and (45) yield a first approximation of the phase and attenuation constants in a curved SELFOC[9] guide.

REFERENCES

1. Tonks, L., "Filamentary Standing-Wave Pattern in Solid State Maser," J. Appl. Phys., *33*, No. 6 (June 1962), pp. 1980–1986.
2. Kogelnik, H., "On the Propagation of Gaussian Beams of Light Through Lens-Like Media Including Those with a Loss or Gain Variation," Appl. Opt., *4*, No. 2 (December 1965), pp. 1562–1569.
3. Tien, P. K., Gordon, J. P., and Whinnery, J. R., "Focusing of a Light Beam of Gaussian Field Distribution in Continuous and Periodic Lens-Like Media," Proc. IEEE, *53*, No. 2 (February 1965), pp. 129–136.
4. Marcatili, E. A. J., "Modes in a Sequence of Thick Astigmatic Lens-Like Focusers," B.S.T.J., *43*, No. 6 (November 1964), pp. 2887–2903.
5. Marcatili, E. A. J., and Miller, S. E., "Improved Relations Describing Direc-

tional Control in Electromagnetic Wave Guidance," B.S.T.J., *48*, No. 7 (September 1969), pp. 2161–2188.
6. Kumagai, N., Kurazono, S., Sawa, S., and Yoshikawa, N., "Surface Waveguide Consisting of Inhomogeneous Dielectric Thin Film," Elec. and Commun. in Japan, *51-B*, No. 3 (March 1968), pp. 50–56.
7. Miller, S. E., "Integrated Optics: An Introduction," B.S.T.J., *48*, No. 7 (September 1969), pp. 2059–2070.
8. Sawa, S., and Kumagai, N., "Surface Wave Along a Circular *H*-Band of an Inhomogeneous Dielectric Thin Film," Elec. and Commun. in Japan, *52-B*, No. 3 (March 1969), pp. 44–50.
9. Uchida, T., Furukawa, M., Kitano, I., Kaizuki, K., and Matsumura, H., "A Light-Focusing Fiber Guide," 1969 IEEE Conference on Laser Eng. and Appl., Washington, D. C.
10. Rawson, E. G., Herriott, D. R., and McKenna, J., "Refractive Index Distributions in Cylindrical, Graded Index Glass Rods (GRIN Rods) Used as Image Relays," Appl. Opt., *9*, No. 3 (March 1970), p. 753–759.
11. Marcuse, D., and Miller, S. E., "Analysis of a Tubular Gas Lens," B.S.T.J., *43*, No. 4, Part 2 (July 1964), pp. 1759–1782.
12. Marcatili, E. A. J., "Bends in Optical Dielectric Guides," B.S.T.J., *48*, No. 7 (September 1969), pp. 2103–2132.
13. Stratton, J. A., *Electromagnetic Theory*, New York: McGraw-Hill, 1941, pp. 360–361.
14. Magnus, W., Oberhettinger, F., and Soni, R. P., *Formulas and Theorems for the Special Functions of Mathematical Physics*, Chelsea Publishing Co., New York: Springer-Verlag, 1966, p. 144 (on page 1660); p. 332 (on page 1659).

Radiation Losses of the Dominant Mode in Round Dielectric Waveguides

By DIETRICH MARCUSE

(Manuscript received March 5, 1970)

The radiation loss theory that has been developed in a series of earlier papers is extended to the dominant mode of the round dielectric waveguide. The theory is applied to the calculation of radiation losses of abrupt steps, gradual tapers, and random wall perturbations of the round dielectric waveguide.

The radiation losses caused by an abrupt step, and consequently the losses of tapers, are far higher for the dominant mode of the round dielectric waveguide than they are for corresponding steps and tapers of the dielectric slab waveguide. However, the losses caused by infinitesimal random wall perturbations of the round waveguide are nearly equal to the random wall losses predicted on the basis of the slab waveguide theory. In fact the losses of the dominate mode as well as the circular electric TE_{01} mode of the round rod due to random wall perturbations are very nearly the same.

The theory is limited to circular symmetric distortions of the round dielectric rod (diameter changes). The radiation losses caused by steps of the round dielectric waveguide that carries the dominate guided mode have been verified by experiments at millimeter wave frequencies.

I. INTRODUCTION

A series of earlier papers was devoted to radiation losses of TE and TM modes in dielectric slab waveguides.[1-3] The radiation losses were assumed to be caused either by random perturbations of the waveguide boundary[1] or by steps and tapers of the slab waveguide.[3] Experiments to verify the radiation loss theory were conducted with millimeter waves in round teflon rods, and the theory was extended to cover this case.[2]

These earlier publications were limited to the simplified case of

electromagnetic fields that are independent of one coordinate. In the case of the slab waveguide we assumed

$$\frac{\partial}{\partial y} = 0 \tag{1}$$

while

$$\frac{\partial}{\partial \phi} = 0 \tag{2}$$

was required of the fields of the round dielectric waveguide. Restrictions (1) and (2) made it possible to separate the fields into transverse electric (TE) or transverse magnetic (TM) modes.

The study of the simple slab waveguide yielded much useful information about the general properties of radiation losses and allowed us to infer the order of magnitude of the radiation losses caused by random wall imperfections. However, the dielectric slab is not a useful practical waveguide and can be used only as a simplified model to obtain information about the behavior of more realistic and more complicated structures. Limitation (2) for the modes of the realistic and practical round dielectric waveguide precludes the application of the theory to the most important dominant mode of this structure.

The present paper is devoted to a study of the radiation losses of the dominant mode of the round dielectric waveguide (optical fiber). To be able to handle the theory we still impose condition (2) on the derivatives related to the geometry of the waveguide but not on the field distribution. The resulting theory is still very complicated so that we must limit ourselves to sketching the theory and stating the final results.

The radiation losses caused by random imperfections [obeying restriction (2)] are very nearly identical to the losses of the corresponding slab waveguide problem. However, the radiation losses of the dominant mode caused by steps and tapers in the waveguide are much higher than the corresponding losses of the TE or TM modes in the slab waveguide. The radiation losses of the dominant mode due to waveguide steps have been found experimentally to be in agreement with the theory.

In order to allow the reader to obtain the information concerning the results of the theory unencumbered by complex mathematical formulas we start the paper with a discussion of the results. The remainder of the paper is devoted to an outline of the theory that was used to obtain these results.

II. NUMERICAL AND EXPERIMENTAL RESULTS

2.1 *Radiation Losses of Waveguide Steps*

We begin the discussion of the consequences of the radiation loss theory of the dominant mode of the round dielectric waveguide by considering the radiation losses caused by the abrupt step of the waveguide diameter shown in Fig. 1. As described in Section II, the radiation losses caused by an abrupt step can be calculated by two different methods. The mode matching technique infers the loss from the transmission coefficient of the guided mode that continues to travel in the waveguide after it has passed the step. The radiation loss method accounts for the lost power by directly calculating the amount of power radiated into space. Both methods involve approximations so that we cannot expect to obtain exactly the same results either way.

Figure 2 shows the results of both methods of calculation. The radiation loss caused by a step with $a_2/a_1 = 0.5$ as a function of ka_1 (as computed by means of the mode matching technique) is shown as the dotted line in the figure, while the solid line represents the result of the radiation loss method. The curve holds for a dielectric rod with index of refraction $n = 1.432$ ($n^2 = 2.05$). This index was chosen since it is representative of teflon at a frequency of 55 GHz. The agreement of the two methods is remarkably good considering the approximations involved in deriving the theoretical expressions.

Even better agreement is obtained by a similar calculation that applies to a dielectric rod with index of refraction $n = 1.01$ as shown in Fig. 3. Both figures are extended over ka_1 values that correspond to single guided mode operation. There are other guided modes possible over part of the range of ka_1 values but these other modes do not couple to the dominant mode of the round dielectric rod because of the restriction on symmetry imposed by equation (2). It is in this sense that the operation of the waveguide is single mode. No other guided mode occurs under the imposed conditions. The shape of the two curves in Figs. 2

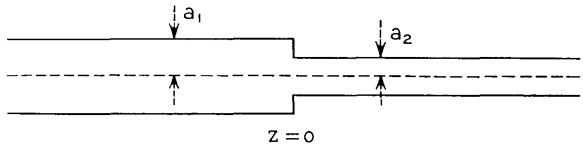


Fig. 1—Step in the round dielectric waveguide.

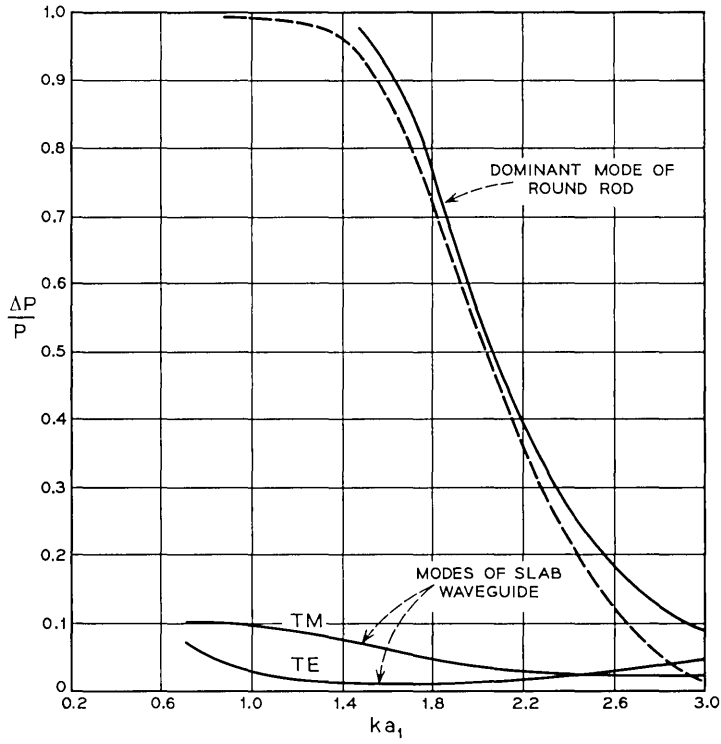


Fig. 2—Relative radiation loss caused by an abrupt step with $a_2/a_1 = 0.5$ of the waveguide. The two curves labeled dominant mode of the round waveguide were obtained by the mode matching technique (dotted line) and by the radiation loss technique (solid line). The two curves at the bottom of the figure labeled TE and TM modes represent the step losses of the slab waveguide. The radius a_1 (appearing in ka_1) belongs to the larger waveguide section. Index of refraction $n = 1.432$.

and 3 is very similar. Both curves reach into high loss regions for small values of ka_1 . The curve of Fig. 3 is applicable to a clad optical fiber with 1 percent index difference between core and cladding. The curves shown on the bottom of Figs. 2 and 3 represent the step losses of TE and TM modes of the slab waveguide.[3] These curves are computed for the same index of refraction. The dimension a_1 (of ka_1) is the half width of the slab in the case of the slab waveguide. It is striking how much lower the radiation losses of the guided modes of the slab waveguide are compared to the dominant mode of the round dielectric rod.

Because of the complexity of the theory and because the step loss results are so different for the round rod and the slab waveguide, it

appeared desirable to confirm the loss predictions of the theory with an experiment. The experiment was conducted with millimeter waves (approximately 55 GHz). A round teflon rod of 0.191 cm diameter was mounted between two metallic reflectors as shown in Fig. 4. The resulting resonant cavity could be excited through small holes in the reflector plates that, simultaneously, acted as supports for the teflon rod. Two teflon sleeves of 0.216 cm and 0.242 cm outer diameter could be slid over the teflon rods to produce a round dielectric waveguide with two steps. The losses caused by the steps could be determined from Q measurements of the cavity with and without the teflon sleeves. The results of these loss measurements (applied to one step) are shown as crosses in Fig. 5. This figure also shows the theoretical loss predictions of the mode matching (dotted line) and the radiation loss approach (solid line) of the theory. Note that the parameter value $ka_2 = 1.1$ of this figure uses the fixed value of the narrower portion of the waveguide as reference.

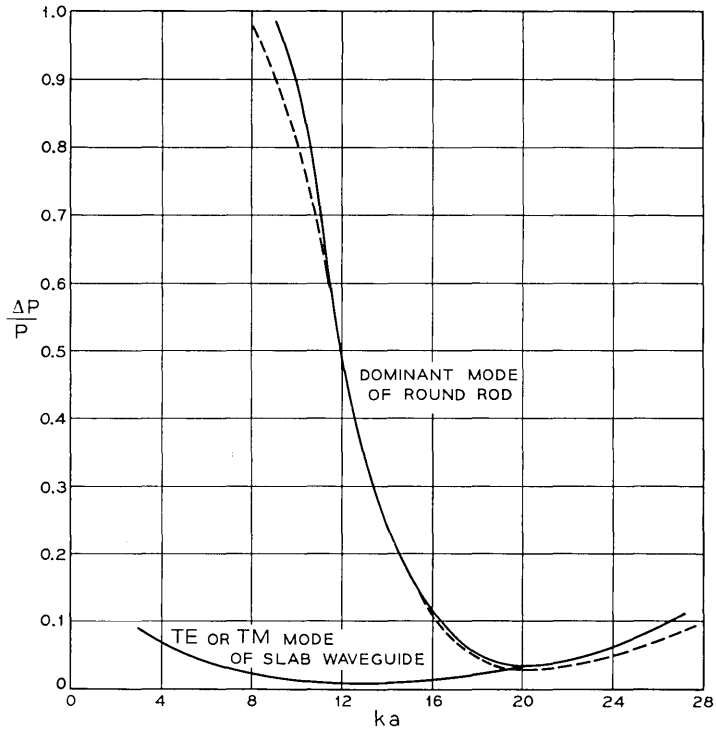


Fig. 3—This curve is similar to Fig. 2 with $n = 1.01$ and $a_2/a_1 = 0.5$.

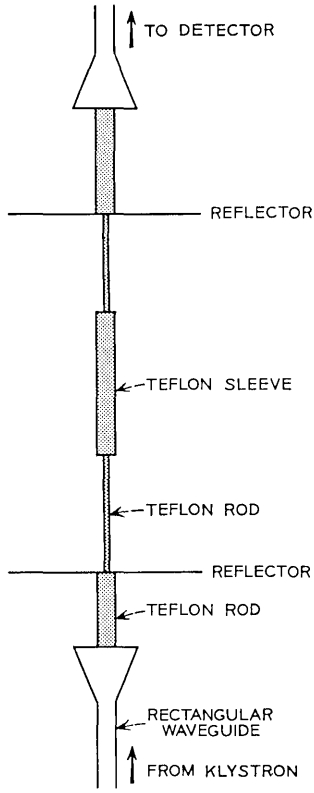


Fig. 4—Experimental resonant cavity set up to measure radiation losses of waveguide steps.

The point $a_2/a_1 = 0.5$ of Fig. 5 corresponds to the point $ka_1 = 2.2$ of Fig. 2. The measurements support the result of the round rod theory. The radiation losses of the slab waveguide even for much larger steps are still far lower than the measured values of these smaller steps of the round rod.

It is not as easy to confirm the loss predictions of the slab theory since a dielectric slab waveguide is somewhat of an idealization. In particular it is hard to excite a slab with a mode that has no field variation in the y-direction. In order to obtain some approximation to the slab waveguide we constructed a dielectric (teflon) ribbon whose dimensions on the narrower portion were 0.380 by 0.095 cm and whose wider dimensions were 0.380 by 0.190 cm. Note that only the narrow side is affected by the step. The losses of this ribbon waveguide with a 2:1 step were measured in the same resonant setup and compared to

the losses of a smooth ribbon with dimensions 0.380 by 0.095 cm. The radiation loss of the ribbon guide was $\Delta P/P = 0.08$ for $kd_2 = 1.1$ (or $kd_1 = 2.2$). This radiation loss value is shown as the circle in Fig. 5. It is apparent that the loss of the ribbon guide is far smaller than the loss of the round waveguide. It is about four times higher than the step loss predicted for the slab waveguide. However, we must keep in mind that the ribbon is only a poor approximation of the slab waveguide. It is therefore not surprising that its radiation loss cannot be predicted by the slab waveguide theory. The slab waveguide apparently can tolerate steps in its width exceptionally well.

2.2 *Radiation Loss of Tapers*

The radiation loss theory that is presented in the theoretical part can be used to determine the loss of round dielectric waveguides with

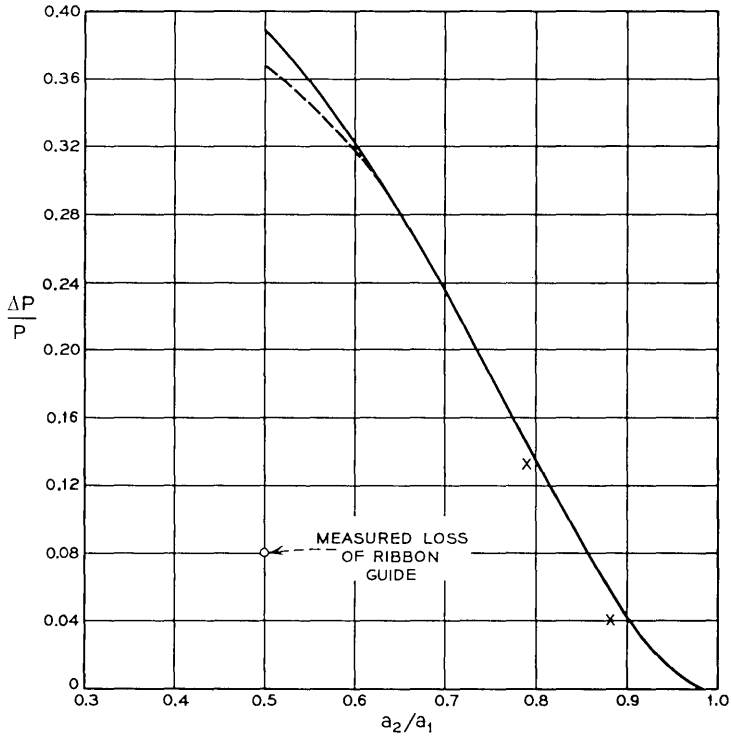


Fig. 5—Comparison of theory and experiment. The crosses are measured step losses of the round dielectric waveguide. The circle is the step loss of a ribbon guide. The curves represent the results of the mode matching theory (dotted line) and the radiation loss theory (solid line). ($n = 1.432$, $ka_2 = 1.1$.) Note that the curve parameter ka_2 uses the radius of the smaller waveguide section.

arbitrary diameter changes. Since the radiation losses of an abrupt step are very high for round dielectric waveguides it is interesting to study the radiation losses of gradual tapers.[6,7]

The calculation of the radiation losses of tapers can be simplified by observing that the dependence of β_0 on the radius of the waveguide is nearly linear over a considerable range of values. Figure 6 shows the ratio of β_0/k as a function of ka for $n = 1.432$. It is apparent that a straight line approximation is possible in the region $1.2 < ka < 2.5$.

We study the radiation losses of two different tapers. The linear taper is the simplest and therefore the most reasonable taper to investigate. However, there are reasons to suspect that the linear taper may have higher radiation losses than other forms of tapers. It is apparent from equation (36) of Section II that the result of the integration (aside from the complicated factor $I(\rho, z)$ which is difficult to evaluate) depends on the product of the derivative of the radius function $a(z)$ with sine and cosine functions of the form $\cos \int_0^z [\beta_0(z) - \beta]dz$. ($\beta_0$ is the propagation constant of the guided mode; β is the z-component of the propagation constant of the radiation modes.) The oscillatory function has the tendency to cancel contributions from those functions that appear multiplied with it under the integrand. The more rapidly the cosine function oscillates, the more effective will be its canceling influence.

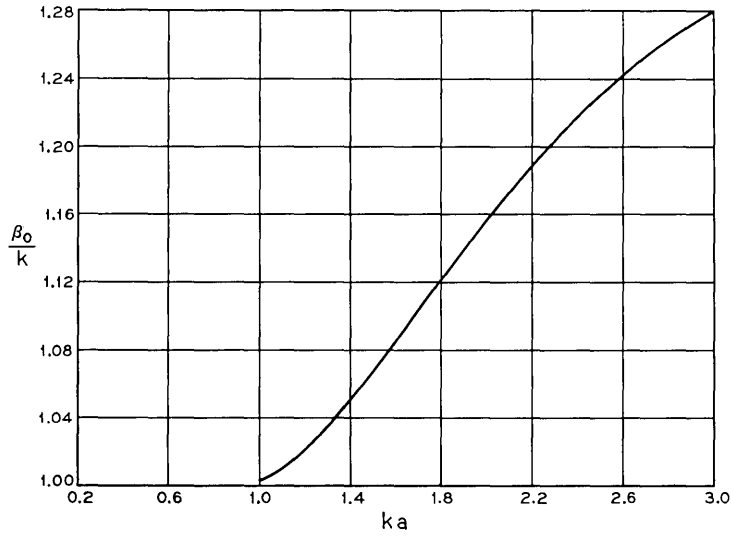


Fig. 6—Plot of the propagation constant β_0 of the dominant mode of the round dielectric waveguide. ($n = 1.432$.)

This consideration shows that we would like to see the values of $\beta_0(z) - \beta$ as large as possible. The smallest possible value, and consequently the most harmful, is the value $\beta_0(z) - k$ that is assumed at the upper end of the integration range in equation (34). However, because of the z dependence of β_0 the values of $\beta_0(z) - k$ are smaller at the narrow portion of the taper than they are on its wider portion. One might expect, therefore, that the narrow region of the linear taper contributes more to the overall radiation loss than its wider portions. It appears that the taper could be optimized if larger values of da/dz appeared at the wider end of the taper where the canceling effect of the sinusoidal functions is still more effective. Following this idea, it is possible to show that an exponential taper should distribute the radiation loss more evenly over its entire length in comparison with the linear taper. A linear taper and an exponential taper are shown in Fig. 7. The exponential taper was calculated from

$$a(z) = a_2 + (a_1 - a_2) \exp\left(-4.6 \frac{z}{L}\right).$$

This taper is designed to equalize the contribution of the integral (36), at least approximately, over the entire length of the taper assuming that $I(\rho, z)$ is constant. The discontinuity of da/dz at $z = 0$ does not contribute to the radiation loss. It would, therefore, be of no advantage to shape the taper such that da/dz is continuous over its entire length.

The radiation losses of the linear and exponential tapers are compared in Fig. 8. Even though the radiation loss of the exponential taper is less than that of the linear taper, in agreement with our expectation, the amount of improvement is insufficient to warrant the greater complexity required to produce such a more complicated taper. Figure 8 also shows that the radiation loss of a taper is far less than the losses caused by an abrupt step. The radiation losses can be made as small as desired with a taper of sufficient length. A linear taper with a length to waveguide radius (on the larger portion of the guide) ratio of $L/a_1 = 400$ reduces the radiation losses, that would occur on an abrupt step, by a factor of 100. With $\lambda = 1$ μm the value $ka_1 = 2.5$ is realized for $a_1 = 0.4$ μm so that the taper would have an actual length of $L = 160$ μm or 0.16 mm. It is apparent that much longer, more effective tapers are feasible.

Figure 8 indicates that there are two distinctly different regions. Below $L/a_1 = 2$ the taper is so short that it acts like an abrupt step. The beneficial effect of the taper makes itself felt only if the taper is long enough. The reduction of the radiation loss of a gradual taper

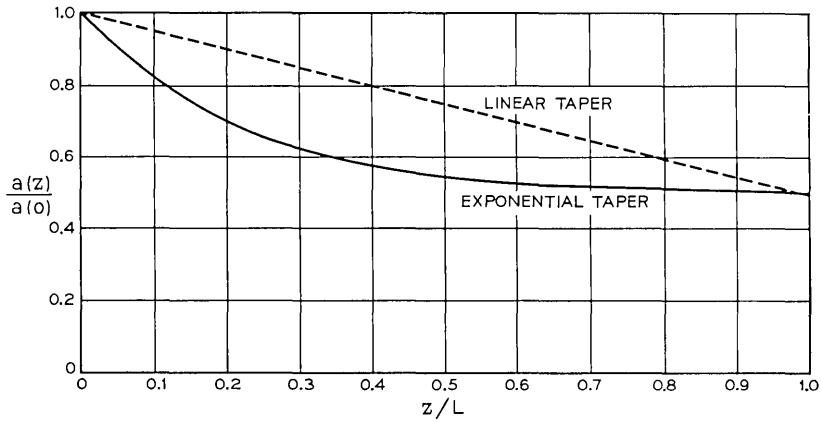


Fig. 7—The profile of the linear (dotted line) and the exponential (solid line) taper.

compared to an abrupt step or steep taper is caused by the canceling influence of the (complex) exponential function in the integral of equation (36).

2.3 *Losses Caused by Random Wall Imperfections*

An important loss contribution is caused by the random deviations of the dielectric waveguide boundary from perfect straightness. These radiation losses have been investigated for the slab waveguide[1] and for the circular electric TE_{01} mode.[2] The theory of radiation losses of the dominant mode of the round dielectric waveguide is sketched in Section III.

We have seen that the radiation losses caused by arbitrary deformations of the waveguide wall can be computed by describing the wall deviation as a series of infinitesimal steps. We have also seen that the single loss for large steps is far higher for the round dielectric waveguide than it is for the slab waveguide. We might thus worry that the losses caused by random wall perturbations may also be far higher for the dominant mode of the round dielectric waveguide. Fortunately, this pessimistic expectation is not true. The radiation losses caused by wall roughness of the round dielectric rod are no worse than they are for the modes of the slab waveguide.

The random wall losses are treated on the basis of a statistical model. The correlation function describing the wall perturbation is assumed to be a simple exponential function that is characterized by two param-

eters, the rms deviation from perfect straightness A and the correlation length B.

Figure 9 shows a series of curves of the normalized relative radiation loss as a function of the ratio of correlation length to waveguide radius B/a for a guide with index of refraction $n = 1.432$ (teflon). The curve parameter is the product of vacuum propagation constant times waveguide radius, ka. Also shown for means of comparison is the loss of the circular electric mode of the round waveguide as a dotted line. It is apparent that the radiation losses of the dominant mode are approximately equal to the radiation loss of the circular electric mode. A comparison with the results of Ref. 1 shows that the losses of Fig. 9 are approximately four times as high as the corresponding losses for the slab waveguide. For a meaningful comparison we must remember, however, that the slab waveguide losses were computed under the assumption that only one of the two slab boundaries was randomly perturbed. It seems reasonable to compare the losses of the round rod to a slab waveguide whose two walls are perturbed in a correlated way. In fact, if we assume that the thickness of the slab waveguide changes in a manner that provides equal but opposite displacement of each side of the guide we would obtain a four times higher loss than is shown in the curves of Ref. 1. The agreement between the radiation losses of the

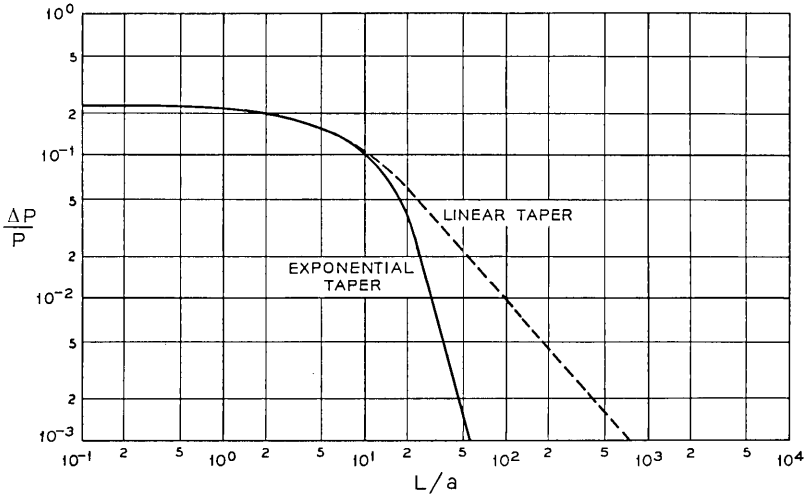


Fig. 8—Relative radiation loss of the linear (dashed line) and the exponential (solid line) taper. ($n = 1.432$, $a_2/a_1 = 0.5$, $ka_1 = 2.5$.)

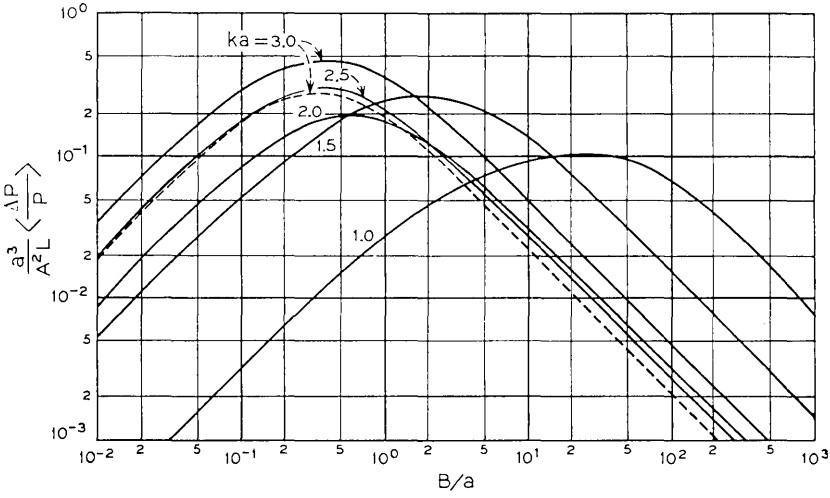


Fig. 9—Normalized radiation losses caused by random wall perturbations. The solid lines correspond to the dominant mode of the round guide, the dashed line represents the TE_{01} mode of this waveguide. (n = 1.432.) The curve parameters are the values of ka.

slab waveguide and the random wall losses of the round dielectric waveguide is quite close.

Figure 10 shows similar loss curves for a round waveguide with index of refraction n = 1.01. These curves too are about four times as high as the corresponding slab waveguide losses for the reason explained above. The curves of Fig. 10 are representative of the wall losses of a clad optical fiber with 1 percent index difference. As an example let us assume that we operate an optical fiber with a vacuum wavelength of λ = 1 μm. The value ka = 15 corresponds to a radius a = 2.4 μm for the inner core of the fiber. If we assume that the correlation length of the exponential correlation function assumes its worst possible value B/a = 2.0, we find from Fig. 10 the normalized loss

$$\frac{a^3}{A^2L}\frac{\Delta P}{P} = 0.04.$$

A loss factor of

$$\alpha = \frac{1}{L}\frac{\Delta P}{P} = 2.3\ \text{km}^{-1} = 10\ \text{dB/km}$$

would be caused by an rms deviation of the waveguide radius = A $9 \cdot 10^{-8}$ cm = 9 Å. This example shows how very stringent the

tolerance requirements can be. In a realistic case there will not only be variations of the waveguide radius. In addition we do not know the statistical model of the correlation function that must be applied in each case. However, comparison of different correlation function models has shown that the peak and its location in Figs. 9 and 10 is not dependent on the assumed statistical model. The decay of the loss curves toward increasing values of B/a is strongly model dependent.

III. THEORY

3.1 *The Dominant Guided Mode*

The field components of an arbitrary guided mode in the waveguide are described by the following equations:[5]

$$E_z = AJ_\nu(\kappa r)\cos\nu\phi \tag{3a}$$

$$H_z = BJ_\nu(\kappa r)\sin\nu\phi \tag{3b}$$

$$E_r = -\frac{i}{\kappa^2}\left[\kappa\beta_0 AJ'_\nu(\kappa r) + \omega\mu B\frac{\nu}{r}J_\nu(\kappa r)\right]\cos\nu\phi \tag{3c}$$

$$E_\phi = \frac{i}{\kappa^2}\left[\beta_0 A\frac{\nu}{r}J_\nu(\kappa r) + \kappa\omega\mu BJ'_\nu(\kappa r)\right]\sin\nu\phi \tag{3d}$$

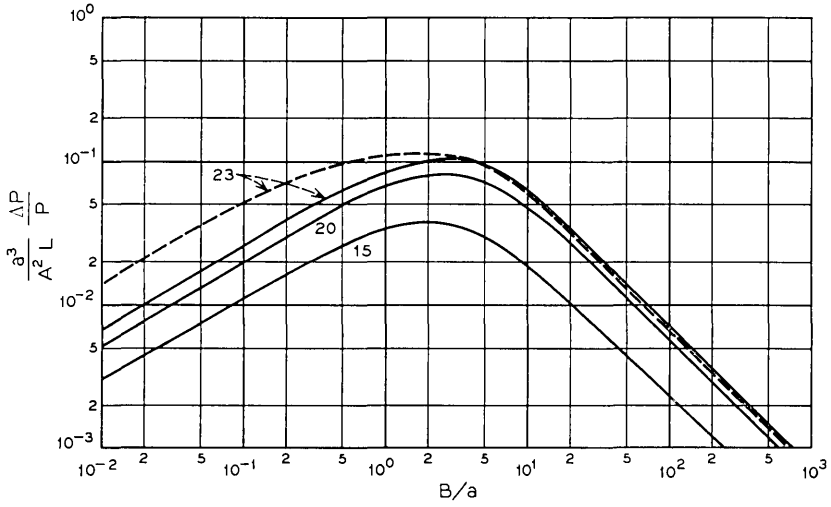


Fig. 10—These curves are similar to Fig. 9 with $n = 1.01$. The curve parameters are the values of ka.

$$H_r = -\frac{i}{\kappa^2}\left[n^2\omega\epsilon_0 A\,\frac{\nu}{r}\,J_\nu(\kappa r) + \kappa\beta_0 B J_\nu'(\kappa r)\right]\sin\nu\phi \tag{3e}$$

$$H_\phi = -\frac{i}{\kappa^2}\left[n^2\kappa\omega\epsilon_0 A J_\nu'(\kappa r) + \beta_0 B\,\frac{\nu}{r}\,J_\nu(\kappa r)\right]\cos\nu\phi. \tag{3f}$$

These equations describe the field inside of the round dielectric rod, $r \leqq a$. The functions J_ν are the Bessel functions of order ν, a prime indicates the derivative with respect to the argument (not with respect to r). The parameter ν must be an integer in order to make sine and cosine periodic functions of the aximuth ϕ with period 2π. The factor

$$e^{i(\omega t-\beta_0 z)} \tag{4}$$

was omitted from equations (3). The propagation constant β_0 is related to the constants κ and the free space propagation constant k by the relations

$$k^2 = \omega^2\epsilon_0\mu_0 \tag{5}$$

and

$$\kappa^2 = n^2k^2 - \beta_0^2\,, \tag{6}$$

where n is the index of refraction of the dielectric material. The constants A and B are not independent of each other. Their mutual dependence is given by the boundary conditions for the field components. The fields on the outside of the dielectric rod $r \geqq a$ are given by the equations

$$E_z = CH_\nu^{(1)}(i\gamma r)\cos\nu\phi \tag{7a}$$

$$H_z = DH_\nu^{(1)}(i\gamma r)\sin\nu\phi \tag{7b}$$

$$E_r = \frac{i}{\gamma^2}\left[i\gamma\beta_0 C H_\nu^{(1)\prime}(i\gamma r) + \omega\mu D\,\frac{\nu}{r}\,H_\nu^{(1)}(i\gamma r)\right]\cos\nu\phi \tag{7c}$$

$$E_\phi = -\frac{i}{\gamma^2}\left[\beta_0 C\,\frac{\nu}{r}\,H_\nu^{(1)}(i\gamma r) + i\gamma\omega\mu D H_\nu^{(1)\prime}(i\gamma r)\right]\sin\nu\phi \tag{7d}$$

$$H_r = \frac{i}{\gamma^2}\left[\omega\epsilon_0 C\,\frac{\nu}{r}\,H_\nu^{(1)}(i\gamma r) + i\gamma\beta_0 D H_\nu^{(1)\prime}(i\gamma r)\right]\sin\nu\phi \tag{7e}$$

$$H_\phi = \frac{i}{\gamma^2}\left[i\gamma\omega\epsilon_0 C H_\nu^{(1)\prime}(i\gamma r) + \beta_0 D\,\frac{\nu}{r}\,H_\nu^{(1)}(i\gamma r)\right]\cos\nu\phi \tag{7f}$$

where $H_\nu^{(1)}$ is the Hankel function of order ν and of the first kind. The prime indicates again its derivative with respect to its argument. The

argument is imaginary in order to ensure that the field distribution decays exponentially at large distance from the rod. The time and z-dependent factor (4) has again been suppressed. The parameter γ is related to the propagation constant β_0 by the equation

$$\gamma^2 = \beta_0^2 - k^2.$$

The field components were written down quite generally for an arbitrary guided mode. The lowest order or dominant mode of the guide follows from these equations with

$$\nu = 1. \tag{8}$$

The following discussion will be limited to the special case $\nu = 1$. The connection between the amplitude coefficients and the determination of the propagation constant follows from the boundary conditions for the field components. The requirement that E_z , E_ϕ , H_z and H_ϕ are continuous at the boundary $r = a$ leads to the following eigenvalue equation for the determination of the propagation constant β_0 of the guided mode

$$\left\{n^2 \frac{a\gamma^2}{\kappa}\left[\frac{J_0(\kappa a)}{J_1(\kappa a)} - \frac{1}{\kappa a}\right] + \left[\gamma a \frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)} - 1\right]\right\}$$
$$\cdot\left\{\frac{a\gamma^2}{\kappa}\left[\frac{J_0(\kappa a)}{J_1(\kappa a)} - \frac{1}{\kappa a}\right] + \left[\gamma a \frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)} - 1\right]\right\} = \left[(n^2 - 1)\frac{\beta_0 k}{\kappa^2}\right]^2. \tag{9}$$

A few numerical values obtained from (9) are shown in Table I. The

TABLE I—SOME NUMERICAL VALUES OF β_0

$n \simeq 1.432$ ($n^2 = 2.05$)		$n = 1.01$	
ka	$\beta_0 a$	ka	$\beta_0 a$
0.5	0.50000013	2.0	2.0000001
0.625	0.62500485	4.0	4.0000011
0.75	0.75006586	5.0	5.0000672
0.875	0.8758141	6.0	6.0006747
1.0	1.0043348	7.0	7.0026448
1.125	1.1387424	8.0	8.0064648
1.25	1.2816903	9.0	9.0121047
1.375	1.434524	10.0	10.019281
1.5	1.5970437	12.0	12.03695
1.75	1.9458015	14.0	14.057344
2.0	2.3149367	16.0	16.07916
2.25	2.6937751	18.0	18.101671
2.5	3.0761411	20.0	20.124481
2.75	3.458978	23.0	23.158808
3.0	3.8409082	24.0	24.170225
		27.0	27.204311

connection between the amplitude coefficients as a consequence of the boundary conditions is stated in the following equations:

$$B = -\left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}} \frac{(ka)(\kappa a)^2}{(\beta_0 a)\left(1 + \frac{\kappa^2}{\gamma^2}\right)} \cdot\left\{\frac{n^2}{\kappa a}\left[\frac{J_0(\kappa a)}{J_1(\kappa a)} - \frac{1}{\kappa a}\right] + \frac{1}{\gamma a}\left[\frac{iH_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)} - \frac{1}{\gamma a}\right]\right\} A \tag{10}$$

$$C = \frac{J_1(\kappa a)}{H_1^{(1)}(i\gamma a)} A \tag{11}$$

$$D = \frac{J_1(\kappa a)}{H_1^{(1)}(i\gamma a)} B. \tag{12}$$

It is necessary to know the relation between the amplitude coefficients and the power P carried by the mode:

$$P = \frac{\pi}{4}\left[\frac{k\beta_0}{\kappa^4}\{(a\kappa)^2[J_0^2(\kappa a) + J_1^2(\kappa a)] - 2J_1^2(\kappa a)\}\left(n^2 + \frac{\mu_0}{\epsilon_0}\frac{B^2}{A^2}\right) + \frac{k\beta}{\gamma^4}\left\{(a\gamma)^2\left[\frac{H_0^{(1)\prime 2}(i\gamma a)}{H_1^{(1)\prime 2}(i\gamma a)} + 1\right] + 2\right\}J_1^2(\kappa a)\left(1 + \frac{\mu_0}{\epsilon_0}\frac{B^2}{A^2}\right) + 2\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\frac{B}{A}\left(\frac{\beta_0^2 + n^2k^2}{\kappa^4} - \frac{\beta_0^2 + k^2}{\gamma^4}\right)J_1^2(\kappa a)\right]\left(\frac{\epsilon_0}{\mu_0}\right)^{\frac{1}{2}} A^2. \tag{13}$$

Equations (3) through (13) provide a complete description of the guided modes of symmetry $\cos\phi$. The lowest order solution of the eigenvalue equation (9) is the dominant mode of the round dielectric rod. This mode does not experience a cutoff. In principle it can be supported by any round dielectric rod of arbitrarily small cross section and arbitrarily low frequency. All other modes of the round dielectric waveguide exist only above their respective cutoff frequencies. All entries in Table I belong to single mode (with $\cos\phi$ symmetry) operation.

3.2 *Radiation Modes of the Round Dielectric Rod*

The number of guided modes that the round dielectric rod can support is finite at any given frequency. In order to obtain a complete set of normal modes of the structure we need to consider also the continuous spectrum of unguided modes.

Any solution of Maxwell's equations that satisfies the boundary condition is called a mode if its z-dependence (and time dependence) is given by equation (4). The guided modes are distinguished from the

unguided or radiation modes by the fact that their field distributions decay exponentially for increasing values of r outside of the waveguide. The radiation modes, on the other hand, extend to infinity. As their name indicates they are necessary to describe the radiation field outside (and inside) of the dielectric waveguide. Since there is no need to limit the functions describing the radiation modes to those that decay exponentially in the limit of large values of r we use a combination of Bessel and Neumann functions to express the unguided modes. However, we must require that the field remains finite on axis at $r = 0$. These considerations allow us to express the unguided solutions of Maxwell's equations as follows: For $r \leqq a$

$$E_z = FJ_\nu(\sigma r) \cos \nu\phi \tag{14a}$$

$$H_z = GJ_\nu(\sigma r) \sin \nu\phi \tag{14b}$$

$$E_r = -\frac{i}{\sigma^2}\left\{\sigma\beta F J'_\nu(\sigma r) + \omega\mu G \frac{\nu}{r} J_\nu(\sigma r)\right\} \cos \nu\phi \tag{14c}$$

$$E_\phi = \frac{i}{\sigma^2}\left[\beta F \frac{\nu}{r} J_\nu(\sigma r) + \sigma\omega\mu G J'_\nu(\sigma r)\right] \sin \nu\phi \tag{14d}$$

$$H_r = -\frac{i}{\sigma^2}\left[n^2\omega\epsilon_0 F \frac{\nu}{r} J_\nu(\sigma r) + \sigma\beta G J'_\nu(\sigma r)\right] \sin \nu\phi \tag{14e}$$

$$H_\phi = -\frac{i}{\sigma^2}\left[n^2\sigma\omega\epsilon_0 F J'_\nu(\sigma r) + \beta G \frac{\nu}{r} J_\nu(\sigma r)\right] \cos \nu\phi. \tag{14f}$$

There is now no restriction to the possible values that the propagation constant β can assume. The relation between β and σ is given by

$$\sigma^2 = n^2k^2 - \beta^2. \tag{15}$$

The field outside of the dielectric rod, $r \geqq a$, is given by

$$E_z = [HJ_\nu(\rho r) + IN_\nu(\rho r)] \cos \nu\phi \tag{16a}$$

$$H_z = [KJ_\nu(\rho r) + MN_\nu(\rho r)] \sin \nu\phi \tag{16b}$$

$$E_r = -\frac{i}{\rho^2}\left\{\rho\beta[HJ'_\nu(\rho r) + IN'_\nu(\rho r)] + \omega\mu \frac{\nu}{r}[KJ_\nu(\rho r) + MN_\nu(\rho r)]\right\} \cos \nu\phi \tag{16c}$$

$$E_\phi = \frac{i}{\rho^2}\left\{\beta \frac{\nu}{r}[HJ_\nu(\rho r) + IN_\nu(\rho r)] + \rho\omega\mu[KJ'_\nu(\rho r) + MN'_\nu(\rho r)]\right\} \sin \nu\phi \tag{16d}$$

$$H_r = -\frac{i}{\rho^2}\left\{\omega\epsilon_0 \frac{\nu}{r}[HJ_\nu(\rho r) + IN_\nu(\rho r)] + \rho\beta[KJ'_\nu(\rho r) + MN'_\nu(\rho r)]\right\} \sin \nu\phi \qquad (16e)$$

$$H_\phi = -\frac{i}{\rho^2}\left\{\rho\omega\epsilon_0[HJ'_\nu(\rho r) + IN'_\nu(\rho r)] + \beta \frac{\nu}{r}[KJ_\nu(\rho r) + MN_\nu(\rho r)]\right\} \cos \nu\phi \qquad (16f)$$

with

$$\rho^2 = k^2 - \beta^2. \qquad (17)$$

The Neumann functions N_ν are here expressed in the notation of Jahnke–Emde.[4] The determination of the coefficients of the radiation modes is complicated by an interesting phenomenon. The boundary conditions provide us with four equations. However, there are six undetermined coefficients in the set of equations (14) and (16). Even allowing for the fact that the power of the mode can be chosen arbitrarily so that one coefficient must remain undetermined by the boundary conditions, we have still one more coefficient than the boundary conditions, combined with the requirement of total power carried by the mode, are able to determine. This situation means physically that the sets of equations (14) and (16) represent a superposition of two modes that could be taken apart. A similar situation would have arisen in the case of the slab waveguide had we not been careful to separate the modes into even and odd field distributions from the very beginning. The present structure does not lend itself to a natural separation of the modes into even and odd ones. However, the formal field expressions (14) and (16) do, nevertheless, represent a superposition of two possible sets of modes. One might try to take arbitrarily either the coefficient F or G appearing in equation (14) equal to zero to try to separate out the two sets of modes. This procedure is mathematically beyond reproach but it suffers from a practical inconvenience. The resulting sets of modes would not be orthogonal. It is very desirable to choose the modes in such a way that they are all mutually orthogonal to each other. It is therefore necessary to determine the coefficients in a way that assures the orthogonality of all the modes. The boundary conditions combined with the requirement of mode orthogonality and a certain amount of power carried by each mode are still not enough to assure a unique solution of our problem. This is not surprising since it is always

possible to combine two arbitrary vectors in an infinite number of ways into two mutually orthogonal vectors.

The boundary conditions alone yield the following relations between the coefficients

$$H = \frac{\pi}{2}(\rho a)\left\{\left[J_\nu(\sigma a)N'_\nu(\rho a) - n^2\frac{\rho}{\sigma}J'_\nu(\sigma a)N_\nu(\rho a)\right]F + \frac{(n^2-1)k^2}{\rho\sigma^2\omega\epsilon_0}\beta\frac{\nu}{a}J_\nu(\sigma a)N_\nu(\rho a)G\right\} \tag{18}$$

$$I = \frac{\pi}{2}(\rho a)\left\{\left[n^2\frac{\rho}{\sigma}J'_\nu(\sigma a)J_\nu(\rho a) - J_\nu(\sigma a)J'_\nu(\rho a)\right]F - \frac{(n^2-1)k^2}{\rho\sigma^2\omega\epsilon_0}\beta\frac{\nu}{a}J_\nu(\sigma a)J_\nu(\rho a)G\right\} \tag{19}$$

$$K = \frac{\pi}{2}(\rho a)\left\{\frac{(n^2-1)k^2}{\rho\sigma^2\omega\mu}\beta\frac{\nu}{a}J_\nu(\sigma a)N_\nu(\rho a)F + \left[J_\nu(\sigma a)N'_\nu(\rho a) - \frac{\rho}{\sigma}J'_\nu(\sigma a)N_\nu(\rho a)\right]G\right\} \tag{20}$$

$$M = \frac{\pi}{2}(\rho a)\left\{-\frac{(n^2-1)k^2}{\rho\sigma^2\omega\mu}\beta\frac{\nu}{a}J_\nu(\sigma a)J_\nu(\rho a)F + \left[\frac{\rho}{\sigma}J'_\nu(\sigma a)J_\nu(\rho a) - J_\nu(\sigma a)J'_\nu(\rho a)\right]G\right\}. \tag{21}$$

Equations (14), (16) and (18) through (21) are sufficient to satisfy Maxwell's equations and the boundary conditions. The coefficients F and G are, so far, completely arbitrary. We consider now two sets of radiation modes. The first set is distinguished by using the coefficients with subscripts F_1 and G_1 while the coefficients of the second set are designated by F_2 and G_2 . The two sets of coefficients must now be adjusted to render the two sets of modes orthogonal. One of the infinitely many solutions of this problem is

$$\frac{F_2}{G_2} = -\frac{F_1}{G_1}. \tag{22}$$

The ratio of F_1/G_1 is now no longer arbitrary but is given by

$$\frac{F_1}{G_1} = \left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}}\left[\frac{(g-b)^2 + (e-d)^2 + (c^2+f^2)}{(g-n^2b)^2 + (e-n^2d)^2 + (c^2+f^2)}\right]^{\frac{1}{2}} \tag{23a}$$

with

$$b = \frac{\rho}{\sigma} J_1'(\sigma a) N_1(\rho a) \tag{23b}$$

$$c = \frac{(n^2 - 1)k}{\rho\sigma^2} \frac{\beta}{a} J_1(\sigma a) N_1(\rho a) \tag{23c}$$

$$d = \frac{\rho}{\sigma} J_1'(\sigma a) J_1(\rho a) \tag{23d}$$

$$e = J_1(\sigma a) J_1'(\rho a) \tag{23e}$$

$$f = \frac{(n^2 - 1)k}{\rho\sigma^2} \frac{\beta}{a} J_1(\sigma a) J_1(\rho a) \tag{23f}$$

$$g = J_1(\sigma a) N_1'(\rho a). \tag{23g}$$

Equation (23) was already specialized to the mode of symmetry $\cos \phi$, taking $\nu = 1$. The power carried by the radiation modes is given by

$$P = \left(\frac{\pi}{2}\right)^3 \frac{a^2\beta}{\rho} \omega\epsilon_0 \left\{ \left[(g - n^2 b) + c\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} \frac{G}{F} \right]^2 + \left[(e - n^2 d) + f\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} \frac{G}{F} \right]^2 + \left[c + (g - b)\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} \frac{G}{F} \right]^2 + \left[f + (e - d)\left(\frac{\mu_0}{\epsilon_0}\right)^{\frac{1}{2}} \frac{G}{F} \right]^2 \right\} F^2. \tag{24}$$

The normalization of the radiation modes involves the delta function in the same way as it did in the case of the slab waveguides.

3.3 *Radiation Losses Caused by a Step*

It has been shown previously[3] that the radiation losses of arbitrary deformations of dielectric waveguides can be calculated from the knowledge of the radiation loss of a step. For simplicity we limit the discussion to waveguide imperfections that do not violate the condition (2). Condition (2) restricts the waveguide deformations to symmetrical changes of the waveguide diameter. More general deformations are far more difficult to calculate.

A step in the round dielectric rod is shown in Fig. 1. We restrict ourselves to a dominant mode waveguide. The radius of the larger part of the waveguide must be small enough to ensure that only the dominant mode of the structure can propagate. Waveguides with larger radii suffer conversion losses to other guided modes in addition to the radiation losses. Such losses have been studied for the case of the slab waveguide[1] and for circular electric modes in round dielectric waveguides.[2]

The radiation field can be expressed as an integral over all the radia-

tion modes. Indicating the modes by script letters with the superscript i for the incident guided mode, r for the reflected guided and radiation modes and t for the transmitted guided and radiation modes we can write the boundary condition at the step as follows:

$$\mathcal{E}_r^{(i)} + a_r \mathcal{E}_r^{(r)} + \int_0^\infty [q_r(\rho)\, \mathcal{E}_{r_1}^{(r)}(\rho) + p_r(\rho)\, \mathcal{E}_{r_2}^{(r)}(\rho)]\, d\rho$$

$$= c_t\, \mathcal{E}_r^{(t)} + \int_0^\infty [q_t(\rho)\, \mathcal{E}_{r_1}^{(t)}(\rho) + p_t(\rho)\, \mathcal{E}_{r_2}^{(t)}(\rho)]\, d\rho \qquad (25)$$

$$\mathcal{E}_\phi^{(i)} + a_r \mathcal{E}_\phi^{(r)} + \int_0^\infty [q_r(\rho)\, \mathcal{E}_{\phi_1}^{(r)}(\rho) + p_r(\rho)\, \mathcal{E}_{\phi_2}^{(r)}(\rho)]\, d\rho$$

$$= c_t\, \mathcal{E}_\phi^{(t)} + \int_0^\infty [q_t(\rho)\, \mathcal{E}_{\phi_1}^{(t)}(\rho) + p_t(\rho)\, \mathcal{E}_{\phi_2}^{(t)}(\rho)]\, d\rho \qquad (26)$$

$$\mathfrak{H}_r^{(i)} + a_r \mathfrak{H}_r^{(r)} + \int_0^\infty [q_r(\rho)\mathfrak{H}_{r_1}^{(r)}(\rho) + p_r(\rho)\mathfrak{H}_{r_2}^{(r)}(\rho)]\, d\rho$$

$$= c_t \mathfrak{H}_\phi^{(t)} + \int_0^\infty [q_t(\rho)\mathfrak{H}_{r_1}^{(t)}(\rho) + p_t(\rho)\mathfrak{H}_{r_2}^{(t)}(\rho)]\, d\rho \qquad (27)$$

$$\mathfrak{H}_\phi^{(i)} + a_r \mathfrak{H}_\phi^{(r)} + \int_0^\infty [q_r(\rho)\mathfrak{H}_{\phi_1}^{(r)}(\rho) + p_r(\rho)\mathfrak{H}_{\phi_2}^{(r)}(\rho)]\, d\rho$$

$$= c_t \mathfrak{H}_\phi^{(t)} + \int_0^\infty [q_t(\rho)\mathfrak{H}_{\phi_1}^{(t)}(\rho) + p_t(\rho)\mathfrak{H}_{\phi_2}^{(t)}(\rho)]\, d\rho. \qquad (28)$$

These equations express the continuity of the transverse electric and magnetic field components at the step. The field components that are shown to be functions of ρ belong to radiation modes while field components that are not explicitly indicated as functions of ρ belong to the dominant guided mode. The amplitude of the incident guided mode is unity. The approximate solution of the equation system (25) through (28) follows the same reasoning that was presented for the case of the slab waveguide.[3] The coefficient c_t can be calculated by using the orthogonality of the waveguide modes to the right of the step. The modes to the right of the step are not orthogonal to the modes to the left of the step because of the different waveguide size. It is thus not possible to separate the coefficients q_r and p_r (which, incidentally, belong to the two orthogonal sets of radiation modes) from the coefficient a_r of the reflected guided mode. This problem makes it impossible to obtain an exact solution of the equation system. We neglect the reflected radiation modes when we calculate the coefficient c_t . This approximation is

justified by the fact that for large steps the radiation favors the forward direction so that q_r and p_r can be assumed to be small. For very small steps where the ratio of forward to backward scattered power can be expected to be more nearly unity we need not worry about the coefficients of the reflected radiation modes since the modes of the two guide sections become more nearly orthogonal to each other.

The transmission and reflection coefficients can thus be determined approximately with the result

$$c_t = \frac{2I_1I_2}{(I_1 + I_2)P} \tag{29}$$

and

$$a_r = \frac{I_1 - I_2}{I_1 + I_2} \tag{30}$$

with

$$\begin{aligned} I_1 = \frac{\pi}{2}\Bigg\{ & \frac{1}{\gamma_2^2}(\beta_1 A_1 - \omega\mu B_1)(\omega\epsilon_0 A_2 - \beta_2 B_2)\frac{J_1(\kappa_2 a_2)}{H_1^{(1)}(i\gamma_2 a_2)} \\ & \cdot\left[\left(\frac{1}{\kappa_1^2} + \frac{1}{\gamma_1^2}\right)J_1(\kappa_1 a_1)H_1^{(1)}(i\gamma_2 a_1) - \frac{1}{\kappa_1^2}J_1(\kappa_1 a_2)H_1^{(1)}(i\gamma_2 a_2)\right] \\ & - \frac{1}{\kappa_1^2\kappa_2^2}(\beta_1 A_1 - \omega\mu B_1)(n^2\omega\epsilon_0 A_2 - \beta_2 B_2)J_1(\kappa_1 a_2)J_1(\kappa_2 a_2) \\ & + \frac{a_2}{\kappa_1\kappa_2(\kappa_1^2 - \kappa_2^2)}(\omega\epsilon_0 n^2\beta_1 A_1 A_2 + \omega\mu\beta_2 B_1 B_2) \\ & \cdot[\kappa_1 J_1(\kappa_1 a_2)J_0(\kappa_2 a_2) - \kappa_2 J_0(\kappa_1 a_2)J_1(\kappa_2 a_2)] \\ & + \frac{1}{\gamma_2}(\omega\epsilon_0\beta_1 A_1 A_2 + \omega\mu\beta_2 B_1 B_2)\frac{J_1(\kappa_2 a_2)}{H_1^{(1)}(i\gamma_2 a_2)} \\ & \cdot\Bigg[\frac{1}{\kappa_1^2 + \gamma_2^2}\Big(ia_2 J_1(\kappa_1 a_2)H_0^{(1)}(i\gamma_2 a_2) - ia_1 J_1(\kappa_1 a_1)H_0^{1}(i\gamma_2 a_1) \\ & + \frac{\gamma_2}{\kappa_1}[a_2 J_0(\kappa_1 a_2)H_1^{(1)}(i\gamma_2 a_2) - a_1 J_0(\kappa_1 a_1)H_1^{(1)}(i\gamma_2 a_1)]\Big) \\ & + \frac{a_1}{\gamma_2^2 - \gamma_1^2}\frac{J_1(\kappa_1 a_1)}{H_1^{(1)}(i\gamma_1 a_1)}\Big(iH_1^{(1)}(i\gamma_1 a_1)H_0^{(1)}(i\gamma_2 a_1) \\ & - i\frac{\gamma_2}{\gamma_1}H_0^{(1)}(i\gamma_1 a_1)H_1^{(1)}(i\gamma_2 a_1)\Big)\Bigg]\Bigg\} \end{aligned} \tag{31}$$

and with

$$
\begin{aligned}
I_2 = \frac{\pi}{2}\Bigg\{ & -\frac{1}{\kappa_1^2}(n^2\omega\epsilon_0 A_1 - \beta_1 B_1)(\beta_2 A_2 - \omega\mu B_2)\frac{J_1(\kappa_2 a_2)}{H_1^{(1)}(i\gamma_2 a_2)} \\
& \cdot\left[\left(\frac{1}{\kappa_2^2}+\frac{1}{\gamma_2^2}\right)J_1(\kappa_1 a_2)H_1^{(1)}(i\gamma_2 a_2) - \frac{1}{\gamma_2^2}J_1(\kappa_1 a_1)H_1^{(1)}(i\gamma_2 a_1)\right] \\
& + \frac{1}{\gamma_1^2\gamma_2^2}(\omega\epsilon_0 A_1 - \beta_1 B_1)(\beta_2 A_2 - \omega\mu B_2)\frac{J_1(\kappa_2 a_2)}{H_1^{(1)}(i\gamma_2 a_2)}J_1(\kappa_1 a_1)H_1^{(1)}(i\gamma_2 a_1) \\
& + \frac{a_1}{\gamma_1\gamma_2(\gamma_2^2-\gamma_1^2)}(\omega\epsilon_0\beta_2 A_1 A_2 + \omega\mu\beta_1 B_1 B_2)\frac{J_1(\kappa_1 a_1)}{H_1^{(1)}(i\gamma_1 a_1)}\frac{J_1(\kappa_2 a_2)}{H_1^{(1)}(i\gamma_2 a_1)} \\
& \cdot[i\gamma_1 H_1^{(1)}(i\gamma_1 a_1)H_0^{(1)}(i\gamma_2 a_1) - i\gamma_2 H_0^{(1)}(i\gamma_1 a_1)H_1^{(1)}(i\gamma_2 a_1)] \\
& + \frac{1}{\kappa_1}(n^2\omega\epsilon_0\beta_2 A_1 A_2 + \omega\mu\beta_1 B_1 B_2) \\
& \cdot\Bigg[\frac{a_2}{\kappa_1^2-\kappa_2^2}\left(\frac{\kappa_1}{\kappa_2}J_1(\kappa_1 a_2)J_0(\kappa_2 a_2) - J_0(\kappa_1 a_2)J_1(\kappa_2 a_2)\right) \\
& + \frac{1}{\kappa_1^2+\gamma_2^2}\frac{J_1(\kappa_2 a_2)}{H_1^{(1)}(i\gamma_2 a_2)}\Big(a_2 J_0(\kappa_1 a_2)H_1^{(1)}(i\gamma_2 a_2) - a_1 J_0(\kappa_1 a_1)H_1^{(1)}(i\gamma_2 a_1) \\
& + \frac{\kappa_1}{\gamma_2}[ia_2 J_1(\kappa_1 a_2)H_0^{(1)}(i\gamma_2 a_2) - ia_1 J_1(\kappa_1 a_1)H_0^{(1)}(i\gamma_2 a_1)]\Big)\Bigg]\Bigg\}. \qquad (32)
\end{aligned}
$$

The indices 1 and 2 attached to the coefficients and parameters indicate that the corresponding quantities belong to the waveguide to the left of the step (index 1) or to the right of the step (index 2). The coefficients A and B are the amplitude coefficients introduced in equations (3), (10) and (13). The factor P in equation (29) is the power carried by the incident guided mode. It was assumed that the power of all the modes is identical. The actual power carried by the mode is accounted for by the expansion coefficients a_r , q_r , p_r , q_t , p_t , and c_t . The power coefficients appearing in equations (13) and (29) are also identical.

The theory of the dominant mode of the round dielectric waveguide is far more complex than the corresponding theory of the slab waveguide. This explains why the slab waveguide is so much more convenient to use for studying the general properties of radiation losses.

The radiation loss caused by the step is obtained from

$$\frac{\Delta P}{P} = 1 - |c_t|^2 - |a_r|^2. \qquad (33)$$

However, the same radiation loss can also be obtained by accounting

for the power carried away in the radiation modes. We can therefore write also

$$\frac{\Delta P}{P} = \int_{-k}^{k} (|\ q\ |^2 + |\ p\ |^2) \frac{|\ \beta\ |}{\rho}\, d\beta. \tag{34}$$

The subscripts r and t have been dropped from the expansion coefficients p and q. Both reflected and transmitted radiation modes are automatically included by extending the integration range from $-k$ to k so that backward as well as forward traveling waves are included. The factor $|\ \beta\ |/\rho$ appearing under the integration sign arose from converting the integration variable ρ to β.

The theory becomes much simpler when we limit the derivation of the p and q coefficients to small steps. It was shown in the work on slab waveguides[3] that arbitrary deformations of the waveguide wall can be treated as a succession of small steps. Even abrupt tapers can be described this way. In the limit of small step height Δa we can write

$$\Delta a = \frac{da}{dz}\, \Delta z. \tag{35}$$

The expansion coefficients q_r and q_t can approximately be obtained from equations (25) through (28) by a method that has been explained in some detail in Ref. 3.

$$q(\rho) = \int_0^L I(\rho_1 z)\, \frac{da}{dz}\, e^{-i\int_0^z (\beta_0 - \beta)\, ds}\, dz. \tag{36}$$

The subscript r or t of q is no longer necessary since q_r corresponds to negative values of β while q_t corresponds to positive values of β. The derivation of q has been simplified by expressing quantities pertaining to the waveguide to the right of the step in terms of the corresponding quantities for the waveguide to the left of the step. This approximation involves an expansion of the field quantities in Taylor series keeping only the first two terms of the expansion

$$F(a_2) = F(a_1) + \left(\frac{\partial F}{\partial a}\right)_{a=a_1} \Delta a. \tag{37}$$

The orthogonality of the modes belonging to the same section of waveguide can be employed to eliminate many terms from the expressions. The resulting expressions for $I(\rho, z)$ is far simpler than it would be had we considered a large step. We obtain

$$I(\rho, z) = \frac{\pi}{4\rho^2\gamma^2 P} J_1(\kappa a)$$
$$\cdot\Bigg\{(\beta_0 + \beta)\gamma\rho\left(\omega\epsilon_0 A \frac{\partial H}{\partial a} + \omega\mu B \frac{\partial K}{\partial a}\right)$$
$$\cdot\left[a \frac{\gamma J_0(\rho a) + i\rho \dfrac{H_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)} J_1(\rho a)}{\gamma^2 + \rho^2} - \frac{1}{\gamma\rho} J_1(\rho a)\right]$$
$$+ (\beta_0 + \beta)\gamma\rho\left(\omega\epsilon_0 A \frac{\partial I}{\partial a} + \omega\mu B \frac{\partial M}{\partial a}\right)$$
$$\cdot\left[a \frac{\gamma N_0(\rho a) + i\rho \dfrac{H_0^{(1)}(i\gamma a)}{H_1^{(1)}(i\gamma a)} N_1(\rho a)}{\gamma^2 + \rho^2} - \frac{1}{\gamma\rho} N_1(\rho a)\right]$$
$$+ (k^2 + \beta_0\beta)\left[\left(A \frac{\partial K}{\partial a} + B \frac{\partial H}{\partial a}\right) J_1(\rho a)\right.$$
$$\left.+ \left(A \frac{\partial M}{\partial a} + B \frac{\partial I}{\partial a}\right) N_1(\rho a)\right]\Bigg\}. \tag{38}$$

The derivatives of the amplitude coefficients H, I, K, and M of equations (18) through (21) are taken by keeping F and G constant. The reason for this prescription is the fact that the terms containing derivatives of F and G disappear from the equations because of mode orthogonality.

$$\frac{\partial H}{\partial a} = \frac{\pi\rho}{2}\left[\left\{a \frac{\sigma^2 - n^2\rho^2}{\sigma} J_0(\sigma a)\left[N_0(\rho a) - \frac{1}{\rho a} N_1(\rho a)\right]\right.\right.$$
$$+ \left[\frac{2}{\rho a} - \rho a + n^2\left(\rho a - \frac{2\rho}{a\sigma^2}\right)\right] J_1(\sigma a) N_1(\rho a)$$
$$\left.+ \left(n^2 \frac{\rho^2}{\sigma^2} - 1\right) J_1(\sigma a) N_0(\rho a)\right\} F + \frac{(n^2 - 1)k^2\beta}{\omega\epsilon_0\rho\sigma^2}$$
$$\left.\cdot\left\{\sigma J_0(\sigma a) N_1(\rho a) + \rho J_1(\sigma a) N_0(\rho a) - \frac{2}{a} J_1(\sigma a) N_1(\rho a)\right\} G\right] \tag{39}$$

$$\frac{\partial I}{\partial a} = -\frac{\pi\rho}{2}\left[\left\{a \frac{\sigma^2 - n^2\rho^2}{\sigma} J_0(\sigma a)\left[J_0(\rho a) - \frac{1}{\rho a} J_1(\rho a)\right]\right.\right.$$

$$+\left[\frac{2}{\rho a}-\rho a+n^2\left(\rho a-\frac{2\rho}{a\sigma^2}\right)\right]J_1(\sigma a)J_1(\rho a)$$
$$+\left(n^2\frac{\rho^2}{\sigma^2}-1\right)J_1(\sigma a)J_0(\rho a)\Big\}F+\frac{(n^2-1)k^2\beta}{\omega\epsilon_0\rho\sigma^2}$$
$$\cdot\left\{\sigma J_0(\sigma a)J_1(\rho a)+\rho J_1(\sigma a)J_0(\rho a)-\frac{2}{a}J_1(\sigma a)J_1(\rho a)\right\}G\Big] \quad (40)$$

$$\frac{\partial K}{\partial a}=\frac{\pi\rho}{2\sigma}(n^2-1)k^2$$
$$\cdot\left[\frac{\beta}{\omega\mu\sigma\rho}\left\{\sigma J_0(\sigma a)N_1(\rho a)+\rho J_1(\sigma a)N_0(\rho a)-\frac{2}{a}J_1(\sigma a)N_1(\rho a)\right\}F\right.$$
$$+\left\{aJ_0(\sigma a)\left[N_0(\rho a)-\frac{1}{\rho a}N_1(\rho a)\right]\right.$$
$$\left.\left.+\frac{2}{\rho\sigma a}J_1(\sigma a)N_1(\rho a)-\frac{1}{\sigma}J_1\,\sigma a)N_0(\rho a)\right\}G\right] \quad (41)$$

$$\frac{\partial M}{\partial a}=-\frac{\pi\rho}{2\sigma}(n^2-1)k^2$$
$$\cdot\left[\frac{\beta}{\omega\mu\sigma\rho}\left\{\sigma J_0(\sigma a)J_1(\rho a)+\rho J_1(\sigma a)J_0(\rho a)-\frac{2}{a}J_1(\sigma a)J_1(\rho a)\right\}F\right.$$
$$+\left\{aJ_0(\sigma a)\left[J_0(\rho a)-\frac{1}{\rho a}J_1(\rho a)\right]\right.$$
$$\left.\left.+\frac{2}{\rho\sigma a}J_1(\sigma a)J_1(\rho a)-\frac{1}{\sigma}J_1(\sigma a)J_0(\rho a)\right\}G\right]. \quad (42)$$

Equation (36) holds for q as well as for p. It is only necessary to insert F_1 and G_1 in equations (38) through (42) to obtain the q coefficients while the p coefficients are obtained by replacing F_1 , G_1 with F_2 , G_2 .

In order to use equation (34) for the relative power loss caused by radiation, it is necessary to calculate q and p with the help of equations (36) and (38). The coefficients appearing in these equations must be obtained from equations (39) through (42), and (10), (13), (22), (23), and (24). It should be apparent that this theory is of considerable complexity and can be handled only on an electronic computer. It is sad that the dominant mode in a round dielectric waveguide leads to such a complicated theory in comparison with the simple treatment of the slab waveguide.

3.4 *Random Wall Perturbations*

An important source of loss is the radiation that is caused by small random perturbations of the waveguide wall. Such radiation losses have been discussed for slab waveguides in Ref. 1 and for round dielectric waveguides operating with the circular electric guided mode in Ref. 2. Equation (36) of our present analysis can be used to calculate the loss of the dominant mode of the round waveguide caused by random wall perturbations. Since the step losses of the dominant mode of the round waveguide are so much higher than the corresponding losses of TE and TM modes of the slab waveguide one might fear that the losses caused by infinitesimal random perturbations of the waveguide wall may also be substantially higher. Fortunately, this is not the case. The losses caused by random wall perturbations are of the same order of magnitude for all types of dielectric waveguides that have been studied so far.

The losses caused by random wall perturbations are calculated with the help of a statistical model. Instead of using equation (34) for a particular waveguide we form the ensemble average $\langle \Delta P/P \rangle$ over many statistically similar systems. For very slight perturbations of the waveguide wall we can assume that $I(\rho, z)$ is independent of the z coordinate and write equation (36), after a partial integration, in the form

$$q(\rho) = +i(\beta_0 - \beta)I(\rho) \int_0^L a(z)e^{-i(\beta_0-\beta)z}\, dz. \tag{43}$$

The argument z has been dropped from $I(\rho)$ since this function is no longer dependent on z. The partial integration had the beneficial effect of causing $a(z)$ instead of its derivative to appear under the integration sign. It was shown in Ref. 1 how substitution of equation (43) in (34) makes the scattering loss dependent on the correlation function

$$R(u) = \langle a(z)a(z - u) \rangle \tag{44}$$

after the expectation value has been taken. It is, therefore, possible to write the average value of the relative radiation loss as

$$\left\langle \frac{\Delta P}{P} \right\rangle = 2L \int_{-k}^{k} (\beta_0 - \beta)^2 [| I^{(1)}(\rho) |^2 + | I^{(2)}(\rho) |^2] F(\beta) \frac{| \beta |}{\rho}\, d\beta \tag{45}$$

with

$$F(\beta) = \int_0^\infty R(u) \cos (\beta_0 - \beta)u\, du. \tag{46}$$

The superscripts 1 and 2 indicate that the function $I(\rho)$ has been com-

puted for both types of radiation modes that are associated with F_1 , G_1 and F_2 , G_2 .

If we use for the correlation function a simple exponential function

$$R(u) = A^2 \exp\left(-\frac{|u|}{B}\right), \tag{47}$$

$F(\beta)$ specializes to[1]

$$F(\beta) = \frac{A^2}{B\left[(\beta_0 - \beta)^2 + \frac{1}{B^2}\right]}. \tag{48}$$

IV. CONCLUSION

We have found that the radiation losses of the dominant mode of a round dielectric waveguide are much higher than the corresponding losses of TE and TM modes of the slab waveguide. The radiation losses of the dominant mode of the round dielectric waveguide with an abrupt step have been verified by a millimeter wave experiment. The step losses of a ribbon waveguide were also measured and found to lie between the losses of the dominant mode of the round waveguide and the TE mode losses of the slab waveguide, but closer to the latter. It is thus apparent that the slab waveguide can tolerate abrupt steps exceptionally well.

The radiation loss of a tapered round waveguide can be minimized by using a gentle taper instead of an abrupt step to accomplish the change of the waveguide radius. The losses of a linear taper are only slightly higher than the losses of a taper that was designed to equalize the loss contributions from different parts of the taper. It appears, therefore, that the design of optimum tapers is not profitable compared to their greater mechanical complexity.

The losses caused by slight random wall imperfections are very similar for the dominant mode and the circular electric TE_{01} mode of the round dielectric rod as well as the TE and TM modes of the dielectric slab waveguide. This result is surprising since the step losses of the dominant mode of the round waveguide are so much higher than the step losses of the slab waveguide. However, this result shows that the radiation losses caused by slight random wall perturbations can be studied with the help of the simple model of the slab waveguide and the results so obtained can be used to evaluate the performance of round dielectric waveguides.

REFERENCES

1. Marcuse, D., "Mode Conversion Caused by Surface Imperfections of a Dielectric Slab Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3177–3215.
2. Marcuse, D., and Derosier, R. M., "Mode Conversion Caused by Diameter Changes of a Round Dielectric Waveguide," B.S.T.J., *48*, No. 10 (December 1969), pp. 3217–3232.
3. Marcuse, D., "Radiation Losses of Tapered Dielectric Slab Waveguides," B.S.T.J., *49*, No. 2 (February 1970), pp. 273–290 and *49*, No. 5 (May-June 1970), p. 919.
4. Jahnke, E., and Emde, F., *Tables of Functions*, New York: Dover, 1945.
5. Collin, R.E., "Field Theory of Guided Waves," New York: McGraw-Hill, 1960.
6. Snyder, A. W., "Coupling of Modes on a Tapered Dielectric Cylinder," IEEE Trans. Microwave Theory and Techniques, *18*, No. 7 (July 1970), pp. 383–392.
7. Snyder, A. W., "Radiation Losses due to Variations of Radius on Dielectric or Optical Fibres," IEEE Trans. Microwave Theory and Techniques, *18*, No. 9 (September 1970).

Excitation of the Dominant Mode of a Round Fiber by a Gaussian Beam

By DIETRICH MARCUSE

(Manuscript received May 4, 1970)

The excitation of the dominant HE_{11} mode of a round optical fiber by a gaussian beam has been calculated. The calculation is based on the assumption that reflected waves can be neglected. It is thus applicable only to fibers with low index difference between core and cladding.

It is found that optimum excitation of the HE_{11} mode is achieved for loosely guided beams if the product of the beam half-width w times the radial decay constant γ of the HE_{11} mode outside of the guide is unity, $\gamma w = 1$. For tightly coupled modes $2^{\frac{1}{2}}w$ must be equal to the core radius in order to achieve optimum excitation. As much as 99 percent of the power can be transferred to the HE_{11} mode.

Also investigated are the effects of an off-set or tilted beam on the mode excitation. The mode excitation drops to 36 percent if the amount of off-set equals the beam half-width. The effect of tilts depends on the parameter kd, free space propagation constant times core radius of the fiber. For small values of kd or loosely guided modes, the mode excitation is very sensitive to tilts of the gaussian beam. As long as the HE_{11} mode is the only mode that can propagate, increasing values of kd lead to less sensitivity with respect to tilts. For multimode operation of the fiber, the sensitivity to tilts increases with increasing values of kd. The minimum of tilt sensitivity coincides with the minimum spot size of the guided mode.

I. INTRODUCTION

Communication by means of optical fibers requires that light energy can be coupled into the fiber in an efficient way. Of the different methods of exciting an optical fiber, the simplest consists of shining a beam of laser light on the end of the fiber. It is the purpose of this paper to investigate the power loss that results at the transition from a laser beam propagating in free space to the lowest order HE_{11} mode of a round optical fiber.

The geometry of the problem is sketched in Fig. 1. It is assumed that the fiber core is embedded in an infinite material, its cladding. For simplicity it is assumed that the value of the refractive index outside of the core is unity. The theory is manageable only if reflections from the end of the fiber are neglected. The transmission coefficients are calculated by matching only the transverse component of the electric or of the magnetic field at $z = 0$. Finally, an average of these two values is taken.

The incident beam is assumed to have a field distribution of the form

$$E_x = A \exp\left[-\left(\frac{r}{w}\right)^2\right] \exp(-ikz) \quad \text{for} \quad z \approx 0 \tag{1}$$

and

$$H_y = \left(\frac{\epsilon_0}{\mu_0}\right)^{1/2} E_x \tag{2}$$

with

$$k = \frac{2\pi}{\lambda_0} = \omega(\epsilon_0\mu_0)^{1/2}. \tag{3}$$

Since the field components of the fiber modes are conveniently expressed in cylindrical polar coordinates r, ϕ and z, it is advantageous to transform the incident field to these coordinates.

$$E_r = E_x \cos\phi; \qquad E_\phi = -E_x \sin\phi; \tag{4}$$

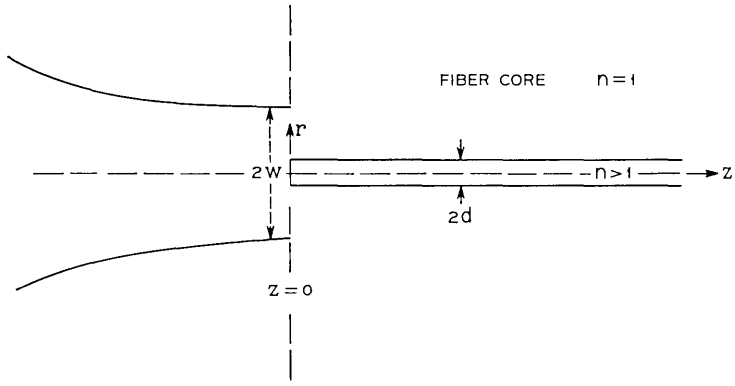


Fig. 1—Gaussian beam incident on the core of a dielectric fiber with refractive index n.

$$H_r = H_y \sin \phi; \qquad H_\phi = H_y \cos \phi. \tag{5}$$

The amplitude coefficient c_t of the HE_{11} mode is approximately determined by the equation

$$c_t = \frac{(I_1 I_2)^{1/2}}{2P} \tag{6}$$

with

$$I_1 = \int (E_r \mathcal{H}_\phi^* - E_\phi \mathcal{H}_r^*) r \, dr \, d\phi \tag{7}$$

and

$$I_2 = \int (\mathcal{E}_r^* H_\phi - \mathcal{E}_\phi^* H_r) r \, dr \, d\phi. \tag{8}$$

P is the power carried by the incident gaussian mode. The script letters indicate the field components of the guided HE_{11} mode,[1] while the other field components belong to the incident gaussian mode.

The r integrations must be carried out numerically while the ϕ integrations can be done analytically even in the more complicated cases of an off-set incident field distribution shown in Fig. 2 or a tilted incident field distribution shown in Fig. 3.

The field components of the guided modes are described by cylinder functions. The arguments of these functions inside of the fiber core at $r < d$ are κr with the radial propagation constant κ determined by

$$\kappa^2 = n^2 k^2 - \beta^2 \tag{9}$$

where β is the propagation constant of the guided mode in z direction. On the outside, $r > d$, the argument of the cylinder functions is γr with

$$\gamma^2 = \beta^2 - k^2. \tag{10}$$

The decay constant γ determines the rate at which the field intensity of the guided mode decays outside of the fiber core. For large values of r the fields behave like

$$\exp(-\gamma r). \tag{11}$$

Equation (6) for the amplitude transmission coefficient is not exact. It was derived under the assumption that reflections at $z = 0$ are negligible. The power transmission coefficient T follows from

$$T = |c_t|^2. \tag{12}$$

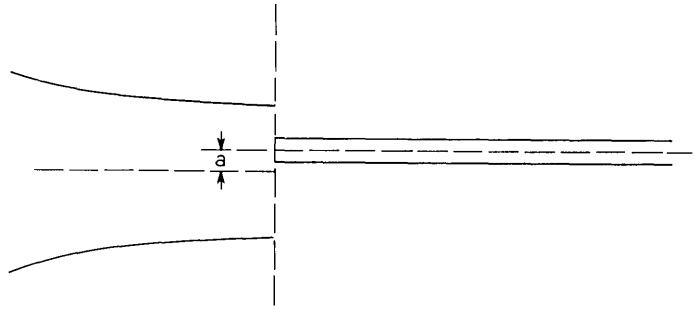


Fig. 2—Fiber excited by a gaussian beam off-set with respect to the fiber axis by an amount a.

II. NUMERICAL RESULTS

We begin the discussion of the dependence of the transmission coefficient T from the incident gaussian field to the guided HE_{11} mode with the simplest case shown in Fig. 1 for a refractive index $n = 1.01$. The gaussian beam is perfectly aligned with its beam waist being coincident with the end of the fiber core at $z = 0$. The transmission coefficient as a function of the product γw is shown in Fig. 4. Each curve belongs to a different value of kd. The normalization of the curves with respect to the radial decay constant γ is convenient since it compresses the dependence of the curve on the horizontal axis. The position of the peaks would differ by two

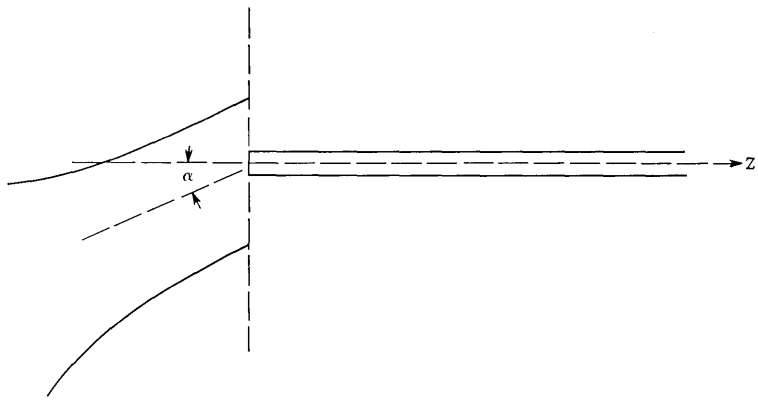


Fig. 3—Fiber excited by a tilted gaussian beam.

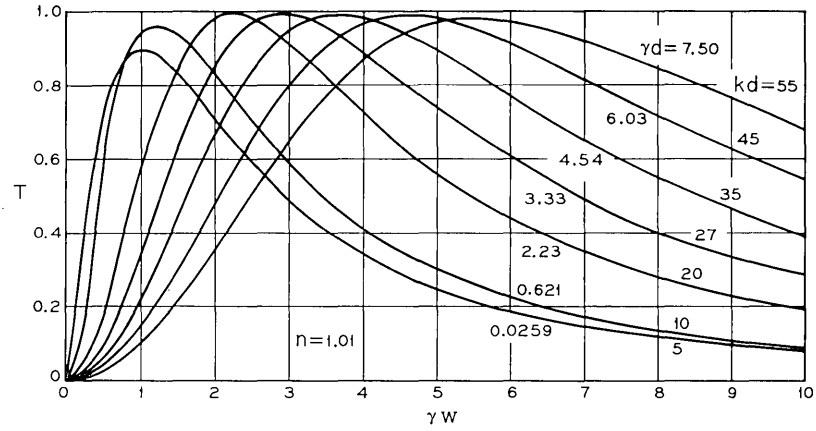


Fig. 4—Transmission coefficient T as a function of γw for several values of kd and $n = 1.01$.

orders of magnitude if the curves were drawn simply as functions of w.

Two remarkable properties can be deduced from Fig. 4. The transmission coefficient approaches extremely close to 100 percent. The dependence of the transmission peaks as a function of kd is shown in more detail in Fig. 5. According to this figure, the transmis-

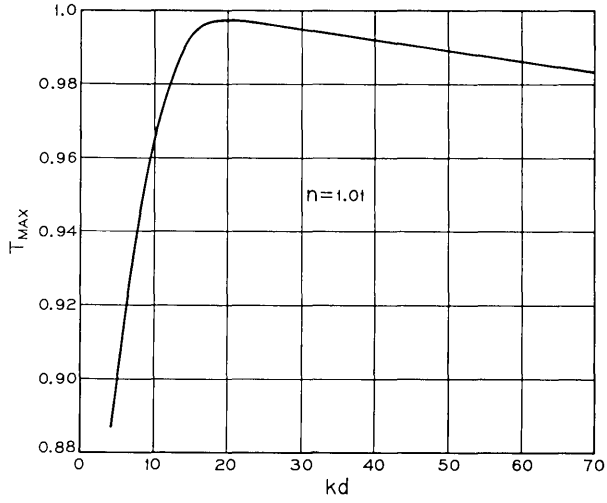


Fig. 5—The peak values of the transmission coefficient as a function of kd.

sion coefficient can be as high as 99.7 percent. These values are probably slightly optimistic as we shall see shortly.

The position of the transmission peaks can be predicted for two regions of operation. For small values of kd the guided mode is only loosely supported by the fiber core. Most of the field is on the outside decaying according to equation (11). In this case the transmission curves peak at

$$\gamma w = 1. \tag{13}$$

This means that the $1/e$ point of the exponential decay of the mode field coincides with the corresponding point of the gaussian curve. For $\lambda = 1\mu$ and $kd = 5$, we have $d = 0.8\mu$ so that for this example $1/\gamma = w = 31\mu$; $kd = 10$ correspond to $1/\gamma = w = 2.6\mu$.

The HE_{11} mode is no longer the only possible guided mode for large values of kd. At the value

$$kd = \frac{2.405}{(n^2 - 1)^{1/2}} \tag{14}$$

the TE_{01} mode begins to propagate. For $n = 1.01$, this point appears for $kd = 17$. For tightly guided modes, most of the field energy is concentrated inside of the fiber core. In this case, the peak of the transmission coefficient occurs at

$$w = d/2^{\frac{1}{2}}. \tag{15}$$

For a very tightly guided mode, the propagation constant approaches $\beta = nk$ so that we obtain from equations (10) and (15)

$$\gamma w = (n^2 - 1)^{\frac{1}{2}} kd/2^{\frac{1}{2}}. \tag{16}$$

For $n = 1.01$, we thus have $\gamma w = 0.1\ kd$. This relationship is indeed apparent in Fig. 4.

For larger refractive indices of the core, our approximation becomes questionable. This breakdown of the approximation is apparent in Fig. 6 where $n = 1.432$. The curve with $kd = 3$ exceeds the value unity very slightly, violating the principle of conservation of power. This shows that our approximate values for T are slightly too large. However, for small values of $n - 1$, it can be expected that the approximation is good because back-scattering of power from the end of the fiber core becomes negligible. This expectation is confirmed by the fact that none of the curves in Fig. 4 exceeds the value unity. It is hard to predict the degree of accuracy of the approximation.

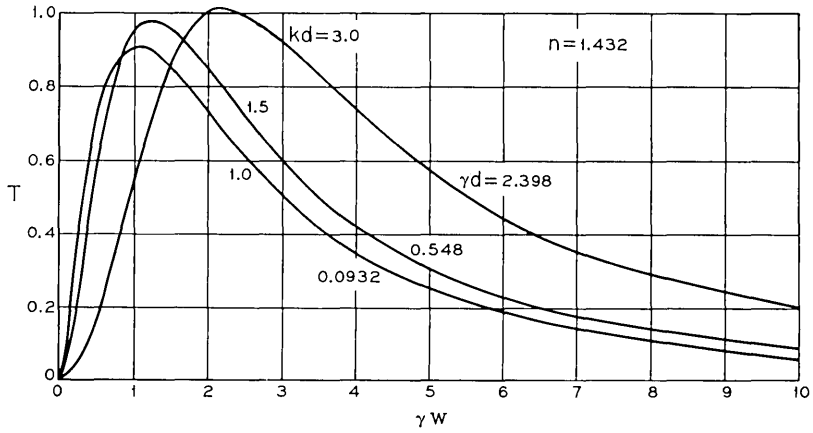


Fig. 6—Transmission coefficient T as a function of γw for $n = 1.432$. The curve with $kd = 3$ exceeds $T = 1$ indicating a breakdown of the approximation.

The values of Fig. 4 are perhaps slightly too high but it is clear that the power transmission from the gaussian mode to the guided HE_{11} mode is very efficient even if it does not quite reach 99.7 percent.

Since perfect beam alignment cannot be achieved, it is important to know how sensitive the transmission coefficient is to misalignments of the beam.

Fig. 7 shows data for the transmission coefficient T as a function of the amount of off-set "a" of the gaussian beam shown in Fig. 2. The independent variable of Fig. 7 is the product γa. Each curve was drawn for its optimum value of γw according to Fig. 4. Fig. 7 shows that the transmission coefficient decreases to 0.36 if $a = w$. This is a simple relationship that apparently holds for all values of kd. An off-set of the gaussian beam is thus not as critical as one might have feared. The direction in which the beam is off-set with respect to the polarization of the input field has been found to be unimportant. The same curves shown in Fig. 7 were obtained for any direction of the off-set.

The dependence of the transmission coefficient on tilts of the input field is shown in Fig. 8. Again w was chosen so that the maximum transmission coefficient is obtained in the absence of a tilt. The trend of these curves is interesting. The transmission coefficient is very sensitive to tilts for small values of kd. This is not surprising since the fields extend far from the fiber core so that a slight tilt causes the two wavefronts of the input field and the guided mode to become

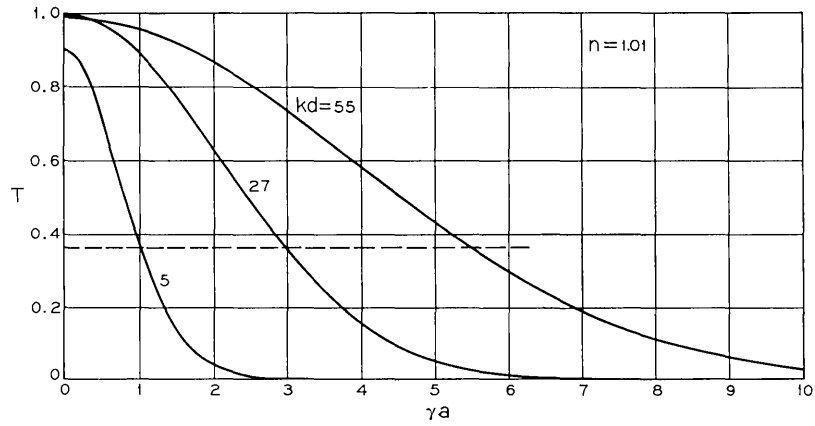


Fig. 7—Peak transmission coefficient T as a function of beam off-set.

seriously misaligned. As the guided mode (and since maximum transmission is assumed also the input field) contracts, the transmission coefficient is far less sensitive to tilts. The least sensitive curve appears for $kd = 20$ in Fig. 8. The next guided mode can be excited by the input field as soon as kd exceeds the value 17. As more and more guided modes appear, the transmission coefficient to the lowest order mode, the HE_{11} mode, becomes more sensitive to tilts. The best operating point as far as sensitivity to tilts is concerned is ap-

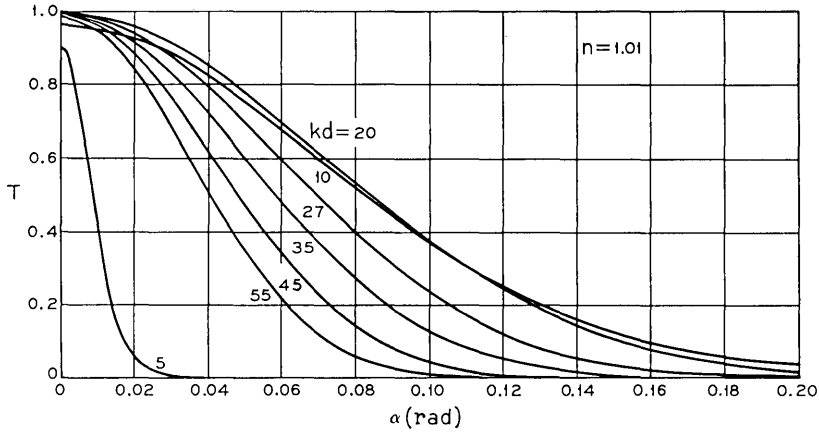


Fig. 8—Peak transmission coefficient as a function of tilt angle α.

parently close to the point where the next guided mode begins to propagate. This behavior can be explained as follows. If the wave length is kept constant and d is increased, the radial extension of the field decreases at first for increasing values of d. However, as d increases further, the field cross-section increases again. The least sensitivity to tilts occurs at the minimum field cross-section.

III. CONCLUSIONS

A numerical study of the excitation of the lowest order HE_{11} mode of the round optical fiber by an incident gaussian mode showed that the achievable transmission coefficient is very high. The predicted optimum value of 99.7 percent may be slightly overoptimistic because of the approximate nature of the calculation. However, Snyder[2] predicts transmission coefficients as high as 80 percent for the case of excitation by a truncated plane wave. The gaussian beam is far better matched to the HE_{11} mode so that a much higher transmission coefficient is not surprising.[3]

An off-set of the peak of the gaussian beam equal to its beam half-width w decreases the transmission coefficient to 36 percent. Tilts of the input field distribution are more serious for small values of the ratio of fiber core radius to wavelength. The least tilt sensitivity is obtained under conditions where the HE_{11} mode is operated close to the cut-off frequency of the TE_{01} mode. The beam cross-section assumes a minimum at this point.

REFERENCES

1. Johnson, C. C., *Field and Wave Electrodynamics,* New York: McGraw Hill, 1965.
2. Snyder, A. W., "Excitation and Scattering of Modes on a Dielectric or Optical Fiber," IEEE Trans. on Microwave Theory and Technique, *17,* No. 12 (December 1969), pp. 1138–1144.
3. Stern, J. R., Peace, M., and Dyott, R. B., "Launching into Optical-Fibre Waveguide," Elec. Letters, *6,* No. 6 (March 19, 1970), pp. 160–162.

The Capacity of the Gaussian Channel with Feedback

By P. M. EBERT

(Manuscript received April 28, 1970)

In this paper we provide a rigorous proof that feedback cannot increase the capacity of the channel with additive colored gaussian noise by more than a factor of two. We also give a tighter bound showing that any increase in capacity is less than the normalized correlation between the signal and noise. It is further shown that gaussian signals and linear feedback processing will achieve capacity.

The practical implications are that (i) feedback should be used to simplify encoding and decoding since there is little to be gained in the way of increased capacity and (ii) the various proposed schemes which use linear feedback are doing the correct thing.

I. INTRODUCTION

When Shannon first showed that feedback could not increase the capacity of a memoryless channel, he mentioned that the capacity could be increased when the channel had memory.[1] One example of such a channel is the additive colored gaussian noise channel with an average power limitation on the transmitted signal. We prove here that the capacity of this channel is never more than twice the capacity without feedback and as the noise becomes white the capacity approaches the forward capacity. The limiting case has been attributed to Shannon for years and has only recently been rigorously proven.[2]

We derive an exact expression for the mutual information between the input and output of the channel. The application of different bounds to this expression produces twice the forward capacity with the weakest bound, or the forward capacity plus the normalized correlation of the signal and noise with a slightly stronger bound. It is shown that a gaussian signal maximizes the information, and consequently the optimum feedback technique is linear.

Our results are based on the model shown in Fig. 1. The added noise

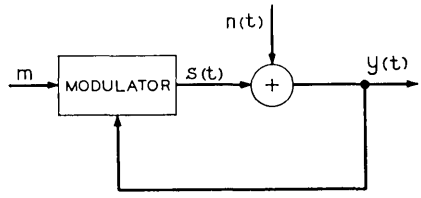


Fig. 1—Channel with noiseless feedback.

spectrum is normalized to 1 at infinite frequency, is bounded, and has an integrable logarithm. This allows us to represent the noise as in Fig. 2. The noise now consists of a white component plus a filtered version of the white noise. The imposed restrictions are for mathematical purposes only and are of no practical significance.

Theorem 1: The mutual information between the input and output of a channel with additive gaussian noise with spectral density $N(\omega)$ and arbitrary causal feedback processing, as shown in Fig. 1, is given by:

$$I(m; Y_T) = \frac{1}{2}\int_0^T E^2[s(t) + z(t) \mid m, Y_t]\, dt - \frac{1}{2}\int_0^T E^2[s(t) + z(t) \mid Y_t]\, dt \qquad (1)$$

where Y_t is $y(\tau)$, $0 \leqq \tau < t$ and the expectations are conditioned on Y_t or Y_t and m. $z(t)$ is a linear causal functional of white noise with the properties that:

$$z(t) = \int_0^t h(t - \tau)\, dw(\tau) + \int_0^\infty h(t + \tau)\, dv(\tau)$$
$$|\, 1 + H(\omega)\,|^2 = N(\omega). \qquad (2)$$

The two functions $w(t)$ and $v(t)$ are independent Wiener processes. The reason for introducing the second term is to make $n(t) = z(t) + \dot{w}(t)$ a stationary process.

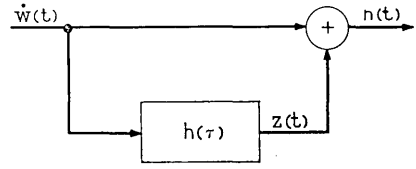


Fig. 2—Model of nonwhite noise.

Proof: We first observe that $\dot{w}(t) + z(t)$ is equivalent to noise with spectral density $N(\omega)$. A causal filter, $h(\tau)$, will exist whenever $N(\omega)$ represents the square magnitude of a causal filter

$$| G(\omega) |^2 = N(\omega)$$

$$H(\omega) = G(\omega) - 1.$$

The logarithm of $G(\omega)$ is

$$\tfrac{1}{2} \ln N(\omega) + iB(\omega)$$

where $B(\omega)$ is the phase characteristic of $G(\omega)$. The conditions of causality, no lower half plane poles, will be met when $B(\omega)$ is one half the Hilbert transform of $\ln N(\omega)$. The conditions on $N(\omega)$ insure that $\ln N(\omega)$ has a Hilbert transform.

Now to prove formula (1) we use a theorem due to Kadota, Zaki and Ziv[2], which we state without proof:

Theorem A: The mutual information between the input parameter m and the output processes Y_T of a finite power system disturbed by additive white gaussian noise is

$$I(m; Y_T) = \tfrac{1}{2}E \int_0^T \phi^2(t, m, Y_t)\, dt - \tfrac{1}{2}E \int_0^T E^2[\phi(t, m, Y_t)/Y_t]\, dt,$$

where $\phi(t, m, Y_t)$ is the causal modulating function.

This result is applied to the non-white noise problem by considering $z(t)$ to be part of the signal. The inclusion is only useful when one is calculating the mutual information; it is not to be included in the calculation of transmitter power. Theorem A cannot be applied directly since the signal, ϕ, which is taken as $s(t) + z(t)$ is not completely determined by m and Y_t , but is also a function of the process $v(t)$. To find $I(m; Y_T)$ we use the decomposition,

$$I(m, V; Y_T) = I(m; Y_T) + I(V; Y_T \mid m), \tag{3}$$

where V is the process $v(\tau)$.

From Theorem A we have,

$$I(m, V; Y_T) = \tfrac{1}{2}E \int_0^T [s(t) + z(t)]^2\, dt - \tfrac{1}{2}E \int_0^T E^2[s(t) + z(t) \mid Y_t]\, dt \tag{4}$$

and

$$I(V; Y_T \mid m) = \tfrac{1}{2}E \int_0^T [s(t) + z(t)]^2 \, dt - \tfrac{1}{2}E \int_0^T E^2[s(t) + z(t) \mid Y_t, m] \, dt,$$

which together with equation (3) proves Theorem 1. $s(t) + z(t)$ has finite energy because $s(t)$ must have finite energy and $z(t)$ will have finite energy whenever the channel has finite capacity without feedback, as we shall see when we evaluate $E[z^2(t)]$. With this basic result we can derive several interesting corollaries concerning the information.

Corollary 1: (*Pinsker*)* *Under the conditions of Theorem 1,*

$$\frac{I(m; Y_T)}{T} \leqq 2C$$

where C is the capacity at the channel without feedback.

First we observe by equation (3) that

$$I(m; Y_T) \leqq I(m, V; Y_T)$$

which is given by equation (4). Furthermore the second term in equation (4) is negative and can be ignored, thus

$$I(m; Y_T) \leqq \tfrac{1}{2}E \int_0^T (s + z)^2 \, dt. \tag{5}$$

$I(m; Y_T)$ can be further bounded by

$$I(m; Y_T) \leqq E \int_0^T s^2 \, dt + E \int_0^T z^2 \, dt \tag{6}$$

since $(s + z)^2 \leqq 2s^2 + 2z^2$.

The next step is to calculate the variance of z, since this enters directly into $I(m; Y_T)$.

$$E \int_0^T z^2(t) \, dt = TE(z^2),$$

$$E(z^2) = \frac{1}{2\pi} \int_\infty^\infty | H(\omega) |^2 \, d\omega = \frac{1}{2\pi} \int_\infty^\infty | G(\omega) - 1 |^2 \, d\omega$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \exp \left[\frac{1}{2} \ln N(\omega) + \frac{i}{2} \widecheck{\ln N(\omega)} \right]^\dagger - 1 \right|^2 d\omega$$

* The factor of 2 has been mentioned earlier by Pinsker but no proof has yet been published.

† Indicates the Hilbert transform.

$$= \frac{-1}{2\pi} \int_{-\infty}^{\infty} [1 - N(\omega)]\, d\omega$$

$$- \frac{\text{Re}}{\pi} \int_{-\infty}^{\infty} \left\{ \exp \left[\frac{1}{2} \ln N(\omega) + \frac{i}{2} \widetilde{\ln N(\omega)} \right] - 1 \right\} d\omega.$$

This latter integral, as chance would have it, is almost identical in structure to an integral which arises in evaluating the spectral density of a single sideband FM wave (at the carrier frequency) which is modulated by a gaussian signal. The quantity $1/2 \ln N(\omega)$ here plays the role of the autocorrelation function of the gaussian signal, and although for our problem $1/2 \ln N(\omega)$ is not in general an autocorrelation function, the integral may be discussed *via* the technique used in the FM problem (see Mazo and Salz)[3].

Define:

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\frac{1}{2} \ln N(\omega) + \frac{i}{2} \widetilde{\ln N(\omega)} \right] e^{i\omega t} = f(t)$$

then

$$\frac{d}{d\omega} [G(\omega) - 1] = G(\omega) \frac{d}{d\omega} F(\omega) = [G(\omega) - 1] \frac{d}{d\omega} F(\omega) + \frac{d}{d\omega} F(\omega).$$

In the time domain this becomes

$$-ith(t) = -itf(t) - i \int_0^t \tau f(\tau) h(t - \tau)\, d\tau$$

because both $h(\tau)$ and $f(\tau)$ are zero for negative τ. Both $f(\tau)$ and $h(\tau)$ are finite for small τ and thus

$$h(\tau = 0) = f(\tau = 0).$$

The integral we are interested in is 2 Re $h(\tau = 0)$ which is equal to

$$2 \text{ Re } f(\tau = 0) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \ln N(\omega)\, d\omega.$$

Thus far we have shown that

$$E \int_0^T s^2\, dt + E \int_0^T z^2\, dt$$

$$= E \int_0^T s^2\, dt - \frac{T}{2\pi} \int_{-\infty}^{\infty} [1 - N(\omega)]\, d\omega - \frac{T}{2\pi} \int_{-\infty}^{\infty} \ln N(\omega)\, d\omega. \quad (7)$$

One more trick is needed to prove the corollary. We have, up to this point, considered only normalized channels which had $N(\infty) = 1$.

This is valid because normalization cannot affect the ratio between capacity without feedback to that with feedback. Some channels cannot be normalized in this manner, i.e., $N(\infty) = \infty$ or $N(\infty) = 0$. The latter case has infinite capacity and thus the corollary applies. The former presents no problems due to the following lemma.

Lemma: Consider the channel without feedback. By the water pouring argument[4] *we know that the signal energy which achieves capacity obeys:*

$$S(\omega) = \begin{cases} K - N(\omega), & N(\omega) \leqq K; \\ 0, & \text{otherwise.} \end{cases}$$

If we define a new noise $N^0(\omega)$

$$N^0(\omega) = \begin{cases} N(\omega), & N(\omega) \leqq K; \\ K, & N(\omega) > K. \end{cases}$$

This new channel has the same capacity without feedback and a larger capacity with feedback.

Proof: The expression for capacity without feedback is the same for $N(\omega)$ and $N^0(\omega)$. The capacity with feedback can only be increased since $N^0(\omega) \leqq N(\omega)$ for all ω. For if the capacity with $N(\omega)$ were larger, one could add a noise with spectrum $N(\omega) - N^0(\omega)$ at the receiver and do just as well as if the noise were $N(\omega)$.

We now normalize the noise, $N^0(\omega)$, in order to apply equation (6), which makes $K = 1$. The capacity without feedback is:

$$\begin{aligned} C &= \frac{1}{4\pi} \int_{-\infty}^{\infty} \ln \frac{1}{N^0(\omega)}\, d\omega, \\ P &= \frac{1}{2\pi} \int_{-\infty}^{\infty} [1 - N^0(\omega)]\, d\omega. \end{aligned} \tag{8}$$

With feedback from equations (6), (7) and (8)

$$I(m; Y_T) \leqq E \int_0^T s^2\, dt - TP + 2TC$$

or

$$\frac{I(m; Y_T)}{T} \leqq 2C.$$

A tighter bound can be obtained by returning to equation (5) and writing:

$$I(m; Y_T) \leqq \frac{1}{2} \left[E \int_0^t s^2\, dt + E \int_0^t z^2\, dt \right] + E \int_0^t sz\, dt,$$

which by the preceding argument is equal to

$$C + E \int_0^t sn^0 \, dt.$$

The correlation Esz^0 is equal to Esn^0 because n^0 and z^0 only differ by a white component. Thus the capacity can be increased only by the correlation of the signal with the noise. The noise n^0 is not the original noise, however the difference occurs only at frequencies not used for signaling without feedback. As $N(w)$ becomes white, the energy in z^0 decreases and consequently Esz^0 must go to zero.

More insight into the problem is supplied by the following theorem.

Theorem 2: Capacity can be attained with a gaussian signal $s(t)$.

Proof: First we observe that

$$E[s(t) + z(t) \mid m, Y_t] = s(t, m, Y_t) + E[z(t) \mid W_t].$$

This is true because $s(t)$ is known given m and Y_t, and $z(t)$ is dependent on W_t which can be calculated given Y_t and $s(t)$. $E[z(t) \mid W_t]$ is a linear functional of $\dot{w}$ because $\dot{w}$ is gaussian.

$$E[z(t) \mid W_t] = \int_0^t K(t, \tau) \, d\omega(\tau).$$

The first term in equation (1) depends only on the correlation properties of $s(t, m, Y_t)$ and $w(\tau)$ and therefore we can use a gaussian s of the appropriate correlation. For the second term we use the property that a least-squares linear estimate has no more energy than the more general least square estimate.

$$Ex^2 = E\hat{x}^2 + E(x - \hat{x})^2 = E\bar{x}^2 + E(x - \bar{x})^2$$

where $\bar{x}$ is the least-square linear estimate of x and $\hat{x}$ is the least-square estimate. Since

$$E(x - \hat{x})^2 \leqq E(x - \bar{x})^2,$$

$$E\hat{x}^2 \geqq E\bar{x}^2.$$

Therefore, since $E[s(t) + z(t) \mid Y_t]$ is the least-squares estimate of $s(t) + z(t)$ given Y_t we have

$$I(m; Y_T) \leqq \frac{1}{2} E \int_0^T E^2[s + z \mid m, Y_t] \, dt - \frac{1}{2} E \int_0^T \widehat{(s + z)}^2 \, dt$$

but for a gaussian signal this inequality is an equality. In addition the

signal power is unchanged and the feedback processor need only be linear. Therefore one need consider only gaussian input and linear processing in calculating capacity.

II. GENERALITY OF THE MODEL

The restrictions on $N(\omega)$ are in fact only needed for $N^0(\omega)$. If a noise spectrum is such that the logarithmic integral of $N^0(\omega)$ is minus infinity then the capacity of the channel is infinite without feedback. Therefore the bound applies to any channel which has a finite capacity without feedback.

The bounds are all valid for noisy feedback as well, however it is not clear that gaussian signals are optimum in that case.

III. ACKNOWLEDGMENT

The author is indebted to J. Salz and J. Mazo for helpful discussion and in particular the evaluation of the integral of $| H(\omega) |^2$.

REFERENCES

1. Shannon, C. E., "The Zero Error Capacity of a Noisy Channel," IRE Trans. Inform. Theory, *IT-2*, No. 3 (September 1956), pp. 8–19.
2. Kadota, T. T., Zakai, M., and Ziv, J., "On the Capacity of Continuous Memoryless Channels with Feedback," to be published in the IEEE Trans. Inform. Theory.
3. Mazo, J. E. and Salz, J., "Spectral Properties of Single-Sideband Angle Modulation," IEEE Trans. on Comm. Tech., *Com-16*, No. 1 (February 1968), pp. 52–62.
4. Shannon, C. E., "A Mathematical Theory of Communication," B.S.T.J., *27*, No. 4 (October 1948), pp. 623–656.

New Theorems on the Equations of Nonlinear DC Transistor Networks

By ALAN N. WILLSON, JR.

(Manuscript received March 26, 1970)

It has long been recognized that equations describing dc transistor networks do not necessarily have unique solutions. The Eccles-Jordan (flip-flop) circuit is an excellent example of one for which the dc equations may have more than one solution.

Only recently, however, has a comprehensive theory concerning matters such as the existence and uniqueness of solutions of the dc equations of general transistor networks begun to take shape. This paper represents another contribution to the evolution of that theory.

A key concept in the development of the recent theory is the concept of a "P_0 matrix." We give a generalization of that concept, showing that one can specify properties possessed by certain pairs of square matrices, analogous to the properties possessed by a single P_0 matrix. Pairs of matrices possessing these properties are called $\mathcal{W}_0$ pairs. Use is made of this $\mathcal{W}_0$ pair concept to prove results which are more general than some of the existing ones. We provide an extension of much of the existing theory in such a manner that a broader class of dc transistor networks may be considered. In particular, the new results provide one with the ability to answer certain questions concerning the existence, uniqueness, boundedness, and so on, of solutions of the equations for any *network which is comprised of transistors, diodes, resistors, and independent sources.*

I. INTRODUCTION

Suppose a network is constructed by connecting in an arbitrary manner any number of transistors, diodes, resistors, and independent voltage and current sources. Without loss of generality, we may consider the network to have the canonical form shown in Fig. 1; that is, we may consider the network to be a multiport containing resistors and inde-

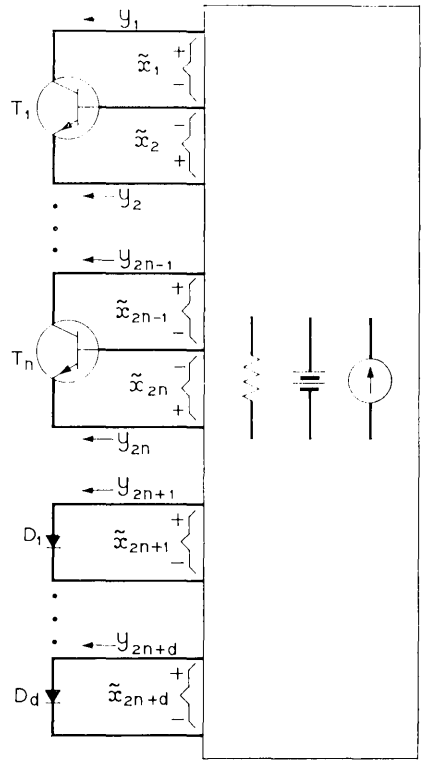


Fig. 1—Canonical form of a transistor network.

pendent sources, with transistors and diodes connected to the ports.*

There are some fundamental questions that one should then, hopefully, be able to answer. For example: Do the equations that describe this dc network have a unique solution? With the exception of certain uniqueness results for a special (but none the less important) class of transistor networks, all of the previous explicit results in Refs. 1, 2, and 3, which have shown methods for obtaining answers to such questions, have been concerned only with the class of transistor networks for which, after setting the value of each independent source to zero, there exists a short-circuit admittance matrix (a G matrix) to characterize the linear

* It will become apparent that the theory can also accommodate many other structures which are of the Fig. 1 type except that the multiport contains additional linear elements (such as controlled sources). We do not stress this point though, since in the present context such elements seem somewhat unnatural.

multiport of Fig. 1. It is the primary purpose of this paper to show how that restriction can be removed. We shall show in fact that almost all of the previous results are but special cases of results that follow from a more general theory in which the assumption of the existence of a G matrix for the linear multiport is unnecessary.*

Section II concerns methods for characterizing a general multiport containing resistors and independent sources. In Section III, we consider the model for a transistor. An equation for dc transistor networks is then developed in Section IV and, after explaining some notation in Section V, we develop the $\mathcal{W}_0$ pair concept in Section VI. Sections VII, VIII and IX show how the $\mathcal{W}_0$ pair concept provides a generalization of the existing results concerning dc transistor networks. Finally, we consider an example network in Section X.

II. LINEAR MULTIPORT CHARACTERIZATION

A multiport having n ports (an n-port) is *characterized* by determining every combination of the $2n$ port voltages and currents that the network admits (see Ref. 4). We discuss here two methods of characterizing multiports that contain resistors and independent sources. The first method makes use of the familiar concept of a hybrid matrix. The second method uses a pair of matrices in a manner that was apparently first suggested—for multiports containing no independent sources—by V. Belevitch.[5]

2.1 *The Hybrid Formalism*

When the value of each independent source is set to zero, for a multiport containing only resistors and independent sources, the multiport becomes, of course, a *resistive* multiport. H. C. So has proved (as a special case of a theorem in Ref. 6) that *any resistive multiport has a hybrid matrix description.* That is, for any resistive n-port, it is always possible to label the port voltage and current variables in such a way that there

* Pragmatists might argue that in any "physical" network, there will always be enough "stray" resistance present which, if taken into account, will guarantee the existence of, say, a G matrix. It seems to this writer, however, that by taking such a point of view, one does not obtain an entirely satisfactory understanding of matters (even *practical* matters). To know that fundamental results do not *depend* (if, in fact, they don't) upon such fortunate occurrences as these (and for many transistor networks this is the case) seems to be the more satisfactory situation. Furthermore, it should be noted that in the analysis of a physical network, to obtain a tractable problem, it often behooves one to neglect the presence of unimportant elements. Thus, it is not necessarily true that such stray resistors will always be present in the model of the network which the analyst desires to consider.

exists an integer m, $0 \leqq m \leqq n$, a pair of n-vectors*

$$x = (i_1 , \cdots , i_m , v_{m+1} , \cdots , v_n)^T,$$
$$y = (v_1 , \cdots , v_m , i_{m+1} , \cdots , i_n)^T,$$

and a real $n \times n$ matrix H, the hybrid matrix, such that the network admits the port variables v_k , i_k as the voltage and current, respectively, at the kth port, for $k = 1, \cdots , n$, if and only if the vectors x and y satisfy

$$y = Hx. \tag{1}$$

Thus, a resistive multiport may always be characterized by a hybrid matrix.

When independent sources whose values are nonzero are present in an otherwise resistive multiport, a hybrid matrix will not generally suffice to characterize the multiport. Clearly the vectors $x = y = (0, 0, \cdots , 0)^T$ which satisfy equation (1) for any matrix H do not always specify an admissible combination of port variables when independent sources are present. One might hope, however, that a characterization of the type

$$y = Hx + c, \tag{2}$$

where c is some constant vector (whose elements are real numbers), might always be possible. Indeed, we are about to show that this is the case. There is one problem, however, that was not present in the consideration of resistive n-ports that must first be dealt with: there are ways to interconnect independent sources and resistors such that the resulting structure doesn't make sense. That is, the independent sources might impose self-contradictory constraints on the network. We rule out such possibilities by agreeing that, when we refer to "a multiport containing resistors and independent sources," we always assume that the multiport possesses the following property:

Assumption: The linear graph that is formed by associating an edge with each resistor, each independent source, and each port, has no cut-sets containing only current source edges for which the values of the current sources cause a violation of Kirchhoff's current law. Similarly, no circuits of voltage source edges for which the values of the voltage sources cause a violation of Kirchhoff's voltage law are present.

This assumption in no way restricts the generality of our work. We

* We use the superscript T to denote the transpose of a vector or a matrix. Thus, the vectors x and y above are both column vectors.

are simply ruling out multiports, like the 2-port of Fig. 2, for which the set of admissible port voltage and current combinations is empty.

We have worded the Assumption so that the presence of, say, a series connection of two 1-ampere current sources in an otherwise resistive multiport does not cause the multiport to be inadmissible. We have done this because no violation of Kirchhoff's laws results from such interconnections of resistors and sources; the network is perfectly legitimate. One should be aware, however, that if "superfluous" sources are present in a network, it will follow that one cannot uniquely determine the value of each branch voltage and current in the network. That is, even though one might be able to uniquely determine the value of the voltage across the *pair* of 1-ampere sources, there is no way to determine the value of the voltage across each individual source. Aside from such ambiguities, it follows (see below and the proof of Theorem 1 in Ref. 6) that the value of all branch voltages and currents can be uniquely determined for a multiport satisfying the Assumption, whenever the values of the "independent" port variables are known.

Theorem: *Any multiport containing resistors and independent sources can be characterized by equation* (2), *where H is a hybrid matrix characterization of the corresponding resistive multiport that is obtained by setting all independent source values to zero, and c is a vector of real numbers.*

A proof of this theorem can be constructed by incorporating a few simple observations and minor modifications into the arguments used by So in Ref. 6. We therefore simply sketch the main ideas: First, if the linear graph mentioned in the Assumption contains any current source cut-sets, then it must be the case (because of that Assumption) that these sources have values such that Kirchhoff's current law is satisfied. That being the case, the port behavior of the multiport will clearly be unaltered if a sufficient number of current sources are removed (by coalescing appropriate nodes) to eliminate such cut-sets. A similar observation applies to voltage source circuits. Therefore without any loss of generality, we may consider the linear graph to have no current source cut-sets and no voltage source circuits. Next, by Lemmas 1 and 2 of Ref. 6, it then follows that there exists a tree* for the linear graph for which all voltage source edges are branches and all current source edges are links. At each port, one of the two port variables is then designated as "independent," the choice depending upon whether the edge corresponding to that port is a branch or a link. The existence of the

* In case the linear graph is not *connected* each reference to the word *tree* should, of course, be changed to *forest*.

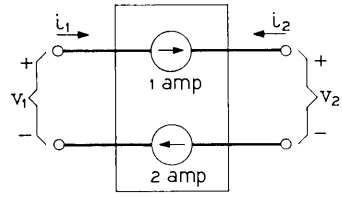


Fig. 2—An inadmissible n-port.

hybrid matrix H and the vector c for the characterization (2) then follows in the same manner as the existence of a hybrid matrix for a resistive multiport follows from So's arguments.

2.2 *Belevitch's Formalism*

For some multiports, it might be that (after setting all independent source values to zero) a hybrid matrix exists such that the vectors x and y in equation (1) satisfy $x = v \equiv (v_1, \cdots, v_n)^T$ and $y = i \equiv (i_1, \cdots, i_n)^T$. In this case the hybrid matrix is given the special name, *admittance matrix*. Similarly, if it happens that H exists such that $x = i$ and $y = v$, then H is called the *impedance matrix*. For many resistive multiports, neither an impedance matrix nor an admittance matrix exists. It is still possible, however, to characterize any n-port for which a hybrid matrix exists in terms of the vectors v and i. Obviously, x and y satisfy equation (1) if and only if v and i satisfy

$$[I_l \,\vdots\, -H_r]v = [H_l \,\vdots\, -I_r]i, \tag{3}$$

where the $n \times m$ matrix H_l and the $n \times (n - m)$ matrix H_r are defined by $H = [H_l \,\vdots\, H_r]$, and similarly $[I_l \,\vdots\, I_r]$ is the $n \times n$ identity matrix.

The characterization (3), being equivalent to equation (1), is perfectly adequate for any resistive n-port. It is, however, but a special case of a more general characterization due to Belevitch, namely:

$$Pv = Qi, \tag{4}$$

where P and Q are $n \times n$ real matrices. Belevitch's characterization can be used for quite a broad class of networks, including some rather pathological ones which require dependent sources, or gyrators and negative resistors to realize, and for which no hybrid characterization exists. For example, the one-port called a *norator*, for which the set of admissible port voltage and current combinations is the set of all pairs of real numbers, may be characterized by $[0]v = [0]i$. We should note, however, that if one allows the aforementioned elements to be present

in an n-port, then even equation (4) cannot always provide a characterization. The *nullator*, for example, a one-port whose only admissible combination of port voltage and current variables is the pair (0, 0), is such an n-port.

When an n-port contains independent sources it can often be characterized by the equation

$$Pv = Qi + c, \tag{5}$$

where P and Q are real $n \times n$ matrices, and c is a constant vector. Clearly, any n-port containing only resistors and independent sources has such a characterization. It is this class of n-ports which is our primary concern in the study of transistor networks. We note, however, that equation (5) is adequate for characterizing a much broader class of n-ports.

III. NONLINEAR TRANSISTOR CHARACTERIZATION

In Fig. 3, a commonly used large signal dc transistor model is displayed. It is easily verified that the voltage and current variables defined in that figure obey the following relationships:

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} 1 & -\alpha_r \\ -\alpha_f & 1 \end{bmatrix} \begin{bmatrix} f_1(v_1) \\ f_2(v_2) \end{bmatrix}, \tag{6}$$

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} \tilde{v}_1 \\ \tilde{v}_2 \end{bmatrix} - \begin{bmatrix} r_e + r_b & r_b \\ r_b & r_c + r_b \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \end{bmatrix}. \tag{7}$$

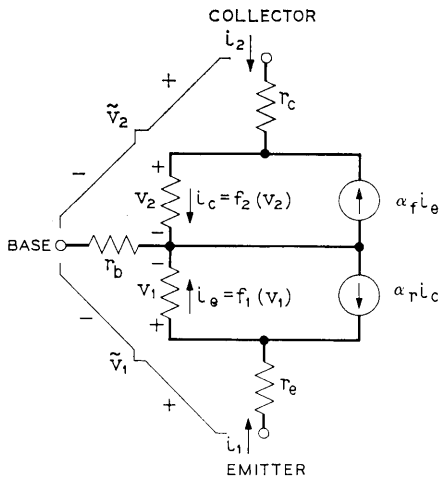


Fig. 3—Large signal dc transistor model.

Each of the parameters α_f and α_r may assume any value in the open interval (0, 1). The parameters r_b , r_c , and r_e , which account for lead resistances, are sometimes omitted by device modelers (their presence is sometimes accounted for by including appropriate additional resistors in the network to which the transistor model is connected). To accommodate these various points of view we specify only, therefore, that the values of the parameters r_b , r_c , and r_e be nonnegative. Thus any or all of them may be zero.

Depending upon whether the transistor being modeled is a pnp or an npn, the graph of each of the functions f_1 and f_2 has one of the general shapes shown in Fig. 4 (at least for values of $|v|$ that are "not too large"). Often these functions are described by an equation of the form

$$f_k(v) = m_k[\exp(n_k v) - 1], \qquad (k = 1, 2), \tag{8}$$

where m_k and n_k are appropriately chosen constants, both being positive for a pnp transistor, and both negative for an npn. On the other hand, for example, a piecewise-linear representation is sometimes specified for f_1 and f_2 .

The nature of the functions f_1 and f_2 for large values of $|v|$ depends upon which assumptions the modeler is willing to make, and which effects he is interested in considering. For large negative (in the pnp case) values of v, for example, the graph of f_k approaches—according to equation (8)—the horizontal asymptote $i = -m_k$. Thus, if the modeler chooses to use equation (8) to describe f_k for all values of v, the range of f_k will not be the entire real line. If, on the other hand, the effect of ohmic surface leakage across the p-n junction is included in the model, the graph of the function f_k will approach asymptotically a straight line having a small, but positive, slope. The range of such a function is, obviously, the whole real line. One might also wish to include the effect of avalanche breakdown in the reverse-biased region. If this is done, the graph of f_k will have a shape reminiscent of that of a Zener diode in the $v < 0$ part of its domain.

In the forward-biased region there are also effects, particularly apparent for large values of v, which the modeler may or may not wish to recognize. For example, there is the so-called high-level injection phenomenon which tends to decrease the value of the forward current and which, using equation (8), is usually accounted for by a decrease in the magnitude of n_k for large values of v. In addition, there is the effect of the ohmic resistance of the crystal which tends to reduce the value of forward current for large values of v.

From the point of view of the device modeler, the question of whether

or not to include some of the effects mentioned above is often a minor issue. For many networks the behavior will be essentially the same whether or not, say, surface leakage is accounted for in the transistor model. From the point of view of the network analyst, however, the situation is somewhat different. For example, the matter of whether or not the functions f_k map the real line *onto* the real line can, in some cases, make the difference between whether or not there exists a solution of the network's equations. Similarly, other results that have been obtained recently (presented later, beginning in Section VII) also seem to depend upon the graphs of the functions f_k having certain special properties.

It seems safe to say that no matter which "special effects" are included (or omitted) in the description of the transistor, the functions f_k may at least be considered to be strictly monotone increasing mappings of the real line into itself. For the purpose of formulating the equations for transistor networks, this is the only hypothesis that we shall make. When additional hypotheses regarding the nature of these functions are needed (to obtain certain results concerning properties of these equations) those hypotheses will be mentioned explicitly. In each case it will be clear that the additional hypotheses are, in some appropriate sense, rather weak.

Similar remarks can be made for the diodes that are shown in Fig. 1, which might also be present in transistor networks. Thus, we assume that each diode is described by an equation of the type $i = f(v)$ where, at this point, we only assume that the function f is a strictly monotone increasing mapping of the real line into itself.

IV. EQUATIONS FOR TRANSISTOR NETWORKS

Suppose we are given a dc network consisting of transistors, diodes, resistors, and independent voltage and current sources, connected together in an arbitrary manner. Let there be n transistors and d diodes. Clearly, there is no loss of generality if we consider the network to be of the type shown in Fig. 1. Using the results of Section III, we may describe the nonlinear devices in the network by the equations

$$y = TF(x), \qquad x = \tilde{x} - Ry, \tag{9}$$

where $T = \text{diag}[T_1, T_2]$, with T_1 a block diagonal matrix with n 2×2 diagonal blocks of the form

$$\begin{bmatrix} 1 & -\alpha_r^{(k)} \\ -\alpha_f^{(k)} & 1 \end{bmatrix}, \quad \text{for} \quad k = 1, \cdots, n, \tag{10}$$

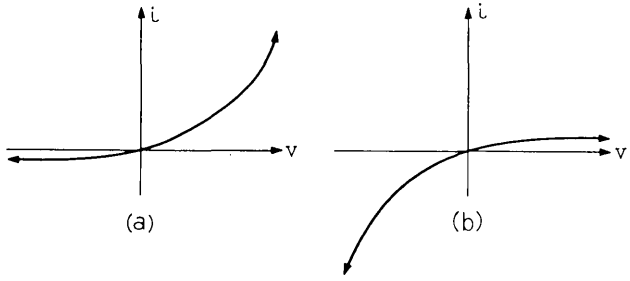


Fig. 4—General shape of the functions f_1 and f_2; (a) pnp transistor, (b) npn transistor.

and T_2 the $d \times d$ identity matrix. Also, $R = \text{diag}\,[R_1, R_2]$, with R_1 a block diagonal matrix with n 2×2 diagonal blocks of the form

$$\begin{bmatrix} r_e^{(k)} + r_b^{(k)} & r_b^{(k)} \\ r_b^{(k)} & r_c^{(k)} + r_b^{(k)} \end{bmatrix}, \qquad k = 1, \cdots, n, \tag{11}$$

and R_2 the $d \times d$ matrix whose elements are all zeros. The function F has the form $F(x) \equiv [f_1(x_1), \cdots, f_{2n+d}(x_{2n+d})]^T$, where each of the f_k is a strictly monotone increasing mapping of the real line into itself.

Using the results of Section II, the effect of the linear multiport in Fig. 1 is to constrain the vectors of port variables, $\tilde{x}$ and y, to obey the relationship

$$P\tilde{x} = -Qy + c, \tag{12}$$

where P and Q are $(2n + d) \times (2n + d)$ real matrices and c is a real $(2n + d)$-vector. The minus sign appears in equation (12) as a consequence of having chosen the reference direction for the port currents (the elements of the vector y) to be opposite to that which is usually assumed.

By using equations (9), we may easily eliminate the variables $\tilde{x}$ and y from equation (12), resulting in the equation

$$(PR + Q)TF(x) + Px = c. \tag{13}$$

The central problem in determining the values of all branch voltages and currents in a dc transistor network is the determination of a solution of equation (13). The rest is relatively straightforward, for if x is a (unique) solution of equation (13), then the (unique) vectors $\tilde{x}$ and y, such that equations (9) and (12) are satisfied, may immediately be computed from equations (9).

Since the matrix T is nonsingular, it follows that whenever either $(PR + Q)$ or P is nonsingular, equation (13) can be transformed into,

respectively, one of the equations

$$F(x) + Ax = b, \tag{14}$$

$$AF(x) + x = b. \tag{15}$$

The first of these equations has been studied rather extensively (see Refs. 1–3 and 7) and for most of the results obtained there, it can be shown that parallel results are possible for equation (15). Both of these equations, however, are but special cases of the equation

$$AF(x) + Bx = c, \tag{16}$$

which accommodates equation (13) directly. It is, therefore, this equation to which we shall now direct our attention. It will be shown that most of the results which have been obtained to date for equation (14) have rather natural (though not obvious) extensions to equation (16). It is important that such extensions be possible because one is often forced to deal with equations like (16) in the analysis of transistor networks. Clearly, this is the case whenever both of the matrices $(PR + Q)$ and P of equation (13) are singular—and this can easily happen (for example, if the matrix R contains all zeros, then it will happen whenever there exists no admittance matrix nor impedance matrix for the linear multiport of Fig. 1).

V. NOTATION

The following notation shall be used throughout the remainder of the paper: For each positive integer n we denote by E^n the n-dimensional Euclidean space, the elements of which are ordered n-tuples of real numbers, which we consider to be column vectors. The origin in E^n is denoted by θ. If $x = (x_1, \cdots, x_n)^T$ and $y = (y_1, \cdots, y_n)^T$ are elements of E^n we denote their inner product by $\langle x, y \rangle = \sum_{k=1}^{n} x_k y_k$. The norm of each $x \in E^n$ is denoted by $||x|| = \langle x, x \rangle^{\frac{1}{2}}$.

If A is an $n \times n$ matrix, then for $k = 1, \cdots, n$, A_k denotes the kth column of A. A principal submatrix of a square matrix A is any square submatrix of A whose main diagonal is contained in the main diagonal of A. A principal minor of A is the determinant of any principal submatrix of A. If D is a diagonal matrix, then $D > 0$ means that each element of the main diagonal is a positive number; similarly, $D \geqq 0$ denotes that each element of the main diagonal is nonnegative. We denote the $n \times n$ identity matrix by either I_n or, when the dimension is unimportant or is clear from the context, simply by I. The direct sum of two matrices A, B is denoted by $A \oplus B$. A square matrix of real

numbers A is said to be strongly row-sum dominant if its elements a_{ij} satisfy $a_{ii} > \sum_{j \neq i} |a_{ij}|$ for $i = 1, \cdots, n$.

If f is a real valued function defined on E^1 then f is said to be monotone increasing if for all $x < y$ it follows that $f(x) \leqq f(y)$. We say that f is strictly monotone increasing if $f(x) < f(y)$ for all $x < y$. For each positive integer n, we denote by $\mathfrak{F}^n$ that collection of mappings of E^n onto itself defined by: $F \in \mathfrak{F}^n$ if and only if there exist, for $i = 1, \cdots, n$, strictly monotone increasing functions f_i mapping E^1 onto E^1 such that for each $x = (x_1, \cdots, x_n)^T \in E^n$, $F(x) = [f_1(x_1), \cdots, f_n(x_n)]^T$.

VI. PAIRS OF MATRICES OF TYPE $\mathfrak{W}_0$

Many of the recent results referred to above, concerning equation (14), have relied heavily upon certain properties that a matrix is known to possess whenever it is a member of a class of matrices that has been given the name P_0. In a similar way the results that follow rely upon useful properties that are possessed by certain *pairs of matrices*. We shall define a class, the elements of which are these pairs of matrices, and give it the name $\mathfrak{W}_0$.

The class of matrices called P_0 was defined by M. Fiedler and V. Pták.[8] They proved that the following properties of a square matrix of real numbers, A, are equivalent:

(*i*) All principal minors of A are nonnegative.

(*ii*) For each vector $x \neq \theta$ there exists an index k such that $x_k \neq 0$ and $x_k(Ax)_k \geqq 0$.

(*iii*) For each vector $x \neq \theta$ there exists a diagonal matrix $D_x \geqq 0$ such that $\langle x, D_x x \rangle > 0$ and $\langle Ax, D_x x \rangle \geqq 0$.

(*iv*) Every real eigenvalue of A, as well as of each principal submatrix of A, is nonnegative.

Sandberg and Willson proved that another property can be added to this list of equivalent properties,[2,3] namely:

(*v*) $\det (D + A) \neq 0$ for every diagonal matrix $D > 0$.

The class of all matrices possessing one (and hence all) of the above properties is called P_0.

We shall now state a theorem which provides a useful generalization of the concept of the class of P_0 matrices.

Definition: For each pair of $n \times n$ matrices (A, B) we shall denote by $\mathfrak{C}(A, B)$ the collection of all the $n \times n$ matrices that can be constructed by juxtaposing columns taken from either A or B while maintaining the original relative ordering of the columns. Thus, $M \in \mathfrak{C}(A, B)$ if and only if for each $k = 1, \cdots, n$, either $M_k = A_k$ or $M_k = B_k$.

Obviously $\mathcal{C}(A, B)$ contains 2^n matrices (for certain pairs (A, B)—namely for those having $A_k = B_k$ for one or more values of k—it can happen that two or more matrices in $\mathcal{C}(A, B)$ are identical).

Definition: The pair of $n \times n$ matrices (M, N) is said to be a complementary pair taken from $\mathcal{C}(A, B)$ if and only if both M and N are members of $\mathcal{C}(A, B)$ and for each $k = 1, \cdots, n$, either $M_k = A_k$ and $N_k = B_k$, or else $M_k = B_k$ and $N_k = A_k$.

It is obvious that (A, B) is a complementary pair taken from $\mathcal{C}(A, B)$. It is also clear that $\mathcal{C}(A, B) = \mathcal{C}(B, A)$ and, moreover, that if (M, N) is any complementary pair taken from $\mathcal{C}(A, B)$, then $\mathcal{C}(M, N) = \mathcal{C}(A, B)$. Furthermore, for each $M \in \mathcal{C}(A, B)$ there exists $N \in \mathcal{C}(A, B)$ such that (M, N) is a complementary pair.

Theorem 1: *The following properties of a pair of $n \times n$ matrices of real numbers (A, B) are equivalent*:

(*i*) *$\det (AD + B) \neq 0$ for every diagonal matrix $D > 0$.*

(*ii*) *There exists a matrix $M \in \mathcal{C}(A, B)$ such that $\det M \neq 0$ and such that $\det M \cdot \det N \geqq 0$ for all $N \in \mathcal{C}(A, B)$.*

(*iii*) *For each vector $x \neq \theta$ there exists an index k such that either $(A^T x)_k \neq 0$ or $(B^T x)_k \neq 0$, and such that $(A^T x)_k (B^T x)_k \geqq 0$.*

(*iv*) *For each vector $x \neq \theta$ there exists a diagonal matrix $D_x \geqq 0$ such that either $\langle A^T x, D_x A^T x \rangle > 0$ or $\langle B^T x, D_x B^T x \rangle > 0$ (that is, such that $\langle A^T x, D_x A^T x \rangle + \langle B^T x, D_x B^T x \rangle > 0$), and such that $\langle A^T x, D_x B^T x \rangle \geqq 0$.*

(*v*) *For each complementary pair of matrices (M, N) taken from $\mathcal{C}(A, B)$, each real value of λ that satisfies $\det(M - \lambda N) = 0$ is nonnegative.*

(*vi*) *There exists a complementary pair of matrices (M, N) taken from $\mathcal{C}(A, B)$ such that $M^{-1}N \in P_0$.*

(*vii*) *There exists a matrix $M \in \mathcal{C}(A, B)$ such that $\det M \neq 0$; and, for any complementary pair of matrices (M, N) taken from $\mathcal{C}(A, B)$ with $\det M \neq 0$, $M^{-1}N \in P_0$.*

In this paper, we do not make use of properties (*iii*), (*iv*), or (*v*) of Theorem 1. The proof that the remaining four properties are equivalent is given in the Appendix. A complete proof of Theorem 1 is given elsewhere.[9]

Definition: The class of all pairs of matrices which possess one (and hence all) of the properties listed in Theorem 1 is called $\mathcal{W}_0$.

To see that properties (*i*) and (*ii*) of Theorem 1 are in fact generalizations of the previously mentioned properties (*v*) and (*i*), respectively, that define P_0 is a simple matter. It happens that for any $n \times n$ matrix B the pair $(I_n, B) \in \mathcal{W}_0$ if and only if $B \in P_0$. (This follows from property (*vii*) of Theorem 1.) With our attention restricted to pairs of matrices of the type (I_n, B), it is clear that property (*i*) of Theorem 1 is equivalent to property (*v*) which determines those matrices B that that are in P_0. Concerning property (*ii*) of Theorem 1, an arbitrary matrix $N \in \mathcal{C}(I_n, B)$ is either the matrix I_n or else, a matrix formed from B by replacing some of the columns of B by the corresponding columns of I_n. Consequently, $\det N = \det B_N$ where B_N is the principal submatrix of B that is formed by removing from B the columns that are not present in N and then removing the corresponding rows. Hence, since $\det I_n \neq 0$, we may take I_n to be the matrix M in property (*ii*) of Theorem 1, and observe that this property then becomes: $\det B_N \geqq 0$ for all $N \in \mathcal{C}(I_n, B)$. It is now clear that this property is equivalent to the property (*i*) that defines the class of P_0 matrices. (Note that there are exactly $2^n - 1$ principal minors for each $n \times n$ matrix, and that the set $\mathcal{C}(I_n, B) \setminus \{I_n\}$ contains exactly $2^n - 1$ members.)

VII. THEOREMS ON EXISTENCE AND UNIQUENESS

7.1 *First Existence and Uniqueness Theorem*

The following theorem, which is proved in Ref. 2, provides a necessary and sufficient condition for the existence of a unique solution of equation (14) for all F that are strictly monotone increasing "diagonal" mappings of E^n *onto* E^n and for all $b \in E^n$.

Theorem 2: *If A is an $n \times n$ matrix of real numbers, then there exists a unique solution of equation* (14) *for each $F \in \mathfrak{F}^n$ and for each $b \in E^n$ if and only if $A \in P_0$.*

Using this theorem along with the results of Section VI we can prove the following (more general) theorem.

Theorem 3: *If A and B are $n \times n$ matrices of real numbers, then there exists a unique solution of equation* (16) *for each $F \in \mathfrak{F}^n$ and each $c \in E^n$ if and only if $(A, B) \in \mathcal{W}_0$.*

Proof: (*if*) Let $(A, B) \in \mathcal{W}_0$. Then, by Theorem 1, there exists a complementary pair (M, N) taken from $\mathcal{C}(A, B)$ such that $M^{-1}N \in P_0$. For each $F \equiv [f_1(\cdot), \cdots, f_n(\cdot)]^T \in \mathfrak{F}^n$ let $G \equiv [g_1(\cdot), \cdots, g_n(\cdot)]^T$ denote the mapping (also in $\mathfrak{F}^n$) defined by

$$g_k(\cdot) = \begin{cases} f_k(\cdot) & \text{if} \quad M_k = A_k \,, \\ f_k^{-1}(\cdot) & \text{if} \quad M_k \neq A_k \,, \end{cases} \quad \text{for} \quad k = 1, \cdots, n.$$

Clearly, the vectors x and y satisfy

$$AF(x) + Bx = MG(y) + Ny$$

if they satisfy the relation

$$y_k = \begin{cases} x_k & \text{if} \quad M_k = A_k \,, \\ f_k(x_k) & \text{if} \quad M_k \neq A_k \,, \end{cases} \quad \text{for} \quad k = 1, \cdots, n, \tag{17}$$

and since this relation defines a homeomorphism of E^n onto itself, it follows that there exists a unique solution of equation (16) for each $c \in E^n$ if there exists a unique solution of the equation

$$MG(y) + Ny = c \tag{18}$$

for each $c \in E^n$. But, that this is so follows immediately from Theorem 2 and from the fact that $M^{-1}N \in P_0$.

(*only if*) Suppose $(A, B) \notin \mathfrak{W}_0$. Then, by Theorem 1, there exists a diagonal matrix $D > 0$ such that $\det(AD + B) = 0$. Choosing $F(x) \equiv Dx$, we have $F \in \mathfrak{F}^n$, while equation (16) does not have, with this choice of F, a unique solution for all $c \in E^n$. □

There are corollaries to Theorem 2, given in Ref. 2, that also may be generalized in a similar manner. For example, the following result is a generalization of an important special case of Corollary 1 of Ref. 2; it shows that the condition $(A, B) \in \mathfrak{W}_0$ is still sufficient to insure the uniqueness of a solution of equation (16) (if a solution exists) even when the mapping F is not *onto*.

Theorem 4: *If* $F(x) \equiv [f_1(x_1), \cdots, f_n(x_n)]^T$, *where each* f_k *is a strictly monotone increasing mapping of* E^1 *into* E^1, *and if* $(A, B) \in \mathfrak{W}_0$, *then there exists at most one solution of equation* (16) *for each* $c \in E^n$.

Proof: Suppose that, for some $c \in E^n$, x^1 and x^2 are solutions of equation (16) with $x^1 - x^2 \neq \theta$. Then, $A[F(x^1) - F(x^2)] + B(x^1 - x^2) = \theta$. But then, since F is a strictly monotone increasing "diagonal" mapping, there exists a diagonal matrix $D > 0$ such that $F(x^1) - F(x^2) = D(x^1 - x^2)$, and hence $(AD + B)(x^1 - x^2) = \theta$. Since $x^1 - x^2 \neq \theta$ it follows that $\det(AD + B) = 0$, which implies that $(A, B) \notin \mathfrak{W}_0$. □

7.2 *A Nonuniqueness Theorem*

From the proof of the "only if" part of Theorem 2 (given in Ref. 2) it follows that whenever $A \notin P_0$, there exists a mapping $F \in \mathfrak{F}^n$ and a

vector $b \in E^n$ such that equation (14) has more than one solution. On the other hand, even if $A \notin P_0$, if the mapping $F \in \mathfrak{F}^n$ is "fixed," then it is easy to see that the nonuniqueness of solutions of equation (14) need not necessarily follow for any $b \in E^n$[take $F(x) \equiv x$ and $Ax \equiv -2x$, for example]. I. W. Sandberg has shown,[10] however, that if one assumes that the "fixed" mapping F has another special property, rather than assuming that $F \in \mathfrak{F}^n$, then the nonuniqueness of solutions of equation (14) follows, for some $b \in E^n$, whenever $A \notin P_0$. Moreover, he has shown that under these hypotheses and for any $\delta > 0$, there exists some $b \in E^n$ such that equation (14) has two solutions, x and y, which satisfy $||x - y|| = \delta$. The special property that F is assumed to have is given in the following definition (in words, the property is: that it be possible to draw a straight line having any given positive slope, and any given length, between some pair of points on the graph of each of the functions f_k).

Definition: For each positive integer n we denote by $\mathcal{E}^n$ that collection of mappings of E^n into itself defined by: $F \in \mathcal{E}^n$ if and only if there exist, for $k = 1, \cdots, n$, continuous functions f_k mapping E^1 into E^1 such that for each $x \in E^n$, $F(x) = [f_1(x_1), \cdots, f_n(x_n)]^T$, with each of the f_k satisfying, for all $\beta > 0$,

$$\inf \{f_k(\alpha + \beta) - f_k(\alpha): -\infty < \alpha < \infty\} = 0,$$

$$\sup \{f_k(\alpha + \beta) - f_k(\alpha): -\infty < \alpha < \infty\} = \infty.$$

By using Theorem 1 it is possible to prove the following generalization of Sandberg's result:

Theorem 5: *Let $F \in \mathcal{E}^n$, let $(A, B) \notin \mathcal{W}_0$ be a pair of real $n \times n$ matrices, and let δ be a positive constant. Then, for some $c \in E^n$ there exist solutions of equation* (16), *x and y, satisfying $||x - y|| = \delta$.*

Proof: Since $(A, B) \notin \mathcal{W}_0$ there exists a diagonal matrix $D = \text{diag}(d_1, \cdots, d_n) > 0$, such that $\det(AD + B) = 0$. Therefore, there exists $x^* \in E^n$, with $||x^*|| = \delta$, such that $(AD + B)x^* = \theta$. Since $F \in \mathcal{E}^n$ there exists $x \in E^n$ such that

$$f_k(x_k) - f_k(x_k - x_k^*) = x_k^* d_k, \qquad \text{for} \qquad k = 1, \cdots, n.$$

Let $c = AF(x) + Bx$, and let $y = x - x^*$. Then

$$\begin{aligned} A[F(x) - F(y)] + B(x - y) &= A[F(x) - F(x - x^*)] + Bx^* \\ &= (AD + B)x^* = \theta. \quad \square \end{aligned}$$

For a mapping F to be a member of $\mathcal{E}^n$, it is not necessary that $F \in \mathfrak{F}^n$. It follows from the above definition of $\mathcal{E}^n$ that $F \in \mathcal{E}^n$ implies that each

of the functions f_k is a monotone increasing function from E^1 onto some interval in E^1 whose length is infinite; the f_k need not, however, be *strictly* monotone increasing, nor *onto* E^1. For those $F \in \mathcal{E}^n$ for which each of the functions f_k is strictly monotone increasing, we have the following corollary to the two preceding theorems.

Corollary: *Let* $F(x) \equiv [f_1(x_1), \cdots, f_n(x_n)]^T \in \mathcal{E}^n$ *and let each of the functions* f_k *be strictly monotone increasing. Then there exists at most one solution of equation* (16) *for each* $c \in E^n$ *if and only if* $(A, B) \in \mathcal{W}_0$.

VIII. RESULTS ON CONTINUITY AND BOUNDEDNESS

For many systems whose behavior is described by an equation having the form (16), the vector c may be regarded as the system's input and the vector x may be regarded as the system's response or output. Those properties that one might expect well-behaved systems to possess are likely to include continuity and boundedness. Thus, one might expect (*i*) "small" changes to result in the value of the system's output when "small" changes are made in the value of the system's input, and (*ii*) a bounded sequence of input vectors to yield a bounded sequence of outputs. We now show that such properties are indeed possessed by the type of system that is the main concern of this paper.

8.1 *Continuity*

When the $n \times n$ matrix A is a member of the class P_0 and the mapping $F \in \mathcal{F}^n$, it follows that the solution x of equation (14) is a continuous function of the (input) vector b.[2] Using this fact, it is easy to prove the following theorem.

Theorem 6: *For each* $F \in \mathcal{F}^n$ *and each pair of* $n \times n$ *matrices* $(A, B) \in \mathcal{W}_0$ *the solution* x *of equation* (16) *is a continuous function of the vector* c.

Proof: Proceeding as in the "if" part of the proof of Theorem 3, we see that the theorem follows immediately from the facts that equation (17) is a homeomorphism and that the aforementioned result guarantees that y, the solution of equation (18), is a continuous function of c. □

8.2 *Boundedness*

In Ref. 2 a theorem (Theorem 5) is proved which shows that, when $F \in \mathcal{F}^n$ and $A \in P_0$, bounds can be obtained for the solution of equation (14) whenever bounds for $b \in E^n$ are given. The proof of a more general theorem concerning equation (16) can be constructed quite easily by using that theorem, and by using the same technique that was used in the proof of the preceding theorem, along with the trivial observations:

(*i*) For any nonsingular $n \times n$ matrix of real numbers, M, and any real numbers $\alpha_i \leqq \beta_i$, $i = 1, \cdots, n$, there exist real numbers, $\alpha_i' \leqq \beta_i'$, $i = 1, \cdots, n$, such that when each of the components c_i of the vector c satisfies $\alpha_i \leqq c_i \leqq \beta_i$, it follows that $\alpha_i' \leqq (M^{-1}c)_i \leqq \beta_i'$, for $i = 1, \cdots, n$.

(*ii*) For any given real numbers $\gamma_i \leqq \delta_i$, $i = 1, \cdots, n$, there exist for the homeomorphism (17), real numbers $\gamma_i' \leqq \delta_i'$, $i = 1, \cdots, n$, such that whenever x, y satisfy equation (17) with $\gamma_i \leqq y_i \leqq \delta_i$, for $i = 1, \cdots, n$, it follows that $\gamma_i' \leqq x_i \leqq \delta_i'$, for $i = 1, \cdots, n$.

The more general theorem, whose quite obvious proof is omitted, is the following:

Theorem 7: *Let* $F \in \mathfrak{F}^n$, *let* $(A, B) \in \mathfrak{W}_0$ *be a pair of* $n \times n$ *matrices, and, for* $i = 1, \cdots, n$, *let* $\alpha_i \leqq \beta_i$ *be given. There exist, for* $i = 1, \cdots, n$, *real numbers* $\gamma_i \leqq \delta_i$ *such that for any* $c = (c_1, \cdots, c_n)^T \in E^n$ *with* $\alpha_i \leqq c_i \leqq \beta_i$ *for* $i = 1, \cdots, n$, *if* x *satisfies equation* (16), *then* $\gamma_i \leqq x_i \leqq \delta_i$ *for* $i = 1, \cdots, n$.

According to Theorem 7, $(A, B) \in \mathfrak{W}_0$ is a sufficient condition for a bounded sequence of vectors c to yield a bounded sequence of solution vectors of equation (16), for all $F \in \mathfrak{F}^n$. The following theorem shows that $(A, B) \in \mathfrak{W}_0$ is also a necessary condition.

Theorem 8: *If* (A, B) *is a pair of real* $n \times n$ *matrices, then* $(A, B) \in \mathfrak{W}_0$ *if and only if for each* $F \in \mathfrak{F}^n$ *and each unbounded sequence of points* $x^1, x^2, x^3, \cdots$ *in* E^n, *the corresponding sequence* $c^1, c^2, c^3, \cdots$ $[c^k = AF(x^k) + Bx^k, k = 1, 2, 3, \cdots]$ *is unbounded.*

This theorem, which is a generalization of Theorem 4 of Ref. 2, can be proved in a manner which is a quite obvious generalization of the proof, given there, of that theorem. Thus, an appeal to Theorem 7 proves the "only if" part, and the "if" part is proved by assuming that $(A, B) \notin \mathfrak{W}_0$ and then choosing the same kind of mapping $F \in \mathfrak{F}^n$ as was chosen in Ref. 2, for which an unbounded sequence of vectors x^k yields a bounded sequence of vectors c^k.

IX. COMPUTATION OF THE SOLUTION

A. Gersho[7] has shown that whenever $F \in \mathfrak{F}^n \cap C^1$ (that is, whenever each of the functions f_k is a continuously differentiable strictly monotone increasing mapping of the real line onto itself), it is possible to compute the solution of equation (14), for any $A \in P_0$ and any $b \in E^n$, by making use of a gradient descent algorithm due to A. A. Goldstein.[11] The following theorem extends this result to the class of equations of the type (16).

Theorem 9: *Let M be an arbitrary positive definite symmetric matrix, and let $Q : E^n \to E^1$ be defined by*

$$Q(x) = [AF(x) + Bx - c]^T M[AF(x) + Bx - c],$$

where $F \in \mathfrak{F}^n \cap C^1$, $(A, B) \in \mathfrak{W}_0$, and $c \in E^n$. For each $x \in E^n$ and each $\gamma \geqq 0$ let

$$g(x, \gamma) = \begin{cases} \dfrac{Q(x) - Q[x - \gamma \nabla Q(x)]}{\gamma \,||\, \nabla Q(x) \,||^2}, & \gamma > 0; \\ 1, & \gamma = 0; \end{cases}$$

where $\nabla Q(x)$ denotes the gradient of Q at the point x. Then, if δ is any real number satisfying $0 < \delta \leqq \frac{1}{2}$, and if x^0 is an arbitrary point in E^n, the sequence $\{x^k : k = 0, 1, 2, \cdots\}$ converges to the solution of equation (16), *where (for $k = 0, 1, 2, \cdots$) the x^k satisfy*

$$x^{k+1} = x^k - \gamma^k \nabla Q(x^k),$$

each γ^k being any real number that satisfies $\delta \leqq g(x^k, \gamma^k) \leqq 1 - \delta$ if $g(x^k, 1) < \delta$, or $\gamma^k = 1$ if $g(x^k, 1) \geqq \delta$.

Proof: This proof uses generalizations of some of the ideas in Ref. 7 and relies ultimately upon the Goldstein algorithm.[11]

We first remark that the sequence $\{x^k\}$ is well-defined: It is easy to show (see the first part of the proof of Theorem 1, p. 31, Ref. 11) that for each $x \in E^n$, $g(x, \cdot)$ is a continuous function on $[0, \infty)$. This being the case, it is clear that if $g(x^k, 1) < \delta$, then for each ξ in the interval $[\delta, 1)$—and, in particular, for each ξ in the interval $[\delta, 1 - \delta]$—there is some γ^k in the interval $(0, 1)$ such that $g(x^k, \gamma^k) = \xi$.

Let $S = \{x \in E^n : Q(x) \leqq Q(x^0)\}$. Using the fact that M is a positive definite symmetric matrix, and using the fact that $F \in \mathfrak{F}^n$, $(A, B) \in \mathfrak{W}_0$ implies that $|| AF(x) + Bx || \to \infty$ if and only if $|| x || \to \infty$ (Theorem 8) we have that the set $S \subset E^n$ is bounded. By continuity of Q, S is closed. Thus, S is compact and, therefore, the gradient ∇Q (which is continuous on E^n, since $F \in C^1$) is uniformly continuous on S, and ∇Q is bounded on S. Also, Q is bounded below on S. [Indeed, we have $Q \geqq 0$ on E^n and by the existence and uniqueness theorem, Theorem 3, there exists exactly one point x^* ($x^* \in S$) at which $Q(x^*) = 0$.]

It is easily verified that, for each $x \in E^n$,

$$\nabla Q(x) = 2(AD_x + B)^T M[AF(x) + Bx - c],$$

where, for $k = 1, \cdots, n$, the kth diagonal element of the diagonal matrix $D_x > 0$ has the value of the derivative of the function f_k, evaluated at the point x_k. Since $(A, B) \in \mathfrak{W}_0$ implies that $\det (AD_x + B) \neq 0$, and

since $\det M \neq 0$, it follows that $\nabla Q(x) = \theta$ if and only if x is the solution of equation (16).

In view of the above, it follows directly from Goldstein's theorem that the sequence $\{x^k\}$ converges to the solution of equation (16). □

Other methods of computing the solution of equation (16), in certain cases, also exist. If one performs a transformation of the type (17) on the independent variable x (in theory this can always be done) then the solution of equation (16) can easily be computed by first computing the solution of an equation of the type $G(y) + M^{-1}Ny = M^{-1}c$, where $G \in \mathfrak{F}^n$ and $M^{-1}N \in P_0$. Methods of computing the solution of certain equations of this type may be found in Refs. 1–3.

X. EXAMPLE

With the aid of the modern computing facilities that are commonly available today, it is clearly a rather routine matter to obtain an equation of the type (16) for any given transistor network. Moreover, it is not unfeasible, even for networks of moderately large size (say, up to 4 or 5 transistors), to consider the straightforward evaluation of the 2^n determinants specified in property (2) of Theorem 1, and thereby resolve the issue of whether or not the matrices involved in the equation are a $\mathfrak{W}_0$ pair. Due regard would of course have to be paid to the matter of performing sufficiently accurate computations.

On the other hand, even without the aid of a computer, it should often be possible to use a little ingenuity and a few devices* to reduce the computations involved in the application of the above theory to many specific problems to a point where they will just about fit onto the back of an envelope. Consider, for example, the following analysis of a three-transistor network:

For the network of Fig. 5, the voltage and current variables defined there must satisfy the following equations:

$$\begin{bmatrix} i_1 \\ i_2 \\ i_3 \\ i_4 \\ i_5 \\ i_6 \end{bmatrix} = \begin{bmatrix} T^{(1)} & & \\ & T^{(2)} & \\ & & T^{(3)} \end{bmatrix} \begin{bmatrix} f_1(v_1) \\ f_2(v_2) \\ f_3(v_3) \\ f_4(v_4) \\ f_5(v_5) \\ f_6(v_6) \end{bmatrix}, \tag{19}$$

* According to R. Bellman: "a device is a trick that works at least twice." [12]

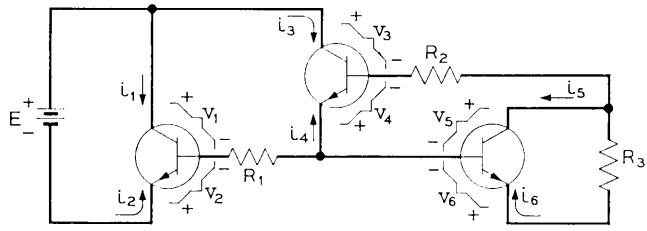


Fig. 5—Example of a three transistor network.

$$\begin{bmatrix} v_1 \\ v_3 \\ v_6 \\ -i_2 \\ -i_4 \\ -i_5 \end{bmatrix} = \left[\begin{array}{ccc:ccc} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & R_1 & 0 & 1 & 1 & 0 \\ 0 & 0 & R_3 & 0 & 0 & 1 \\ \hdashline -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & G_2 & G_2 \\ 0 & 0 & -1 & 0 & G_2 & G_2 \end{array}\right] \begin{bmatrix} -i_1 \\ -i_3 \\ -i_6 \\ v_2 \\ v_4 \\ v_5 \end{bmatrix} + \begin{bmatrix} E \\ E \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{20}$$

where (we are using the transistor model of Fig. 3, with $r_b = r_c = r_e = 0$) each of the 2×2 matrices $T^{(k)}$, $k = 1, 2, 3$, is of the form (10). A hybrid characterization has been used for the linear part of the network. As indicated in equation (3), this hybrid characterization can easily be converted into a characterization of the Belevitch type. Thus, denoting the 3×3 blocks of the hybrid matrix in equation (20) by H_{11}, H_{12}, H_{21}, H_{22}, in the usual manner, one obtains

$$\begin{bmatrix} I & -H_{12} \\ 0 & -H_{22} \end{bmatrix} v = -\begin{bmatrix} H_{11} & 0 \\ H_{21} & -I \end{bmatrix} i + c, \tag{21}$$

where $v = (v_1, v_3, v_6, v_2, v_4, v_5)^T$ and i is similarly defined. We could now simply reorder the columns of each matrix in equation (21) in such a way that the resulting equation would have the same form, except that the subscripts on the components of the vectors v and i would occur in the natural order (1, 2, 3, 4, 5, 6) and then use that equation, along with equation (19), to produce an equation of the type (16) for our network. In this example, though, it's probably easier to reorder the rows and columns of the matrix T (recall, $T = T^{(1)} \oplus T^{(2)} \oplus T^{(3)}$) to obtain from equation (19) an equation that is compatible with equation (21). Thus,

$$i = \begin{bmatrix} I & -P \\ -Q & I \end{bmatrix} F(v), \tag{22}$$

where

$$F(v) \equiv [f_1(v_1), f_3(v_3), f_6(v_6), f_2(v_2), f_4(v_4), f_5(v_5)]^T,$$

and

$$P = \text{diag}\,[\alpha_r^{(1)}, \alpha_r^{(2)}, \alpha_f^{(3)}], \qquad Q = \text{diag}\,[\alpha_f^{(1)}, \alpha_f^{(2)}, \alpha_r^{(3)}].$$

Eliminating i from equations (21) and (22), we obtain

$$\begin{bmatrix} H_{11} & 0 \\ H_{21} & -I \end{bmatrix} \begin{bmatrix} I & -P \\ -Q & I \end{bmatrix} F(v) + \begin{bmatrix} I & -H_{12} \\ 0 & -H_{22} \end{bmatrix} v = c. \tag{23}$$

Note that since $\det H_{11} = \det H_{22} = 0$, it is impossible to put this equation into either of the forms (14) or (15). Clearly this would be the same situation no matter which ordering of subscripts was chosen for the components of v. The cause of the difficulty is simply the fact that neither an impedance matrix nor an admittance matrix exists for the linear part of our network.

Let us determine whether or not the pair of matrices

$$\left(\begin{bmatrix} H_{11} & 0 \\ H_{21} & -I \end{bmatrix} \begin{bmatrix} I & -P \\ -Q & I \end{bmatrix}, \begin{bmatrix} I & -H_{12} \\ 0 & -H_{22} \end{bmatrix} \right)$$

is a $\mathfrak{W}_0$ pair. We shall try to verify property (1) of Theorem 1. Let $\delta_1, \cdots, \delta_6$ denote arbitrary positive real numbers, and let $\Delta_I = \text{diag}\,(\delta_1, \delta_2, \delta_3)$, $\Delta_{II} = \text{diag}\,(\delta_4, \delta_5, \delta_6)$. We wish to show that

$$\det \left\{ \begin{bmatrix} H_{11} & 0 \\ H_{21} & -I \end{bmatrix} \begin{bmatrix} I & -P \\ -Q & I \end{bmatrix} \begin{bmatrix} \Delta_I^{-1} & 0 \\ 0 & \Delta_{II} \end{bmatrix} + \begin{bmatrix} I & -H_{12} \\ 0 & -H_{22} \end{bmatrix} \right\} \neq 0.$$

By multiplying the above matrix on the left by the (nonsingular) matrix diag $(I_3, -I_3)$ and then multiplying on the right by diag (Δ_I, I_3), we obtain the equivalent statement:

$$\det \left[\begin{array}{c|c} H_{11} + \Delta_I & -H_{12} - H_{11}P\,\Delta_{II} \\ \hline -H_{21} - Q & H_{22} + (I + H_{21}P)\,\Delta_{II} \end{array} \right] \neq 0.$$

The 3×3 submatrix in the upper left corner is nonsingular and diagonal. The 3×3 submatrix in the lower left corner can be diagonalized by performing a single elementary row operation on the matrix; namely, by subtracting $1/(\delta_2 + R_1)$ times the second row from the fourth row. Having done this, our problem reduces to one of showing that

$$\det \left[\begin{array}{ccc|ccc} \delta_1 & 0 & 0 & -1 & 0 & 0 \\ 0 & \delta_2 + R_1 & 0 & -1 & -(1 + \alpha_r^{(2)} R_1\, \delta_5) & 0 \\ 0 & 0 & \delta_3 + R_3 & 0 & 0 & -(1 + \alpha_f^{(3)} R_3\, \delta_6) \\ \hline 1 - \alpha_f^{(1)} & 0 & 0 & (1 - \alpha_r^{(1)})\, \delta_4 + \dfrac{1}{\delta_2 + R_1} & \dfrac{1 - \alpha_r^{(2)}\, \delta_2\, \delta_5}{\delta_2 + R_1} & 0 \\ 0 & 1 - \alpha_f^{(2)} & 0 & 0 & G_2 + (1 - \alpha_r^{(2)})\, \delta_5 & G_2 \\ 0 & 0 & 1 - \alpha_r^{(3)} & 0 & G_2 & G_2 + (1 - \alpha_f^{(3)})\, \delta_6 \end{array}\right] \neq 0.$$

It is easy to verify that whenever $\det A_{11} \neq 0$, then

$$\det \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \neq 0$$

if and only if $\det (A_{22} - A_{21}A_{11}^{-1}A_{12}) \neq 0$. In our case both A_{11} and A_{21} are diagonal and hence we can immediately reduce our problem to:

$$\det \begin{bmatrix} (1-\alpha_r^{(1)})\delta_4 + \dfrac{1}{\delta_2+R_1} + \dfrac{1-\alpha_f^{(1)}}{\delta_1} & \dfrac{1-\alpha_r^{(2)}\delta_2\delta_5}{\delta_2+R_1} & 0 \\ \dfrac{1-\alpha_f^{(2)}}{\delta_2+R_1} & G_2+(1-\alpha_r^{(2)})\delta_5+\dfrac{1-\alpha_f^{(2)}}{\delta_2+R_1}(1+\alpha_r^{(2)}R_1\delta_5) & G_2 \\ 0 & G_2 & G_2+(1-\alpha_f^{(3)})\delta_6+\dfrac{1-\alpha_r^{(3)}}{\delta_3+R_3}(1+\alpha_f^{(3)}R_3\delta_6) \end{bmatrix} \neq 0.$$

It is obvious that the above determinant is always positive. First, note that every term in the matrix is nonnegative except, possibly, the (1, 2) term, which may be either positive or negative (or zero). In the event that the (1, 2) term is positive (or zero), we have $1/(\delta_2 + R_1) \geqq (1 - \alpha_r^{(2)}\delta_2\delta_5)/(\delta_2 + R_1)$, and hence we observe that the matrix is strongly row-sum dominant. This implies that its determinant is positive.

In the event that the (1, 2) term is negative, we do not necessarily have dominance; however, considering an expansion of the determinant along its first row we see that, because of the assumption that the (1, 2) term is negative, the value of the determinant is computed as the sum of two *positive* terms.

We have thus shown that, no matter which (positive) values are assigned to R_1, R_2, R_3, or which values the transistor's current gains assume $[0 < \alpha_f^{(k)} < 1, 0 < \alpha_r^{(k)} < 1]$, the pair of 6×6 matrices that appear in equation (23) is a $\mathfrak{W}_0$ pair. Thus, all of the results concerning a solution's existence, uniqueness, continuity, boundedness, and so on, hold for this equation.

XI. ACKNOWLEDGMENT

The author has benefited from discussions with his colleagues I. W. Sandberg and H. C. So.

APPENDIX

Proof of Part of Theorem 1

In this appendix we prove the equivalence of properties (*i*), (*ii*), (*vi*), and (*vii*) of Theorem 1, which define the class of pairs of matrices $\mathfrak{W}_0$. We omit the proof of the equivalence of the three remaining properties, since those properties are not referred to in this paper. A complete proof of Theorem 1 is given elsewhere.[9] We begin by proving a useful lemma:

Lemma 1: *For each positive integer n the polynomial*

$$(c_0)\, d_1\, d_2 \cdots d_n + (c_1)\, d_1\, d_2 \cdots d_{n-1} + \cdots + (c_n)\, d_2 \cdots d_n$$

$$+ (c_{n+1})\, d_1\, d_2 \cdots d_{n-2} + \cdots + (c_{n(n+1)/2})\, d_3 \cdots d_n + \cdots + (c_{2^n-1})$$

in the n variables $d_1, d_2, \cdots, d_n$ *is nonzero for all positive values of the variables if and only if at least one of the coefficients* $c_0, \cdots, c_{2^n-1}$ *is nonzero, and all nonzero coefficients have the same sign.*

Proof: (By induction) For $n = 1$ the statement is obviously true. Let N be a positive integer. Then any polynomial of the above type in $N + 1$ variables, $(c_0)d_1 \cdots d_{N+1} + \cdots + (c_{2^{N+1}-1})$, can be written as $P(d_1, \cdots, d_N)\cdot d_{N+1} + Q(d_1, \cdots, d_N)$ where P and Q are both polynomials of the above type in N variables. Then, assuming that the statement is true for $n = N$, $P + Q \not\equiv 0$ and $P\cdot Q \geqq 0$ for all positive values of the variables $d_1, \cdots, d_N$ if and only if at least one of the coefficients $c_0, \cdots, c_{2^{N+1}-1}$ is nonzero and all nonzero coefficients have the same sign. But, we know that $P\cdot d_{N+1} + Q \not\equiv 0$ for all $d_{N+1} > 0$ if and only if $P + Q \not\equiv 0$ and $P\cdot Q \geqq 0$. □

A.1 *Property* (*i*) *is Equivalent to* (*ii*)

Let $D = \text{diag}\,(d_1, \cdots, d_n)$. By expanding $\det\,(AD + B)$ along the first column we have

$$\det\,(AD + B) = d_1\cdot\det P + \det Q,$$

where the first columns of P and Q satisfy $P_1 = A_1$, $Q_1 = B_1$, and for $k = 2, \cdots, n$, $P_k = Q_k = (AD + B)_k$. Both P and Q are independent of d_1. We now expand $\det P$ and $\det Q$ along their second columns, resulting in

$$\det P = d_2\cdot\det R + \det S,$$

$$\det Q = d_2\cdot\det U + \det V,$$

and hence,

$$\det (AD + B) = d_1 d_2 \cdot \det R + d_1 \cdot \det S + d_2 \cdot \det U + \det V,$$

where

$$\begin{aligned} R_1 &= A_1 , & R_2 &= A_2 , \\ S_1 &= A_1 , & S_2 &= B_2 , \\ U_1 &= B_1 , & U_2 &= A_2 , \\ V_1 &= B_1 , & V_2 &= B_2 , \end{aligned}$$

and for $k = 3, \cdots, n$,

$$R_k = S_k = U_k = V_k = (AD + B)_k .$$

Proceeding in this manner until all columns of $(AD + B)$ have been encountered, we obtain an expansion of $\det (AD + B)$ as a polynomial in the variables $\{d_1, d_2, \cdots, d_n\}$ whose coefficients are the determinants of the matrices in $\mathcal{C}(A, B)$. By using Lemma 1 it thus follows that (*i*) and (*ii*) are equivalent.

A.2 Property (vi) Follows from (i) and (ii)

According to (*ii*) there exists a complementary pair of matrices (M, N) taken from $\mathcal{C}(A, B)$ such that $\det M \neq 0$. Let $D = \operatorname{diag} (d_1, \cdots, d_n) > 0$, then $\det (M^{-1}N + D) \neq 0$ if and only if $\det (MD + N) \neq 0$. But, using property (*i*), $\det (MD + N) = \det (A\hat{D} + B) \cdot \det \tilde{D} \neq 0$, where the matrices $\hat{D} = \operatorname{diag} (\hat{d}_1, \cdots, \hat{d}_n) > 0$ and $\tilde{D} = \operatorname{diag} (\tilde{d}_1, \cdots, \tilde{d}_n) > 0$ are defined by $\hat{d}_k = d_k$ and $\tilde{d}_k = 1$ if $M_k = A_k$, and $\hat{d}_k = 1/d_k$, $\tilde{d}_k = d_k$ otherwise (for $k = 1, \cdots, n$). Thus, $M^{-1}N \in P_0$.

A.3 Property (i) Follows from (vi)

Using the notation above, it is clear that for each diagonal matrix $D > 0$, $\det (AD + B) = \det (M\hat{D} + N) \cdot \det \tilde{D}$. Thus, if $M^{-1}N \in P_0$ it follows that $\det (AD + B) \neq 0$.

A.4 Property (vii) is Equivalent to (vi)

Clearly property (*vi*) follows from property (*vii*). Thus, we need only prove that (*vi*) implies (*vii*). Let (M, N) and (P, Q) both be complementary pairs taken from $\mathcal{C}(A, B)$ with $M^{-1}N \in P_0$ and $\det P \neq 0$. For any $D = \operatorname{diag} (d_1, \cdots, d_n) > 0$, $\det (P^{-1}Q + D) \neq 0$ if and only if $\det (PD + Q) \neq 0$. But $\det (PD + Q) = \det (M\hat{D} + N) \cdot \det \tilde{D} \neq 0$, where the matrices $\hat{D} = \operatorname{diag} (\hat{d}_1, \cdots, \hat{d}_n) > 0$ and $\tilde{D} = \operatorname{diag} (\tilde{d}_1, \cdots, \tilde{d}_n) > 0$ are defined by $\hat{d}_k = d_k$ and $\tilde{d}_k = 1$ if $P_k = M_k$, and $\hat{d}_k = 1/d_k$, $\tilde{d}_k = d_k$ otherwise (for $k = 1, \cdots, n$). Thus, $P^{-1}Q \in P_0$. □

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. Kuh, E. S., and Rohrer, R. A., *Theory of Linear Active Networks*, San Francisco: Holden-Day, Inc., 1967, p. 2.
5. Belevitch, V., "Four-Dimensional Transformations of 4-Pole Matrices with Applications to the Synthesis of Reactance 4-Poles," IRE Trans. Circuit Theory, *CT-3*, No. 2 (June 1956), pp. 105–111.
6. So, H. C., "On the Hybrid Description of a Linear n-Port Resulting from the Extraction of Arbitrarily Specified Elements," IEEE Trans. Circuit Theory, *CT-12*, No. 3 (September 1965), pp. 381–387.
7. Gersho, A., "Solving Nonlinear Network Equations Using Optimization Techniques," B.S.T.J., *48*, No. 9 (November 1969), pp. 3135–3138.
8. Fiedler, M., and Pták, V., "Some Generalizations of Positive Definiteness and Monotonicity," Numer. Math., *9*, No. 2 (1966), pp. 163–172.
9. Willson, A. N., Jr., "A Useful Generalization of the P_0 Matrix Concept," to be published.
10. Sandberg, I. W., "Theorems on the Analysis of Nonlinear Transistor Networks," B.S.T.J., *49*, No. 1 (January 1970), pp. 95–114.
11. Goldstein, A. A., *Constructive Real Analysis*, New York: Harper and Row, 1967, p. 31.
12. Bellman, R., *Stability Theory of Differential Equations*, New York: Dover, 1969, p. ix.

Theorems on the Computation of the Transient Response of Nonlinear Networks Containing Transistors and Diodes

By I. W. SANDBERG

(Manuscript received June 15, 1970)

We consider in detail the nonlinear equations encountered at each time step when certain implicit numerical-integration algorithms are used. In terms of only the properties of the Jacobian matrix of the pertinent set of differential equations, we present necessary and sufficient conditions for the existence and uniqueness of the solution of the nonlinear equations for all continuous forcing functions and any given step size. Since engineers often think about dynamic nonlinear transistor network problems in terms of the eigenvalues of the relevant Jacobian matrix, the results described are of immediate conceptual value. In particular, it is possible to carry out the algorithms whenever the conditions presented are satisfied.

Several other types of results are also presented. For example, for a special but significant and useful numerical-integration formula, theorems are proved concerning properties of the computed sequence such as the extent to which the sequence is relatively immune to small local errors introduced at each step as a result of the fact that it is ordinarily not possible to compute the solution of a certain equation exactly.

All of the results are concerned with network models that are often used in computer simulations. In fact, we heavily exploit some special properties possessed by the nonlinear functions associated with such models.

I. INTRODUCTION

The set P_0 of all real square matrices each with all principal minors nonnegative plays a key role in the study[1-3] of nonlinear equations of the form $F(x) + Ax = B$, and more generally[4] of equations of the form $CF(x) + Ax = B$, in which $F(\cdot)$ is a "diagonal monotone-nondecreasing mapping" of real Euclidean n-space E^n into itself, A and C are real

$n \times n$ matrices and B is an element of E^n. Such equations arise in the dc analysis of transistor networks, the computation of the transient response of transistor networks, and the numerical solution of certain nonlinear partial-differential equations.

In Ref. 3 a nonuniqueness theorem is proved which focuses attention on a simple special property of transistor-type nonlinearities. It shows that for any transistor-type exponential $F(\cdot)$ the equation $F(x) + Ax = B$ has at least two solutions x for some $B \in E^n$ whenever $A \notin P_0$. The theorem shows that some earlier conditions[1,2] for the existence of a unique solution cannot be improved by taking into account more information concerning the nonlinearities, and therefore makes more clear that the set of matrices P_0 plays a basic role in the theory of nonlinear transistor networks. Ref. 3 also contains material concerned with the convergence of algorithms for computing the solution of $F(x) + Ax = B$ as well as of more general equations, and some related problems concerning the numerical integration of the ordinary differential equations which govern the transient response of nonlinear transistor networks are considered briefly.

The primary purpose of this paper is to present the results of a continuation of the numerical integration study initiated in Ref. 3. Here we further exploit the special property of transistor-type exponential nonlinearities used in Ref. 3.

We consider in detail the nonlinear equations encountered at each time step when certain implicit numerical-integration algorithms are used, and, in terms of only the properties of the Jacobian matrix of the pertinent set of differential equations, we present necessary and sufficient conditions for the existence and uniqueness of the solution of the nonlinear equations for all continuous forcing functions and any given step size. Since engineers often think about dynamic nonlinear transistor network problems in terms of the location of the eigenvalues of the relevant Jacobian matrix, the results described in Section 2.2 are of immediate conceptual value. In particular, these results are of a very different character than those that appear in the literature, and whenever the conditions presented are satisfied, it is possible to carry out the algorithms. Under the assumption that the conditions are satisfied, we also show that there are convergent algorithms for solving the nonlinear equations, and that the Jacobian matrix of the nonlinear equations is essentially always at least weakly well-conditioned in a significant sense.

A part of Section 2.3 reports on a general result concerning conditions under which it is possible to invert nonlinear mappings in E^n. More

explicitly, we show that a proposition proved by G. H. Meyer enables us to give a short proof of a new theorem which is a considerably stronger result than that described and used in Ref. 11.

We also present a set of results concerning properties of an important class of transistor-diode networks for which certain implicit numerical-integration algorithms can be carried out for all values of the step size, and, for a special but significant and useful numerical-integration formula, theorems are proved concerning some properties of the computed sequence such as the extent to which the sequence is relatively immune to small local errors introduced at each step as a result of the fact that it is ordinarily not possible to compute the solution of a certain equation exactly.

Finally, in addition to other results, we present new theorems concerning the existence of solutions of the nonlinear dc equation under very realistic assumptions from the viewpoint of models often used in computer simulations.†

Section II contains a detailed discussion of the results and their significance.

II. TRANSIENT RESPONSE OF TRANSISTOR-DIODE NETWORKS AND IMPLICIT NUMERICAL-INTEGRATION FORMULAS

2.1 *Introduction*

We shall consider explicitly only networks containing transistors, diodes, and resistors. However, the material to be presented can be extended to take into account other types of elements as well. In addition, we shall focus attention on the use of linear multipoint integration formulas of closed (i.e., of implicit) type, since such formulas are of considerable use in connection with the typically "stiff systems" of differential equations encountered.

A very large class of networks containing resistors, transistors, and diodes modeled in a standard manner is governed by the equation[5,‡]

$$\frac{du}{dt} + TF[C^{-1}(u)] + GC^{-1}(u) = B(t), \qquad t \geqq 0 \tag{1}$$

† Results concerning the dc equation are directly relevant to the problem of computing the transient response to the extent that in order to numerically integrate the differential equations it is ordinarily necessary to first solve a dc problem to determine the initial conditions.

‡ As a practical matter, the models of transistors and diodes employed here are often used in computer simulations. Of course in some cases it is necessary to use more complicated models.

with $G = \hat{G}(I + R\hat{G})^{-1}$ and where, assuming that there are q diodes and p transistors,

(*i*) $T = T_1 \oplus T_2 \oplus \cdots \oplus T_p \oplus I_q$, the direct sum of the identity matrix of order q and p 2×2 matrices T_k in which

$$T_k = \begin{bmatrix} 1 & -\alpha_r^{(k)} \\ -\alpha_f^{(k)} & 1 \end{bmatrix}$$

with $0 < \alpha_r^{(k)} < 1$ and $0 < \alpha_f^{(k)} < 1$ for $k = 1, 2, \cdots, p$.

(*ii*) $R = R_1 \oplus R_2 \oplus \cdots \oplus R_p \oplus R_0$, the direct sum of a diagonal matrix $R_0 = \text{diag}\ (r_1, r_2, \cdots, r_q)$ with $r_k \geqq 0$ for $k = 1, 2, \cdots, q$ and p 2×2 matrices R_k in which for all $k = 1, 2, \cdots, p$

$$R_k = \begin{bmatrix} r_e^{(k)} + r_b^{(k)} & r_b^{(k)} \\ r_b^{(k)} & r_c^{(k)} + r_b^{(k)} \end{bmatrix}$$

with $r_e^{(k)} \geqq 0$, $r_b^{(k)} \geqq 0$, and $r_c^{(k)} \geqq 0$. (The matrix R takes into account the presence of bulk resistance in series with the diodes and the emitter, base, and collector leads of the transistors.)

(*iii*) $\hat{G}$ is the short-circuit conductance matrix associated with the resistors of the network. (It does not take into account the bulk resistances of the semiconductor devices.)

(*iv*) $F(\cdot)$ is a mapping of $E^{(2p+q)}$ into $E^{(2p+q)}$ defined by the condition that

$$F(x) = [f_1(x_1), f_2(x_2), \cdots, f_{2p+q}(x_{2p+q})]^{\text{tr}}$$

for all $x \in E^{(2p+q)}$ with each $f_i(\cdot)$ a continuously-differentiable mapping of E^1 into E^1 such that $f_i'(\alpha) > 0$ for all $\alpha \in E^1$.

(*v*) $C^{-1}(\cdot)$ is the inverse of the mapping $C(\cdot)$, of $E^{(2p+q)}$ into itself, defined by

$$C(x) = cx + \tau F(x)$$

for all $x \in E^{(2p+q)}$ with $c = \text{diag}\ (c_1, c_2, \cdots, c_{(2p+q)})$, $\tau = \text{diag}\ (\tau_1, \tau_2, \cdots, \tau_{(2p+q)})$, and with each τ_i and each c_i a positive constant.

(*vi*) $B(t)$ is a $(2p + q)$-vector which takes into account the voltage and current generators present in the network, and

(*vii*) u is related to v the vector of ideal-junction voltages of the semiconductor devices (v does not take in account the voltage drops across the bulk resistors) through $C(v) = u$ for all $v \in E^{(2p+q)}$.

Equation (1) is equivalent to[†]

[†] In Ref. 5 it is shown if $B(\cdot)$ is a continuous mapping of $[0, \infty)$ into $E^{(2p+q)}$, then for any initial condition $u^{(0)} \in E^{(2p+q)}$ there exists a unique continuous $(2p + q)$-vector-valued function $u(\cdot)$ such that $u(0) = u^{(0)}$ and (1) is satisfied for all $t > 0$.

$$\dot{u} + f(u, t) = \theta_{(2p+q)}, \qquad t \geqq 0 \tag{2}$$

in which of course

$$f(u, t) = TF[C^{-1}(u)] + GC^{-1}(u) - B(t) \tag{3}$$

and $\theta_{(2p+q)}$ is the zero vector of order $(2p + q)$.

It is well known that certain specializations of the general multipoint formula[6,7]

$$y_{n+1} = \sum_{k=0}^{r} a_k y_{n-k} + h \sum_{k=-1}^{r} b_k \tilde{y}_{n-k} \tag{4}$$

in which

$$\tilde{y}_{n-k} = -f(y_{n-k}, (n - k)h) \tag{5}$$

can be used as a basis for computing the solution of equation (2). Here h, a positive number, is the step size, the a_k and the b_k are real numbers, and of course y_n is the approximation to $u(nh)$ for $n \geqq 1$.

In the literature dealing with formulas of the type (4) in connection with systems of equations of the type (2), information concerning the location of the eigenvalues of the Jacobian matrix J_u of $f(u, t)$ with respect to u plays an important role in determining whether or not a given formula will be (in some suitable sense) stable. In particular, an assumption often made is that all of the eigenvalues of J_u lie in the strict right-half plane for all $t \geqq 0$ and all u. For $f(u, t)$ given by equation (3), we have

$$J_u = T \operatorname{diag} \left\{ \frac{f_i'[g_i(u_i)]}{c_i + \tau_i f_i'[g_i(u_i)]} \right\} + G \operatorname{diag} \left\{ \frac{1}{c_i + \tau_i f_i'[g_i(u_i)]} \right\} \tag{6}$$

in which for $j = 1, 2, \cdots, (2p + q)$ $g_j(u_j)$ is the jth component of $C^{-1}(u)$. Thus here J_u is a matrix of the form

$$TD_1 + GD_2 \tag{7}$$

where D_1 and D_2 are diagonal matrices with positive diagonal elements. A simple result concerning (7), Theorem 4 of Ref. 3, asserts that if there exists a diagonal matrix D with positive diagonal elements such that†

(*i*) DT is strongly column-sum dominant, and

(*ii*) DG is weakly column-sum dominant,

then for all diagonal matrices D_1 and D_2 with positive diagonal elements,

† The terms "strongly-column sum dominant" and "weakly column-sum dominant" are reasonably standard. However, they are defined in Section III.

all eigenvalues of (7) lie in the strict right-half plane. This condition on T and G is often satisfied.†

The subclass of numerical integration formulas (4) defined by the condition that $b_{-1} > 0$ are of considerable use[8-10] in applications involving the typically "stiff systems" of differential equations encountered in the analysis of nonlinear transistor networks. With $b_{-1} > 0$, y_{n+1} is defined *implicitly* through

$$y_{n+1} + hb_{-1}f(y_{n+1}, (n+1)h) = \sum_{k=0}^{r} a_k y_{n-k} + h \sum_{k=0}^{r} b_k \tilde{y}_{n-k}$$

in which the right side depends on y_{n-k} only for $k \in \{0, 1, 2, \cdots, r\}$, and for $f(u, t)$ given by equation (3), we have

$$y_{n+1} + hb_{-1}\{TF[C^{-1}(y_{n+1})] + GC^{-1}(y_{n+1})\} = q_n \tag{8}$$

in which

$$q_n = \sum_{k=0}^{r} a_k y_{n-k} + h \sum_{k=0}^{r} b_k \tilde{y}_{n-k} + hb_{-1}B[(n+1)h].$$

Obviously, the numerical integration formula (8) makes sense only if there exists for each n a $y_{n+1} \in E^{(2p+q)}$ such that (8) is satisfied.

2.2 *The Jacobian Matrix J_u and Necessary and Sufficient Conditions for the Existence of a Unique Solution y_{n+1} of (8) for All $q_n \in E^{(2p+q)}$*

Here we shall make the additional assumption that the functions $f_i(\cdot)$ are such that the mapping $F(\cdot)$ belongs to the set $\mathfrak{F}_0^{(2p+q)}$ defined in Section 3.1. This assumption is satisfied whenever the $f_i(\cdot)$ are the usual Ebers–Moll exponential-type nonlinearities. That is, $\mathfrak{F}_0^{(2p+q)}$ contains all of the mappings $F(\cdot)$ such that for each j

$$f_j(x_j) = a_j[\exp(b_j x_j) - 1] \quad \text{or} \quad f_j(x_j) = a_j[1 - \exp(-b_j x_j)]$$

for all $x_j \in E^1$ with a_j and b_j positive constants.

Our first result, Theorem 1 of Section III, is a rather strong result concerning the relation between properties of the Jacobian matrix J_u and properties of equation (8). Let Ξ denote the set of all real numbers σ such that $\det(\sigma I + J_u) = 0$ for some $u \in E^{(2p+q)}$. In other words, let Ξ denote the set of all real numbers σ such that $-\sigma$ is an eigenvalue of J_u at some point u. According to Theorem 1, equation (8) possesses a unique solution y_{n+1} for each $q_n \in E^{(2p+q)}$ (and hence each $B[(n+1)h] \in E^{(2p+q)}$) if and only if $(hb_{-1})^{-1} \notin \Xi$, and also if $(hb_{-1})^{-1} \in \Xi$ then equation (8) possesses at least two solutions for some $q_n \in E^{(2p+q)}$ (and hence for

† See Ref. 5 for examples.

some $B[(n+1)h] \in E^{(2p+q)}$). Therefore, in particular, equation (8) possesses a unique solution for all $q_n \in E^{(2p+q)}$ and all $h \in (0, \bar{h}]$, in which $\bar{h}$ is an arbitrary positive constant, if and only if the intersection of the interval $[(\bar{h}b_{-1})^{-1}, \infty)$ and Ξ is the null set, and equation (8) possesses a unique solution for all $q_n \in E^{(2p+q)}$ and all $h > 0$ if and only if Ξ contains no points of the interval $(0, \infty)$. Finally, as a somewhat peripheral matter, according to Theorem 1, the dc equation $TF(v) + Gv = B$ has at most one solution v for each $B \in E^{(2p+q)}$ if and only if $0 \notin \Xi$.

The statements made in the preceding paragraph are surprising to the extent that on the one hand they are rather definitive and on the other hand they involve only the location of the real eigenvalues of J_u.† Since engineers often find it helpful to think about nonlinear systems in terms of the location of the eigenvalues of a pertinent Jacobian matrix, it is also of interest to note here that equation (8) can possess more than one solution y_{n+1} for some q_n and some $h > 0$ only if the transistor-diode network is locally exponentially unstable at some operating point, that is, only if at some operating point u, $-J_u$ has a real positive eigenvalue.

2.3 *Existence of Convergent Algorithms for Computing the Solution of* (8)

Throughout this section we assume that the $f_j(\cdot)$ are such that the additional condition that $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ is satisfied.

Whenever $(hb_{-1})^{-1}$ is not contained in the set Ξ of Section 2.2, equation (8), which we shall write as $Q(y_{n+1}) = q_n$, possesses a unique solution y_{n+1} for any $q_n \in E^{(2p+q)}$. We show here that when $(hb_{-1})^{-1} \notin \Xi$ and each $f_j(\cdot)$ is twice continuously differentiable on E^1,‡ there exist steepest descent as well as Newton-type algorithms each of which generates a sequence in $E^{(2p+q)}$ which converges to y_{n+1}.

Assume that $(hb_{-1}) \notin \Xi$. The Jacobian matrix $(I + hb_{-1}J_{y_{n+1}})$ of $Q(\cdot)$ satisfies

$$\det (I + hb_{-1}J_{y_{n+1}}) \neq 0 \quad \text{for all} \quad y_{n+1} \in E^{(2p+q)}. \tag{9}$$

Hence $Q(\cdot)$ is a local homeomorphism on $E^{(2p+q)}$ and since there exists a unique $y_{n+1} \in E^{(2p+q)}$ such that $Q(y_{n+1}) = q_n$ for each $q_n \in E^{(2p+q)}$, $Q(\cdot)$

† Indeed, while we can write (8) as $Q(y_{n+1}) = q_n$ with $Q(\cdot)$ a continuously-differentiable mapping of $E^{(2p+q)}$ into itself with Jacobian matrix $(I + hb_{-1}J_{y_{n+1}})$ recall that for $R(\cdot)$ a general continuously-differentiable mapping of E^n into itself with Jacobian matrix J, $\det J \neq 0$ throughout E^n does not imply that (and is not implied by the statement that) for each $x \in E^n$ there exists a unique $y \in E^n$ such that $R(y) = x$, even for $n = 1$.

‡ This differentiability condition is obviously satisfied if the $f_j(\cdot)$ are the usual exponential functions.

is a homeomorphism of $E^{(2p+q)}$ onto itself. Thus, with $||\cdot||$ any norm on $E^{(2p+q)}$,

$$|| Q(y) || \to \infty \quad \text{as} \quad || y || \to \infty .^{\dagger}$$

Let $R(\cdot)$ be defined by the condition that $R(y) = Q(y) - q_n$ for all $y \in E^{(2p+q)}$. Then $R(\cdot)$ satisfies $|| R(y) || \to \infty$ as $|| y || \to \infty$ and the determinant of the Jacobian matrix of $R(\cdot)$ does not vanish throughout $E^{(2p+q)}$. Therefore, assuming that $R(\cdot)$ is twice continuously differentiable on $E^{(2p+q)}$, it follows (see the Appendix) that the solution y_{n+1} of $R(y_{n+1}) = \theta_{(2p+q)}$ can be computed by using certain steepest descent or Newton-type algorithms.

2.4 *The Jacobian Matrix* $(I + hb_{-1}J_{y_{n+1}})$, *and Inversion of Nonlinear Operators on* E^n *and Jacobian Matrices*

As in Section 2.3, let the additional condition that $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ be satisfied and let $Q(\cdot)$ be the mapping of $E^{(2p+q)}$ into itself with the property that equation (8) can be written as $Q(y_{n+1}) = q_n$. According to Theorem 2 of Section III the Jacobian matrix $(I + hb_{-1}J_{y_{n+1}})$ possesses the property that there exists a constant $\epsilon > 0$ such that

$$\det (I + hb_{-1}J_{y_{n+1}}) \geqq \epsilon \quad \text{for all} \quad y_{n+1} \in E^{(2p+q)} \tag{10}$$

if and only if the matrix

$$[(hb_{-1})^{-1}\tau + T]^{-1}[(hb_{-1})^{-1}c + G],$$

which we shall call S, belongs to the set P of all real square matrices each with all principal minors positive. Thus when $S \in P$ the matrix $(I + hb_{-1}J_{y_{n+1}})$ is well conditioned in at least the weak sense of (10). This fact is of some interest for two reasons. First, certain standard algorithms require that the matrix $(I + hb_{-1}J_{y_{n+1}})$ be inverted along a sequence of points $\{y_{n+1}^{(k)}\}$ in order to compute the solution y_{n+1} of equation (8), and, secondly, Theorem 3 of Section III shows that if $\det [(hb_{-1})^{-1}I + J_u] \neq 0$ for all $u \in E^{(2p+q)}$ and all $(hb_{-1})^{-1} \in \mathcal{I}'$ in which $\mathcal{I}'$ denotes either $(0, \infty)$ or any interval contained in $(0, \infty)$, then $S \in P$ for all but at most a finite number of points $(hb_{-1})^{-1}$ contained in $\mathcal{I}'$. Therefore, referring to the material of Section 2.2, if $Q(y_{n+1}) = q_n$ possesses a unique solution y_{n+1} for all $q_n \in E^{(2p+q)}$ and all $(hb_{-1})^{-1} \in \mathcal{I}'$, then $(I + hb_{-1}J_{y_{n+1}})$ is at least weakly well conditioned at all but at most a finite number of points contained in $\mathcal{I}'$.

---

† Since $Q(\cdot)$ is a homeomorphism of $E^{(2p+q)}$ onto itself, $Q(\cdot)^{-1}$ exists and is continuous. Therefore, the image of any closed ball in $E^{(2p+q)}$ under $Q(\cdot)^{-1}$ is contained in some closed ball in $E^{(2p+q)}$, and hence $|| Q(y) || \to \infty$ as $||y|| \to \infty$.

Since the elements of $(I + hb_{-1}J_{y_{n+1}})$ are bounded on $y_{n+1} \in E^{(2p+q)}$, it follows from a theorem described by M. Vehovec[11] that for each $q_n \in E^{(2p+q)}$ there exists a unique $y_{n+1} \in E^{(2p+q)}$ such that $Q(y_{n+1}) = q_n$ if $S \in P$. More explicitly, the theorem described† by Vohovec asserts that if $R(\cdot)$ is a continuously-differentiable mapping of E^n into E^n with $J(R)_q$ the Jacobian matrix of $R(\cdot)$ at an arbitrary point $q \in E^n$, if the elements of $J(R)_q$ are bounded on E^n, and if there exists a positive constant ϵ such that $\det J(R)_q \geqq \epsilon$ for all $q \in E^n$, then $R(\cdot)$ is a homeomorphism. Thus, using the theorem of Ref. 11 and Theorems 2 and 3 of Section III, we are able to show that if $\det [(hb_{-1})^{-1}I + J_u] \neq 0$ for all $u \in E^{(2p+q)}$ and all $(hb_{-1})^{-1} \in g'$, then for all but at most a finite number of points $(hb_{-1})^{-1} \in g'$, (8) possesses a unique solution y_{n+1} for each $q_n \in E^{(2p+q)}$. Although this result is obviously much weaker than the existence proposition presented in Section 2.2, it shows that the theorem of Ref. 11 can be exploited to provide some insight in connection with the specific problem considered here.

The theorem of Ref. 11 is of interest primarily because the key hypothesis concerns only the determinant $\det J(R)_q$ (as opposed to the condition of Palais‡ that $\| R(q) \| \to \infty$ as $\| q \| \to \infty$). Theorem 4 of Section III is a general result which is considerably stronger than the theorem of Ref. 11. It shows that the condition of the theorem of Ref. 11 that there exist a positive constant ϵ such that $\det J(R)_q \geqq \epsilon$ for all q can be replaced with the condition that there exist real constants $a > 0$ and $b \geqq 0$ such that

$$\det J(R)_q \geqq \frac{1}{a + b \| q \|} \quad \text{for all} \quad q \in E^n.$$

2.5 *A Class of Networks for Which* (8) *Possesses a Unique Solution for All Values of the Step Size*

There is an interesting class of transistor-diode-resistor networks with the property that for each network in the class, equation (8) possesses a unique solution for all $h > 0$ (i.e., for all $h > 0$, all $q_n \in E^{(2p+q)}$, and all diagonal matrices c and τ with positive diagonal elements). In order to define and discuss that class, consider the dc equation $TF(v) + Gv = B$ in which v is the $(2p + q)$-vector of semiconductor ideal-junction voltages and $B \in E^{(2p+q)}$. If $p > 0$ and the matrix R of Section 2.1 is the zero matrix, v_1 is the emitter-to-base voltage of transistor one, v_2 is the collector-to-base voltage of transistor one, and so forth. By port

† According to Vehovec, the theorem was recently proved by I. Vidar, and the proof is expected to appear in the journal Glasnik Matematicki.

‡ See Ref. 12 and the appendix of Ref. 13. Here $\| \cdot \|$ denotes any norm on E^n.

j of the transistor-diode-resistor network we mean the terminal pair between which the voltage v_j appears. Again we shall make the assumption that $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$.

In Ref. 3 it is proved that $TF(v) + Gv = B$ possesses at most one solution v for each $B \in E^{(2p+q)}$ if and only if $T^{-1}G \in P_0$. It is also proved in Ref. 3 that equation (8) possesses a unique solution y_{n+1} for each $q_n \in E^{(2p+q)}$ and each $h > 0$ if $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$ in which here $\mathfrak{I}(T)$ denotes the set of all real matrices having the same form as T and with the "α's" of M not larger than those of T.† In other words, it was also proved in Ref. 3 that equation (8) possesses a unique solution y_{n+1} for each $q_n \in E^{(2p+q)}$ and each $h > 0$ if the dc equation possesses at most one solution for each $B \in E^{(2p+q)}$ for "the original set of α's as well as for an arbitrary set of not-larger α's." Before proceeding, and for the sake of completeness, we mention here that the same result can be obtained by way of the approach of Section 2.2; a direct corollary of Theorem 5 of Section III, Corollary 1, shows that if $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$, then $\det(\sigma I + J_u) \neq 0$ for all real $\sigma \geqq 0$ and all $u \in E^{(2p+q)}$.

Theorem 5 of Section III provides considerable information concerning the nature of the class of networks for which $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$. In particular, the theorem shows that $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$ if and only if $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}_0(T)$ in which $\mathfrak{I}_0(T)$ is the set of all 2^{2p} real square matrices M having the same form as T and with each "α" of M either zero or the corresponding "α" of T.‡ The theorem also shows that "$M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$" is equivalent to each of six other statements involving T and G. For example, according to Theorem 5, we have $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$ if and only if either $T^{-1}(G + D) \in P_0$ for all diagonal matrices D with positive diagonal elements, which has an obvious network interpretation in terms of the addition of resistors to the network characterized by G, or $T^{-1}G \in P_0$ and $(T_w)^{-1}G_w \in P_0$ for all pairs of matrices T_w and G_w obtained from T and G, respectively, by deleting an arbitrary set w of rows, and the same set of columns, of both T and G.

When the matrix R of Section 2.1 is the zero matrix, the last condition on T and G of the preceding paragraph also has a simple network interpretation: Given T and G, we have $T^{-1}G \in P_0$, and any network obtained from the network characterized by T and G by short-circuiting an arbitrary set w of at most all but one of the $(2p + q)$ semiconductor junctions possesses the following property. With respect to the voltage vector v_w associated with the junctions not short-circuited, and with

† See Definition 4 of Section III for a precise definition of $\mathfrak{I}(T)$.
‡ See Definition 5 of Section III for a precise definition of $\mathfrak{I}_0(T)$.

the components of v_w taken in the same order as those of v, the "new T and G" matrices† T_w and G_w satisfy $(T_w)^{-1}G_w \in P_0$. As reasonable as this condition or any of the other seven equivalent conditions of Theorem 5 might seem, and even though, as Theorem 6 of Section III shows, $T^{-1}G \in P_0$ implies that $(T_w)^{-1}G_w \in P_0$ whenever w has the property that if the port number associated with one junction of a given transistor is contained in w, then the port number associated with the other junction of that transistor is also contained in w, it is the case that there are transistor-diode-resistor networks for which $T^{-1}G \in P_0$ and $M^{-1}G \notin P_0$ for some $M \in \mathfrak{I}(T)$. In fact, Ref. 14 presents an example in which $p = 3$, $q = 0$, $T^{-1}G \in P_0$, and $T^{-1}(G + D) \notin P_0$ for some diagonal matrix D with positive diagonal elements. However, the class of networks for which $T^{-1}G \in P_0$ implies that $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$ is clearly quite large; it obviously includes all networks in which $p = 0$, it includes all networks in which the base terminals of all transistors are connected to a common point, and as Theorem 7 of Section III shows, the class includes all networks in which $T^{-1}G \in P_0$ and $p = 1$ or $p = 2$.††

2.6 *Results Concerning the Numerical-Integration Formula* $y_{n+1} = y_n + h\tilde{y}_{n+1}$

The general multipoint formula (4) reduces to the well-known implicit numerical-integration formula $y_{n+1} = y_n + h\tilde{y}_{n+1}$ when $a_0 = b_{-1} = 1, b_0 = 0$, and $a_k = b_k = 0$ for $k = 1, 2, \cdots, r$. For that important special case, and with $\tilde{y}_{n+1}$ given by equations (3) and (5), $\{y_{n+1}\}$ is defined implicitly through

$$y_{n+1} + h\{TF[C^{-1}(y_{n+1})] + GC^{-1}(y_{n+1})\} = y_n + hB_n \tag{11}$$

for all $n \geqq 0$, in which $B_n = B[(n + 1)h]$. Here we describe some detailed results concerning the relation between the sequences $\{y_{n+1}\}$ and $\{B_n\}$. We assume throughout this section that G is such that there exists a diagonal matrix D with positive diagonal elements with the property that both DT and DG are strongly column-sum dominant. This condition, which is often satisfied,§ guarantees that there exists a unique solution‡ y_{n+1} of equation (11) for each $(y_n + hB_n) \in E^{(2p+q)}$.

† It is a simple matter to show that the "new T and G" matrices are T_w and G_w.

†† It is proved in Ref. 14 that if $q = 0$ and if $p = 1$ or $p = 2$, then $T^{-1}G \in P_0$ implies that $T^{-1}(G + D) \in P_0$ for all diagonal matrices with positive diagonal elements. Thus, by the equivalence of statements (i) and (v) of Theorem 5 of Section III, it follows at once that if $T^{-1}G \in P_0$ then $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$ if $q = 0$ and $p = 1$ or $p = 2$. The proof of essentially the same end result given here is of a very different nature and is quite short.

§ See Ref. 5 for examples.

‡ A result mentioned in Section 2.1 implies that if DT and DG are both strongly column-sum dominant, then $\det [(h)^{-1}I + J_u] \neq 0$ for all $u \in E^{(2p+q)}$ and all $h > 0$.

Let $||\cdot||_1$ be defined by the condition that $||\ v\ ||_1 = \sum_{i=1}^{(2p+q)} |\ v_i\ |$ for all $v \in E^{(2p+q)}$. According to Theorem 8 of Section III, there exists a positive constant δ depending only on the c_i , the τ_i , T, G, and D such that

$$||\ Dy_n\ ||_1 \leqq (1 + \delta h)^{-n}\ ||\ Dy_0\ ||_1 + h \sum_{k=1}^{n} (1 + \delta h)^{-k}\ ||\ DB_{(n-k)}\ ||_1$$

for all $n \geqq 1$. Therefore, it follows that *for all* $h > 0$, the sequence $y_1 , y_2 , \cdots$ is bounded whenever the sequence $B_1 , B_2 , \cdots$ is bounded, and $y_1 , y_2 , \cdots$ approaches $\theta_{(2p+q)}$ the zero vector of $E^{(2p+q)}$ whenever $B_1 , B_2 , \cdots$ approaches $\theta_{(2p+q)}$.

Typically at each step an iterative algorithm is employed to compute the solution y_{n+1} of equation (11). Since it is ordinarily not possible to compute y_{n+1} with infinite precision, it is important to consider the effects of the errors which are introduced. While, ideally, we would like to determine the sequence $\{y_{n+1}\}$ defined by equation (11) and some initial-condition vector y_0 , suppose that we determine instead a sequence $\{\hat{y}_{n+1}\}$ such that, with ϵ an arbitrary positive constant, $||\ D(\hat{y}_n - y_n^*)\ ||_1 \leqq \epsilon$ for all $n \geqq 1$ and

$$y_{n+1}^* + h\{TF[C^{-1}(y_{n+1}^*)] + GC^{-1}(y_{n+1}^*)\} = \hat{y}_n + hB_n \tag{12}$$

for all $n \geqq 0$. That is, suppose that at each step the *local* error $||\ D(\hat{y}_n - y_n^*)\ ||_1$ in solving for "y_{n+1}" is at most ϵ. Then, according to Theorem 8, and with δ the positive constant referred to above,

$$||\ D(y_n - \hat{y}_n)\ ||_1 \leqq (1 + \delta h)^{-n}\ ||\ D(y_0 - \hat{y}_0)\ ||_1 + \epsilon \sum_{k=0}^{n} (1 + \delta h)^{-k} \quad \text{for all} \quad n \geqq 1$$

in which $\hat{y}_0$ is the approximation to y_0 . Therefore, given an arbitrarily small positive constant ρ, *for any* $h > 0$ it is possible to choose $\hat{y}_0$ and $\epsilon > 0$ such that the accumulated-error vector $(y_n - \hat{y}_n)$ satisfies $||\ y_n - \hat{y}_n\ ||_1 \leqq \rho$ *for all* $n \geqq 1$.

Finally, Theorem 9 of Section III provides us with a conceptually interesting uniform bound on the norm of the difference between corresponding elements of the sequences $\{y_n\}$ and $\{u_n\}$ in which $u_n = u(nh)$ for all $n \geqq 0$ and $u(\cdot)$ satisfies the differential equation (1). According to Theorem 9, there exist positive constants δ and ρ, both independent of h, such that

$$||\ D(u_n - y_n)\ ||_1 \leqq (1 + \delta h)^{-n}\ ||\ D(u_0 - y_0)\ ||_1 + \rho h$$

for all $n \geqq 1$, assuming that the elements of $B(\cdot)$ and $(d/dt)B(\cdot)$ are

bounded and continuous on $[0, \infty)$. In particular, if $y_0 = u_0$ we see that there exists a positive constant ρ', independent of h, such that $|| u_n - y_n ||_1 \leqq \rho' h$ for all $n \geqq 1$, provided only that the assumptions of this section are satisfied and that $B(\cdot)$ and $(d/dt)B(\cdot)$ are bounded and continuous on $[0, \infty)$.

2.7 *Conditions Which Imply That* $T^{-1}\hat{G}(I + R\hat{G})^{-1} \in P_0$

In this section and in Section 2.8 we present some results concerning properties of the dc equation $TF(v) + Gv = B$. These results are directly relevant to the problem of computing the transient response of transistor-diode networks to the extent that in order to numerically integrate the differential equation (1) it is ordinarily necessary to first solve a dc problem to determine the initial conditions.

As indicated in Section 2.1, $G = \hat{G}(I + R\hat{G})^{-1}$ in which R takes into account the bulk resistances associated with the semiconductor devices. Here we present some material concerning conditions which imply that $T^{-1}\hat{G}(I + R\hat{G})^{-1}$ belongs to P_0.

Let $p > 0$. Theorem 10 of Section III asserts that $T^{-1}\hat{G}(I + R\hat{G})^{-1} \in P_0$ whenever $T^{-1}\hat{G} \in P_0$ and R satisfies

$$\alpha_r^{(k)}(1 - \alpha_r^{(k)})^{-1} r_e^{(k)} = r_b^{(k)}$$
$$\alpha_f^{(k)}(1 - \alpha_f^{(k)})^{-1} r_c^{(k)} = r_b^{(k)}$$

for $k = 1, 2, \cdots, p$. This rather special result shows that if $F(\cdot)$ satisfies the additional condition that $F(\cdot)$ belongs to the set $\mathfrak{F}_0^{(2p+q)}$ defined in Section 3.1, and if the network associated with T and $\hat{G}$ possesses the property that there is at most one solution v of the dc equation $TF(v) + \hat{G}v = B$ for each $B \in E^{(2p+q)}$, then it is always possible to add certain resistors of positive value in series with each transistor lead such that the dc equation of the resulting network possesses at most one solution.

Theorem 11 of Section III directs attention to the fact that there is a nontrivial class of transistor networks for which $T^{-1}\hat{G}(I + R\hat{G})^{-1} \in P_0$ for all R. According to Theorem 11, if $p > 0$ and $\hat{G}$ is such that $T^{-1}\hat{G} \in P_0$ for all "α's" (i.e., for all $\alpha_r^{(k)}$ and $\alpha_f^{(k)}$ belonging to $(0, 1)$), then for any particular set of "α's" $T^{-1}\hat{G}(I + R\hat{G})^{-1} \in P_0$ for all R.†

Given T, an interesting characterization of the class of short-circuit-conductance matrices $\hat{G}$ such that $M^{-1}\hat{G} \in P_0$ for all $M \in \mathfrak{I}(T)$ is provided by Theorem 12 of Section III.‡ According to Theorem 12, $M^{-1}\hat{G} \in P_0$ for all $M \in \mathfrak{I}(T)$ if and only if $T^{-1}\hat{G}(I + R\hat{G})^{-1} \in P_0$ for all R satisfying certain inequality-type conditions. In particular, if the base-lead

† A similar result is proved in Ref. 2 under the assumption that $\hat{G}$ is not singular.
‡ The set $\mathfrak{I}(T)$ is described in Section 2.5.

resistance of each transistor is taken to be zero, then $M^{-1}\hat{G} \in P_0$ for all $M \in \mathfrak{I}(T)$ implies that $T^{-1}\hat{G}(I + R\hat{G})^{-1} \in P_0$ *for all* nonnegative values of each emitter-lead resistor and each collector-lead resistor.

2.8 *Ebers-Moll Models and the Existence of a Solution of* $TF(v) + Gv = B$

In Section III, a set $\mathfrak{F}_3$ of mappings $F(\cdot)$ is defined such that each element of $\mathfrak{F}_3$ possesses certain important properties possessed by an arbitrary $F(\cdot)$ of the type that arises when an Ebers-Moll exponential-nonlinear-function model is used for each transistor and diode. In contrast with the set of all $F(\cdot)$ such that each $f_j(\cdot)$ is a strictly-monotone-increasing mapping of E^1 *onto* E^1 , an arbitrary element $F(\cdot)$ of $\mathfrak{F}_3$ possesses the properties that for each j, $f_j(\cdot)$ is bounded on either $[0, \infty)$ or $(-\infty, 0]$, and the two nonlinear functions associated with the same transistor are *both* bounded on either $[0, \infty)$ or $(-\infty, 0]$. The set $\mathfrak{F}_3$ is contained in $\mathfrak{F}_0^{(2p+q)}$ and contains every Ebers-Moll exponential-nonlinear-function-type $F(\cdot)$.

The first part of Theorem 13 of Section III asserts that the equation $TF(v) + Gv = B$ possesses a unique solution v for each $F(\cdot) \in \mathfrak{F}_3$ and each $B \in E^{(2p+q)}$ if and only if $T^{-1}G \in P_0$ *and* $\det G \neq 0$. It is the "only if" part of this proposition which is the new result presented here. The proof exploits some special properties of transformerless resistor networks; it shows that if $T^{-1}G \in P_0$ but $\det G = 0$, then there are functions $t(\cdot)$ and $d(\cdot)$, both functions taking on only the values 1 or -1, such that there is no solution v of $TF(v) + Gv = B$ for some $B \in E^{(2p+q)}$ *for any* set of Ebers-Moll-modeled transistors and diodes with the property that for all k transistor k is a pnp device (as opposed to a npn device) if and only if $t(k) = 1$, and for all j diode j is a p-n junction if and only if $d(j) = 1$.†

The discussion of the preceding paragraph concerning the proof of Theorem 13 shows that it is not possible to make stronger assertions concerning the existence of a unique solution of $TF(v) + Gv = B$ for all $B \in E^{(2p+q)}$ for Ebers-Moll-modeled transistors and diodes unless we take into account more information about the nature of the semiconductor junctions. A good deal of progress in this direction has recently been made, and we state here without proof the following complete result dealing with diode-resistor networks.

Theorem 14:‡ *Let* $p = 0$ *and* $q > 0$. *Let* $F(\cdot) \in \mathfrak{F}_3$ (*see Definition* 12 *of*

† In contrast, the proof of the "only if" part of Theorem 3 of Ref. 1 shows that if $A \notin P_0$ then there is a mapping $F(\cdot)$ with each $f_j(\cdot)$ a *linear* function such that $F(x) + Ax = B$ does not possess a unique solution for all $B \in E^n$.

‡ The proof of Theorem 14 will be presented in a subsequent paper.

Section 3.31), *and for* $j = 1, 2, \cdots, q$ *let* s_j *equal either* 1 *or* -1 *depending on whether* $f_j(\cdot)$ *is bounded on* $[0, \infty)$ *or* $(-\infty, 0]$, *respectively. Then, with* A *any real symmetric nonnegative-definite matrix of order* q, *there exists a unique solution* v *of* $F(v) + Av = B$ *for all* $B \in E^q$ *if and only if there is no real* q*-vector* η *such that* $\eta \neq \theta_q$, $A\eta = \theta_q$, *and* $\eta \in S$, *in which*

$$S = \{y: y \in E^q \text{ and } y_j s_j \geqq 0 \text{ for } j = 1, 2, \cdots, q\}.^{\ddagger}$$

III. THEOREMS AND PROOFS

3.1 *Notation and Definitions*

Throughout Section III,

(*i*) unless stated otherwise, p and q denote nonnegative integers such that $(p + q) > 0$, and n denotes an arbitrary positive integer;

(*ii*) the set of all real n-vectors is denoted by E^n, θ is the zero element of E^n, and if $v \in E^n$ and j is an integer such that $1 \leqq j \leqq n$, then v_j denotes the jth component of v;

(*iii*) $\| v \| = (\sum_{i=1}^n v_i^2)^{1/2}$ and $\| v \|_1 = \sum_{i=1}^n | v_i |$ for all $v \in E^n$; for any real $n \times n$ matrix M, $\| M \|$ denotes sup $\{m: \| Mx \| \leqq m \| x \|, x \in E^n\}$;

(*iv*) the transpose of an arbitrary (not necessarily square) matrix M is denoted by M^{tr};

(*v*) I_n denotes the identity matrix of order n, and I denotes the identity matrix of order determined by the context in which the symbol is used; if $Q_1, Q_2, \cdots, Q_n$ are square matrices, then $Q_1 \oplus Q_2 \oplus \cdots \oplus Q_n$ denotes the direct sum of $Q_1, Q_2, \cdots, Q_n$, in the order indicated;

(*vi*) if D is a real diagonal matrix, then $D > 0 (D \geqq 0)$ means that the diagonal elements of D are positive (nonnegative); and

(*vii*) we say that a real $n \times n$ matrix M is strongly (weakly) column-sum dominant if and only if for $j = 1, 2, \cdots, n$

$$m_{jj} > (\geqq) \sum_{i \neq j} | m_{ij} |.$$

Definition 1: The set of all real square matrices M such that every principal minor of M is nonnegative (positive) is denoted by $P_0(P)$.

Definition 2: Let $\mathfrak{F}_0^{(2p+q)}$ denote that collection of mappings of $E^{(2p+q)}$ into itself defined by: $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ if and only if there exist for $j =$

‡ In the network case, $A = G$, and it is often possible to determine by inspection whether or not there exists an $\eta \neq \theta_q$ such that $G\eta = \theta_q$ and $\eta \in S$.

1, 2, $\cdots$, $(2p + q)$ continuous functions $f_i(\cdot)$ mapping E^1 into E^1 such that for each $x \in E^{(2p+q)}$, $F(x) = [f_1(x_1), f_2(x_2), \cdots, f_{(2p+q)}(x_{(2p+q)})]^{tr}$, and

$$(i) \quad \inf_{\alpha \in (-\infty, \infty)} [f_i(\alpha + \beta) - f_i(\alpha)] = 0,$$

$$(ii) \quad \sup_{\alpha \in (-\infty, \infty)} [f_i(\alpha + \beta) - f_i(\alpha)] = +\infty$$

for all $\beta > 0$ and all $j = 1, 2, \cdots, (2p + q)$.

Definition 3: Let $\mathfrak{Z}$ denote the set of all real matrices M such that $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p \oplus I_q$ with

$$M_k = \begin{bmatrix} 1 & -\alpha_r^{(k)} \\ -\alpha_f^{(k)} & 1 \end{bmatrix},$$

$0 \leqq \alpha_r^{(k)} < 1$, and $0 \leqq \alpha_f^{(k)} < 1$ for all $k = 1, 2, \cdots, p$. As suggested, if $q = 0$, then $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p$, while if $p = 0$, then $M = I_q$.

Assumption 1: Throughout Section III, G denotes a real nonnegative-definite matrix of order $(2p + q)$.

A tool that we shall use often is:

Lemma 1: A real square matrix M is an element of P_0 if and only if $\det (D + M) \neq 0$ *for all real diagonal matrices $D > 0$.*

Lemma 1 is proved in Ref. 2.

3.2 *Theorem 1: Let $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ with each $f_i(\cdot)$ continuously differentiable on $(-\infty, \infty)$ and $f_i'(\alpha) > 0$ for all $\alpha \in (-\infty, \infty)$. Let $T \in \mathfrak{Z}$, let $C(\cdot)$ [that is, $c + \tau F(\cdot)$], G, and J_u be as defined in Section* 2.1, *and let σ be a real nonnegative constant. Then*

$$\sigma y + TF[C^{-1}(y)] + GC^{-1}(y) = r \tag{13}$$

possesses at most one solution y for each $r \in E^{(2p+q)}$ if and only if

$$\det (\sigma I + J_u) \neq 0 \quad \text{for all} \quad u \in E^{(2p+q)}, \tag{14}$$

and if $\sigma > 0$ and condition (14) *is satisfied then for each $r \in E^{(2p+q)}$ there exists a solution y of* (13).

3.3 *Proof of Theorem 1*

We have

$$\det(\sigma I + J_u)$$
$$= \det(\sigma I + TF'[g(u)]\{c + \tau F'[g(u)]\}^{-1} + G\{c + \tau F'[g(u)]\}^{-1})$$
$$= \det\{c + \tau F'[g(u)]\}^{-1} \cdot \det\{\sigma c + \sigma\tau F'[g(u)] + TF'[g(u)] + G\},$$

in which $g(\cdot)$ is the mapping of $E^{(2p+q)}$ onto itself defined by $g(u) = C^{-1}(u)$ for all $u \in E^{(2p+q)}$, and $F'[g(u)] = \text{diag}\{f_i'[g_i(u_i)]\}$. Since $\det\{c + \tau F'[g(u)]\} > 0$ for all u, $\det(\sigma I + J_u) \not\equiv 0$ for all u if and only if

$$\det\{(\sigma\tau + T)F'[g(u)] + (\sigma c + G)\} \not\equiv 0 \quad \text{for all } u.$$

For each j $g_j(\cdot)$ maps E^1 onto E^1, and since $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ with each $f_i(\cdot)$ continuously differentiable on $(-\infty, \infty)$ and $f_i'(\alpha) > 0$ for all $\alpha \in (-\infty, \infty)$, the image of E^1 under the mapping $f_j'[g_j(\cdot)]$ is $(0, \infty)$† for all j. Thus, by Lemma 1 (since $\det(\sigma\tau + T) \not\equiv 0$)$(\sigma\tau + T)^{-1}(\sigma c + G) \in P_0$ if and only if

$$\det(\sigma I + J_u) \not\equiv 0 \quad \text{for all} \quad u. \tag{15}$$

The equation

$$\sigma y + TF[C^{-1}(y)] + GC^{-1}(y) = r$$

possesses a solution y if and only if $x = C^{-1}(y)$ satisfies

$$\sigma C(x) + TF(x) + Gx = r,$$

that is, if and only if

$$(\sigma\tau + T)F(x) + (\sigma c + G)x = r. \tag{16}$$

But equation (16) possesses at most one solution for each $r \in E^{(2p+q)}$ if and only if $(\sigma\tau + T)^{-1}(\sigma c + G) \in P_0$ (see pp. 105–107 of Ref. 3) and hence if and only if condition (15) is met.

Suppose now that $\sigma > 0$. Since G is nonnegative definite, $\det(\sigma c + G) \not\equiv 0$. If condition (15) is satisfied then $(\sigma\tau + T)^{-1}(\sigma c + G) \in P_0$ and hence for each $r \in E^{(2p+q)}$, equation (16) possesses a solution x (see p. 99 of Ref. 3). □

3.4 *Theorem 2: Let* $T \in \mathfrak{I}$, *and let* $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ *with each* $f_i(\cdot)$ *continuously differentiable on* $(-\infty, \infty)$ *and* $f_i'(\alpha) > 0$ *for all* $\alpha \in (-\infty, \infty)$. *Then for each* $\sigma \geqq 0$ *there exists a positive constant* ϵ *such that* $\det(\sigma I + J_u) \geqq \epsilon$ *for all* $u \in E^{(2p+q)}$ *if and only if* $(\sigma\tau + T)^{-1}(\sigma c + G) \in P$.

---

† For any $\beta > 0$ and any $\alpha \in (-\infty, \infty)$, $f_j(\alpha + \beta) - f_j(\alpha) = \beta f_j'(\delta)$ for some $\delta \in [\alpha, \alpha + \beta]$.

3.5 *Proof of Theorem 2*

We have

$$\begin{aligned}\det\,&(\sigma I + J_u)\\ &= \det\,(\sigma I + TF'[g(u)]\{c + \tau F'[g(u)]\}^{-1} + G\{c + \tau F'[g(u)]\}^{-1})\\ &= \det\,\{c + \tau F'[g(u)]\}^{-1}\cdot\det\,\{(\sigma\tau + T)F'[g(u)] + (\sigma c + G)\}\\ &= \det\,(\sigma\tau + T)\,\frac{\det\,(F'[g(u)] + A)}{\prod_{j=1}^{(2p+q)}(c_j + \tau_j f_j'[g_j(u_j)])} \qquad (17)\end{aligned}$$

in which $A = (\sigma\tau + T)^{-1}(\sigma c + G)$.

For each sequence $e_1, e_2, \cdots, e_{(2p+q)}$ with each e_j either zero or unity and $e_1, e_2, \cdots, e_{(2p+q)}$ not the sequence 1, 1, $\cdots$, 1: let $m_{e_1,e_2,\cdots,e_{(2p+q)}}$ denote the determinant obtained from A by deleting rows $\rho_1, \rho_2, \cdots, \rho_l$ and columns $\rho_1, \rho_2, \cdots, \rho_l$ in which $\{\rho_1, \rho_2, \cdots, \rho_l\} = \{j: e_j = 1\}$. Thus for each sequence $e_1, e_2, \cdots, e_{(2p+q)}$ other than the sequence 1, 1, $\cdots$, 1 $m_{e_1,e_2,\cdots,e_{(2p+q)}}$ is a principal minor of A. Let $m_{1,1,\cdots,1} = 1$, and let $d_j = f_j'[g_j(u_j)]$ for all j. Then by a standard expression[15] for the determinant of the sum of two matrices

$$\det\,(F'[g(u)] + A) = \sum{}' d_1^{e_1}d_2^{e_2}\cdots d_{(2p+q)}^{e_{(2p+q)}}m_{e_1,e_2,\cdots,e_{(2p+q)}}$$

in which $\sum'$ denotes a summation over all $2^{(2p+q)}$ sequences $e_1, e_2, \cdots, e_{(2p+q)}$ and $d_j^0 = 1$ for all j. It is clear that

$$\prod_{j=1}^{(2p+q)}(c_j + \tau_j f'[g_j(u_j)]) = \sum{}' d_1^{e_1}d_2^{e_2}\cdots d_{(2p+q)}^{e_{(2p+q)}}c_{e_1,e_2,\cdots,e_{(2p+q)}}$$

in which each $c_{e_1,e_2,\cdots,e_{(2p+q)}}$ is a positive constant. Thus with $\eta = \det(\sigma\tau + T)$,

$$\eta^{-1}\det\,(\sigma I + J_u) = \frac{\sum' d_1^{e_1}d_2^{e_2}\cdots d_{(2p+q)}^{e_{(2p+q)}}m_{e_1,e_2,\cdots,e_{(2p+q)}}}{\sum' d_1^{e_1}d_2^{e_2}\cdots d_{(2p+q)}^{e_{(2p+q)}}c_{e_1,e_2,\cdots,e_{(2p+q)}}}. \qquad (18)$$

Suppose that all principal minors of A are positive. Then there is a positive constant δ such that

$$m_{e_1,e_2,\cdots,e_{(2p+q)}} \geqq \delta c_{e_1,e_2,\cdots,e_{(2p+q)}}$$

for all $e_1, e_2, \cdots, e_{(2p+q)}$ and hence (since $d_j > 0$ for all j) $\det(\sigma I + J_u) \geqq \eta\delta$ for all $u \in E^{(2p+q)}$.

As in the proof of Theorem 1, the range of each $d_j = f_j'[g_j(u_j)]$ is $(0, \infty)$, and for any positive constants $p_1, p_2, \cdots, p_{(2p+q)}$ there exists a $u \in E^{(2p+q)}$ such that $d_j = p_j$ for all j. If $A \notin P$ then at least one principal

minor of A is not positive. If $A \notin P_0$, then $\det(F'[g(u) + A]) = 0$ for some u. Therefore to complete the proof it is sufficient to show that if $A \in P_0$ but $A \notin P$ then there is no constant $\epsilon > 0$ such that $\det(\sigma I + J_u) \geq \epsilon$ for all u.

With $A \in P_0$ and $A \notin P$, for at least one sequence $e'_1, e'_2, \cdots, e'_{(2p+q)}$

$$m_{e'_1, e'_2, \cdots, e'_{(2p+q)}} = 0.$$

If $\det A = m_{0,0,\cdots,0} = 0$ we have

$$\inf_{u \in E^{(2p+q)}} \det(\sigma I + J_u) = 0$$

since $\det(\sigma I + J_u) \to 0$ as $d_j \to 0$ for all j. Suppose now that $\det A > 0$ and that $m_{e'_1, e'_2, \cdots, e'_{(2p+q)}} = 0$ for some sequence $e'_1, e'_2, \cdots, e'_{(2p+q)}$. Then with $d_j = d$ for all j for which $e'_j = 1$ and $d_j = d^{-1}$ for all j for which $e'_j = 0$, we have [see equation (18)] $\det(\sigma I + J_u) \to 0$ as $d \to \infty$. □

3.6 *Theorem 3: Let* $T \in \mathfrak{I}$, *let* $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$ *with each* $f_j(\cdot)$ *continuously differentiable on* $(-\infty, \infty)$ *and* $f'_j(\alpha) > 0$ *for all* $\alpha \in (-\infty, \infty)$, *and let* $\mathcal{I}$ *denote* $[0, \infty)$ *or an interval contained in* $[0, \infty)$. *Then for all but at most a finite number of points* σ *contained in* $\mathcal{I}$, *there is a real constant* $\epsilon_\sigma > 0$ *such that* $\det(\sigma I + J_u) \geqq \epsilon_\sigma$ *for all* $u \in E^{(2p+q)}$ *if and only if* $\det(\sigma I + J_u) \neq 0$ *for all* $\sigma \in \mathcal{I}$ *and all* $u \in E^{(2p+q)}$.

3.7 *Proof of Theorem* 3

As in the proof of Theorem 1, $(\sigma T + T)^{-1}(\sigma C + G) \in P_0$ for all $\sigma \in \mathcal{I}$ if and only if $\det(\sigma I + J_u) \neq 0$ for all $\sigma \in \mathcal{I}$ and all u. We shall also use the fact that since $\det(\sigma T + T) > 0$ for all $\sigma \geqq 0$, each principal minor of $(\sigma T + T)^{-1}(\sigma C + G)$ is a finite-valued rational function of σ for all $\sigma \geqq 0$.

(if) If $\det(\sigma I + J_u) \neq 0$ for all u and all $\sigma \in \mathcal{I}$, then $(\sigma T + T)^{-1}(\sigma C + G) \in P_0$ for all $\sigma \in \mathcal{I}$. It is clear that $(\sigma T + T)^{-1}(\sigma C + G) \in P$ for all sufficiently large $\sigma > 0$. Thus each principal minor of $(\sigma T + T)^{-1}(\sigma C + G)$ is nonnegative for all $\sigma \in \mathcal{I}$ and is positive for all sufficiently large $\sigma > 0$. They are therefore positive for all but at most a finite number of values of $\sigma \in \mathcal{I}$. Thus, by Theorem 2, if $\det(\sigma I + J_u) \neq 0$ for all $\sigma \in \mathcal{I}$ and all u there exist for all but at most a finite number of points $\sigma \in \mathcal{I}$ a positive constant ϵ_σ such that $\det(\sigma I + J_u) \geqq \epsilon_\sigma$ for all u.

(only if) If $\det(\sigma I + J_u) = 0$ for some $\sigma \in \mathcal{I}$ and some u, then, for that σ, $(\sigma T + T)^{-1}(\sigma C + G) \notin P_0$. That is, for that σ at least one principal minor of $(\sigma T + T)^{-1}(\sigma C + G)$ is negative. This means that $(\sigma T + T)^{-1}(\sigma C + G) \notin P_0$ for all σ contained in some interval $\mathcal{I}' \subset \mathcal{I}$, and by Theorem 2, for all $\sigma \in \mathcal{I}'$ there is no $\epsilon_\sigma > 0$ such that $\det(I + J_u) \geqq \epsilon_\sigma$ for all u. □

3.8 *Theorem 4: Let $R(\cdot)$ be a continuously differentiable mapping of E^n into E^n, and let $J(R)_q$ denote the Jacobian matrix of $R(\cdot)$ at an arbitrary point $q \in E^n$. If the elements of $J(R)_q$ are bounded on E^n, and if there exist real constants $a > 0$ and $b \geqq 0$ such that* $\det J(R)_q \geqq (a + b \,||\, q \,||)^{-1}$ *for all $q \in E^n$, then $R(\cdot)$ is a homeomorphism of E^n onto E^n.*

3.9 *Proof of Theorem* 4

If Ref. 16 Meyer proves† that $R(\cdot)$ is a homeomorphism of E^n onto E^n if $J(R)_q^{-1}$ exists for all $q \in E^n$ and there exist real constants $\alpha > 0$ and $\beta \geq 0$ such that $|| J(R)_q^{-1} || \leq \alpha + \beta || q ||$ for all $q \in E^n$.

With q an arbitrary element of E^n, let $\lambda_1, \lambda_2, \cdots, \lambda_n$ denote the eigenvalues of $J(R)_q^{tr} J(R)_q$, and let $\lambda_1 = \min_i \{\lambda_i\}$. Then $\lambda_1 \lambda_2 \cdots \lambda_n = [\det J(R)_q]^2 \geqq (a + b \,||\, q \,||)^{-2}$, and since the elements of $J(R)_q$ are bounded on E^n, there is a constant $\lambda > 0$ such that $\lambda_i \leqq \lambda$ for all j and all $q \in E^n$. Thus

$$(\lambda_1)^{1/2} \geqq \lambda^{-(1/2)(n-1)} (a + b || q ||)^{-1} \tag{19}$$

for all q. For any $x \in E^n$ and any $q \in E^n$, $x^{tr} J(R)_q^{tr} J(R)_q x \geqq \lambda_1 x^{tr} x$; that is,

$$|| J(R)_q x || \geqq (\lambda_1)^{1/2} || x || \geqq \lambda^{-(1/2)(n-1)} (a + b || q ||)^{-1} || x ||.$$

With $x = J(R)_q^{-1} y$ in which y is an arbitrary element of E^n, we have

$$||J(R)_q^{-1} y|| \leqq \lambda^{(1/2)(n-1)} (a + b || q ||) \, ||y||,$$

which shows that our hypothesis concerning $\det J(R)_q$ ensures that Meyer's condition on $||J(R)_q^{-1}||$ is satisfied. □

3.10 *Some Further Definitions*

Definition 4: For each $T \in \mathfrak{I}$, let $\mathfrak{I}(T)$ denote the set of all matrices M such that $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p \oplus I_q$ with

$$M_k = \begin{bmatrix} 1 & -\delta_r^{(k)} \\ -\delta_f^{(k)} & 1 \end{bmatrix}$$

and

$$0 < \delta_r^{(k)} \leqq \alpha_r^{(k)} \quad \text{if} \quad \alpha_r^{(k)} > 0 \quad \text{and} \quad \delta_r^{(k)} = 0 \quad \text{if} \quad \alpha_r^{(k)} = 0,$$

$$0 < \delta_f^{(k)} \leqq \alpha_f^{(k)} \quad \text{if} \quad \alpha_f^{(k)} > 0 \quad \text{and} \quad \delta_f^{(k)} = 0 \quad \text{if} \quad \alpha_f^{(k)} = 0,$$

for all $k = 1, 2, \cdots, p$. As suggested, if $q = 0$, then $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p$, while if $p = 0$, then $M = I_q$.

† Meyer's result is a generalization of a well-known result of Hadamard.[17] Hadamard proved that $R(\cdot)$ is a homeomorphism if $J(R)_q^{-1}$ exists for all $q \in E^n$ and satisfies $|| J(R)_q^{-1} || \leqq \alpha$ for all $q \in E^n$ for some positive constant α.[17]

Definition 5: For each $T \in \mathfrak{I}$, let $\mathfrak{I}_0(T)$ denote the set of all 2^{2p} matrices M such that $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p \oplus I_q$ with

$$M_k = \begin{bmatrix} 1 & -\delta_r^{(k)} \\ -\delta_f^{(k)} & 1 \end{bmatrix}$$

and

$$\delta_r^{(k)} = \alpha_r^{(k)} \quad \text{or} \quad \delta_r^{(k)} = 0,$$
$$\delta_f^{(k)} = \alpha_f^{(k)} \quad \text{or} \quad \delta_f^{(k)} = 0,$$

for all $k = 1, 2, \cdots, p$. As suggested, if $q = 0$, then $M = M_1 \oplus M_2 \oplus \cdots \oplus M_p$, while if $p = 0$, then $M = I_q$.

Definition 6: Let $Q_{(2p+q)}$ denote the family of all $2^{(2p+q)} - 1$ sets $w = \{i_1, i_2, \cdots, i_r\}$, including the null set, such that $r < (2p + q)$ and $w \subset \{1, 2, \cdots, (2p + q)\}$.

Definition 7: For M an arbitrary square matrix of order $(2p + q)$, and for each $w \in Q_{(2p+q)}$, let M_w denote the principal submatrix obtained from M by deleting rows $i_1, i_2, \cdots, i_r$ and columns $i_1, i_2, \cdots, i_r$. (If w is the null set, then $M_w = M$.)

Definition 8: For each $j \in \{1, 2, \cdots, (2p + q)\}$, let U_j denote the $(2p + q)$-column-vector with unity in the jth position and zeros in all other positions.

Definition 9: For each $T \in \mathfrak{I}$ and each $w \in Q_{(2p+q)}$, let T^w denote the matrix obtained from T by replacing the jth column of T with U_j for all $j \in w$.

3.11 *Theorem 5: Let $T \in \mathfrak{I}$. Then the following statements are equivalent.*

(*i*) $M^{-1}G \in P_0$ *for all* $M \in \mathfrak{I}(T)$.
(*ii*) $(D_a + T)^{-1}(D_b + G) \in P_0$ *for all diagonal* $D_a \geqq 0$ *and all diagonal* $D_b \geqq 0$.
(*iii*) $T^{-1}(G + D) \in P_0$ *for all diagonal* $D \geqq 0$.
(*iv*) $(D_a + T)^{-1}(D_b + G) \in P_0$ *for all diagonal* $D_a > 0$ *and all diagonal* $D_b > 0$.
(*v*) $T^{-1}(G + D) \in P_0$ *for all diagonal* $D > 0$.
(*vi*) $(T_w)^{-1}G_w \in P_0$ *for all* $w \in Q_{(2p+q)}$.
(*vii*) $[(T^w)^{-1}G]_w \in P_0$ *for all* $w \in Q_{(2p+q)}$.
(*viii*) $M^{-1}G \in P_0$ *for all* $M \in \mathfrak{I}_0(T)$.

3.12 *Proof of Theorem 5*

[(*i*) and (*ii*) are equivalent]

By Lemma 1, $(D_a + T)^{-1}(D_b + G) \in P_0$ if and only if $\det [(D_a + T)^{-1}(D_b + G) + D] \neq 0$ for all diagonal $D > 0$. Thus $(D_a + T)^{-1}(D_b + G) \in P_0$ for all $D_a \geqq 0$ and all $D_b \geqq 0$ if and only if

$$\det [(D_b D^{-1} + D_a + T)D + G] \neq 0$$

for all $D_a \geqq 0$, all $D_b \geqq 0$, and all $D > 0$, and hence if and only if

$$\det [(\Lambda + T)D + G] \neq 0$$

for all diagonal $\Lambda \geqq 0$ and $D > 0$. Let $T_\Lambda = (\Lambda + T)(I + \Lambda)^{-1}$. Then $(D_a + T)^{-1}(D_b + G) \in P_0$ for all $D_a \geqq 0$ and all $D_b \geqq 0$ if and only if

$$\det [T_\Lambda (I + \Lambda)D + G] \neq 0$$

for all $\Lambda \geqq 0$ and all $D > 0$, and hence if and only if $\det (T_\Lambda \hat{D} + G) \neq 0$ for all diagonal $\hat{D} > 0$ and all $\Lambda \geqq 0$. By Lemma 1, this means that $T_\Lambda^{-1} G \in P_0$ for all $\Lambda \geqq 0$ if and only if $(D_a + T)^{-1}(D_b + G) \in P_0$ for all $D_a \geqq 0$ and all $D_b \geqq 0$. We observe that $T_\Lambda = (T_\Lambda)_1 \oplus (T_\Lambda)_2 \oplus \cdots \oplus (T_\Lambda)_p \oplus I_q$ in which, with $\Lambda = \operatorname{diag} (\lambda_1, \lambda_2, \cdots, \lambda_{(2p+q)})$,

$$(T_\Lambda)_k = \begin{bmatrix} 1 & \dfrac{-\alpha_r^{(k)}}{1 + \lambda_{2k}} \\ \dfrac{-\alpha_f^{(k)}}{1 + \lambda_{2k-1}} & 1 \end{bmatrix}$$

for $k = 1, 2, \cdots, p$. Thus for each $\Lambda \geqq 0$, $T_\Lambda \in \mathfrak{I}(T)$; and if M is an arbitrary element of $\mathfrak{I}(T)$, there is a $\Lambda \geqq 0$ such that $M = T_\Lambda$. Therefore $(D_a + T)^{-1}(D_b + G) \in P_0$ for all $D_a \geqq 0$ and all $D_b \geqq 0$ if and only if $M^{-1} G \in P_0$ for all $M \in \mathfrak{I}(T)$.

[(*i*) and (*iii*) are equivalent]

Repeat the proof of "(*i*) is equivalent to (*ii*)" with each statement that $D_a \geqq 0$ replaced with $D_a = \operatorname{diag} (0, 0, \cdots, 0)$.

[(*ii*) and (*iv*) are equivalent and (*iii*) and (*v*) are equivalent]

Suppose that (*ii*) and (*iv*) are not equivalent. Then $(D_a + T)^{-1}(D_b + G) \in P_0$ for all $D_a > 0$ and all $D_b > 0$, and for some $D_a^* \geqq 0$ and some $D_b^* \geqq 0$, with $D_a^* \not> 0$ or $D_b^* \not> 0$ or $D_a^* \not> 0$ and $D_b^* \not> 0$, $(D_a^* + T)^{-1}(D_b^* + G) \notin P_0$. Thus some principal minor of $(D_a^* + T)^{-1}(D_b^* + G)$, and hence of $(D_a^* + T)^{-1}(D_b^* + G) \det (D_a^* + T)$, is negative. Let

$m(D_a^*, D_b^*)$ be some negative principal minor of $(D_a^* + T)^{-1}(D_b^* + G)$ det $(D_a^* + T)$, and let $m(D_a^* + \epsilon I, D_b^* + \epsilon I)$ be the corresponding principal minor of $(D_a^* + \epsilon I + T)^{-1}(D_b^* + \epsilon I + G)$ det $(D_a^* + \epsilon I + T)$ for all real $\epsilon \geqq 0$. Thus $m(D_a^* + \epsilon I, D_b^* + \epsilon I)$ is a polynomial $p(\epsilon)$ in ϵ for $\epsilon \geqq 0$, and $p(\epsilon) \geqq 0$ for all $\epsilon > 0$. Therefore $p(0) \geqq 0$, which contradicts $m(D_a^*, D_b^*) < 0$.

A proof that (*iii*) and (*v*) are equivalent can be obtained by modifying the previous paragraph in an obvious manner.

[(*vi*) is equivalent to (*v*)]

By Lemma 1, $T^{-1}(G + D) \in P_0$ for all diagonal $D > 0$ if and only if det $[T^{-1}(G + D) + D^*] \neq 0$ for all diagonal $D^* > 0$ and $D > 0$, and hence if and only if det $(G + TD^* + D) \neq 0$ for all $D^* > 0$ and all $D > 0$. Therefore, by Lemma 1, $T^{-1}(G + D) \in P_0$ for all $D > 0$ if and only if $(G + TD^*) \in P_0$ for all $D^* > 0$, that is, if and only if det $[G_w + (TD^*)_w] \geqq 0$ for all $w \in Q_{(2p+q)}$ and all $D^* > 0$. Since $(TD^*)_w = T_w D_w^*$, we see that $T^{-1}(G + D) \in P_0$ for all $D > 0$ if and only if

$$\det [(T_w)^{-1}G_w + D_w^*] \geqq 0 \quad \text{for all} \quad w \in Q_{(2p+q)} \quad \text{and all} \quad D^* > 0. \tag{20}$$

But, by Lemma 2 (which follows) condition (20) is equivalent to the condition that det $[(T_w)^{-1}G_w + D_w^*] > 0$, and hence that det $[(T_w)^{-1}G_w + D_w^*] \neq 0$, for all $w \in Q_{(2p+q)}$ and all $D^* > 0$. Thus by Lemma 1, $T^{-1}(G + D) \in P_0$ for all $D > 0$ if and only if $(T_w)^{-1}G_w \in P_0$ for all $w \in Q_{(2p+q)}$.

Lemma 2: If A is a real square matrix of order n such that det $(D + A) = 0$ *for some diagonal $D > 0$, then* det $(D + A) < 0$ *for some diagonal $D > 0$.*

Proof: Using the notation of the proof of Theorem 2,

$$\det (D + A) = \sum{}' d_1^{e_1} d_2^{e_2} \cdots d_n^{e_n} m_{e_1, e_2, \cdots, e_n} \tag{21}$$

for all $D > 0$. Since $m_{1,1,\cdots,1} = 1$, if det $(D + A) = 0$ for some $D > 0$, then for at least one sequence $e_1', e_2', \cdots, e_n'$ we have $m_{e_1', e_2', \cdots, e_n'} < 0$. If $m_{0,0,\cdots,0} = \det A < 0$, then there exists a positive constant σ_1 such that det $(D + A) < 0$ whenever $0 < d_i < \sigma_1$ for all j. If det $A \geqq 0$, then, with $d_j = d$ for all j such that $e_j' = 1$ and $d_i = d^{-1}$ for all j such that $e_j' = 0$, there exists a positive constant σ_2 such that det $(D + A) < 0$ for all $d > \sigma_2$ [see (21)]. □

[(*vi*) and (*vii*) are equivalent]

We shall prove that

$$[(T^w)^{-1}G]_w = (T_w)^{-1}G_w \quad \text{for all} \quad w \in Q_{(2p+q)}. \tag{22}$$

Obviously the equality of (22) is satisfied if w is the null set.

It is convenient to introduce the following notation. Let u denote the 1×1 matrix containing the entry 1. Let φ denote what might be called the empty matrix, a matrix with no rows or columns; by this we mean that φ is to be interpreted in the following manner: $\varphi \oplus \varphi = \varphi$, $I_s = \varphi$ when $s = 0$, $\varphi^{-1} = \varphi$, and if M_1 and M_2 are any two (ordinary) matrices, then $\varphi \oplus M_1 = M_1$, $M_1 \oplus \varphi = M_1$, and $M_1 \oplus \varphi \oplus M_2 = M_1 \oplus M_2$.

Let $w \in Q_{(2p+q)}$ and let w not be the null set. The matrix T can be written as the direct sum $T_1 \oplus T_2 \oplus \cdots \oplus T_p \oplus I_q$. In terms of u and φ, $T_w = t_1 \oplus t_2 \oplus \cdots \oplus t_p \oplus I_s$, in which $s = q - \bar{q}$ where $\bar{q}$ is the number of elements contained in the intersection of the sets w and $\{2p + 1, 2p + 2, \cdots, 2p + q\}$, and for $k = 1, 2, \cdots, p$: $t_k = T_k$ if both $(2k - 1)$ and $2k$ are not elements of w, $t_k = \varphi$ if both $(2k - 1)$ and $2k$ are elements of w, and $t_k = u$ if either $(2k - 1) \in w$ and $2k \notin w$ or $(2k - 1) \notin w$ and $2k \in w$. Thus $(T_w)^{-1} = t_1^{-1} \oplus t_2^{-1} \oplus \cdots \oplus t_p^{-1} \oplus I_s$. But $(T^w)^{-1} = \hat{T}_1^{-1} \oplus \hat{T}_2^{-1} \oplus \cdots \oplus \hat{T}_p^{-1} \oplus I_q$, in which for $k = 1, 2, \cdots, p$: $\hat{T}_k = T_k$ if both $(2k - 1)$ and $2k$ are not elements of w,

$$\hat{T}_k^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

if both $(2k - 1)$ and $2k$ are contained in w,

$$\hat{T}_k^{-1} = \begin{bmatrix} 1 & \alpha_r^{(k)} \\ 0 & 1 \end{bmatrix}$$

if $(2k - 1) \in w$ and $2k \notin w$, and

$$\hat{T}_k^{-1} = \begin{bmatrix} 1 & 0 \\ \alpha_f^{(k)} & 1 \end{bmatrix}$$

if $(2k - 1) \notin w$ and $2k \in w$. Thus we see that $[(T^w)^{-1}]_w = (T_w)^{-1}$. Let ${}_{(w)}(T^w)^{-1}$ denote the $(2p + q - r) \times (2p + q)$ matrix obtained from $(T^w)^{-1}$ by deleting rows $i_1, i_2, \cdots, i_r$. But all elements of columns $i_1, i_2, \cdots, i_r$ of ${}_{(w)}(T^w)^{-1}$ are zeros, and hence, with $G_{(w)}$ the matrix obtained from G by deleting columns $i_1, i_2, \cdots, i_r$,

$$\begin{aligned} [(T^w)^{-1}G]_w &= {}_{(w)}(T^w)^{-1}G_{(w)} \\ &= [(T^w)^{-1}]_w G_w = (T_w)^{-1}G_w. \end{aligned}$$

[(*viii*) and (*i*) are equivalent]

If $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}_0(T)$, then $[(T^w)^{-1}G]_w \in P_0$ for all $w \in Q_{(2p+q)}$. Thus, statement (*viii*) implies statement (*vii*). Since we have proved that (*vii*) is equivalent to (*i*), it suffices to prove that (*i*) implies (*viii*).

Suppose that $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$. Let $\hat{M}$ be an arbitrary element of $\mathfrak{I}_0(T)$. Then $[\hat{M} + \delta(T - \hat{M})] \in \mathfrak{I}(T)$ for all $\delta \in (0, 1]$, and therefore $[\hat{M} + \delta(T - \hat{M})]^{-1}G \in P_0$ for all $\delta \in (0, 1]$. At this point a continuity-type argument similar to that used in the proof of [(*ii*) and (*iv*) are equivalent] shows that $\hat{M}^{-1}G \in P_0$. □

3.13 *Corollary 1 (Corollary to Theorem 5):*

If $T \in \mathfrak{I}$ *and* $M^{-1}G \in P_0$ *for all* $M \in \mathfrak{I}(T)$, *then* $\det(\sigma I + J_u) \neq 0$ *for all* $\sigma \geqq 0$ *and all* $u \in E^{(2p+q)}$ *provided that for all* j $f_j(\cdot)$ *is continuously differentiable on* $(-\infty, \infty)$ *and* $f_j'(\alpha) > 0$ *for all* $\alpha \in (-\infty, \infty)$.

3.14 *Proof of Corollary 1.*

If $T \in \mathfrak{I}$ and $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$, then, by the equivalence of (*i*) and (*ii*) of Theorem 5, $(\sigma\tau + T)^{-1}(\sigma c + G) \in P_0$ for all $\sigma \geqq 0$. The first portion of the proof of Theorem 1 shows that if $(\sigma\tau + T)^{-1}(\sigma c + G) \in P_0$ for all $\sigma \geqq 0$ and if for all j $f_j(\cdot)$ is continuously differentiable on $(-\infty, \infty)$ and $f_j'(\alpha) > 0$ for all $\alpha \in (-\infty, \infty)$, then $\det(\sigma I + J_u) \neq 0$ for all $\sigma \geqq 0$ and all $u \in E^{(2p+q)}$.

3.15 *Definition 10:* For $p > 0$ let $Q'_{(2p+q)}$ denote the subset of $Q_{(2p+q)}$ containing all sets w belonging to $Q_{(2p+q)}$ such that w is not the null set and $2k \in w$ if and only if $(2k - 1) \in w$ for $k = 1, 2, \cdots, p$. For $p = 0$, let $Q'_{(2p+q)}$ denote the family of all sets contained in $Q_{(2p+q)}$ with the exception of the null set.

3.16 *Theorem 6: If* $T \in \mathfrak{I}$ *and* $T^{-1}G \in P_0$, *then* $(T_w)^{-1}G_w \in P_0$ *for all* $w \in Q'_{(2p+q)}$.

3.17 *Proof of Theorem 6*

Let $T \in \mathfrak{I}$, and let $T^{-1}G \in P_0$. By Lemma 1, $\det(TD + G) \neq 0$ (and hence $\det(TD + G) > 0$) for all diagonal $D > 0$. Let $w = \{i_1, i_2, \cdots, i_r\} \in Q'_{(2p+q)}$, and let $d_{i_k} = d$ for $k = 1, 2, \cdots, r$.

It may be the case that $(TD + G)$ is a block matrix of the form

$$\begin{bmatrix} (TD + G)_w & H_{12} \\ H_{21} & (d\hat{T} + H_{22}) \end{bmatrix} \tag{23}$$

in which $\hat{T}$ is a direct sum of all 2×2 and 1×1 block matrices on the diagonal of T which do not appear in T_w, and H_{12}, H_{21}, and H_{22} are independent of D. Clearly $\det \hat{T} > 0$. If $(TD + G)$ is not of the form (23), then by a sequence of interchanges of rows and corresponding columns of $(TD + G)$ we obtain a matrix of that form.

Thus, for some $\hat{T}$ of the form indicated above and for the corresponding constant matrices H_{12}, H_{21}, and H_{22} whose elements are elements of G,

$$\det (TD + G) = \det \begin{bmatrix} (TD+G)_w & H_{12} \\ H_{12} & (d\hat{T} + H_{22}) \end{bmatrix}$$

for all $d_i > 0$ for $j \notin w$. For all sufficiently large $d > 0$, $\det (d\hat{T} + H_{22}) > 0$, and then

$$0 < \det (TD + G) = \det (d\hat{T} + H_{22}) \cdot \det [(TD + G)_w - H_{12}(d\hat{T} + H_{22})^{-1}H_{21}]$$

for all $d_i > 0$ for $j \notin w$. Since $H_{12}(d\hat{T} + H_{22})^{-1}H_{21}$ approaches the zero matrix of order $(2p + q - r)$ as $d \to \infty$, we must have $\det (TD + G)_w \geqq 0$ for all $d_i > 0$ for $j \notin w$. Therefore, since $(TD)_w = T_w D_w$, we must have $\det (T_w D_w + G_w) \geqq 0$ for all $D_w > 0$. But this means (see Lemma 2) that $\det (T_w D_w + G_w) \not\equiv 0$ for all $D_w > 0$. Thus, by Lemma 1, $(T_w)^{-1}G_w \in P_0$. □

3.18 *Theorem 7: If $T \in \mathfrak{I}$ with $p = 1$ or $p = 2$, and if $T^{-1}G \in P_0$ with G the short-circuit conductance matrix of a transformerless positive-element resistance network, then $(T_w)^{-1}G_w \in P_0$ for all $w \in Q_{(2p+q)}$.*

3.19 *Proof of Theorem 7*

Suppose that $T^{-1}G \in P_0$ with $p = 2$. Theorem 6 asserts that $(T_w)^{-1}G_w \in P_0$ for all $w \in Q'_{(2p+q)}$. But, aside from the null set, the sets $w = \{i_1, i_2, \cdots, i_r\}$ that are contained in $Q_{(2p+q)}$ but not in $Q'_{(2p+q)}$ possess the property that $T_w = T_1 \oplus I_{(2+q-r)}$, or $T_w = u \oplus T_2 \oplus I_{(1+q-r)}$ where u is the 1×1 matrix containing the element 1, or $T_w = I_{(4+q-r)}$.

If $T_w = I_{(4+q-r)}$, then obviously $(T_w)^{-1}G_w \in P_0$. If $T_w = T_1 \oplus I_{(2+q-r)}$, then for any $D_w = \text{diag}\,[D_2 \oplus D_{(2+q-r)}]$ with $D_2 > 0$ and $D_{(2+q-r)} > 0$ diagonal matrices of order 2 and $(2 + q - r)$ respectively,

$$\det (T_w D_w + G_w) = \begin{bmatrix} T_1 D_2 + G_{11} & G_{12} \\ G_{21} & D_{(2+q-r)} + G_{22} \end{bmatrix} \tag{24}$$

in which G_{11}, G_{12}, G_{21}, and G_{22} are the appropriate block matrices of G_w. Since $\det [D_{(2+q-r)} + G_{22}] > 0$, we have

$$\det (T_w D_w + G_w) = \det [D_{(2+q-r)} + G_{22}] \cdot \det \{T_1 D_2 + G_{11} - G_{12}[D_{(2+q-r)} + G_{22}]^{-1}G_{21}\}.$$

But $G_{11} - G_{12}[D_{(2+q-r)} + G_{22}]^{-1}G_{21}$ is the short-circuit conductance matrix of a transformerless common-ground 2-port network; it is of the form

$$\begin{bmatrix} g_{11} & -g_{12} \\ -g_{12} & g_{22} \end{bmatrix}$$

with $g_{11} \geqq 0$, $g_{22} \geqq 0$, $g_{12} \geqq 0$, $g_{11} \geqq g_{12}$, and $g_{22} \geqq g_{12}$. Therefore[1]

$$\det \{T_1 D_2 + G_{11} - G_{12}[D_{(2+q-r)} + G_{22}]^{-1}G_{21}\} > 0$$

for all $D_2 > 0$ and all $D_{(2+q-r)} > 0$, $\det (T_w D_w + G_w) \not\equiv 0$ for all $D_w > 0$, and hence, by Lemma 1, $(T_w)^{-1}G_w \in P_0$. Finally, the case in which $T_w = u \oplus T_2 \oplus I_{(1+q-r)}$ can be treated in a manner similar to that used to show that $(T_w)^{-1}G_w \in P_0$ when $T_w = T_1 \oplus I_{(2+q-r)}$, since, with w such that $T_w = u \oplus T_2 \oplus I_{(1+q-r)}$, and with D an arbitrary diagonal matrix of order $(4 + q - r)$, a sequence of interchanges of rows and corresponding columns of $(T_w D + G_w)$ can be performed to obtain a matrix of the type that appears on the right side of equation (24). Therefore $(T_w)^{-1}G_w \in P_0$ for all $w \in Q_{(2p+q)}$.

When $p = 1$, aside from the null set, the sets $w = \{i_1, i_2, \cdots, i_r\}$ that are contained in $Q_{(2p+q)}$ but not in $Q'_{(2p+q)}$ possess the property that $T_w = I_{(2+q-r)}$ and obviously when $T_w = I_{(2+q-r)}$, $(T_w)^{-1}G_w \in P_0$. □

3.20 *Theorem 8: Let $T \in \mathfrak{I}$ and let G possess the property that for some diagonal matrix $D > 0$, both DT and DG are strongly-column-sum dominant. For each $j = 1, 2, \cdots, (2p + q)$ let $f_j(\cdot)$ be a continuous monotone-nondecreasing mapping of E^1 into itself such that $f_j(0) = 0$, let $h \in (0, \infty)$, and, with $F(\cdot)$ and $C(\cdot)$ defined relative to the $f_j(\cdot)$ as in Section 2.1, suppose that the sequences $\{y_n\}$ and $\{w_n\}$ in $E^{(2p+q)}$ satisfy*

$$y_{n+1} + h\{TF[C^{-1}(y_{n+1})] + GC^{-1}(y_{n+1})\} = y_n + w_n$$

for all $n \geqq 0$. Then there exists a positive constant δ depending only on the c_j, the τ_j, T, G, and D such that

$$(i) \qquad \| Dy_n \|_1 \leqq (1 + \delta h)^{-n} \| Dy_0 \|_1 + \sum_{k=1}^{n} (1 + \delta h)^{-k} \| Dw_{(n-k)} \|_1$$

for all $n \geqq 1$, and

$$(ii) \qquad \| D(y_n - \tilde{y}_n) \|_1 \leqq (1 + \delta h)^{-n} \| D(y_0 - \tilde{y}_0) \|_1 + \epsilon \sum_{k=0}^{n} (1 + \delta h)^{-k}$$

for all $n \geqq 1$, in which $\{\tilde{y}_n\}$ is any sequence in $E^{(2p+q)}$ with the property that $\| D(\tilde{y}_n - y_n^*) \|_1 \leqq \epsilon$ for all $n \geqq 1$ with ϵ a positive constant and the

sequence $\{y_n^*\}$ *such that*

$$y_{n+1}^* + h\{TF[C^{-1}(y_{n+1}^*)] + GC^{-1}(y_{n+1}^*)\} = \tilde{y}_n + w_n$$

for all $n \geqq 0$.

3.21 *Proof of Theorem 8*

We shall first prove part (*ii*). With D such that DT and DG are strongly-column-sum dominant, we have for all $n \geqq 0$

$$Dy_{n+1} + h\{DTF[C^{-1}(y_{n+1})] + DGC^{-1}(y_{n+1})\} = Dy_n + Dw_n$$

and

$$Dy_{n+1}^* + h\{DTF[C^{-1}(y_{n+1}^*)] + DGC^{-1}(y_{n+1}^*)\} = Dy_n^* + D(\tilde{y}_n - y_n^*) + Dw_n$$

in which we shall take y_0^* to be $\tilde{y}_0$. As in the proof of Theorem 2 of Ref. 3, we write

$$F[C^{-1}(y_{n+1})] - F[C^{-1}(y_{n+1}^*)] = \operatorname{diag}\left(\frac{r(n)_j}{c_j + \tau_j r(n)_j}\right)(y_{n+1} - y_{n+1}^*) \quad (25)$$

and

$$C^{-1}(y_{n+1}) - C^{-1}(y_{n+1}^*) = \operatorname{diag}\left(\frac{1}{c_j + \tau_j r(n)_j}\right)(y_{n+1} - y_{n+1}^*) \quad (26)$$

in which $r(n)_j$ depends on the jth components of y_{n+1} and y_{n+1}^*, and $r(n)_j \geqq 0$ for all $n \geqq 0$ and all j.

Thus, with $Q = DTD^{-1}$ and $R = DGD^{-1}$,

$$\left\{I + hQ \operatorname{diag}\left(\frac{r(n)_j}{c_j + \tau_j r(n)_j}\right) + hR \operatorname{diag}\left(\frac{1}{c_j + \tau_j r(n)_j}\right)\right\} D(y_{n+1} - y_{n+1}^*)$$
$$= D(y_n - y_n^*) - D(\tilde{y}_n - y_n^*)$$

for all $n \geqq 0$. At this point we shall use the proposition that if M is any real matrix of order $(2p + q)$ with the property that there exists a positive constant η such that $m_{jj} - \sum_{i \neq j} |m_{ij}| \geqq \eta$ for all j, then $\| Mx \|_1 \geqq \eta \| x \|_1$ for all $x \in E^{(2p+q)}$. Now let

$$M = \left\{I + hQ \operatorname{diag}\left(\frac{r(n)_j}{c_j + \tau_j r(n)_j}\right) + hR \operatorname{diag}\left(\frac{1}{c_j + \tau_j r(n)_j}\right)\right\}$$

for arbitrary $n \geqq 0$. Then for arbitrary j

$$m_{jj} - \sum_{i \neq j} |m_{ij}| = 1 + hq_{jj}\left(\frac{r(n)_j}{c_j + \tau_j r(n)_j}\right) + hr_{jj}\left(\frac{1}{c_j + \tau_j r(n)_j}\right)$$
$$- h \sum_{i \neq j} \left| q_{ij} \frac{r(n)_j}{c_j + \tau_j r(n)_j} + r_{ij} \frac{1}{c_j + \tau_j r(n)_j} \right|$$

$$\geqq 1 + h\Big(q_{ii} - \sum_{i \neq j} |q_{ii}|\Big) \frac{r(n)_i}{c_i + \tau_i r(n)_i}$$

$$+ h\Big(r_{ii} - \sum_{i \neq j} |r_{ii}|\Big) \frac{1}{c_i + \tau_i r(n)_i}$$

$$\geqq 1 + \delta h,$$

in which

$$\delta = \min\left\{ \min_i c_i^{-1}\Big(r_{ii} - \sum_{i \neq j} | r_{ii} |\Big), \quad \min_i \tau_i^{-1}\Big(q_{ii} - \sum_{i \neq j} | q_{ii} |\Big)\right\}.$$

Therefore

$$\begin{aligned} &|| D(y_{n+1} - y^*_{n+1}) ||_1 \\ &\qquad \leqq (1 + \delta h)^{-1} || D(y_n - y^*_n) - D(\tilde{y}_n - y^*_n) ||_1 \\ &\qquad \leqq (1 + \delta h)^{-1} || D(y_n - y^*_n) ||_1 + (1 + \delta h)^{-1} || D(\tilde{y}_n - y^*_n) ||_1 \\ &\qquad \leqq (1 + \delta h)^{-1} || D(y_n - y^*_n) |_{11} + \epsilon(1 + \delta h)^{-1} \end{aligned}$$

for all $n \geqq 0$, and hence

$$|| D(y_n - y^*_n) ||_1 \leqq (1 + \delta h)^{-n} || D(y_0 - y^*_0) ||_1 + \epsilon \sum_{k=1}^{n} (1 + \delta h)^{-k}$$

for all $n \geqq 1$. Finally, since $|| D(y_n - \tilde{y}_n) ||_1 \leqq || D(y_n - y^*_n) ||_1 + || D(y^*_n - \tilde{y}_n) ||_1 \leqq || D(y_n - y^*_n) ||_1 + \epsilon$, and since $y^*_0 = \tilde{y}_0$,

$$|| D(y_n - \tilde{y}_n) ||_1 \leqq (1 + \delta h)^{-n} || D(y_0 - \tilde{y}_0) ||_1 + \epsilon \sum_{k=0}^{n} (1 + \delta h)^{-k}$$

for all $n \geqq 1$, which completes the proof of part (*ii*) of the theorem.

The proof of part (*i*) is similar to that of part (*ii*). Using

$$Dy_{n+1} + h\{DTF[C^{-1}(y_{n+1})] + DGC^{-1}(y_{n+1})\} = Dy_n + Dw_n$$

for all $n \geqq 0$, and equations (25) and (26) with $y^*_{n+1} = \theta$ for all n, we find that

$$|| Dy_{n+1} ||_1 \leqq (1 + h\,\delta)^{-1} || Dy_n ||_1 + (1 + h\,\delta)^{-1} || Dw_n ||_1$$

for all $n \geqq 0$. Therefore

$$|| Dy_n ||_1 \leqq (1 + h\delta)^{-n} || Dy_0 ||_1 + \sum_{k=1}^{n} (1 + h\delta)^{-k} || Dw_{(n-k)} ||_1$$

for all $n \geqq 1$. □

3.22 *Theorem 9:* *Let $T \in \mathfrak{I}$ and let G possess the property that for some diagonal matrix $D > 0$, both DT and DG are strongly-column-sum dominant. Let $B(\cdot)$ denote a real continuously-differentiable $(2p + q)$-vector-valued function of t for $t \in [0, \infty)$ such that both $B(\cdot)$ and $(d/dt)B(\cdot)$ are bounded on $[0, \infty)$. With $F(\cdot)$ such that each $f_j(0) = 0$, and with $C(\cdot)$ defined relative to $F(\cdot)$ as in Section 2.1, let $u(\cdot)$ satisfy*

$$\frac{du}{dt} + TF[C^{-1}(u)] + GC^{-1}(u) = B(t), \qquad t \geqq 0$$

and, with h an arbitrary positive constant, let u_n denote $u(nh)$ for all $n \geqq 0$. Let $\{y_n\}$ be a sequence in $E^{(2p+q)}$ such that

$$y_{n+1} + h\{TF[C^{-1}(y_{n+1})] + GC^{-1}(y_{n+1})] = y_n + hB[(n + 1)h], \quad n \geqq 0.$$

Then there exist positive constants δ and ρ, both independent of h, such that

$$|| D(u_n - y_n) ||_1 \leqq (1 + \delta h)^{-n} || D(u_0 - y_0) ||_1 + \rho h$$

for all $n \geqq 1$.

3.23 *Proof of Theorem 9*

The sequence $\{u_n\}$ satisfies

$$u_{n+1} + h\{TF[C^{-1}(u_{n+1})] + GC^{-1}(u_{n+1})\} = u_n + B[(n + 1)h] + \xi_n, \qquad n \geqq 0$$

in which ξ_n is often referred to as "the local-truncation error at step n." We shall first bound ξ_n.

Since $B(\cdot)$ is bounded on $[0, \infty)$, and since for some $D > 0$, both DT and DG are strongly-column-sum dominant, a direct modification of the proof of Theorem 1 of Ref. 5 shows that $u(\cdot)$ is bounded on $[0, \infty)$; and hence since

$$\frac{d^2u}{dt^2} = J_u\{TF[C^{-1}(u)] + GC^{-1}(u)\} - J_uB(t) + \frac{d}{dt}B(t), \qquad t \geqq 0 \qquad (27)$$

with $(d/dt)B(\cdot)$ and the elements of the Jacobian matrix J_u bounded, it is clear that (d^2u/dt^2) is bounded on $[0, \infty)$. By the usual Taylor-series-type argument we can show that for arbitrary $n \geqq 0$, $\xi_n = \frac{1}{2}h^2U_n$ in which for each j the jth component of U_n is the jth component of (d^2u/dt^2) evaluated at some point contained in the interval $[nh, (n + 1)h]$. Thus there exists a positive constant ρ_1 such that

$$|| D\xi_n ||_1 \leqq \tfrac{1}{2}h^2\rho_1 \quad \text{for all} \quad n \geqq 0. \qquad (28)$$

Therefore, using (28) and the equations

$$u_{n+1} + h\{TF[C^{-1}(u_{n+1})] + GC^{-1}(u_{n+1})\} = u_n + B[(n+1)h] + \xi_n, \qquad n \geqq 0$$

$$y_{n+1} + h\{TF[C^{-1}(y_{n+1})] + GC^{-1}(y_{n+1})\} = y_n + B[(n+1)h], \qquad n \geqq 0$$

by an argument similar to that used in the proof of part (*ii*) of Theorem 8, and with δ as defined there, we find that

$$|| D(u_{n+1} - y_{n+1}) ||_1 \leqq (1 + \delta h)^{-1} || D(u_n - y_n) ||_1 + (1 + \delta h)^{-1}\tfrac{1}{2}h^2\rho_1$$

for all $n \geqq 0$, and hence that

$$\begin{aligned} || D(u_n - y_n) ||_1 &\leqq (1 + \delta h)^{-n} || D(u_0 - y_0) ||_1 + \tfrac{1}{2}h^2\rho_1 \sum_{k=1}^{n} (1 + \delta h)^{-k} \\ &\leqq (1 + \delta h)^{-n} || D(u_0 - y_0) ||_1 + \tfrac{1}{2}h^2\rho_1 \sum_{k=1}^{\infty} (1 + \delta h)^{-k} \\ &\leqq (1 + \delta h)^{-n} || D(u_0 - y_0) ||_1 + \tfrac{1}{2}h\delta^{-1}\rho_1 \end{aligned}$$

for all $n \geqq 1$. □

3.24 *Definition 11:* Let $R = R_1 \oplus R_2 \oplus \cdots \oplus R_p \oplus R_0$ in which $R_0 = \text{diag}\,(r_1, r_2, \cdots, r_q)$ with $r_i \geqq 0$ for $j = 1, 2, \cdots, q$ and

$$R_k = \begin{bmatrix} r_e^{(k)} + r_b^{(k)} & r_b^{(k)} \\ r_b^{(k)} & r_c^{(k)} + r_b^{(k)} \end{bmatrix}$$

with $r_e^{(k)} \geqq 0$, $r_b^{(k)} \geqq 0$, and $r_c^{(k)} \geqq 0$ for all $k = 1, 2, \cdots, p$. As suggested, if $q = 0$, then $R = R_1 \oplus R_2 \oplus \cdots \oplus R_p$, while if $p = 0$, then $R = R_0$.

3.25 *Theorem 10: Let* $T \in \mathfrak{I}$. *If* $p > 0$ *and if* R *satisfies*

$$\alpha_r^{(k)}(1 - \alpha_r^{(k)})^{-1} r_e^{(k)} = r_b^{(k)}$$

$$\alpha_f^{(k)}(1 - \alpha_f^{(k)})^{-1} r_c^{(k)} = r_b^{(k)}$$

for $k = 1, 2, \cdots, p$, *then* $T^{-1}G(I + RG)^{-1} \in P_0$ *whenever* $T^{-1}G \in P_0$.

3.26 *Proof of Theorem 10*

By Lemma 1, $T^{-1}G(I + RG)^{-1} \in P_0$ if and only if

$$\det [T^{-1}G(I + RG)^{-1} + D^*] \neq 0 \tag{29}$$

for all diagonal $D^* > 0$. But (29) is satisfied if and only if

$$\det (T^{-1}G + D^*RG + D^*) \neq 0.$$

Here, since

$$\alpha_r^{(k)}(1 - \alpha_r^{(k)})^{-1} r_e^{(k)} = r_b^{(k)}$$
$$\alpha_f^{(k)}(1 - \alpha_f^{(k)})^{-1} r_c^{(k)} = r_b^{(k)}$$

for $k = 1, 2, \cdots, p$ we have $R = DT^{-1}$ for some diagonal matrix $D \geqq 0$. Thus (29) is satisfied if and only if

$$\det [(I + DD^*)T^{-1}G + D^*] \neq 0.$$

When $T^{-1}G \in P_0$ we have

$$\det (T^{-1}G + \tilde{D}) \neq 0$$

for all diagonal $\tilde{D} > 0$. Thus (29) is satisfied for all $D^* > 0$ whenever $T^{-1}G \in P_0$. □

3.27 *Theorem 11: If $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}$, then for any $T \in \mathfrak{I}$, $T^{-1}G(I + RG)^{-1} \in P_0$ for all R.*

3.28 *Proof of Theorem 11*

Let $T \in \mathfrak{I}$. As in the proof of Theorem 10, $T^{-1}G(I + RG)^{-1} \in P_0$ if and only if

$$\det [(T^{-1} + D^*R)G + D^*] \neq 0$$

for all diagonal $D^* > 0$. It is a simple matter to verify that for each $D^* > 0$ and each R there exists an $\tilde{M} \in \mathfrak{I}$ and a diagonal matrix $D > 0$ such that $(T^{-1} + D^*R) = D\tilde{M}^{-1}$. Since $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}$, we have (by Lemma 1)

$$\det (D\tilde{M}^{-1}G + D^*) \neq 0$$

for all $D^* > 0$. □

3.29 *Theorem 12: Let $T \in \mathfrak{I}$ with $p > 0$ and $q \geqq 0$. Then $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$ if and only if $T^{-1}G(I + RG)^{-1} \in P_0$ for all R such that*

$$\alpha_r^{(k)}(1 - \alpha_r^{(k)})^{-1} r_e^{(k)} \geqq r_b^{(k)}$$
$$\alpha_f^{(k)}(1 - \alpha_f^{(k)})^{-1} r_c^{(k)} \geqq r_b^{(k)}$$

for $k = 1, 2, \cdots, p$ and $r_i \geqq 0$ for all j such that $1 \leqq j \leqq q$.

3.30 *Proof of Theorem 12*

As in the proof of Theorem 10, $T^{-1}G(I + RG)^{-1} \in P_0$ if and only if

$$\det (T^{-1}G + D^*RG + D^*) \neq 0 \tag{30}$$

for all diagonal $D^* > 0$. The inequalities $r_j \geqq 0$ for all j such that $1 \leqq j \leqq q$ and

$$\alpha_r^{(k)}(1 - \alpha_r^{(k)})^{-1} r_e^{(k)} \geqq r_b^{(k)}$$
$$\alpha_f^{(k)}(1 - \alpha_f^{(k)})^{-1} r_c^{(k)} \geqq r_b^{(k)}$$

for $k = 1, 2, \cdots, p$ are equivalent to the condition that $R = D_1 T^{-1} + D_2$ for some diagonal matrix $D_2 \geqq 0$ and some diagonal matrix $D_1 \in S$, in which S is the set of all diagonal matrices $D \geqq 0$ such that DT^{-1} is symmetric. Hence $T^{-1}G(I + RG)^{-1} \in P_0$ for all such R if and only if

$$\det \{[(I + D_1 D^*)T^{-1} + D^* D_2]G + D^*\} \neq 0 \tag{31}$$

for all diagonal $D^* > 0$, $D_2 \geqq 0$, and $D_1 \in S$.

Let $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_{(2p+q)})$ be such that

$$D_2 = D^{*-1} \Delta^{-1} \Lambda (I + D_1 D^*)$$

in which

$$\Delta = \operatorname{diag}(\delta_1, \delta_1, \delta_2, \delta_2, \cdots, \delta_p, \delta_p) \oplus I_q$$

if $q > 0$, $\Delta = \operatorname{diag}(\delta_1, \delta_1, \delta_2, \delta_2, \cdots, \delta_p, \delta_p)$ if $q = 0$, and

$$\delta_k = 1 - \alpha_f^{(k)} \alpha_r^{(k)} \quad \text{for} \quad k = 1, 2, \cdots, p.$$

The left side of (31) is

$$\det [(I + D_1 D^*)(T^{-1} + \Delta^{-1}\Lambda)G + D^*]$$

which can be written as

$$\det [(I + D_1 D^*)\, \Delta^{-1}(I + \Lambda)\, \Delta_\Lambda T_\Lambda^{-1} G + D^*] \tag{32}$$

with

$$T_\Lambda^{-1} = \Delta_\Lambda^{-1}\, \Delta(I + \Lambda)^{-1}(T^{-1} + \Delta^{-1}\Lambda)$$

and

$$\Delta_\Lambda = \operatorname{diag}(\delta_1', \delta_1', \delta_2', \delta_2', \cdots, \delta_p', \delta_p') \oplus I_q$$

if $q > 0$ and $\Delta_\Lambda = \operatorname{diag}(\delta_1', \delta_1', \delta_2', \delta_2', \cdots, \delta_p', \delta_p')$ if $q = 0$, in which for $k = 1, 2, \cdots, p$

$$\delta_k' = 1 - \alpha_f^{(k)} \alpha_r^{(k)} (1 + \lambda_{(2k-1)})^{-1} (1 + \lambda_{2k})^{-1}.$$

But (32) vanishes if and only if $\det (T_\Lambda^{-1} G + \tilde{D})$ vanishes, in which $\tilde{D} = \Delta_\Lambda^{-1}(I + \Lambda)^{-1}\, \Delta(I + D_1 D^*)^{-1} D^*$. We observe that $\tilde{D}$ is a positive diagonal matrix and that given any diagonal $\tilde{D}' > 0$ and given any

$\Lambda \geqq 0$ we can choose $D^* > 0$ and $D_1 \in S$ so that $\tilde{D} = \tilde{D}'$. Thus $T^{-1}G(I + RG)^{-1} \in P_0$ for all $R = (D_1T^{-1} + D_2)$ with $D_1 \in S$ and $D_2 \geqq 0$ if and only if

$$\det (T_\Lambda^{-1}G + \tilde{D}) \neq 0$$

for all $\Lambda \geqq 0$ and $\tilde{D} > 0$, that is, if and only if $T_\Lambda^{-1}G \in P_0$ for all $\Lambda \geqq 0$ (see Lemma 1 of Section 3.1). But

$$T_\Lambda = T_1 \oplus T_2 \oplus \cdots \oplus T_p \oplus I_q \quad \text{if} \quad q > 0$$

and

$$T_\Lambda = T_1 \oplus T_2 \oplus \cdots \oplus T_p \quad \text{if} \quad q = 0$$

with

$$T_k = \begin{bmatrix} 1 & \dfrac{-\alpha_r^{(k)}}{1 + \lambda_{2k-1}} \\ \dfrac{-\alpha_f^{(k)}}{1 + \lambda_{2k}} & 1 \end{bmatrix}$$

for all $k = 1, 2, \cdots, p$. Therefore $T^{-1}G(I + RG)^{-1} \in P_0$ for all $R = (D_1T^{-1} + D_2)$ with $D_2 \geqq 0$ and $D_1 \in S$ if and only if $M^{-1}G \in P_0$ for all $M \in \mathfrak{I}(T)$. □

3.31 *Definition 12:* Let $\mathfrak{F}_3$ denote the set of all $F(\cdot)$ such that

(*i*) $F(\cdot) \in \mathfrak{F}_0^{(2p+q)}$, and
(*ii*) for each $j = 1, 2, \cdots, (2p + q)$ there exists a real constant β_j such that $f_j(\cdot)$ is a strictly-monotone-increasing mapping of E^1 onto either (β_j, ∞) or $(-\infty, \beta_j)$, and
(*iii*) whenever $p > 0$, $f_{(2k-1)}(\cdot)$ and $f_{2k}(\cdot)$ are both bounded on either $[0, \infty)$ or $(-\infty, 0]$ for $k = 1, 2, \cdots, p$.

3.32 *Theorem 13: Let $T \in \mathfrak{I}$, and, referring to the network of Fig. 1 in which it is assumed that R (see Section 2.1) is the zero matrix, let G denote the short-circuit conductance matrix of the linear portion of the network. (The linear portion is assumed to contain only sources and linear resistors of nonnegative resistance.) Then the equation $F(x) + T^{-1}Gx = B$ possesses a unique solution x for each $F(\cdot) \in \mathfrak{F}_3$ and each $B \in E^{(2p+q)}$ if and only if $T^{-1}G \in P_0$ and $\det G \neq 0$. If $T^{-1}G \in P_0$ and $\det G = 0$, then there exists a real $(2p + q)$-vector η such that (i) $\eta \neq \theta$, and for some $F(\cdot) \in \mathfrak{F}_3$ all of the components of $F(\alpha\eta)$ are bounded on $\alpha \in [0, \infty)$, and (ii) for any $F(\cdot) \in \mathfrak{F}_3$ with the property that all of the components of $F(\alpha\eta)$ are bounded on $\alpha \in [0, \infty)$ the equation $F(x) + T^{-1}Gx = B$ does not possess a solution for some $B \in E^{(2p+q)}$.*

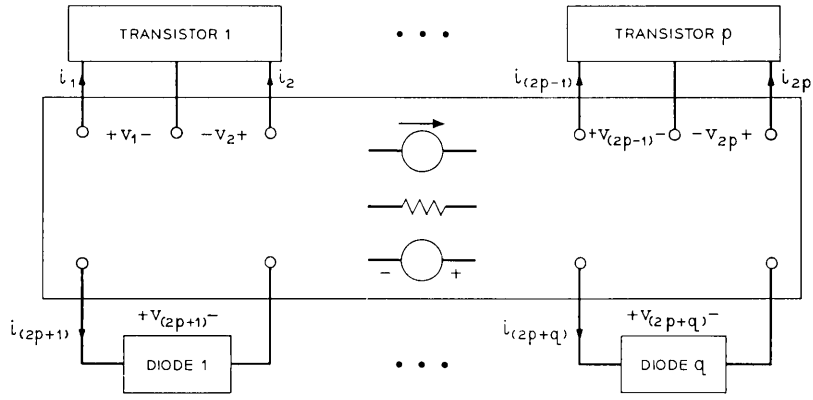


Fig. 1—General network containing transistors, diodes, resistors, and sources.

3.33 *Proof of Theorem 13*

(if) If $T^{-1}G \in P_0$ with $\det T^{-1}G \neq 0$, and if $F(\cdot) \in \mathfrak{F}_3$, then, since each $f_i(\cdot)$ is a strictly-monotone-increasing mapping of E^1 onto (β_i, ∞) or $(-\infty, \beta_i)$ for some real constant β_i, by Theorem 4 of Ref. 2, the equation $F(x) + T^{-1}Gx = B$ possesses a unique solution x for each $B \in E^{(2p+q)}$.

(only if) Assume that $T^{-1}G \notin P_0$. Then since $\mathfrak{F}_3$ is contained in $\mathfrak{F}_0^{(2p+q)}$, by Theorem 1 of Ref. 3, for each $F(\cdot) \in \mathfrak{F}_3$ there exists a $B \in E^{(2p+q)}$ such that there are at least two solutions x of $F(x) + T^{-1}Gx = B$.

Assume now that $T^{-1}G \in P_0$ and that $\det G = 0$. We shall use the proposition that if $R(\cdot)$ is any continuous mapping of $E^{(2p+q)}$ into itself, then $R(\cdot)$ is a homeomorphism of $E^{(2p+q)}$ onto itself if and only if $R(\cdot)$ is a local homeomorphism on $E^{(2p+q)}$ and $\| R(x) \| \to \infty$ as $\| x \| \to \infty$.†

Let $R(\cdot)$ be defined by the condition that $R(x) = F(x) + T^{-1}Gx$ for all $x \in E^{(2p+q)}$. For any $F(\cdot) \in \mathfrak{F}_3$ the operator $R(\cdot)$ is a local homeomorphism on $E^{(2p+q)}$, since with $F(\cdot)$ such that each $f_i(\cdot)$ is a strictly-monotone-increasing mapping of E^1 *onto* E^1 the mapping $[F(\cdot) + T^{-1}G]$ is a homeomorphism of $E^{(2p+q)}$ onto itself.[1] In addition, for any $F(\cdot) \in \mathfrak{F}_3$ and any $B \in E^{(2p+q)}$, there is at most one $x \in E^{)2p+q)}$ such that $R(x) = B$.[1]

Let us suppose that for each $B \in E^{(2p+q)}$ and each $F(\cdot) \in \mathfrak{F}_3$ there exists a solution x of $R(x) = B$. Then for all $F(\cdot) \in \mathfrak{F}_3$, $R(\cdot)$ is a homeomorphism of $E^{(2p+q)}$ onto itself, and hence for all $F(\cdot) \in \mathfrak{F}_3$ $\| R(x) \| \to \infty$ as $\| x \| \to \infty$. But, by Lemma 3 (which appears below) $E^{(2p+q)}$ contains a vector η such that $\eta \neq \theta$, $\eta_i \in \{0, +1, -1\}$ for all j, and $G\eta = \theta$; and if

† See Ref. 12 and the appendix of Ref. 13.

$p > 0$, η satisfies $\eta_{(2k-1)}\eta_{2k} \geqq 0$ for all $k = 1, 2, \cdots, p$. Let $\mathfrak{F}_3(\eta)$ denote the subset of $\mathfrak{F}_3$ containing all elements $F(\cdot)$ with the property that $f_i(\alpha\eta_i)$ is bounded on $\alpha \in [0, \infty)$ for all $j = 1, 2, \cdots, (2p + q)$. Since $\eta_{(2k-1)}\eta_{2k} \geqq 0$ for all $k = 1, 2, \cdots, p$ when $p > 0$, it is clear that $\mathfrak{F}_3(\eta)$ is not empty. However, for any $F(\cdot) \in \mathfrak{F}_3(\eta)$ we have $\| R(\alpha\eta) \| = \| F(\alpha\eta) \|$ with $\| F(\alpha\eta) \|$ bounded on $\alpha \in [0, \infty)$, which contradicts the assumption that there exists a solution x of $R(x) = B$ for each $F(\cdot) \in \mathfrak{F}_3$ and each $B \in E^{(2p+q)}$.

Lemma 3: Let G be the short-circuit conductance matrix of the linear portion of the network of Fig. 1. If $\det G = 0$, *then there exists a vector* $\eta \in E^{(2p+q)}$ *such that* $G\eta = \theta$, $\eta \neq \theta$, *and* $\eta_j \in \{0, +1, -1\}$ *for all* $j = 1, 2, \cdots, (2p + q)$; *and if* $p > 0$ η *also satisfies* $\eta_{(2k-1)}\eta_{2k} \geqq 0$ *for* $k = 1, 2, \cdots, p$.

Proof of Lemma 3:

Let N denote the $(2p + q)$-port resistor network obtained from the network of Fig. 1 by removing all transistors and diodes and by setting the value of each source to zero. The short-circuit conductance matrix G possesses the property that if $v \in E^{(2p+q)}$ denotes the vector of port voltages of N and $i \in E^{(2p+q)}$ denotes the corresponding vector of port currents (with polarities as indicated in Fig. 1), then $i = -Gv$.

Let $\det G = 0$. Then the open-circuit resistance matrix of N does not exist. Therefore there exists a port ℓ of N such that there is no path through resistors of N that connects the two terminals of port ℓ when all other ports are open-circuited. Let a one-volt source be placed at port ℓ so that $v_\ell = 1$. Then when all ports j of N with $j \neq \ell$ are open-circuited, $i_\ell = 0$ and there is zero current in every resistor of N. Let S denote a set of port numbers of N with the following properties. The number ℓ is not contained in S and when all ports j with $j \in S$ are short-circuited and all ports j with $j \notin S \cup \{\ell\}$ are open-circuited then zero current flows through the one-volt source; when any port $j_1 \notin S \cup \{\ell\}$ and all ports j with $j \in S$ are short-circuited and all ports j with $j \notin S \cup \{\ell, j_1\}$ are open-circuited then nonzero current flows through the one-volt source. It is clear that such a set S exists (with the understanding that S might be the null set). In general S contains r port numbers where $0 \leqq r \leqq (2p + q - 1)$.

If $r = (2p + q - 1)$, then with $v_\ell = 1$ and with all remaining components of v equal to zero, we have $Gv = \theta$. Obviously in this case we can take the vector η of the statement of Lemma 3 to be v.

If $r \neq (2p + q - 1)$, then, with $v_\ell = 1$, with $v_j = 0$ for all $j \in S$,

and with all ports $j \notin S \cup \{\ell\}$ open-circuited, there exists for each $j \notin S \cup \{\ell\}$ some path through the one-volt source and the resistors of N that connects the two terminals of port j. Therefore when $r \neq (2p + q - 1)$, when all ports $j \notin S \cup \{\ell\}$ are open circuited, when $v_\ell = 1$, and when $v_i = 0$ for all $j \in S$, the open-circuit voltage v_i at each port j with $j \notin S \cup \{\ell\}$ is well defined and nonzero. Since no current flows in any resistor of N when $v_\ell = 1$, $v_i = 0$ for all $j \in S$, and all ports $j \notin S \cup \{\ell\}$ are open-circuited, it follows that $v_i \in \{-1, +1\}$ for all $j \notin S$. With $v_\ell = 1$, with $v_i = 0$ for all $j \in S$, and with v_i the corresponding open-circuit voltage for each $j \notin S \cup \{\ell\}$, we have $Gv = \theta$. When $p > 0$, the vector v also satisfies the condition that $v_{(2k-1)}v_{2k} \geqq 0$ for all $k = 1, 2, \cdots, p$ since if $v_{(2k-1)}v_{2k}$ were negative for some k, then for that k $v_{(2k-1)} = 1$ and $v_{2k} = -1$ or $v_{(2k-1)} = -1$ and $v_{2k} = 1$; in either case $| v_{(2k-1)} - v_{2k} | = 2$ which contradicts the proposition that a network of nonnegative resistors can have no voltage gain. □

APPENDIX*

A theorem due to R. S. Palais† asserts that if $R(\cdot)$ is a continuously-differentiable mapping of E^n into itself with values $R(q)$ for $q \in E^n$, then $R(\cdot)$ is a diffeomorphism‡ of E^n onto itself if and only if

(*i*) $\det J_q \neq 0$ for all $q \in E^n$, in which J_q is the Jacobian matrix of $R(\cdot)$ with respect to q, and

(*ii*) $|| R(q) || \to \infty$ as $|| q || \to \infty$.

If $R(\cdot)$ is any twice-continuously-differentiable mapping of E^n into itself such that conditions (*i*) and (*ii*) of Palais' theorem are satisfied, then E^n contains a unique element x such that $R(x) = \theta$ in which θ is the zero element of E^n, and there are steepest decent as well as Newton-type algorithms each of which generates a sequence in E^n that converges to x. To show this, let [18] $f(y) = || R(y) ||^2$ for all $y \in E^n$ in which $||\cdot||$ denotes the usual Euclidean norm (i.e., the square-root of the sum of squares). Since condition (*i*) of Palais' theorem is satisfied, the gradient ∇f of $f(\cdot)$ satisfies $(\nabla f)(y) \neq \theta$ unless $f(y) = 0$,§ and since condition (*ii*) of Palais' theorem is satisfied, the set $S = \{y \in E^n : f(y) \leqq f(x^{(0)})\}$ is bounded for any $x^{(0)} \in E^n$. Therefore we may appeal to, for example, the theorem of page 43 of Ref. 18 according to which for any $x^{(0)} \in E^n$, for any member of a certain class of mappings $\varphi(\cdot)$ of S

* The material of this appendix together with some misprints appears in Ref. 3.

† See Ref. 12 and the appendix of Ref. 13.

‡ A diffeomorphism of E_n onto itself is a continuously differentiable mapping of E_n into E_n which possesses a continuously differentiable inverse.

§ Here we have used the fact that $(\nabla f)(y) = 2J_y{}^{tr}R(y)$ for all $y \in E_n$.[18]

into E^n, and for suitably chosen constants γ_0 , γ_1 , $\cdots$, the sequence $x^{(0)}$, $x^{(1)}$, $\cdots$ defined by

$$x^{(k+1)} = x^{(k)} + \gamma_k \varphi(x^{(k)}) \quad \text{for all} \quad k \geqq 0$$

belongs to S and is such that $\| R(x^{(k)}) \| \to 0$ as $k \to \infty$. However, since $R^{-1}(\cdot)$ exists and is continuous, it follows from

$$x^{(k)} = R^{-1}[R(x^{(k)})] \quad \text{for all} \quad k \geqq 0$$

and the fact that $R(x^{(k)}) \to \theta$ as $k \to \infty$, that $\lim_{k\to\infty} x^{(k)}$ exists and

$$\lim_{k\to\infty} x^{(k)} = R^{-1}(\theta),$$

which means that $\lim_{k\to\infty} x^{(k)}$ is the unique solution x of $R(y) = \theta$.

REFERENCES

1. 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.
2. 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. 1293–1312.
3. Sandberg, I. W., "Theorems on the Analysis of Nonlinear Transistor Networks," B.S.T.J., *49*, No. 1 (January 1970), pp. 95–114.
4. Willson, A. N., Jr., "New Theorems on the Equations of Nonlinear DC Transistor Networks," B.S.T.J., this issue, pp. 1713–1738.
5. Sandberg, I. W., "Some Theorems on the Dynamic Response of Nonlinear Transistor Networks," B.S.T.J., *48*, No. 1 (January 1969), pp. 35–54.
6. Hamming, R. W., *Numerical Methods for Scientists and Engineers*, New York: McGraw-Hill Book Co., (1962).
7. Ralston, A. A., *A First Course in Numerical Analysis*, New York: McGraw-Hill Book Co., (1965).
8. Hachtel, G. D., and Rohrer, R. A., "Techniques for the Optimal Design and Synthesis of Switching Circuits," Proc. of the IEEE, *55*, No. 11 (November 1967), pp. 1864–1876.
9. Sandberg, I. W., and Shichman, H., "Numerical Integration of Systems of Stiff Nonlinear Differential Equations," B.S.T.J., *47*, No. 4 (April 1968), pp. 511–527.
10. Calahan, D. A., "Efficient Numerical Analysis of Non-Linear Circuits," Proc. Sixth Ann. Allerton Conf. on Circuit and System Theory, University of Illinois, 1968, pp. 321–331.
11. Vehovec, M., "Simple Criterion for the Global Regularity of Vector-Valued Functions," Elec. Letters, *5*, No. 26 (December 1969), pp. 680–681.
12. Palais, R. S., "Natural Operations on Differential Forms," Trans. Amer. Math. Soc., *92*, No. 1 (1959), pp. 125–141.
13. Holzmann, C. A., and Liu, R., "On the Dynamical Equations of Nonlinear Networks with n-coupled elements," Proc. Third Ann. Allerton Conf. on Circuit and System Theory, University of Illinois, 1965, pp. 536–545.
14. Mitra, D., Sandberg, I. W., and Gopinath, B., "A Note on a Curious Property of the Equations of Nonlinear Networks Containing Transistors," to be published.
15. Muir, T., *A Treatise on the Theory of Determinants*, New York: Dover Publications, Inc., (1960), pp. 31–33.
16. Meyer, G. H., "On Solving Nonlinear Equations with a One-Parameter Operator Imbedding," Tech. Rep. 67–50, Comp. Science Center, University of Maryland, College Park, 1967.
17. Hadamard, J., "Sur Les Transformations Ponctuelles," Bull. Soc. Math. France, 48, (1920), pp. 13–27.
18. Goldstein, A. A., *Constructive Real Analysis*, New York: Harper and Row (1967), pp. 41–45.

Characterization of Second-Harmonic Effects in IMPATT Diodes

By C. A. BRACKETT

(Manuscript received May 20, 1970)

We discuss characterization of the tuned-harmonic mode of operation in IMPATT oscillators, and introduce an equivalent circuit which incorporates the large-signal, "single-frequency" oscillator admittances at the fundamental and second-harmonic frequencies. Complete characterization of this mode is equivalent to specifying the behavior of each of the four elements of the equivalent circuit as functions of the oscillation state variables: fundamental voltage and frequency, second-harmonic voltage and relative phase. Using the approximate large-signal analysis of Blue,[1] the values of the equivalent circuit elements are presented, as an example, for a 6-GHz IMPATT diode under a variety of oscillation conditions. This equivalent circuit is used to clarify the role played by the fundamental and second-harmonic, single-frequency oscillator admittances in the tuned-harmonic mode.

Using an approximation to the equivalent circuit, we investigate the criteria for stable oscillation of the tuned-harmonic mode. It is found that the stability criteria are in general quite restrictive. For the same 6-GHz germanium diode, the range of stable phase is investigated, as a function of the RF parameters, for certain special cases. It is found to be possible to satisfy the stability criteria for the phase which gives an optimum enhancement of the fundamental power output if certain conditions on the external RF circuit are satisfied.

I. INTRODUCTION

It was found by Swan[2] that the introduction of a trapped resonance at the second harmonic of the oscillation frequency in a 6-GHz Ge IMPATT diode oscillator provided dramatic increases in the output power and efficiency, as compared with the results obtained with the ordinary single quarter-wave transformer coaxial circuit. Since that time several authors[1,3-8] have reported both theoretical and experi-

mental examinations of the effect. It appears that the addition of a properly phased second-harmonic voltage improves the phasing of the RF current relative to the fundamental voltage so as to increase the negative conductance and (at least at lower frequencies) give an increase in the power output at the fundamental frequency. The circuit conditions required for the observation of this effect have been incompletely understood.

The purpose of this paper is to present the results of an analytical study of the interaction of an IMPATT diode with a circuit having resonances at two harmonically related frequencies. The analysis is begun by the introduction of an equivalent circuit for the diode by which these two-frequency oscillators may be characterized. A stability theory is then developed along the lines taken by Kurokawa which examines whether a particular circuit, even though matching the impedances required by the diode at both frequencies, will or will not provide a stable oscillation.[9,10] The stability theory is examined in some generality, and three special cases are studied for which tractable analytical results can be obtained. It is found that in the case of zero fundamental or second-harmonic voltage, the theory reduces to the single-frequency stability criteria derived by Kurokawa. In more general cases, the theory indicates that by designing (or adjusting) the circuit carefully one can obtain stable operation at phase angles which enhance the fundamental power. However, the theory also indicates that stable operation may be impossible if the circuit-diode interaction is not just right, even though the diode and circuit are matched to each other at the two frequencies.

In a final section, a numerical example is given in which the theory is applied to a model of a 6-GHz germanium IMPATT diode, using the approximate large-signal analysis of Blue.[1]

II. TWO-FREQUENCY CHARACTERIZATION

The IMPATT oscillator is truly a single-frequency oscillator only at very small ac voltages and currents. At larger signal levels the nonlinearity is very strong, and therefore there should be strong interactions between harmonically related signals. However, by operating the diode in a well-designed single-frequency circuit, the power output may be limited to a single frequency. This may be done, for example, by presenting short-circuit, open-circuit, or reactive loads at the harmonic frequencies. In the case of short circuited harmonics, the harmonic voltage amplitudes V_k are zero, and only the fundamental voltage V_1 is nonzero. It is then common practice to calculate a large-

signal diode admittance as a function of V_1 and to use this admittance to describe device behavior. On the other hand, for the case of open-circuited harmonics, the harmonic currents, I_k are zero, and only the fundamental current I_1 is nonzero. It is then preferable to characterize the diode by a large-signal impedance which is a function of the RF current amplitude I_1. Both of these conditions constitute tunings at the harmonic frequencies, albeit ones that are particularly useful and simple to express analytically.

To consider other, more general, loading conditions at the harmonic frequencies, one must introduce two more variables (amplitude and phase) for each additional frequency for which the amplitude is nonzero. One of the most important points is that the input admittance (for example) at the fundamental frequency is no longer a unique function of V_1 and the frequency f; but instead defining the state of oscillation requires a vector whose components are $V_1, \cdots, V_N, f, \varphi_2, \cdots, \varphi_N$ where N is the maximum harmonic number of interest and φ_k is the phase of the kth harmonic voltage relative to the fundamental. This vector does uniquely describe the state of oscillation, and for every such vector, there exists a set of complex admittances $y_1 \cdots y_N$ which are uniquely determined. If this is not so, it simply means we have inadequately described the system and must include more component signals, either harmonics or subharmonics.

We shall limit the discussion to include only two harmonically related frequencies and consider that $V_k = 0$ for $k > 2$. This also means that we will only discuss the admittance characterization and not the impedance characterization.

A convenient way of utilizing the information already known about the large-signal single-frequency admittance of the diode is to separate the input admittances at the two frequencies as shown in Fig. 1. This equivalent circuit shows a fundamental port and a second-harmonic

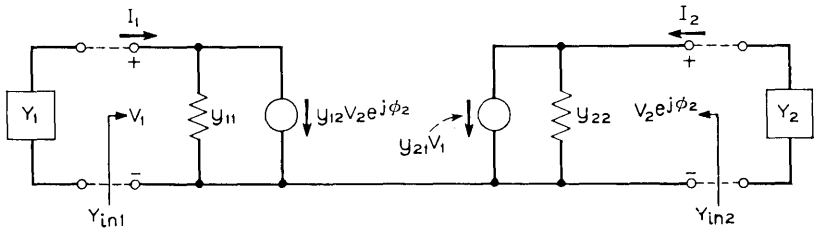


Fig. 1—Equivalent circuit of the IMPATT diode which includes nonzero voltages at two harmonically related frequencies. Port 1 is the fundamental port and port 2 is the second-harmonic port; y_{11} and y_{22} are the large-signal single-frequency diode admittances at the fundamental and second-harmonic frequencies, respectively.

port. The admittances $y_{11}(V_1)$ and $y_{22}(V_2)$ are the large-signal single-frequency admittances that would be measured at the fundamental if there were no harmonic (or subharmonic) voltages present. That is, they are just the ordinary large-signal admittances $y(V)$ at the frequencies f and $2f$.

The admittances $y_{12}(V_1, V_2, f, \varphi_2)$ and $y_{21}(V_1, V_2, f, \varphi_2)$ account for the conversion of current between the two frequencies and it is the study of their effects that is the main subject of this paper. The phase φ_2 is defined by the assumed voltage waveforms

$$v_1(t) = V_1 \cos \omega_0 t$$

and

$$v_2(t) = V_2 \cos (2\omega_0 t + \varphi_2).$$

The input admittances are

$$Y_{in1} = y_{11} + y_{12} \frac{V_2 \exp (j\varphi_2)}{V_1} \tag{1}$$

and

$$Y_{in2} = y_{22} + y_{21} \frac{V_1}{V_2 \exp (j\varphi_2)} \tag{2}$$

at the fundamental and second-harmonic frequencies respectively. Since y_{11} and y_{22} are independent of the phase φ_2 by definition, equations (1) and (2) show that the input admittance loci for fixed V_1 and V_2 will be counter rotating closed curves as a function of φ_2. These curves will enclose the admittance points y_{11} and y_{22} separately providing that y_{12} and y_{21} are not strong functions of φ_2. If, for example, y_{12} and y_{21} are independent of φ_2, Y_{in1} and Y_{in2} will be circles centered about y_{11} and y_{22} respectively, the radii of which depend upon the ratio V_2/V_1. They generally turn out to be somewhat elliptical in shape[8] although, in many cases, of very low eccentricity.

Figure 2 is the calculated[1] large-signal, single-frequency, admittance plane plot for a 6-GHz germanium diode, from which y_{11} and y_{22} may be obtained directly. Figures 3 and 4 show Y_{in1} and Y_{in2} for various fundamental frequencies when the voltages are held constant, demonstrating the elliptical and circular behavior noted above. Note that in Fig. 4 the second-harmonic input admittance has a positive real part for some ranges of the phase φ_2. To operate at such phase angles and RF voltages, the external circuit must supply power to the diode at the second-harmonic frequency, and thus these conditions are un-

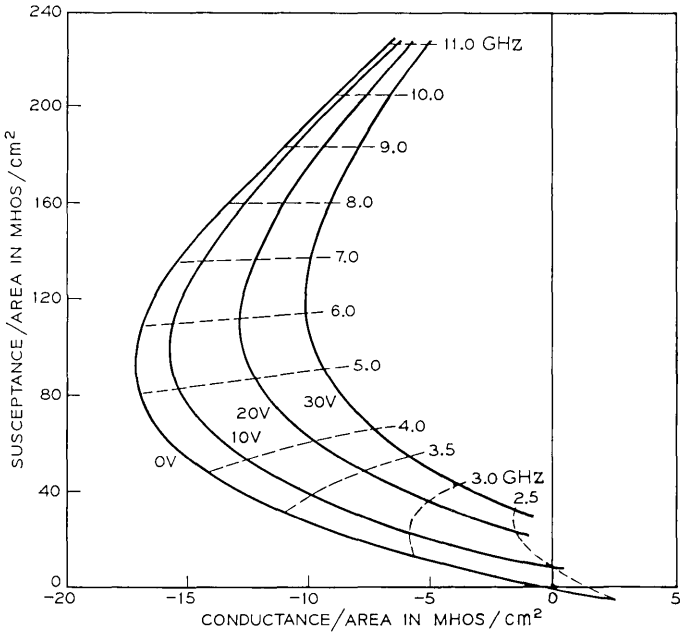


Fig. 2—The calculated large-signal single-frequency admittance of a 6-GHz germanium IMPATT diode at a bias current density $J_0 = 340$ A/cm².

realizable when operating into a passive circuit. The diameter of these admittance contours is inversely proportional to the second-harmonic voltage amplitude V_2, however, so that at higher values of V_2, the entire contour may lie in the left-half plane.

The rather simple structure of the Y_{in1} and Y_{in2} loci of Figs. 3 and 4 suggests that y_{12} and y_{21} might be rather insensitive functions of φ_2. This is borne out by the plots of Fig. 5 in which y_{12} and y_{21} are shown at constant fundamental voltage V_1 and several values of V_2, with φ_2 ranging $0 \leqq \varphi_2 \leqq 2\pi$. This figure also establishes that y_{12} and y_{21} do not change drastically as a function of V_2. It was also found that y_{12} and y_{21} depend upon V_1 in an approximately linear fashion. This is shown in Fig. 6 where y_{12}/V_1 and y_{21}/V_1 are plotted versus V_1 for several values of φ_2 with V_2 constant. Thus, for moderate values of V_1 and V_2, we can make the approximation that y_{12} and y_{21} are both proportional to V_1 and independent of φ_2 and V_2. To demonstrate this analytically, let the phase of the fundamental voltage $\varphi_1 \neq 0$, and consider a power series expansion of the currents $i_{12} = y_{12}V_2 \exp(j\varphi_2)$ and

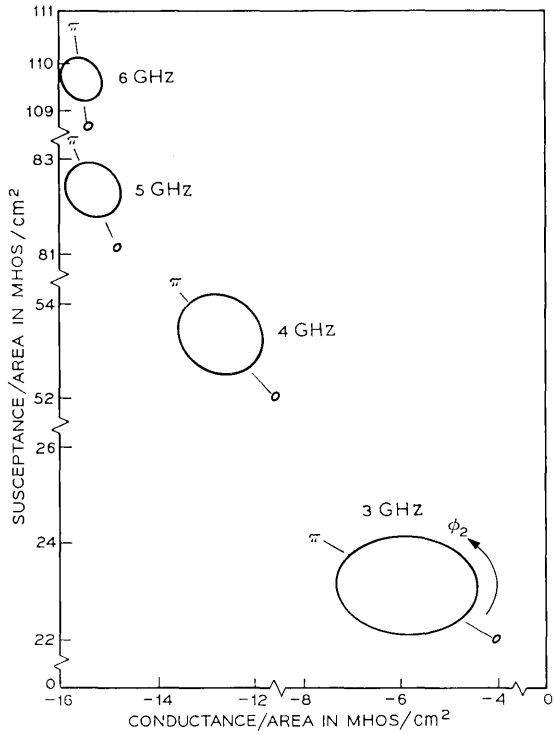


Fig. 3—The input admittance, Y_{in1}, at 3, 4, 5 and 6 GHz as it is modified by the presence of a second-harmonic voltage for $V_1 = 10$ volts, $V_2 = 1$ volt and $J_0 = 340$ A/cm².

$i_{21} = y_{21}V_1 \exp(j\varphi_1)$. Selecting the lowest-order terms having the appropriate frequencies, we find that

$$y_{12} \propto V_1 \exp(-j\varphi_1)$$

and (3)

$$y_{21} \propto V_1 \exp(j\varphi_1)$$

which confirms the approximate linear dependence on V_1 and gives the appropriate form of the φ_1 dependence. It will be convenient later to approximate y_{12} and y_{21} by the quantities

$$\begin{aligned} \bar{y}_{12} &= K_1 V_1 \exp(-j\varphi_1) = \kappa_1 V_1 \exp[-j(\varphi_1 - \psi_1)], \\ \bar{y}_{21} &= K_2 V_1 \exp(j\varphi_1) = \kappa_2 V_1 \exp[j(\varphi_1 + \psi_2)], \end{aligned} \tag{4}$$

where $\kappa_1 = |K_1|$, $\kappa_2 = |K_2|$, $\psi_1 = \arg(K_1)$ and $\psi_2 = \arg(K_2)$. Note that for $\varphi_1 = 0$ (only the phase $\varphi_2 - \varphi_1$ is important), $\psi_1 = \arg(y_{12})$ and $\psi_2 = \arg(y_{21})$ which is what will usually be assumed.

The quantities $\bar{y}_{12}$ and $\bar{y}_{21}$ may be defined as the average of y_{12} and y_{21} over the phase φ_2. For the 6-GHz oscillator example, the calculated values of $\bar{y}_{12}$ and $\bar{y}_{21}$ as a function of frequency are shown in Figs. 7 and 8 and the phases ψ_1, ψ_2 and $\psi_1 + \psi_2$ are shown in Fig. 9. Obviously these are only first-order approximations, but the complexity of the stability analysis requires some suitable approximation to obtain qualitative understanding.

The interaction of the diode equivalent circuit of Fig. 1 with an external circuit can be visualized by connecting an admittance Y_2 to the second-harmonic port. The fundamental input admittance is then

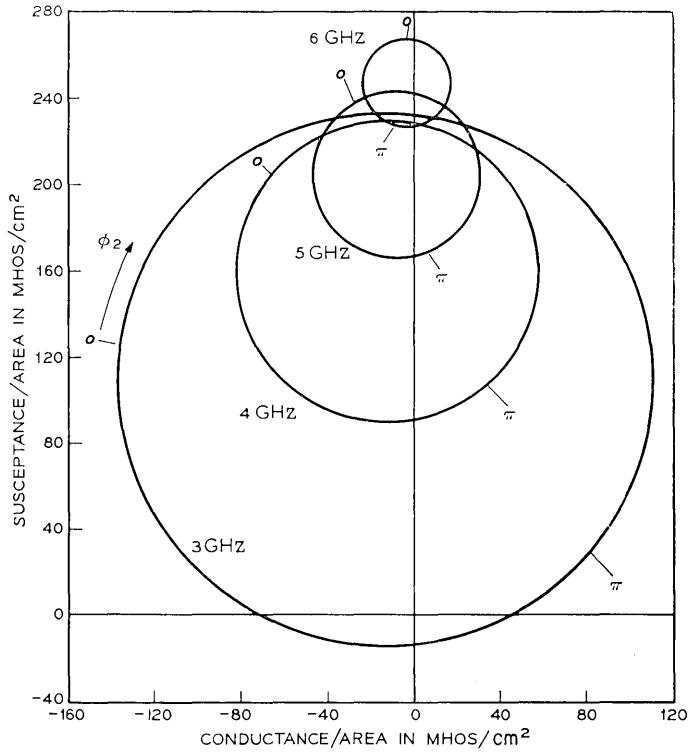


Fig. 4—The input admittance, Y_{in2}, at the second harmonic of 3, 4, 5 and 6 GHz for the same conditions as Fig. 3.

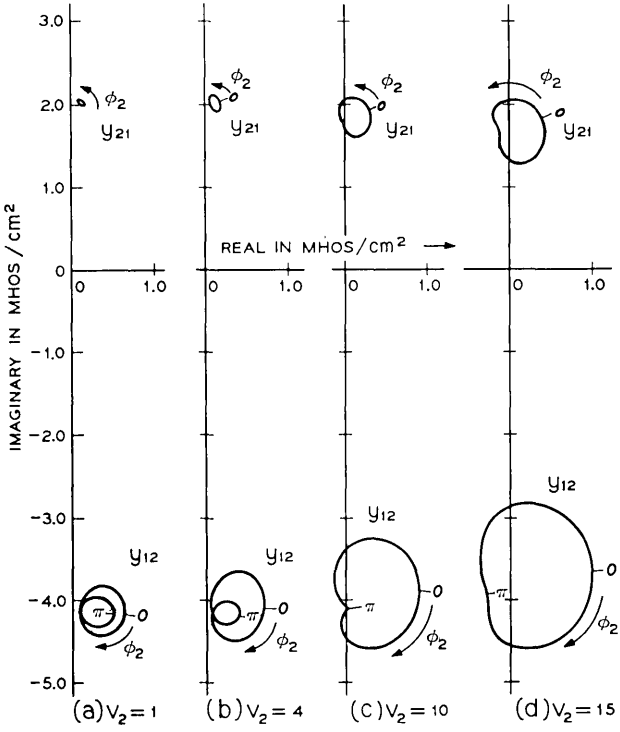


Fig. 5—Complex plane plot of y_{12} and y_{21} for $V_1 = 10$ volts and $V_2 = 2, 6, 10, 14$ volts at 6 GHz, showing the relative insensitivity of y_{12} and y_{21} to changes in V_2 and φ_2 for moderate values of V_2.

$$Y_{in1} = y_{11} - \frac{y_{12}y_{21}}{y_{22} + Y_2}. \tag{5}$$

Tuning the second harmonic by adjusting Y_2 provides the possibility of almost any input admittance Y_{in1}. In particular, $| Y_2 | = \infty$ gives the short-circuit termination and $Y_{in1} = y_{11}$. Equation (5) also predicts a pole in Y_{in1} at the frequency for which $y_{22} + Y_2 = 0$. This is not an ordinary pole as in linear circuit theory however for two reasons: (*i*) y_{22} may have a negative real part because it is an active device, and (*ii*) y_{22} is a function of V_2 so that the "pole" at $y_{22} + Y_2 = 0$ moves with changing V_2. This means that a resonance type of behavior should be observed, but that the only condition where $y_{22} + Y_2 = 0$ is for $V_1 \equiv 0$, which is just the single-frequency oscillator condition at $2f$.

III. STABILITY OF THE OSCILLATION STATE

Given an oscillation state which prescribes the admittances at the two frequencies, there are two requirements on the circuit that must be met in order that this be an obtainable state of steady oscillations. These are the requirements of circuit realizability and oscillation-state stability. The realizability criterion is simply that the required circuit have admittances whose real parts are greater than zero. The stability criterion is that any perturbation away from the given state will asymptotically return to the original state.

The stability problem has been recently discussed by Kurokawa[9,10] for the single-frequency negative-resistance oscillator. By following the approach used by Kurokawa and extending it to two-frequency interactions, the equations governing the stability of the harmonically tuned oscillator are derived in Appendix A. In this section, they are

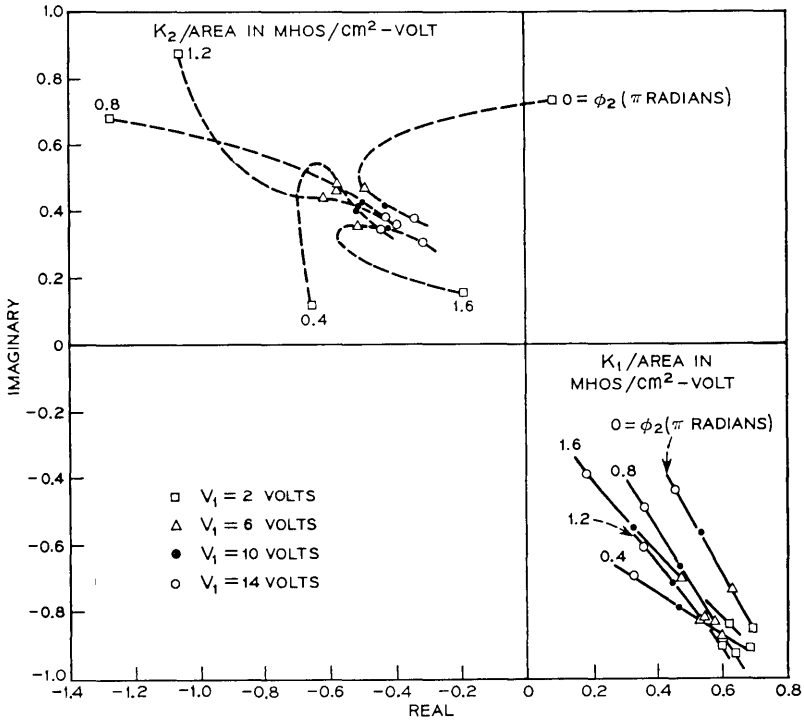


Fig. 6—Complex plane plot of $K_1 = y_{12}/V_1$ and $K_2 = y_{21}/V_1$ as a function of V_1 for various values of second-harmonic phase φ_2, at 4 GHz.

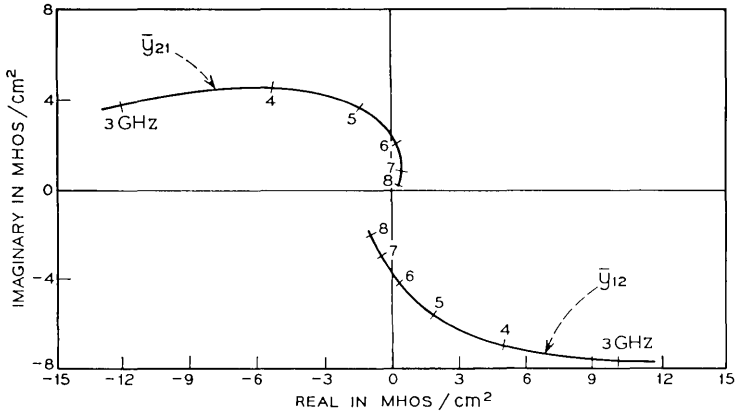


Fig. 7—Complex plane plot of $\bar{y}_{12}$ and $\bar{y}_{21}$ as a function of the fundamental frequency for $V_1 = 10$ volts and $V_2 = 1$ volt.

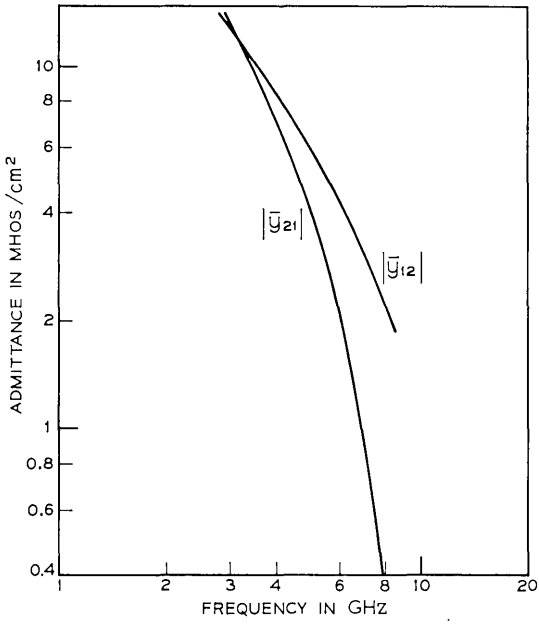


Fig. 8—$|\bar{y}_{12}|$ and $|\bar{y}_{21}|$ versus frequency for $V_1 = 10$ volts and $V_2 = 1$ volt.

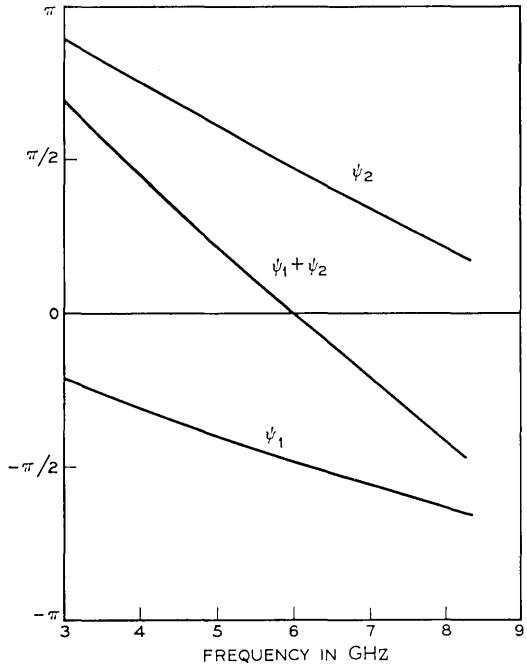


Fig. 9—The arguments ψ_1, ψ_2 and $\psi_1 + \psi_2$ versus frequency, showing a nearly linear dependence.

applied to several special cases, and theoretical examples of their use with the 6-GHz germanium oscillator model of Blue are given in Section IV.

In Appendix A, it is shown that the stability of an oscillation-state for small perturbations is determined by the solution of the system of equations

$$\frac{d\epsilon}{dt} + B\epsilon = 0 \tag{6}$$

where the vector ϵ is defined as

$$\epsilon = \begin{bmatrix} \delta a_1/V_1 \\ \delta a_2/V_2 \\ \delta(\varphi_2 - 2\varphi_1) \end{bmatrix} \tag{7}$$

and the matrix B is given by

$$B = \begin{bmatrix} \dfrac{\sqrt{r^2+s^2}\,G_{10}}{|Y_1'|}\sin(\alpha_1-\gamma_1) & \dfrac{\kappa_1 V_2}{|Y_1'|}\sin(\alpha_1+\theta_{10}) & -\dfrac{\kappa_1 V_2}{|Y_1'|}\cos(\alpha_1+\theta_{10}) \\ \dfrac{\kappa_2 V_1}{|Y_2'|}\sin(\alpha_2+\theta_{20}) & \dfrac{\sqrt{u^2+v^2}\,G_{20}}{|Y_2'|}\sin(\alpha_2-\gamma_2) & \dfrac{\kappa_2 V_1}{|Y_2'|}\cos(\alpha_2+\theta_{20}) \\ \left\{\begin{matrix}\dfrac{-2\sqrt{r^2+s^2}\,G_{10}}{|Y_1'|}\cos(\alpha_1-\gamma_1) \\ +\dfrac{\kappa_2 V_1}{|Y_2'|}\cos(\alpha_2+\theta_{20})\end{matrix}\right\} & \left\{\begin{matrix}\dfrac{\sqrt{u^2+v^2}\,G_{20}}{|Y_2'|}\cos(\alpha_2-\gamma_2) \\ -\dfrac{2\kappa_1 V_2}{|Y_1'|}\cos(\alpha_1+\theta_{10})\end{matrix}\right\} & \left\{\begin{matrix}-\dfrac{2\kappa_1 V_2}{|Y_1'|}\sin(\alpha_1+\theta_{10}) \\ -\dfrac{\kappa_2 V_1}{|Y_2'|}\sin(\alpha_2+\theta_{20})\end{matrix}\right\} \end{bmatrix}. \tag{8}$$

As discussed in the Appendix, δa_1, δa_2 and $\delta(\varphi_2 - 2\varphi_1)$ are the perturbations in the fundamental and second-harmonic voltage amplitudes and the relative phase, respectively. V_1 and V_2 are the unperturbed values of fundamental and second-harmonic voltage amplitudes.

The remaining quantities in the B matrix are defined as follows. The fundamental and second-harmonic external circuit admittances are $Y_1(\omega_0) = G_{10} + jB_{10}$ and $Y_2(2\omega_0) = G_{20} + jB_{20}$, respectively. The primes on Y_1 and Y_2 in equation (8) denote differentiation with respect to frequency at ω_0 and $2\omega_0$ respectively. κ_1 and κ_2 are defined in equation (4).

The saturation parameters s, r and u, v are defined by equations (55) through (58) in the Appendix. They relate to the nonlinear saturation of the diode's conductance and susceptance at the fundamental and second harmonic frequencies, respectively. The significance of s and r is shown schematically in Fig. 10, with u and v interpreted by a similar diagram for the second-harmonic admittance.

We have also introduced the angles α_1 and α_2 which give the slope on the complex plane of the circuit admittances at ω_0 and $2\omega_0$

$$\cos \alpha_1 = \frac{G'_{10}}{\sqrt{G'^2_{10} + B'^2_{10}}}, \qquad \sin \alpha_1 = \frac{B'_{10}}{\sqrt{G'^2_{10} + B'^2_{10}}} \tag{9}$$

$$\cos \alpha_2 = \frac{G'_{20}}{\sqrt{G'^2_{20} + B'^2_{20}}}, \qquad \sin \alpha_2 = \frac{B'_{20}}{\sqrt{G'^2_{20} + B'^2_{20}}} \tag{10}$$

and the angles γ_1 and γ_2 which measure the slope of the admittance curves $y_{11}(V_1)$ and $y_{22}(V_2)$;

$$\cos \gamma_1 = \frac{s}{\sqrt{r^2 + s^2}}, \qquad \sin \gamma_1 = \frac{r}{\sqrt{r^2 + s^2}}; \tag{11}$$

$$\cos \gamma_2 = \frac{u}{\sqrt{u^2 + v^2}}, \qquad \sin \gamma_2 = \frac{v}{\sqrt{u^2 + v^2}}. \tag{12}$$

Also, θ_{10} and θ_{20} are defined as in equations (48) and (49) of the Appendix but with the phase φ_1 set to zero. That is

$$\theta_{10} = -\varphi_2 - \psi_1$$

and (13)

$$\theta_{20} = \varphi_2 - \psi_2.$$

Note that

$$\theta_{10} + \theta_{20} = -\psi_1 - \psi_2. \tag{14}$$

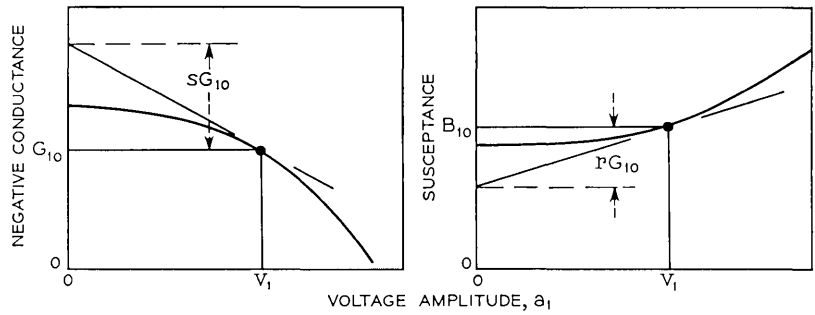


Fig. 10—Interpretation of the saturation parameters s and r for the fundamental admittance y_{11}. Similar definitions hold for u and v for the second-harmonic admittance y_{22}.

For the Ge oscillator considered here, the direct relationship between φ_2, θ_{20} and θ_{10}, as determined from equation (13), is shown in Fig. 11 for several frequencies.

The angles α_1, γ_1 and θ_{10} are shown in Fig. 12 which is a plot of the negative of an assumed circuit admittance $-Y_1(\omega)$ and the diode single-frequency admittance $y_{11}(V_1)$ in the neighborhood of the fundamental frequency. The point of intersection at ω_s gives the frequency and amplitude of the fundamental oscillation with zero second-harmonic voltage. As the voltage V_2 is increased by presenting an appropriate value of $Y_2(2\omega_0)$, the frequency will shift to some new value ω_0 generally accompanied by a change in voltage to V_1. This shows that the current injected into the fundamental circuit by the $y_{12}V_2 \exp(j\varphi_2)$ current source of Fig. 1 is just that sufficient to obtain the difference between the admittances $-Y_1(\omega_0)$ and $y_{11}(V_1)$. This additional admittance may be considered as a vector pointing from $y_{11}(V_1)$ to $-Y_1(\omega_0)$, and it is the angle θ_{10} measured clockwise about the $y_{11}(V_1)$ point that determines the orientation of this vector. Its length is given by $|y_{12}|\ V_2/V_1$. The angle α_1 gives the slope of the circuit curve at $-Y_1(\omega_0)$, and the angle γ_1 gives the relative change in reactive to real part of $y_{11}(V_1)$ with increasing voltage V_1 at the operating point. The angles α_2, γ_2 and θ_{20} may be defined in a similar manner in the second-harmonic admittance plane.

The solution of equation (6) subject to a small initial perturbation has a decreasing amplitude with increasing time if the eigenvalues of the stability matrix B all have real parts greater than zero. Suitable tests have been devised to determine this property.[11] The general case is difficult to do analytically and generally difficult to interpret if done

numerically because of the large number of parameters of the system. This is done however for the 6-GHz oscillator example given in Section IV, and the results are compared with the simplified results of this section.

In the remainder of this section, three special cases are examined which are severe approximations to the general case, but which yield interesting information. The first of these is that of a single-frequency oscillator, $V_2 = 0$. The second is the fictitious weak-coupling case which does not apply to the germanium diodes modeled here, but is included because of simplicity and for completeness. The third case is that of a strongly coupled small-signal approximation which gives qualitatively most of the features observed from the complete study

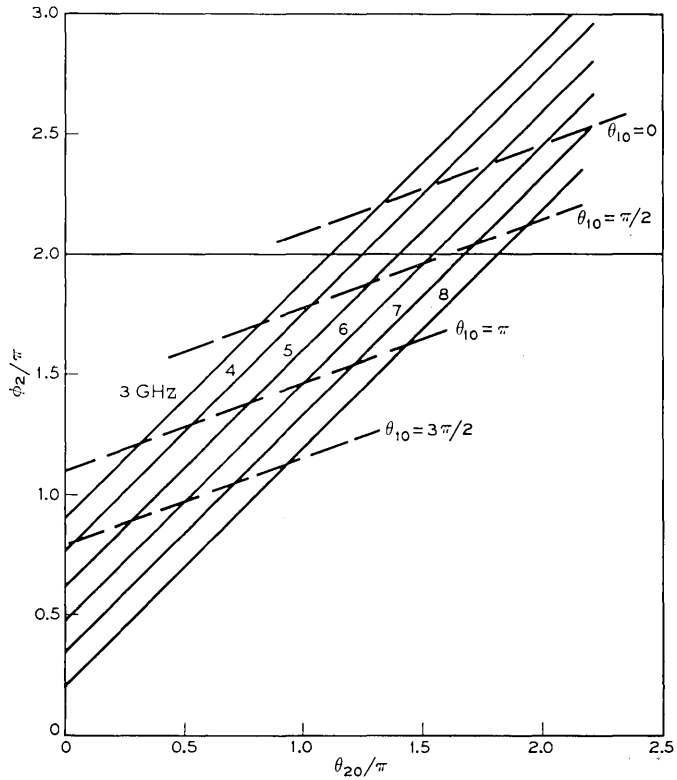


Fig. 11—Oscillator phase relations for the 6-GHz germanium example; φ_2 versus θ_{20} with loci of constant θ_{10} at 3, 4, 5, 6, 7 and 8 GHz.

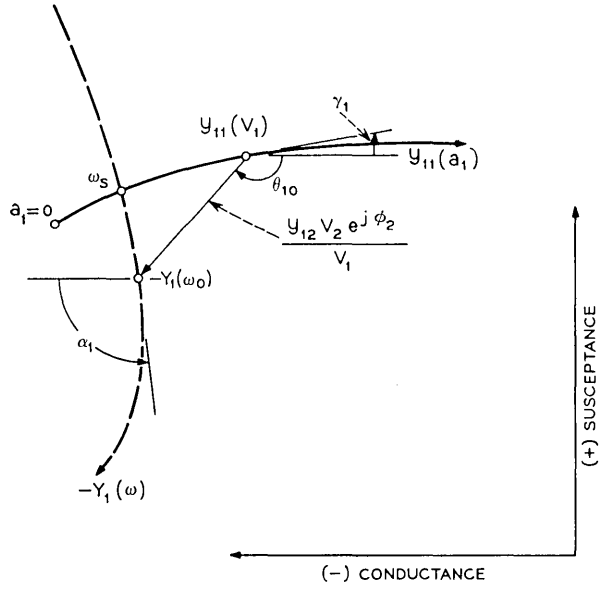


Fig. 12—An assumed fundamental admittance plane plot showing the angles α_1, γ_1 and θ_{10}. The device admittance is $y_{11}(a_1)$ and the negative of the circuit admittance is $-Y_1(\omega)$. A similar diagram defines α_2, γ_2 and θ_{20} in the neighborhood of $2\omega_0$.

of the eigenvalues of B, which is carried out in Section IV for the germanium diode case.

3.1 *Single-Frequency Limit*

In the very special case of $V_2 = 0$, only the first and third parts of equation (6) remain and they give the conditions

$$\sin(\alpha_1 - \gamma_1) > 0 \tag{15}$$

and

$$\frac{\kappa_2 V_1}{|Y_2'|} \sin(\alpha_2 + \theta_{20}) < 0. \tag{16}$$

These are simply the conditions required for stability of a single-frequency oscillator [equation (15)] with the added condition (16) due to the coupling to the harmonic. If the coupling to the harmonic, κ_2, is zero for $V_2 = 0$, equation (16) does not apply. Thus, for the single frequency oscillator with $V_2 = 0$, the familiar stability relation is recovered.[10]

3.2 *Weak-Coupling Limit*

For an oscillator having very small κ_1 and κ_2, the first two parts of equation (6) decouple. This gives

$$\sin(\alpha_j - \gamma_j) > 0 \qquad (j = 1, 2)$$

which are the single-frequency stability conditions at ω_0 and $2\omega_0$ for $j = 1$ and 2, respectively. The third equation then requires

$$\sin(\alpha_1 + \theta_{10}) + \mu \sin(\alpha_2 + \theta_{20}) < 0 \tag{17}$$

where the parameter μ is defined by

$$\mu = \frac{\kappa_2 V_1 \mid Y_1' \mid}{2\kappa_1 V_2 \mid Y_2' \mid}. \tag{18}$$

We may write equation (17) as

$$\sin(\varphi_2 + \xi) < 0 \tag{19}$$

where ξ is defined by the equations

$$\rho \sin \xi = -[\sin(\psi_1 - \alpha_1) + \mu \sin(\psi_2 - \alpha_2)] \tag{20}$$

and

$$\rho \cos \xi = -[\cos(\psi_1 - \alpha_1) - \mu \cos(\psi_2 - \alpha_2)]. \tag{21}$$

For a given pair of V_1, V_2 and for a fixed circuit, equation (19) thus gives the range of φ_2 for stable operation in the weak coupling limit.

3.3 *Small-Signal, Strong-Coupling Limit*

For very small signals the admittances y_{11} and y_{22} are independent of V_1 and V_2 so that $s \equiv r \equiv u \equiv v \equiv 0$ provides another approximation of some interest, providing that the coupling is still significant. In this limit, we obtain four constraints which are necessary and sufficient[11] to insure that the matrix B have positive eigenvalues. These are

$$k_1 = -\sin(\alpha_1 + \theta_{10}) - \mu \sin(\alpha_2 + \theta_{20}) > 0, \tag{22}$$

$$k_2 = -\sin(\alpha_1 + \theta_{10}) \cdot \sin(\alpha_2 + \theta_{20}) + 3 \cos(\alpha_1 + \theta_{10}) \cdot \cos(\alpha_2 + \theta_{20}) > 0, \tag{23}$$

$$k_3 = \sin(\alpha_2 + \theta_{20}) + \mu \sin(\alpha_1 + \theta_{10}) > 0, \tag{24}$$

$$k_4 = k_1 k_2 - k_3 > 0, \tag{25}$$

where μ is defined by equation (18).

The significance of this case is that for $\mu = 1$, conditions $k_1 > 0$ and $k_3 > 0$ are contradictory. This implies that $\mu = 1$ is a critical value and is indeed unstable, whereas for μ approaching zero or infinity stable states of oscillation do exist. These $\mu \ll 1$ and $\mu \gg 1$ stable states are exclusive of each other so that, as the conditions of oscillation are changed, if μ passes through the value unity a discontinuity in the oscillation will occur wherein the phase, the power and the frequency may all jump suddenly to new values.

To demonstrate the existence and exclusive nature of the $\mu \ll 1$ and $\mu \gg 1$ limits, consider equations (22) through (25) Note first of all that if a solution is obtained for a given value of μ, the solution for the reciprocal of that value of μ is obtained by interchanging the subscripts 1 and 2 on the angles α and θ. Thus, we need only consider the limit $\mu \ll 1$; the limit $\mu \gg 1$ being obtained from symmetry. For $\mu \ll 1$, equations (22) and (24) yield [Using equation (13)]

$$\pi - \psi_1 + \alpha_1 < \varphi_2 < 2\pi - \psi_1 + \alpha_1 \quad (k_1 > 0) \tag{26}$$

and

$$\psi_2 - \alpha_2 < \varphi_2 < \pi + \psi_2 - \alpha_2 \quad (k_3 > 0) \tag{27}$$

respectively. For purposes of illustration we consider $\alpha_1 = \alpha_2 = \bar{\alpha}$. Then the regions defined by equations (26) and (27) may be plotted in the φ_2, $\bar{\alpha}$ plane. From equation (25), if k_1, k_3 and k_4 are > 0, $k_2 > 0$ is automatically satisfied. Consider the constraint $k_4 > 0$, which may be written

$$-\cos(\alpha_1 + \theta_{10})[2 \sin(\alpha_1 + \theta_{10} + \alpha_2 + \theta_{20}) + \sin(\alpha_1 + \theta_{10} - \alpha_2 - \theta_{20})] > 0. \tag{28}$$

We see that $\cos(\alpha_1 + \theta_{10}) = 0$ is a critical condition, on either side of which the term in the brackets must also change sign. Thus, the lines

$$\varphi_2 = \psi_1 - \bar{\alpha} \pm \pi/2 \quad (k_4 = 0) \tag{29}$$

in the φ_2, $\bar{\alpha}$ plane are critical lines. Further, consider $\cos(\alpha_1 + \theta_{10}) > 0$, then

$$\sin(\psi_1 + \psi_2 - 2\bar{\alpha}) > -\sin(2\varphi_2 + \psi_1 - \psi_2)/2 \quad (k_4 > 0). \tag{30}$$

Equation (30) represents a curved boundary in the φ_2, $\bar{\alpha}$ plane and must be computed numerically. In Fig. 13 the regions bounded by equations (26), (27), (29) and (30) are plotted. The data used for this figure (ψ_1 and ψ_2) were taken from the Ge IMPATT example at a fre-

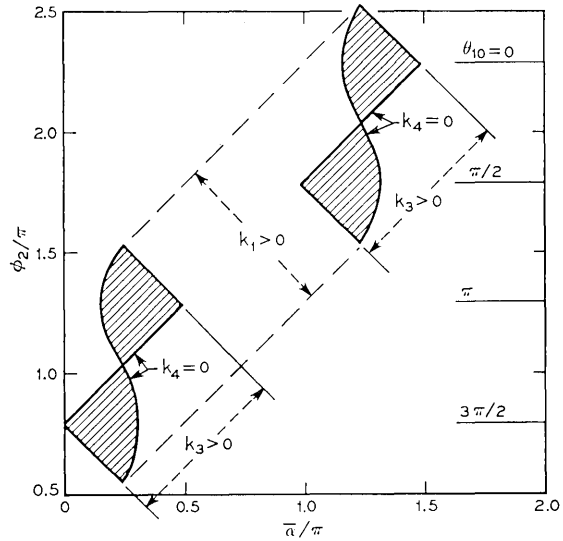


Fig. 13—Regions of stable φ_2 versus $\bar{\alpha}(\alpha_1 = \alpha_2 = \bar{\alpha})$ in the strongly coupled small-signal limit at 4 GHz; $\mu \ll 1$.

quency of 4 GHz from Fig. 9. Figure 13 shows that, for $\mu \ll 1$, there are two disjoint regions. Also indicated are the values of φ_2 for which $\theta_{10} = 0, \pi/2, \pi, 3\pi/2$. The angle θ_{10} (Fig. 12) measures the relative location of the diode's actual input conductance with respect to the single-frequency large-signal negative conductance, at the fundamental frequency. For $-\pi/2 < \theta_{10} < \pi/2$, $\cos \theta_{10}$ is positive and the input conductance is less negative than it would be for zero harmonic voltage. For this range of θ_{10} then, the fundamental output power is degraded by harmonic tuning. On the other hand, for $\pi/2 < \theta_{10} < 3\pi/2$, the input conductance is more negative than for $V_2 = 0$, and the fundamental output power is enhanced by the presence of harmonic tuning. These relationships can readily be seen by rewriting equation (1)

$$\text{Re}\,(Y_{in1}) = -g_1 + |\,y_{12}\,|\,\frac{V_2 \cos \theta_{10}}{V_1}.$$

Indeed, $\theta_{10} = \pi$ maximizes the fundamental output power for the particular values of V_1, V_2 being studied. We see that at 4 GHz, the maximum fundamental power point exists within a stable region for $\mu \ll 1$. It is also interesting that the minimum fundamental power phase ($\theta_{10} = 0$) is in a separate region which requires a considerably different circuit.

To obtain the similar diagram for $\mu \gg 1$, the same considerations can be reapplied to k_1 through k_4, or the subscripts on φ and α can be interchanged. Either way, Fig. 14 shows the result. Comparison of Figs. 13 and 14 shows indeed the disjointed, mutually exclusive behavior of the $\mu \ll 1$ and $\mu \gg 1$ regions of stability. Additionally, it shows that for a given circuit (i.e., a given $\bar{\alpha}$), there are two stable ranges of phase φ_2 (if any at all) depending on the value of μ relative to unity. One of these encompasses the $\theta_{10} = \pi$ maximum power phase and the other encompasses the $\theta_{10} = 0$ minimum power phase. A change in the bias current, which does not alter significantly the circuit variable $\bar{\alpha}$, may well change the relative value of μ from >1 to <1 or vice versa, and such a change would necessitate a change of phase to a different branch. Thus, which branch of the stability diagram the oscillation state is in is determined by the history of tuning and bias current changes. This type of behavior would be observed experimentally as a hysteresis in frequency or power or both, which if analyzed would indicate that the input admittance of the diode at the fundamental frequency is a nonunique function of the fundamental RF voltage. The presence of this effect would be indicated if one were able to obtain two different values of power output for the same frequency

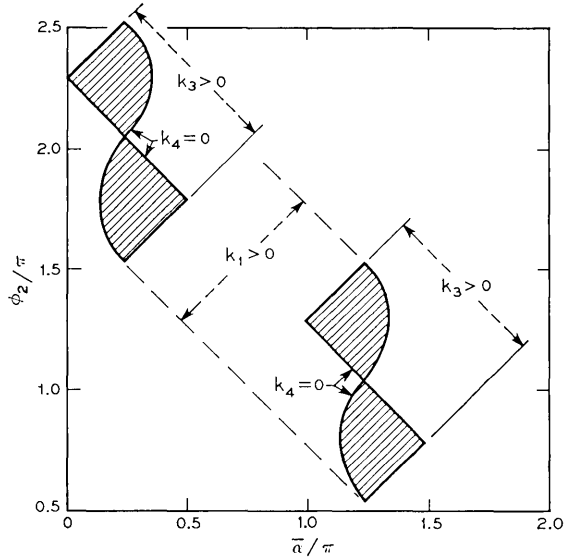


Fig. 14—Regions of stable φ_2 versus $\bar{\alpha}(\alpha_1 = \alpha_2 = \bar{\alpha})$ in the strongly coupled small-signal limit at 4 GHz; $\mu \gg 1$.

by changing the bias current only, without retuning the RF circuit in any respect. Observation at a single frequency is required in order to rule out the possibility of multiple-valued circuit admittances.[10]

In the next section, we compute the regions of stability for the germanium IMPATT example in full generality; that is, we use the complete form of the matrix B, equation (8). This must be done numerically so a limited number of cases can be examined, and the results are compared with the approximate forms of this section.

IV. 6-GHz Germanium Oscillator Example

Using Blue's approximate large-signal analysis,[1] the equivalent circuit parameters of Fig. 1 have been calculated for a germanium diode of depletion layer width 4.75 microns with an assumed avalanche zone width of 1.5 microns. This gives a critical field $E_c = 1.87 \times 10^5$ V/cm for a bias current density $J_0 = 340$ A/cm^2, which agrees quite well with the value obtained from a more exact numerical treatment. The design of this model was an attempt to model the germanium diodes reported by Swan[2] and by Gewartowski and Morris.[12] Because the Read theory is slightly incorrect in its reactive effects, the frequency of maximum negative conductance was at about 6 GHz for the model but appeared to be at about 8 or 9 GHz for the actual diodes. In comparing the results of this work with those of the experiments, it therefore seems most useful to discuss frequency relative to f_{max}, at which maximum output power is obtained. Thus, 4 GHz in this analytical work is roughly equivalent to 6 GHz in Swan's experiments. Table I lists the large-signal information obtained from Figs. 2, 8 and 9 that is needed for the solution of the stability constraints. This information was obtained for $V_1 = 10$ volts and $V_2 = 10$ volts, and a dc bias current density $J_0 = 340$ A/cm^2.

It is known that at resonance in a low-loss circuit where the real part of the admittance is constant or nearly so, the external Q can be written

$$Q_{ext} = \frac{\omega_0}{2G_0} \frac{dB}{d\omega} \bigg|_{\omega=\omega_0}$$

where G_0 is the real part of the admittance at ω_0 and B is the susceptance. Resonance is defined by the vanishing of $B(\omega_0)$. It is useful here to extend this definition to define the slope parameters

$$D_1 = \frac{\omega_0}{2G_{10}} \left| \frac{dY_1}{d\omega} \right|_{\omega=\omega_0}$$

TABLE I—DIODE LARGE-SIGNAL PARAMETERS AT $V_1 = V_2 = 10$ VOLTS

	3 GHz	4 GHz	5 GHz	6 GHz
g_{11} (mhos/cm²)	5.8	12.7	15.4	15.6
g_{22} (mhos/cm²)	15.6	12.7	8.5	4.3
$\frac{\partial g_1}{\partial V_1}$ (mhos/cm²-volt)	0.0	0.22	0.24	0.21
$\frac{\partial b_1}{\partial V_1}$ (mhos/cm²-volt)	1.1	0.65	0.30	0.20
$\frac{\partial g_2}{\partial V_2}$ (mhos/cm²-volt)	0.21	0.125	0.065	0.035
$\frac{\partial b_2}{\partial V_2}$ (mhos/cm²-volt)	0.20	0.0	0.0	0.0
ψ_1 (π radians)	−0.2089	−0.3056	−0.395	−0.477
ψ_2 (π radians)	0.9031	0.7742	0.6181	0.4798
κ_1 (mhos/cm²-volt)	1.25	0.86	0.59	0.42
κ_2 (mhos/cm²-volt)	1.25	0.70	0.385	0.205

at the fundamental frequency and

$$D_2 = \frac{\omega_0}{G_{20}} \left| \frac{dY_2}{d\omega} \right|_{\omega = 2\omega_0}$$

at the second harmonic. If, at $\omega = \omega_0$ and $\omega = 2\omega_0$, G'_{10} and G'_{20} vanish respectively, then D_1 and D_2 reduce to the external Q's of the circuit at these two frequencies, particularly since the major portion of the diode's susceptance is considered to be part of the external circuit.

Since, at an equilibrium point, from equations (44) and (46) of the Appendix

$$G_{10} = g_1 - \kappa_1 V_2 \cos \theta_1$$

and

$$G_{20} = g_2 - \kappa_2 V_1 \cos \theta_2 ,$$

specification of the parameters D_1 and D_2 permits the calculation of $| Y'_1 |$ and $| Y'_2 |$ from the information of Table I.

The general stability criteria for the matrix B are as follows: Let

B be represented

$$B = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}.$$

The condition that the eigenvalues of B all be positive implies that

$$k_1 = a + e + i > 0,$$

$$k_2 = ae + ei + ai - bd - fh - gc > 0,$$

$$k_3 = \det B > 0$$

and

$$k_4 = k_1 k_2 - k_3 > 0. \tag{31}$$

These conditions must be checked numerically, and the number of independent variables for a general study is quite large. In the calculations done here, the circuit variables have been restricted to $\alpha_1 = \alpha_2 = \bar{\alpha}$, with two sets of slope parameters; (*i*) $D_1 = 50$, $D_2 = 500$ and (*ii*) $D_1 = 50$, $D_2 = 10$. The restriction on α_1 and α_2 is quite artificial but allows comparison with the approximately determined regions of Section III. The two sets of slope parameters D_1, D_2 are an attempt to model (*i*) a high Q and (*ii*) a low Q second-harmonic circuit, respectively, and to thereby approximate the two conditions $\mu \ll 1$ and $\mu \gg 1$ for the same set of diode data.

The results of these calculations are shown in Figs. 15 and 16 for the frequencies 3, 4, 5 and 6 GHz. These show the values of stable second-harmonic phase φ_2 as functions of the circuit angles, $\alpha_1 = \alpha_2 = \bar{\alpha}$. These regions repeat themselves with a periodicity of 2π in both φ_2 and $\bar{\alpha}$. Only the principle branches are shown but it should be understood that wherever one of these regions extends across the boundaries chosen, it should be reflected back into the region at the opposite boundary. Figure 15 is for the case $D_1 = 50$, $D_2 = 500$, and corresponds to a value of $\mu \leq 0.4$ everywhere. Figure 16, for which $D_1 = 50$, $D_2 = 10$, corresponds to values of μ from near or slightly less than unity, to greater than 4 to 8 (the only exception is in Fig. 16a where one region appears having a value of $\mu \sim 0.02$). It should be noted that the value of $\mu \equiv 1$ is no longer a critical value, inasmuch as stable states may now exist for which $\mu = 1$. They do not appear to be large in number, however, and one may think of $\mu = 1$ as a transition value for which the area of the stable regions in the φ_2, $\bar{\alpha}$ plane becomes small.

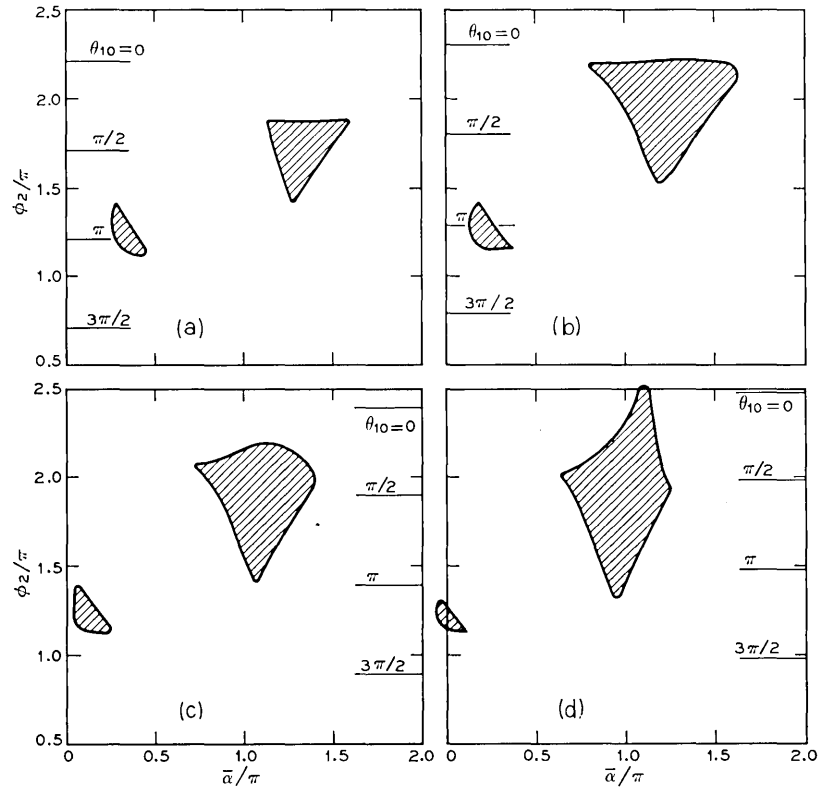


Fig. 15—Large-signal regions of stable φ_2 versus $\bar{\alpha}(\alpha_1 = \alpha_2 = \bar{\alpha})$ as obtained from the eigenvalues of the complete B matrix for the germanium oscillator example at (a) 3 GHz, (b) 4 GHz, (c) 5 GHz, and (d) 6 GHz; circuit variables $D_1 = 50$, $D_2 = 500$; diode variables $V_1 = V_2 = 10$ volts, $J_0 = 340$ A/cm^2. This figure has $\mu < 1$ everywhere.

Consider the 4-GHz results and compare Figs. 15b and 16b with Figs. 13 and 14. The locations of the stable regions in the φ_2, $\bar{\alpha}$ plane show a one-to-one correspondence but with greatly distorted shapes. It therefore appears that the strongly coupled small-signal approximation used in Figs. 13 and 14, together with the $\mu \ll 1$ and $\mu \gg 1$ cases, does give useful information about the general location of these stable regions for more realistic cases. The general properties of disjointedness and mutual exclusiveness are no longer strictly true (for example, there is some overlap of the regions centered at $\bar{\alpha} = \pi$ in Figs. 15d and 16d). However, it is easy to see that tuning discontinuities may still occur, and that the circuit angles $\bar{\alpha}$ must be considerably different

to obtain oscillation at $\theta_{10} = \pi$, for example, for the two different sets of values of slope parameters considered.

It is interesting that the angles α_1 and α_2 (and therefore, $\bar{\alpha}$) are equal to $\pi/2$ for simple shunt resonant circuits at both ω and 2ω, and that the stability diagrams show no cases of stable operation for this condition. Because of the approximations of this analysis, this cannot be construed to be a general conclusion, even for the diode modeled. It does show however, that such conditions may arise and that obtaining just the correct phase relations for maximum output power with a given circuit may be extremely difficult.

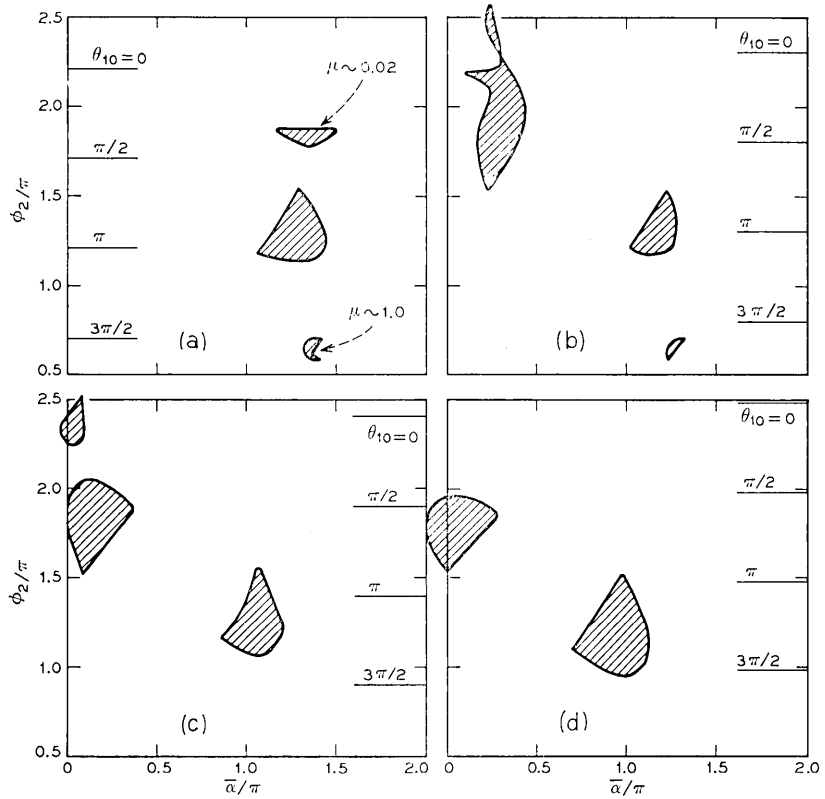


Fig. 16—Large-signal regions of stable φ_2 versus $\bar{\alpha}(\alpha_1 = \alpha_2 = \bar{\alpha})$ as obtained from the eigenvalues of the complete B matrix for the germanium oscillator example at (a) 3 GHz, (b) 4 GHz, (c) 5 GHz, and (d) 6 GHz; circuit variables $D_1 = 50$, $D_2 = 10$; diode variables $V_1 = V_2 = 10$ volts, $J_0 = 340$ A/cm^2. This figure has $\mu > 1$ everywhere except as noted.

Another observation is that the angle $\theta_{10} = \pi$ for maximum fundamental power output does have a stable realization in almost every case examined, even with the restriction $\alpha_1 = \alpha_2$.

If the points of operation along the circuit admittance curves $Y_1(\omega_0)$, $Y_2(2\omega_0)$ are near minima of their real parts, the angles α_1 and α_2 are restricted to lie in the range $0 < \alpha_j < \pi$, $j = 1, 2$. Such a limitation seems to imply different possibilities at the four frequencies calculated. At 3 GHz, stability is obtained in the neighborhood of $\theta_{10} = \pi$ and only for the $D_2 = 500$ case ($\mu < 1$). At 4 GHz, stability near $\theta_{10} = \pi$ is only obtained for the $D_2 = 500$ ($\mu < 1$) case, but there are additional stable states at or near $\theta_{10} = 0$ for both the $D_2 = 500$ ($\mu < 1$) and $D_2 = 10$ ($\mu > 1$) cases. Also, at 4 GHz, Fig. 16b shows a region which encompasses the $\theta_{10} = \pi/2$ point which is a crossover between enhanced and degraded fundamental power. The 5-GHz cases are very similar to those at 4 GHz except that there are more enhanced-power stable states for the $D_2 = 10$ ($\mu > 1$) case than at the lower frequencies. At 6 GHz, this shift is more advanced with roughly an equal number of stable states in the enhanced power region for the $D_2 = 10$ ($\mu > 1$) and $D_2 = 500(\mu < 1)$ cases.

V. SUMMARY AND CONCLUSIONS

An analysis of the stability of the tuned-harmonic mode in IMPATT oscillators has been presented using a simplified model of the frequency conversion in the avalanche diode. It has been shown that the stability constraints are generally quite restrictive and difficult to satisfy, particularly for diodes showing strong harmonic interactions. The goal of this work has not been to present a set of design curves which insure stable tuned-harmonic operation, but rather to consider the difficulties which the stability constraints present.

When the circuit restricts the voltage across the diode to be largely sinusoidal, this analysis reduces to that of the stability of a "single-frequency" oscillator. For nonzero fundamental and second-harmonic voltages V_1 and V_2, a characteristic parameter μ has been defined [equation (18)] which is dependent upon both diode and circuit characteristics and degree of excitation. The value of $\mu = 1$ appears to be somewhat critical in that the stable regions for $\mu > 1$ and $\mu < 1$ are usually separate. Any tuning or bias changes which force μ to pass through unity are very likely to produce sudden changes in the output variables, i.e., power and frequency. For example, the single-frequency oscillator is destined to have $\mu \gg 1$ because of the small value of V_2.

However, for equal V_1 and V_2 and $D_2/D_1 \sim 10$, $\mu < 0.4$. Thus the single-frequency oscillator and the tuned-harmonic oscillator (high Q, $2\omega_0$ circuit) are likely to operate in different regions of stability.

The numerical treatment of the stability criteria have been restricted to the case where the circuit angles α_1 and α_2 are equal. Thus the results presented here cannot be considered complete. However, in the example studied, it was found that at an operating frequency two-thirds the frequency of maximum output power, the phase φ_2 for maximum power is indeed stable and also corresponds to a realizable circuit. It was also found that it is possible to degrade the output power, and therefore, harmonic interactions when improperly adjusted can severely lower a diode's output power from that which would exist with no harmonic voltage at all.

As a necessary part of this instability analysis, a two-port model for the interaction was introduced and characterized for the 6-GHz germanium IMPATT model presented. This characterization illustrates the role of the second harmonic in introducing a "pseudo-pole" into the nonlinear admittance of the fundamental, and it clarifies the relevance of the single-frequency admittance plane characterization for the tuned-harmonic mode of operation.

This analysis also has assumed that y_{12} and y_{21} may be described by equation (4). If, on the other hand, y_{12} and y_{21} are assumed constant, then this analysis becomes identical with that of two nonlinear oscillators coupled through a linear circuit. That analysis can be carried through in the same manner as presented here. In such a case, the weakly coupled case becomes of considerable interest and has been treated by Schlosser.[13]

It is not necessary, of course, to introduce the two-port model of Fig. 1 at all, with its attendant assumptions and approximations, but it is possible to consider the perturbation of the oscillation-state directly from the numerical solution of the IMPATT equations. This would be a more accurate method to pursue; however, it is felt that the approach presented in this paper provides insight that might be obscured in a more complicated approach.

VI. ACKNOWLEDGMENTS

The author would like to acknowledge many helpful discussions with J. W. Gewartowski and K. Kurokawa in the development of this work, and the use of the approximate large-signal analysis computer program originally written by J. L. Blue.

APPENDIX A

Derivation of the Stability Matrix

In this appendix, the stability of the oscillation-state is considered using a linearized perturbation treatment about any general large-signal operating state. The result of this appendix is the derivation of the state-equation (6) and the stability matrix B, equation (8).

Consider a prescribed state of oscillation satisfying the two conditions

$$Y_1(\omega_0) + Y_{in1}(V_1 , V_2 , \varphi_1 , \varphi_2) = 0 \tag{32}$$

and

$$Y_2(2\omega_0) + Y_{in2}(V_1 , V_2 , \varphi_1 , \varphi_2) = 0, \tag{33}$$

where $Y_1(\omega_0)$ and $Y_2(2\omega_0)$ are the circuit admittances at ω_0 and $2\omega_0$ respectively. An approximation is made that the input admittances of the diode, Y_{in1} and Y_{in2}, are slowly varying functions of frequency as compared with the circuit admittances $Y_1(\omega_0)$ and $Y_2(2\omega_0)$. This is facilitated by considering the depletion layer capacitance, for example, to be a part of the external circuit. Generally speaking, equations (32) and (33) prescribe a functional dependence of ω, the frequency of oscillation, upon the voltage amplitudes and phases for small variations. For small variations in ω we can approximate

$$Y_1(\omega_0 + \delta_1) \approx Y_1(\omega_0) + \frac{dY_1}{d\omega}\bigg|_{\omega_0} \cdot \delta_1$$

and

$$Y_2(2\omega_0 + \delta_2) \approx Y_2(2\omega_0) + \frac{dY_2}{d\omega}\bigg|_{2\omega_0} \cdot \delta_2 .$$

The δ_k can be determined by allowing the voltage amplitudes and phases to be slowly varying functions of time

$$v_1(t) = a_1(t) \cos [\omega_0 t + \varphi_1(t)] \tag{34}$$

and

$$v_2(t) = a_2(t) \cos [2\omega_0 t + \varphi_2(t)]. \tag{35}$$

Differentiating with respect to time gives

$$\frac{dv_1}{dt} = \mathrm{Re}\left\{\left[j\omega_0 + j\frac{d\varphi_1}{dt} + \frac{1}{a_1}\frac{da_1}{dt}\right]a_1 \exp [j(\omega_0 t + \varphi_1)]\right\} \tag{36}$$

and

$$\frac{dv_2}{dt} = \mathrm{Re}\left\{\left[2j\omega_0 + j\frac{d\varphi_2}{dt} + \frac{1}{a_2}\frac{da_2}{dt}\right]a_2 \exp\left[j(2\omega_0 t + \varphi_2)\right]\right\}. \tag{37}$$

Thus, we can identify[10]

$$\delta_1 = \frac{d\varphi_1}{dt} - j\frac{1}{a_1}\frac{da_1}{dt}$$

and

$$\delta_2 = \frac{d\varphi_2}{dt} - j\frac{1}{a_2}\frac{da_2}{dt},$$

and therefore

$$Y_1(\omega_0 + \delta_1) \approx Y_1(\omega_0) + \frac{dY_1}{d\omega}\bigg|_{\omega_0} \cdot \left(\frac{d\varphi_1}{dt} - j\frac{1}{a_1}\frac{da_1}{dt}\right) \tag{38}$$

and

$$Y_2(2\omega_0 + \delta_2) \approx Y_2(2\omega_0) + \frac{dY_2}{d\omega}\bigg|_{2\omega_0} \cdot \left(\frac{d\varphi_2}{dt} - j\frac{1}{a_2}\frac{da_2}{dt}\right) \tag{39}$$

are the circuit admittances related to slow variations of the amplitudes and phases.

From the equivalent circuit of Fig. 1, the currents at the fundamental and second harmonic are

$$i_1(t) = \mathrm{Re}\left\{[y_{11}a_1 \exp(j\varphi_1) + y_{12}a_2 \exp(j\varphi_2)] \cdot \exp(j\omega_0 t)\right\}$$

and

$$i_2(t) = \mathrm{Re}\left\{[y_{21}a_1 \exp(j\varphi_1) + y_{22}a_2 \exp(j\varphi_2)] \cdot \exp(j2\omega_0 t)\right\},$$

which may be rewritten using the assumptions (4) as

$$\begin{aligned} i_1(t) = {} & [-g_1a_1 + \kappa_1 a_1 a_2 \cos(2\varphi_1 - \varphi_2 - \psi_1)] \cos(\omega_0 t + \varphi_1) \\ & + [-b_1a_1 + \kappa_1 a_1 a_2 \sin(2\varphi_1 - \varphi_2 - \psi_1)] \sin(\omega_0 t + \varphi_1) \end{aligned} \tag{40}$$

and

$$\begin{aligned} i_2(t) = {} & [-g_2a_2 + \kappa_2 a_1^2 \cos(\varphi_2 - 2\varphi_1 - \psi_2)] \cos(2\omega_0 t + \varphi_2) \\ & + [-b_2a_2 + \kappa_2 a_1^2 \sin(\varphi_2 - 2\varphi_1 - \psi_2)] \sin(2\omega_0 t + \varphi_2). \end{aligned} \tag{41}$$

Here we have introduced

$$y_{11} = -g_1 + jb_1$$

and

$$y_{22} = -g_2 + jb_2 .$$

Kirchoff's laws for the nonequilibrium case are

$$i_1(t) + \text{Re}\{Y_1(\omega_1)a_1 \exp(j\varphi_1) \exp(j\omega_0 t)\} = 0 \tag{42}$$

and

$$i_2(t) + \text{Re}\{Y_2(\omega_2)a_2 \exp(j\varphi_2) \exp(j2\omega_0 t)\} = 0, \tag{43}$$

where ω_1 and ω_2 are the perturbed fundamental and second-harmonic frequencies.

Equations (40) and (41) with (42) and (43) give the following four differential equations for the quantities $a_1(t)$, $a_2(t)$, $\varphi_1(t)$ and $\varphi_2(t)$

$$G_1 - g_1 + G_1' \frac{d\varphi_1}{dt} + B_1' \frac{1}{a_1}\frac{da_1}{dt} = -\kappa_1 a_2 \cos\theta_1 , \tag{44}$$

$$-(B_1 + b_1) - B_1' \frac{d\varphi_1}{dt} + G_1' \frac{1}{a_1}\frac{da_1}{dt} = -\kappa_1 a_2 \sin\theta_1 , \tag{45}$$

$$G_2 - g_2 + G_2' \frac{d\varphi_2}{dt} + B_2' \frac{1}{a_2}\frac{da_2}{dt} = -\kappa_2 a_1 \cos\theta_2 , \tag{46}$$

$$-(B_2 + b_2) - B_2' \frac{d\varphi_2}{dt} + G_2' \frac{1}{a_2}\frac{da_2}{dt} = -\kappa_2 a_1 \sin\theta_2 . \tag{47}$$

Here we have defined $Y_1 = G_1 + jB_1$, $Y_2 = G_2 + jB_2$ and the primes denote differentiation with respect to ω. Also

$$\theta_1 = 2\varphi_1 - \varphi_2 - \psi_1 \tag{48}$$

and

$$\theta_2 = \varphi_2 - 2\varphi_1 - \psi_2 . \tag{49}$$

Equations (44) through (47) may be rewritten so as to contain only a single time derivative in each

$$B_1'(G_1 - g_1) - G_1'(B_1 + b_1) + |Y_1'|^2 \frac{1}{a_1}\frac{da_1}{dt} = -\kappa_1 a_2 [B_1' \cos\theta_1 + G_1' \sin\theta_1], \tag{50}$$

$$G_1'(G_1 - g_1) + B_1'(B_1 + b_1) + |Y_1'|^2 \frac{d\varphi_1}{dt} = -\kappa_1 a_2 [G_1' \cos\theta_1 - B_1' \sin\theta_1], \tag{51}$$

$$B_2'(G_2 - g_2) - G_2'(B_2 + b_2) + |\ Y_2'\ |^2 \frac{1}{a_2}\frac{da_2}{dt} = -\kappa_2 a_1[B_2' \cos\theta_2 + G_2' \sin\theta_2], \quad (52)$$

$$G_2'(G_2 - g_2) + B_2'(B_2 + b_2) + |\ Y_2'\ |^2 \frac{d\varphi_2}{dt} = -\kappa_2 a_1[G_2' \cos\theta_2 - B_2' \sin\theta_2]. \quad (53)$$

Since φ_1 is an aribitrary quantity with no physical significance, it can be eliminated in favor of the difference phase $\varphi_2 - 2\varphi_1$ since this appears in both θ_1 and θ_2. This is done by multiplying equation (51) by $2/|\ Y_1'\ |^2$, equation (53) by $1/|\ Y_2'\ |^2$ and subtracting equation (51) from (53), giving

$$\frac{d}{dt}(\varphi_2 - 2\varphi_1) + \frac{G_2'(G_2 - g_2) + B_2'(B_2 + b_2)}{|\ Y_2'\ |^2} - 2\frac{G_1'(G_1 - g_1) + B_1'(B_1 + b_1)}{|\ Y_1'\ |^2} = -\kappa_2 a_1\left(\frac{G_2' \cos\theta_2 - B_2' \sin\theta_2}{|\ Y_2'\ |^2}\right) + 2\kappa_1 a_2\left(\frac{G_1' \cos\theta_1 - B_1' \sin\theta_1}{|\ Y_1'\ |^2}\right). \quad (54)$$

Equations (50), (52) and (54) form the set of differential equations for $a_1(t)$, $a_2(t)$ and $\varphi_2(t) - 2\varphi_1(t)$ which will be linearized for small perturbations around the oscillation state. These perturbations take the form

$$a_1 = V_1 + \delta a_1\ ,$$

$$a_2 = V_2 + \delta a_2\ ,$$

and

$$\varphi_2 - 2\varphi_1 = \varphi_{20} - 2\varphi_{10} + \delta(\varphi_2 - 2\varphi_1),$$

where V_1, V_2, φ_{10} and φ_{20} are the unperturbed values of $a_1(t)$, $a_2(t)$, $\varphi_1(t)$ and $\varphi_2(t)$. The perturbations in the voltage amplitudes will change g_1, b_1, g_2, b_2 away from their values $\bar{g}_1$, $\bar{b}_1$, $\bar{g}_2$, $\bar{b}_2$ which correspond to $\delta a_1 = \delta a_2 = \delta(\varphi_2 - 2\varphi_1) \equiv 0$. Thus, we define the saturation parameters s, r, u, v which describe the linearized variation of g_1 around $\bar{g}_1$, etc., by the equations (see Fig. 10)

$$s = \frac{V_1}{G_{10}}\frac{\delta(G_{10} - g_1)}{\delta a_1}\ , \quad (55)$$

$$r = \frac{V_1}{G_{10}} \frac{\delta(B_{10} + b_1)}{\delta a_1}, \tag{56}$$

$$u = \frac{V_2}{G_{20}} \frac{\delta(G_{20} - g_2)}{\delta a_2}, \tag{57}$$

and

$$v = \frac{V_2}{G_{20}} \frac{\delta(B_{20} + b_2)}{\delta a_2}, \tag{58}$$

where the zero subscript on the circuit variables denotes their evaluation at ω_0 or $2\omega_0$ as appropriate.

Equations (50), (52) and (54) may now be cast in a simple matrix form

$$\frac{d\epsilon}{dt} + B\epsilon = 0 \tag{59}$$

where the vector ϵ is defined as

$$\epsilon = \begin{bmatrix} \delta a_1/V_1 \\ \delta a_2/V_2 \\ \delta(\varphi_2 - 2\varphi_1) \end{bmatrix} \tag{60}$$

and the matrix B is given by equation (8) of Section III. Equation (59) indicates that the perturbations decay with time, giving a stable state of oscillation, if the eigenvalues of the matrix B are all positive.

APPENDIX B

List of Symbols

a_1 , a_2	Slowly varying amplitudes of the fundamental and second-harmonic voltages; equations (34) and (35).
B	Stability matrix; equation (8).
B_1 , B_2	Fundamental and second-harmonic external circuit susceptances; following equation (47).
b_1 , b_2	Imaginary parts of y_{11} and y_{22}, the susceptances of the single-frequency oscillator admittances; following equation (41).
D_1 , D_2	Fundamental and second-harmonic external circuit slope parameters; Section IV.
G_1 , G_2	Fundamental and second-harmonic external circuit conductances; following equation (47).

g_1 , g_2 Negative of the conductances of the single-frequency oscillator admittances; following equation (41).

K_1 , K_2 Complex normalized form of y_{12} and y_{21}; equation (4).

s, r Saturation parameters for the admittance y_{11}; equations (55) and (56).

u, v Saturation parameters for the admittance y_{22}; equations (57) and (58).

V_1 , V_2 Fundamental and second-harmonic voltage amplitudes; preceding equation (1).

Y_1 , Y_2 Fundamental and second-harmonic external circuit admittances; Fig. 1.

Y_{in1} , Y_{in2} Fundamental and second-harmonic IMPATT diode input admittances; equations (1) and (2) and Fig. 1.

y_{11} , y_{22} Fundamental and second-harmonic "single-frequency" oscillator admittances; Fig. 1.

y_{12} , y_{21} Conversion transfer admittances between fundamental and second harmonic; Fig. 1.

$\bar{y}_{12}$, $\bar{y}_{21}$ Approximate form of y_{12} and y_{21}; equation (4).

α_1 , α_2 Fundamental and second-harmonic circuit admittance slope angles; Fig. 12.

γ_1 , γ_2 Fundamental and second-harmonic single-frequency diode admittance slope angles; Fig. 12.

θ_1 , θ_2 phase variables; equations (48) and (49).

θ_{10} , θ_{20} θ_1 and θ_2 for $\varphi_1 \equiv 0$, equation (13).

κ_1 , κ_2 Magnitudes of K_1 and K_2 ; equation (4).

μ Stability parameter, equation (18).

φ_1 , φ_2 Fundamental and second-harmonic voltage phases; preceding equation (1).

ψ_1 , ψ_2 Arguments of K_1 and K_2 ; equation (4).

ω_0 Fundamental radian frequency.

REFERENCES

1. Blue, J. L., "Approximate Large-Signal Analysis of IMPATT Oscillators," B.S.T.J., *48*, No. 2 (February 1969), pp. 383–396.
2. Swan, C. B., "IMPATT Oscillator Performance Improvement with Second-Harmonic Tuning," Proc. IEEE (Letter), *56*, No. 9 (September 1968), pp. 1616–1617.
3. Lee, T. P., and Standley, R. D., "Frequency Modulation of a Millimeter-Wave IMPATT Diode Oscillator and Related Harmonic Generation Effects," B.S.T.J., *48*, No. 1 (January 1969), pp. 143–161.
4. Claassen, M., and Harth, W., "Analogue-Computer Model for an Avalanche-Diode Oscillator," Electronics Letters, *5*, No. 10 (May 15, 1969), pp. 218–219.
5. Giblin, R. A., Hambleton, K. G., and Tearle, C. A., "Octave Tuning and the Effect of Second-Harmonic Loading on Avalanche-Diode Oscillators," Electronics Letters, *5*, No. 16 (August 7, 1969), pp. 361–363.

6. Mouthaan, K., and Rijpert, H. P. M., "Second-Harmonic Tuning of the Avalanche Transit-Time Oscillator," Proc. IEEE (Letters), *57*, No. 8 (August 1969), pp. 1449–1450.
7. Mouthaan, K., "Nonlinear Analysis of the Avalanche Transit-Time Oscillator," IEEE Trans. Electron Devices, *ED-16*, No. 11 (November 1969), pp. 935–944.
8. Schroeder, W. E., Greiling, P. T., and Haddad, G. I., "Multifrequency Operation of IMPATT Diodes," Int. Electron Devices Meeting, Washington, D. C., October 31, 1969.
9. Kurokawa, K., "Noise in Synchronized Oscillators," IEEE Trans. Microwave Theory and Techniques, *MTT -16*, No. 4 (April 1968), pp. 234–240.
10. Kurokawa, K., "Some Basic Characteristics of Broadband Negative Resistance Oscillator Circuits," B.S.T.J., *48*, No. 6 (July–August 1968), pp. 1937–1956.
11. Bellman, *Introduction to Matrix Analysis*, New York: McGraw-Hill, 1960, p. 245.
12. Gewartowski, J. W., and Morris, J. E., "Active Diode Parameters Obtained by Computer Reduction of Experimental Data," IEEE Trans. Microwave Theory and Techniques, *MTT-18*, No. 3 (March 1970), pp. 157–161.
13. Schlosser, W. O., "Noise in Mutually Synchronized Oscillators," IEEE Trans. Microwave Theory and Techniques, *MTT-16*, No. 9 (September 1968), pp. 732–737.

An Analysis of Adaptive Retransmission Arrays in a Fading Environment

By Y. S. YEH

(Manuscript received December 3, 1969)

We analyze in this paper the performance of adaptive retransmission for improving two-way communication between antenna arrays in a randomly fading environment.

For a stationary environment, S. P. Morgan has shown that complex conjugate retransmission reaches a stable state and maximizes the signal-to-noise ratio of a maximal ratio diversity reception system. We show that a simpler system using phase conjugate retransmission will also stabilize and maximize the signal-to-noise ratio of an equal gain diversity reception system.

Where the fading is slow in comparison to the system settling-down time, both systems provide a significant improvement in transmission.

Subject to Rayleigh fading, we have obtained the average signal strength and its cumulative probability distribution for various combinations of numbers of antennas in the two arrays for each of the above mentioned systems. This information is useful in choosing an optimal division of diversity branches for the two antenna arrays. It is further observed that although the phase conjugate retransmission system is much simpler to implement, its performance is only slightly inferior to the corresponding complex conjugate system.

I. INTRODUCTION

Adaptive antenna arrays have been the subject of numerous investigations.[1–3] In an adaptive transmitting array, the individual element is excited according to information derived from the incident pilot field. For example, in a *complex conjugate* system, the excitation currents are proportional to the complex conjugate of the incident voltages while the total power radiated is kept constant. In a *phase conjugate* system, the currents are kept constant while the phases are adjusted according to the conjugate phase of the incident voltages.

In a free-space environment, that is, plane wave incident from a particular direction, it is well known that phase reversal would steer the radiated beam toward the source antenna. Cutler and others[2] have shown how phase reversal can be achieved by frequency conversion of the pilot signal.

The role of adaptive retransmission in a multipath fading environment, for example, mobile radio, tropscatter communication, and so on, has received far less attention. Still unanswered is the question of whether the phase conjugate or the complex conjugate retransmission schemes could improve the communication link and reach a stable state. In his work, S. P. Morgan has shown that, in a stationary arbitrary environment, stable state and maximal power transfer can be achieved by complex conjugate retransmission.[3]

In this paper, we show that the much simpler phase conjugate system will also reach a stable state. Furthermore, assuming equal amplitude transmitting currents on the antenna elements, the summation of voltages received at one array is equal to that of the other array and is maximized. Consequently, the phase conjugate retransmission system will maximize the signal-to-noise ratio (S/N) of an equal gain diversity reception system.[4]

In general, the fundamental differences of the two retransmission schemes are that the phase conjugate retransmission maximizes the sum of the amplitudes of the voltages received and the complex conjugate retransmission maximizes the total power received.

Where fading is slow in comparison to the time required to reach an equilibrium state, both systems could be used to improve the quality of a fading communication link.

We investigate the performance of these two systems in actual fading environments. In particular, we want to know how these two systems differ in average S/N, what the S/N probability distributions are, how much they improve fading statistics over a single branch system and, finally, what the optimal division of number of antennas would be between the two antenna arrays.

In order to answer these questions, we must first establish the characteristics of the medium which links the two antenna arrays. For example, in a mobile radio the signal received by a single antenna is rapid varying and can be characterized by Rayleigh statistics over distances of a few hundred wavelengths.[5] However, over an extended range of observations, other large-scale phenomena such as distance variations, shadowing, and channeling by streets will produce slow variations of the average signal strength received. The adaptive

retransmission system per se can reduce the rapid fluctuations but will be of little help in reducing those long-term variations. Consequently, the comparison of the performance of adaptive retransmission arrays will be based on their relative effectiveness in reducing the rapid Rayleigh fading.

The Rayleigh fading is also an excellent approximation in other communication systems such as long-range UHF and SHF tropospheric transmission,[4] and so on. Furthermore, results obtained from Rayleigh fading can give significant insight into the performance of adaptive antenna arrays under other fading conditions.

Based on Rayleigh fading statistics, we investigated the cumulative probability distribution (CPD) of the signal strength of an $m{:}n$ array system. By $m{:}n$ we mean that there are m antennas at station 1 and n antennas at station 2. The analysis is done by the Monte Carlo method on a digital computer. The 99 percent reliability level* as well as the average signal strength for a unity transmitter power are obtained. It is interesting to note that with the help of interpolation, in most cases, only 96 computer samples are sufficient to yield a CPD which is accurate up to a few tenths of a dB for all the information we need.

The average S/N of the two retransmission schemes are compared. It is observed that although the phase conjugate system is much simpler to build, it is only slightly inferior to the complex conjugate retransmission system.

For other types of fading distributions, the techniques described here can readily be applied.

II. ANALYSIS OF THE PHASE CONJUGATE RETRANSMISSION

The configuration of the arrays is depicted in Fig. 1. The open circuit voltages and the transmitting currents in each array are represented by column vectors with the time factor exp ($j\omega t$) suppressed. The mutual couplings are neglected and the antennas in each array are assumed to be identical, with input resistance R during transmission and admittance G during reception.

The transmitting current vector I_2 at array 2 produces the received voltage vector at array 1,

$$V_1 = C\Gamma I_2 \tag{1}$$

where Γ is an $m \times n$ matrix whose elements are proportional to the

* The 99 percent reliability level is defined such that for 99 percent of the time the signal strength is above this level.

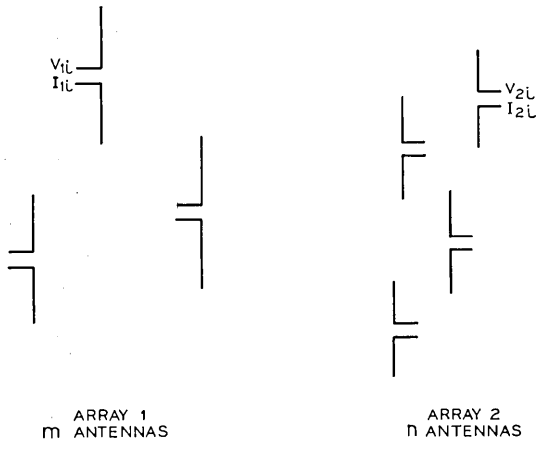


Fig. 1—Arrays in adaptive retransmission system.

transmission between a particular pair of antennas. The real constant C stands for the average transmission loss.

By reciprocity, the received voltage at array 2 is,

$$V_2 = C\Gamma^t I_1 \tag{2}$$

where the superscript t stands for the transpose of the Γ matrix.

Here according to our definition of phase conjugate retransmission, the elements of I_1 and I_2 are of unity amplitudes although their phases could be different. Multiplying equations (1) and (2) by I_1 and I_2, respectively, we obtain the following

$$\langle V_1 , I_1 \rangle = C\langle \Gamma I_2 , I_1 \rangle, \tag{3}$$

$$\langle V_2 , I_2 \rangle = C\langle \Gamma^t I_1 , I_2 \rangle \tag{4}$$

where the brackets $\langle\ \rangle$ stand for inner product. Equations (3) and (4) are equal, and we obtain the following reciprocity relation

$$\langle V_1 , I_1 \rangle = \langle V_2 , I_2 \rangle. \tag{5}$$

2.1 *Stabilization of the Phase Conjugate Retransmission System*

Let array 1 be excited initially with current I_1 which produces V_2 at array 2. And let array 2 be excited with I_2 which produces V_1 at array 1. Equation (5) holds and we have the following

$$\sum_{i=1}^{m} V_{1i} I_{1i} = \sum_{i=1}^{n} V_{2i} I_{2i} \tag{6}$$

where the subscript i stands for the ith element of the array.

Consider now the excitation at array 2. Since the I_{2i}'s are of unity amplitude, the quantity $\sum_{i=1}^{n} V_{2i}I_{2i}$ can be maximized by choosing I'_{2i} to be phase conjugate to V_{2i} . We shall call this real maximum quantity λ. Let V'_1 be the voltage vector produced by I'_{2i} ; then we have

$$\sum_{i=1}^{m} V'_{1i}I_{1i} = \sum_{i=1}^{n} V_{2i}I'_{2i} = \sum_{i=1}^{n} |V_{2i}| = \lambda. \tag{7}$$

Let us now consider the excitation of array 1. Obviously the quantity $\sum_{i=1}^{m} V'_{1i}I_{1i}$ can be maximized if we choose I'_{1i} to be the phase conjugate of V'_{1i} . It then follows that

$$\sum_{i=1}^{m} V'_{1i}I'_{1i} = \sum_{i=1}^{m} |V'_{1i}| = \lambda' \geqq \lambda. \tag{8}$$

Let V'_{2i} be the voltages produced by I'_{1i} . We obtain, by applying equation (6), the following,

$$\sum_{i=1}^{m} V'_{1i}I'_{1i} = \sum_{i=1}^{n} V'_{2i}I'_{2i} = \lambda' \geqq \lambda. \tag{9}$$

Now I''_{2i} can again be chosen to be phase conjugate to V'_{2i} and we obtain

$$\sum_{i=1}^{m} V'_{2i}I''_{2i} = \sum_{i=1}^{n} |V'_{2i}| = \lambda'' \geqq \lambda' \geqq \lambda. \tag{10}$$

This process continues with each new choice of I representing the actual retransmission adjustment made by the antenna system. It is obvious from equation (10) that each retransmission yields a new value of λ which is real and bigger than or equal to the previous value. However, because of the finite number of antennas involved, λ cannot increase indefinitely. The iteration process must therefore finally settle down to a value λ_f which no longer changes. If this is so, we have

$$\sum_{i=1}^{m} V^f_{1i}I^f_{1i} = \sum_{i=1}^{n} V^f_{2i}I^f_{2i} = \lambda_f. \tag{11}$$

The fact that λ_f is real, and also that we cannot vary the phase of I^f_{2i} and I^f_{1i} to make λ_f larger automatically guarantees that I^f_{1i} and I^f_{2i} are phase conjugate to V^f_{1i} and V^f_{2i} , respectively. In this case, our phase conjugate retransmission apparatus will no longer change the phases of I^f_{1i} and I^f_{2i} because they have already reached their proper value. Therefore, we have arrived at a stable state. In this case equation (11) can be further simplified to

$$\sum_{i=1}^{m} |V^f_{1i}| = \sum_{i=1}^{n} |V^f_{2i}| = \lambda_f. \tag{12}$$

So far we have demonstrated that each retransmission tends to increase λ and a stable state must finally be reached. It still remains to be shown that this stable state yields the absolute maximum λ. It is quite possible that several pairs of I_1 and I_2 exist such that they are phase conjugate to V_1 and V_2 but their corresponding λ_f's are different. This is similar to the existence of different eigenstates in matrix analysis. As is well known in matrix algebra, unless the initial vector is orthogonal to the maximum eigenstate, we would invariably obtain the maximum eigenstate through iterations.

Since the phase conjugate operation on V to produce I is a nonlinear operation, an analytical analysis along the above lines is extremely difficult, if not impossible. However, in the next section we show with computer simulation that the phase conjugate retransmission process converges rapidly and the probability of ending up in a nonmaximum state of λ_f is practically zero.

2.2 *Computer Simulation*

The convergence test was done by choosing a 3 : 4 array system as a particular trial case. We started by arbitrarily choosing a Γ matrix, which was defined by $\Gamma_{IJ} = I/1.2 + J/2 - 1 + j[I/2.3 + 2 - J/1.2]$. The initial values of I_1 were chosen such that,

$$I_1 = [1, \exp(j\theta), \exp(j\phi)]. \tag{13}$$

The phase angles θ and ϕ were allowed to run through 0 to 2π, in 10 equal steps. Therefore, we had 100 different initial trial values of I_1. For each initial set of I_1, we calculated V_2 produced and formed I_2 which produced V_1. I_1 was then readjusted according to the V_1 just produced. In each retransmission, we also computed the quantity λ. It was observed that in all these one hundred trials, the currents and λ approached their specific final values within a few retransmissions. For this particular choice of Γ, $\lambda_f = 31.3719$. The first value of λ obtained, that is, $\sum_{i=1}^{4} |V_{2i}|$, was always smaller than λ_f but after the first retransmission, it invariably came very close to λ_f. For example, in one case the first λ was 10.72; after retransmission at array 2 we obtained a λ of 30.73 at array 1. After this array retransmitted back to array 2, the value agreed with λ_f to the fourth decimal place.

Next we tried to determine if λ_f is the absolute maximum. In other words, we wanted to check if λ_f was bigger than the λ, that is, $\sum_{i=1}^{4} V_{2i}$, produced by any arbitrary I_1. This survey was done by varying θ and ϕ in 50 steps from 0 to 2π. Computation indicated that all the 2500 values of λ produced were smaller than λ_f and that λ_f was indeed the real maximum.

A similar test was performed on a 4 : 5 array system and we obtained similar results as reported for the 3 : 4 system. In the 4 : 5 array system, the Γ_{IJ} were defined as $(I - J)/3 + I^2J/6 - 5 + j[(I - I^2 + J)/1.4 + 3.5]$.

III. SIGNAL-TO-NOISE RATIO

Let V_{1i} be the voltage response at the ith elementary antenna. Furthermore, let η_{1i} be the corresponding noise voltage which satisfies,

$$\langle \eta_{1i}\eta_{1j} \rangle_{\text{av}} = \begin{cases} N^2 & i = j, \\ 0 & i \neq j \end{cases} \tag{14}$$

where the $\langle\ \rangle_{\text{av}}$ stand for time average.

3.1 *S/N of Phase Conjugate System Using Equal Gain Diversity Combining Technique*

The S/N of an m-branch diversity equal gain system is,

$$\text{S/N} = \left[\sum_{i=1}^{m} |V_{1i}|\right]^2 \Big/ mN^2 = \lambda_f^2/mN^2. \tag{15}$$

Recall that there are n elements at the other array, which radiates a total power to the amount of nR, therefore the S/N of the received signal per unit power radiated is,

$$\text{S/N} = \lambda_f^2/nmN^2R. \tag{16}$$

It is therefore obvious that the S/Ns at both arrays are identical.

3.2 *S/N of Complex Conjugate System Using Maximal Ratio Diversity Combining Technique*

The excitation currents of a complex conjugate retransmission system are related to the incoming voltages by,

$$I_2 = K_2V_2^*, \tag{17}$$

$$I_1 = K_1V_1^* \tag{18}$$

where K_1 and K_2 are scalars to keep the total radiated power constant. For unity transmitter power, the received power at arrays 1 and 2 are maximized and are equal,[3]

$$P_{1R} = P_{2R} = \frac{G}{R} C^2\lambda_m \tag{19}$$

where λ_m is the maximum eigenvalue of the hermitian matrix $\Gamma\Gamma^+$.

The validity of equation (19) is subject to the constraint that when the adaptive retransmission array starts operation, its current vector should not be orthogonal to the maximum eigenvector of the $\Gamma\Gamma^*$ matrix. The S/N of a multibranch maximal ratio reception system then is,

$$\mathrm{S/N} = \frac{C^2}{\mathrm{RN}^2}\lambda_m. \tag{20}$$

It can be seen that the S/Ns at both arrays are equal.

IV. EVALUATION OF THE CUMULATIVE PROBABILITY DISTRIBUTION

The complexity of the quantities λ_m and λ_f makes a closed form solution of the CPD extremely difficult, if not impossible. Therefore, we try instead the Monte Carlo method and aim at a numerical solution. The essence of the method is to choose for each element of the Γ matrix a random variable of the form $u + jv$. The variables u and v, according to our assumption of independent Rayleigh fading statistics, are normalized independent gaussian variables. For a particular $m{:}n$ array system, we can therefore evaluate the maximum eigenvalue λ_m by repeated matrix multiplication.[6] The value λ_f is evaluated by iterations according to the retransmission schemes defined in Section 2.2.

The computed values of λ_m *and* λ_f are stored. Then we start the whole process again by choosing elements for another Γ matrix and evaluate the corresponding λ_m and λ_f. The CPD curves are developed after a sufficient number of calculations.

Two tests of convergence are made. The first is the comparison of the calculated CPD curves of variables $| u_1 + jv_1 |^2$ or $| u_1 + jv_1 |^2 + | u_2 + jv_2 |^2$ to that of the known theoretical curves. It is understood here that u's and v's refer to independent normalized gaussian random variables. Hence, these curves represent respectively the CPD of maximal reception of single or two-channel Rayleigh signals.[4]

The results are presented in Fig. 2. A close look at Fig. 2 indicates that as far as the 99 percent reliability and the average signal levels are concerned, 900 sample points are sufficient for a single Rayleigh and 300 sample points for two Rayleighs.

A second test is made on the 2:2 and 2:4 antenna system and is shown in Fig. 3. The dB scale is chosen such that the average S/N of a single Rayleigh variable, that is, the received S/N of a 1:1 array system, is at 0 dB. It is observed that 96 samples are already sufficient

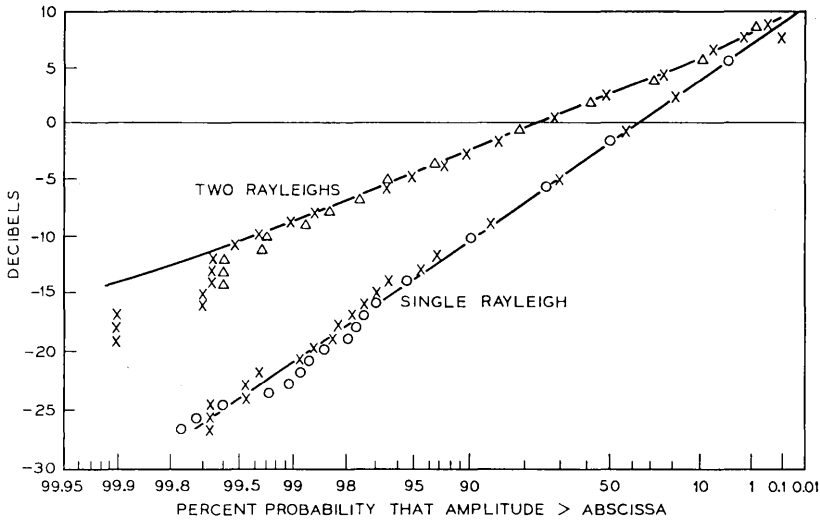


Fig. 2—Comparison of Monte Carlo method and theoretical calculation, △, 300 samples; x, 900 samples; o, 1800 samples; ————, theoretical curve.

to yield what we want since these points lie very close to the curve drawn through the points computed from 900 samples. With the required sample points greatly reduced to this number, it is possible to make a fast and inexpensive check of an extensive combination of m:n arrays.

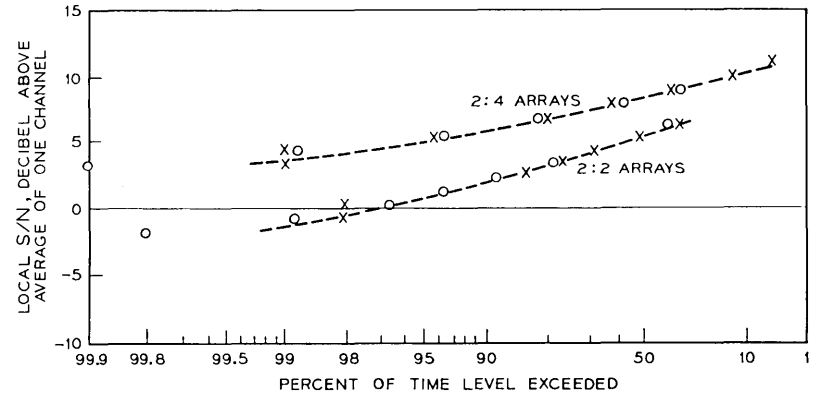


Fig. 3—Comparison of 96 and 960 samples. Complex conjugate retransmission maximal ratio reception. o, 960 points; x, 96 points; — — —, curve fitted to 960 points.

V. DISCUSSION OF NUMERICAL RESULTS

We look at the complex conjugate retransmission system first. Incorporated with maximal ratio diversity reception, this system provides the best S/N performance obtainable from a particular $m:n$ array system.

The average S/N is presented in Fig. 4. It is seen that for small numbers of n, there do exist appreciable improvements in average signal level as m changes from 1 to 4. However, as n increases the advantage diminishes. For example, a 1:50 array has the same average signal level as 2:44, 3:39, and 4:35 arrays. This is in sharp contrast to the case of adaptive arrays with nonfading signals. In that case, plane wave incidence is assumed and an $m:n$ array would have the same S/N as a $1:mn$ array (Fig. 4).

A simple explanation of the difference between the fading and the nonfading arrays is the following: In both cases, the $1:mn$ adaptive retransmission system guarantees that the voltages produced by the mn elements at the single array add in phase. In the $m:n$ system, the

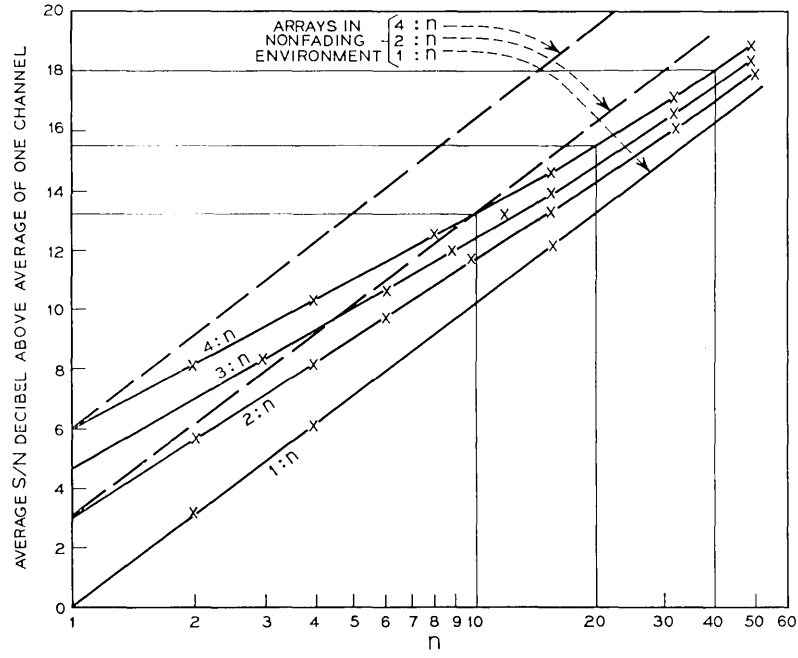


Fig. 4—Average S/N of complex conjugate retransmission arrays.

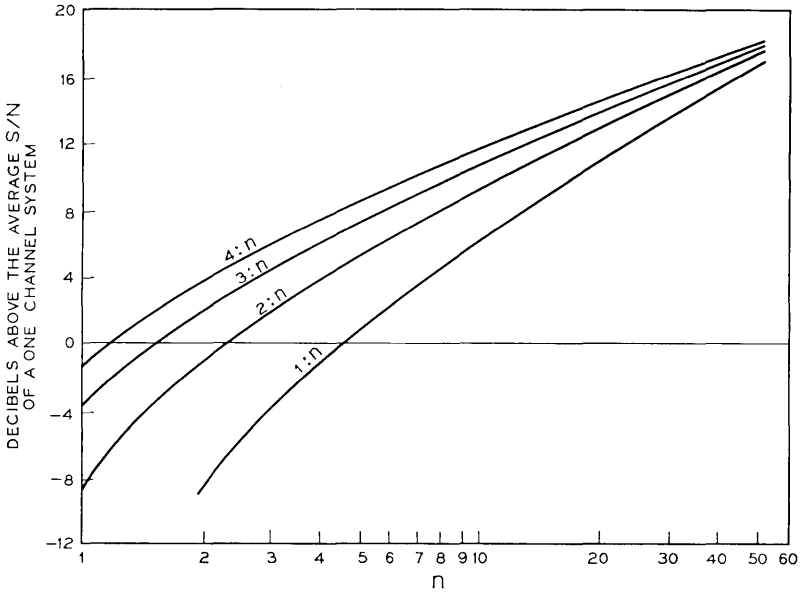


Fig. 5—99 percent reliability level. Complex conjugate retransmission maximal ratio diversity reception.

voltage components produced by the n antennas again add in phase at each antenna of the "m" array if plane wave incidence is assumed. Consequently, the power received is identical to that of the $1{:}mn$ array. However, in a random environment the n voltages components at each antenna element in the m array no longer add in phase; therefore, the $m{:}n$ system receives less power than that of the $1{:}mn$ system.

With reference to Fig. 2, we notice that for 99 percent of the time, the single Rayleigh signal has a value above −20.6 dB; we will designate −20.6 dB as the 99 percent reliability level. Hence the difference in dB values of two antenna systems for a particular reliability indicates their difference in signal threshold or their difference in the required transmitter power. The 99 percent reliability level is presented in Fig. 5. We next define fading range as the dB difference between the average S/N and the 99 percent reliability level. Therefore, fading range should provide a good indication of the smoothness of the received signal. The fading range is presented in Fig. 6. It is seen that as n increases, the 99 percent reliability level approaches the average signal level. In other words this means that as

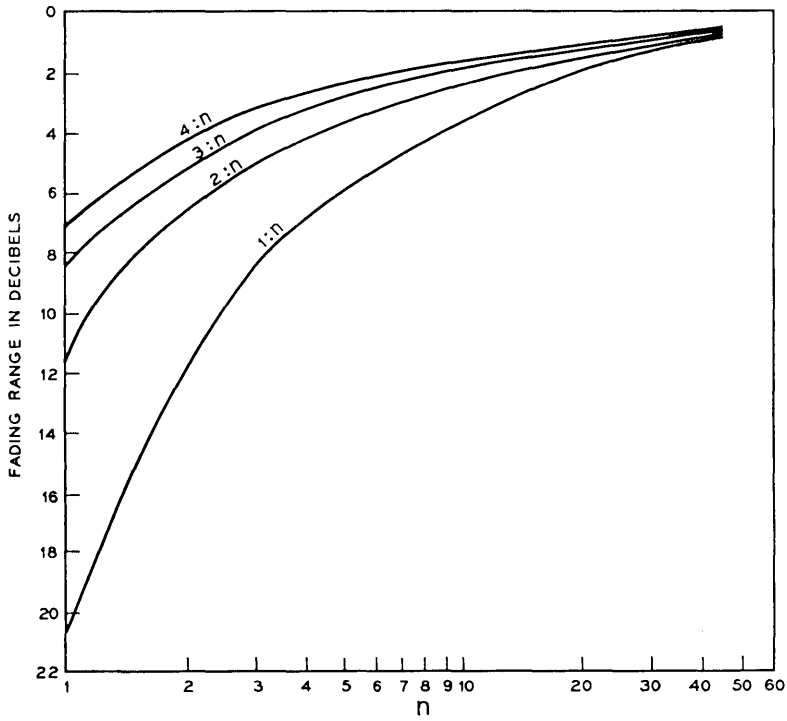


Fig. 6—Fading range of an $m:n$ array system. Complex conjugate retransmission maximal ratio diversity reception.

the number of diversity branches increases, the fading range starts to diminish. Figure 7 presents the CPD of a 4:32 array system. We note that the CPD curve is extremely flat and the signal level varies within a ± 1 dB range, indicating a greatly reduced fading range as compared to either Figs. 2 or 3.

We discuss now results obtained from the phase conjugate retransmission system. In this system, as was discussed in Section II, the S/N, of an equal gain diversity reception system is maximized. It is observed that because of this maximization effect, the performance of the phase conjugate system is not much inferior to that of the complex conjugate system. For example, the CPDs of the S/N for both systems in the case of a 2:4 array system are presented in Fig. 8. The CPD curves of the two systems differ approximately by the average S/N difference. Therefore, the difference in average S/N

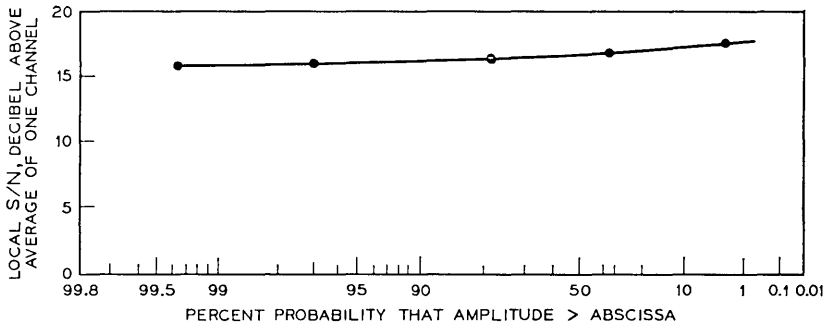


Fig. 7—CPD of a 4:32 array system. Complex conjugate retransmission maximal ratio diversity reception.

of the two systems is also a good indication of their difference in percentile reliability levels.

The average S/N of the two systems is shown in Fig. 9 for 2:n and 4:n array systems. It is seen that for the same m:n array, the difference of the two systems is small, that is, within a dB or so.

VI. CONCLUSIONS

We observed that in a fading environment, both complex conjugate retransmission and phase conjugate retransmission systems are capable of reaching a stable state and yield optimum results by greatly increasing the S/N at the receiving stations.

The performance of these two systems differs little. Therefore the

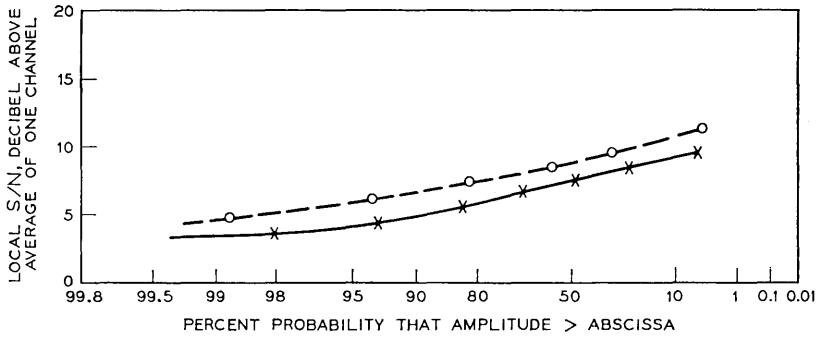


Fig. 8—CPD curves of a 2:4 array system. o, complex conjugate retransmission with maximal ratio reception; x, phase conjugate retransmission with equal gain reception.

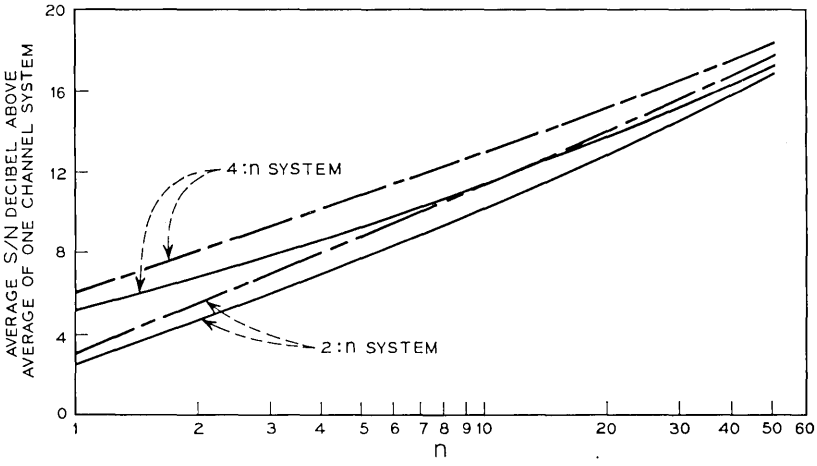


Fig. 9—Average S/N of antenna systems $2{:}n$ and $4{:}n$ _ _ _ _ _ _, complex conjugate maximal ratio reception; ————, phase conjugate retransmission equal gain reception.

choice of a particular scheme should be based on practical considerations. For example, in the phase conjugate system, the total power is divided equally among all the antenna elements. On the other hand, the complex-conjugate retransmission system requires that the total power be distributed in a complicated fashion. In practice this means that each antenna-feeding appartus must be equipped to handle power far exceeding that of the phase conjugate system.

In view of the simplicity of the phase conjugate retransmission compared to the complex conjugate retransmission (which must keep the total power transmitted constant), and only slightly inferior performance, the former appears to be a more attractive system.

As far as the division of diversity branches is concerned, it can be seen from Fig. 4 that for small numbers of antennas, an $m{:}n$ array would have similar performance to an $mn{:}1$ array. However, as the number of elements involved becomes larger, this relation no longer holds. For example the performance of a $4{:}n$ array would approach a $1{:}n$ array as n increases indefinitely.

VII. ACKNOWLEDGMENT

The author wishes to express his appreciation to W. C. Jakes for suggesting the problem and M. J. Gans for many helpful discussions.

REFERENCES

1. Special Issue on Active and Adaptive Arrays, IEEE Trans. on Antenna and Propagation, *AP-12,* No. 2 (March 1964), pp. 140–233.
2. Cutler, C. C. Kompfner, R., and Tillotson, L. C., "A Self-Steering Array Repeater," B.S.T.J., *42,* No. 5 (September 1963), pp. 2013–2032.
3. Morgan, S. P., "Interaction of Adaptive Antenna Arrays in an Arbitrary Environment," B.S.T.J., *44,* No. 1 (January 1965), pp. 23–47.
4. Brennan, D. G., "Linear Diversity Combining Techniques," Proc. of IRE, *47,* No. 6 (June 1959), pp. 1075–1102.
5. Clark, R. H., "A Statistical Theory of Mobile Radio Reception," B.S.T.J., *47,* No. 6 (July–August 1968), pp. 957–1000.
6. Aitken, A. C., *Determinants and Matrices,* New York: Inter-Science, 1954.

Microwave Line-of-Sight Propagation With and Without Frequency Diversity

By W. T. BARNETT

(Manuscript received May 5, 1970)

Amplitude measurements were made for 68 days in 1966 for seven 4-GHz and 6-GHz signals on a typical radio relay path. Identical measurements were also made for one 4-GHz signal on a second path having a common reception point with the first path. We present the results from an analysis centered on the fade-depth distribution for fades exceeding 20 dB. The more significant results are:

(i) The fade-depth distribution for all single (nondiversity) channels in a 5-10 percent band on the same path are essentially the same. Further, the distribution has the Rayleigh slope.

(ii) The single-channel fade-depth distributions differ for 4 and 6 GHz on the same path; the distributions also differ for the same 4-GHz frequency on adjacent paths with a common reception point.

(iii) One-for-one frequency diversity can be characterized during multipath fading periods for either the 4- or 6-GHz bands by the ratio of two quantities. The first is the precent frequency separation between diversity components. The second is the nondiversity fade-depth distribution.

I. INTRODUCTION

Line-of-sight microwave systems are affected by multipath propagation. When this phenomenon is present, the output from a receiving antenna can be practically zero for seconds at a time. Experimental data are difficult to obtain because long time periods of continuous coverage are needed to observe sufficient fading activity at the fade depths (30–40 dB) of interest for high performance systems. The literature is extensive on this general topic[1–7] but limited and in some cases contradictory[8] for these fade depths. The results available regarding frequency diversity are even more limited[9]. For these and other reasons, an extensive experimental program was undertaken in 1966.

Continuous amplitude measurements were made for 68 days at a rate of 5 samples per second per channel for seven 4-GHz and six 6-GHz signals on a radio relay path at West Unity, Ohio. Identical measurements were also made for one 4-GHz signal on a second path having a common reception point with the first path. Here a 68-day summer period (July 22 to September 28) in 1966 has been subjected to detailed analysis.

We present the results of the data analysis and their interpretation along with pertinent background information. Briefly the order of presentation is (*i*) experiment description, (*ii*) determination of the reference values used for calibration, (*iii*) nondiversity results, (*iv*) frequency diversity results, (*v*) a mathematical description of pairwise fading which is used to interpret the improvement obtained from frequency diversity, (*vi*) 4/6 GHz crossband results, (*vii*) adjacent hop results, and (*viii*) a comparison of space and frequency diversity.

II. SUMMARY

New results have been obtained from the data concerning 4- and 6-GHz propagation on line-of-sight paths. The present analysis was centered on the fade-depth distribution for fades of 20 dB or more. A simplified listing of the significant findings follows.

(*i*) During nonfading conditions, the received microwave signal power was constant for the entire test period to within ±1 dB including equipment variations.

(*ii*) The fade-depth distributions for all single (nondiversity) channels in a 5-10 percent band are essentially the same and have a Rayleigh slope.

(*iii*) The single-channel fade-depth distributions differ for 4 and 6 GHz on the same path; the distributions also differ for the same 4-GHz frequency on adjacent paths with a common reception point.

(*iv*) The performance of a one-for-one frequency diversity system can be specified for either the 4- or 6-GHz bands by the ratio of two quantities. The first is the percent frequency separation ($100\ \Delta f/f$) between in-band diversity signal components. The second is the experimental nondiversity fade-depth distribution $P(L)$. In these terms the improvement (I) of a diversity system relative to the nondiversity system as obtained from the data is simply

$$I = 0.13 \frac{\Delta f}{f} \Big/ P(L).$$

This model is based upon the in-band frequency diversity data and is in agreement therewith.

The factor I characterizes frequency diversity during multipath fading periods. As such, it should be applicable to different climates and terrains for path lengths of approximately 28 miles.

(*v*) The improvement from 4/6 GHz crossband diversity was not significantly better than in-band diversity of 2 percent or more separation.

(*vi*) Adjacent section diversity with a common point (as based on data on a single channel) was not significantly better than in-band frequency diversity. This raises some provoking (unanswered) questions about the correlation of selective fading on adjacent routes, for example, limitations on the maximum possible diversity improvement to values less than those expected from independent fading.

(*vii*) The performance of space diversity[10] is comparable to that of one-for-one frequency diversity on the same hop.

(*viii*) The polarization of the radio signals had no noticeable effect on the amount of fading.

These results are presented in detail along with the necessary background information in the following sections.

III. EXPERIMENT DESCRIPTION

The transmitted power in microwave radio systems is constant. Propagation data can therefore be obtained from in-service systems without interfering with their operation by using suitable monitoring equipment. Such equipment (MIDAS*) was installed at West Unity, Ohio, to monitor and record the received envelope voltages of standard TD-2 (4 GHz) and TH(6 GHz) signals. A list of the channels is given on Table I. Briefly there were seven 4-GHz, six 6-GHz, and two space-diversity channels on one hop and one 4-GHz channel on a second hop. A functional block diagram is shown on Fig. 1.

West Unity, Ohio, was chosen as the measuring site for this experiment because it lies along a major route in an area with a reputation for considerable fading. Further, the hops measured have average lengths (28.5 and 29.4 miles) with negligible ground reflections. The two paths differ in azimuth by 68 degrees and their profiles are given on Figs. 2 and 3; clearance is adequate even for the extreme case of equivalent earth radius (k) equal to two-thirds.

* An acronym for Multiple Input Data Acquisition System.

TABLE I—RADIO CHANNELS MEASURED AT WEST UNITY, OHIO

From Pleasant Lake, Indiana (28.5 mi)

Channel No.	Frequency	Antenna	Polarization
4-7	3750	Horn Reflector	V
4-1	3770		H
4-8	3830		V
4-2	3850		H
4-9	3910		V
4-11	4070		V
4-6	4170		H
6-11	5945.2		H
6-13	6004.5		H
6-14	6034.2		V
6-15	6063.8		H
6-17	6123.1		H
6-18	6152.8		V
6-UD	6152.8	Upper Dish	V
6-LD	6152.8	Lower Dish	V
From Paulding, Ohio (29.4 mi)			
4-6	4170	Horn Reflector	V

Note: The 4-*X* channels correspond to standard TD-2 radio system signals; 6-*X* corresponds to TH.

The MIDAS equipment derived received signal strength information by sampling the voltage of the 70-MHz IF signal at a point where it was linearly related to the RF signal. At any instant the particular channel being measured was selected automatically by MIDAS. A common detector then converted the IF amplitude measurement to a dc voltage which was quantized into one of 32 contiguous steps over a 45-dB range. The MIDAS input-output curve is given as Fig. 4.

The data were recorded on paper tape along with the necessary timing information. Measurements were made throughout the 68-day period at a rate of 5 samples per second on each channel. The information was recorded for all channels at rates of either 1 sample per 30 seconds, 1 sample per 2 seconds and 5 samples per second (normal, intermediate, and fast rates) depending on the fading activity of the channels under test. The recording rate was automatically selected by MIDAS so as to record all significant fading. During computer processing of the data, the amplitude value at a sampling instant was assumed to hold until the next sampling instant.

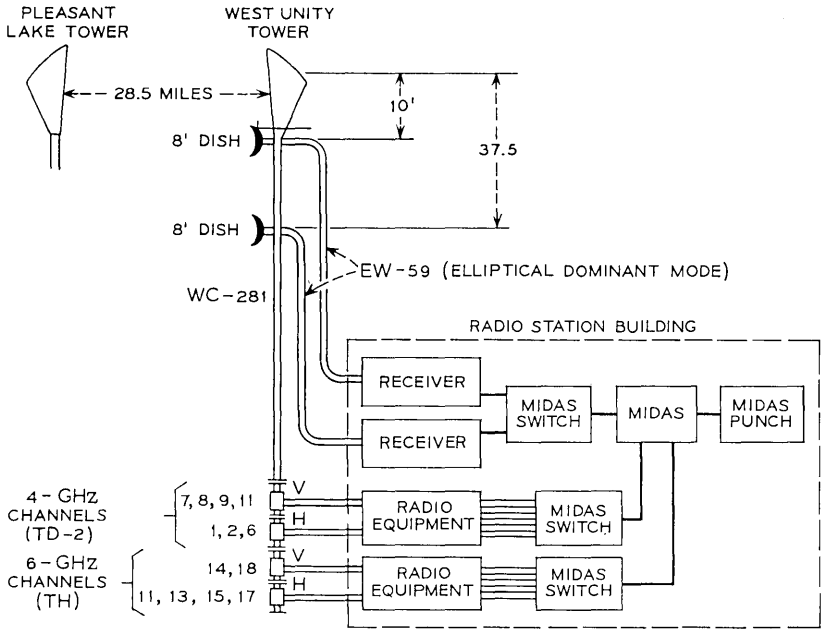


Fig. 1—Experimental layout, Pleasant Lake to West Unity (Paulding to West Unity not shown).

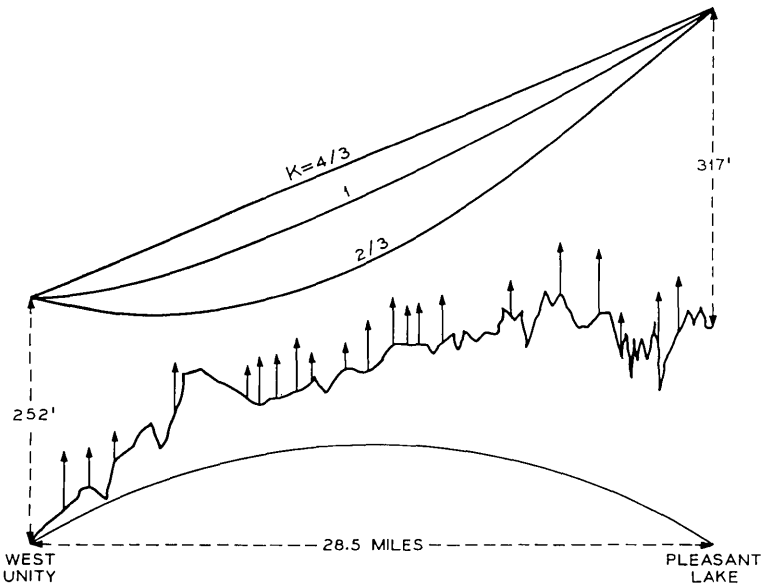


Fig. 2—West Unity—Pleasant Lake path profile.

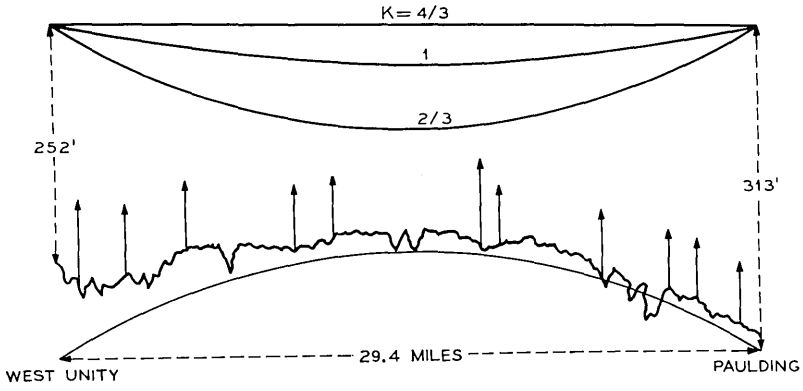


Fig. 3—West Unity—Paulding, Ohio, path profile.

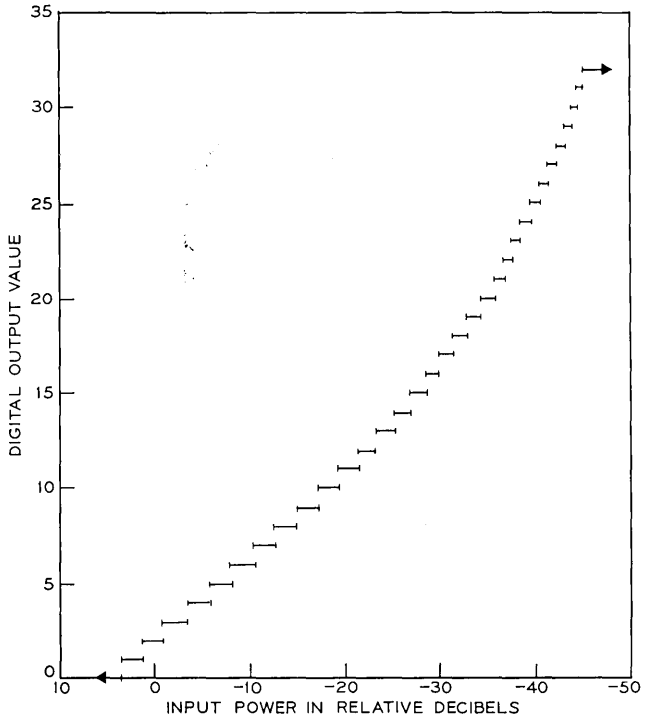


Fig. 4—MIDAS calibration curve.

An important feature of the experiment was long-term continuous coverage. Deep fades are rare events occurring at unpredictable times; the test equipment had to be on-line continuously to obtain an adequate sample.

The required test equipment reliability and measurement uniformity was obtained by maximum use of common equipment. The essentially continuous coverage was obtained by recording mainly the significant fading data. Even so the subsequent processing was a formidable task, even with the computer, because of the high volume of raw data.

IV. NONFADING SIGNAL VALUES

A fade is defined as a decrease in the envelope of the received signal voltage with respect to a reference or free-space value. Thus before fading data can be quantified, the reference or nonfading value must be determined.

If the atmosphere between the transmitting and receiving antennas was homogeneous (that is, no vertical or horizontal variations in the index of refraction), then the single frequency RF power at the output of the receiving antenna would be invariant for a fixed transmitted power.* Its value (called the free-space value) could be calculated in a straightforward manner. However, even during nonfading periods, there are small time-varying random deviations in the refractive index which cause small scintillations in the received power even when the average value remains constant. There are also long-term variations in the received RF power due to equipment variation. For our purposes we must determine the nonfaded received power as a function of time and, if possible, quantify the scintillations.

Inspection of the data showed that the midday hours had the least amount of fading. Here the differentiation between fading and free-space scintillations is made on the basis of the magnitude of the effect. Fading causes variations of one or more quantizing levels in the envelope from hour-to-hour on most of the 15 channels.

To establish a reference value, midday periods were sought which had no fading with respect to either time or frequency. It was easy to find a total of 129 midday hours simultaneously for all channels on 30 different days scattered throughout the entire 68-day period.

Table II gives the summaries for the midday values. The table shows the average signal in terms of quantizing level for five consecutive time periods of from 9 to 20 days duration. Several points can be made about

* Assuming adequate ground clearance and no ground reflections.

TABLE II—AVERAGE NONFADED VALUES
(in terms of quantizing levels)

	Time Periods				
	1	2	3	4	5
Pleasant Lake					
4- 7	4	4	4.5	4	4.5
- 1	2	2	2	1.5	1.5
- 8	3	3	3	2.5	2.5
- 2	3	3	2.5	3	3
- 9	4	4	4	4	4
-11	4	4	4	4	3.5
- 6	4	3.5	3.5	3.8	3.5
6-11	2	1.5	2	1.5	1.4
-13	2	2	2	1	1
-14	2	2	2	2	2
-15	1	1	1.5	1	1
-17	2	2	2	2	2
-18	2	1.5	2	2	1.5
6-UD	4	3.5	3.5	3.8	3.5
6-LD	5	4.5	5	4.5	4.2
Paulding					
4- 6	2	2	2	3	2.8
Total Hours in Period	254	375	482	264	263
Hours Used	31	22	24	25	27
Total Days in Period	9	16	20	11	12
Days Used	7	7	7	5	4

Note: Quantizing level 4.X means that the average value was 0.X of a level offset from the center of level 4 in the direction of level 5.

this data. First the maximum peak-to-peak variation on any channel is one level or about 2 dB while the average variation is $\pm\frac{1}{4}$ of a level or about ± 0.5 dB. Further some of the channels, for example, 4-1 and 6-13, exhibit a definite trend over the 129 hours. The belief is that these long-term effects are attributable to the radio equipment.

In any case, the average deviation of ± 0.5 dB is small enough so that a single reference value for each channel can be used for the entire time period. This simplifies data reduction considerably.

Now consider the statistics of small scintillations in the received signal

power. Table III gives the percent distributions by level for all the channels for the 129 midday hours. Of course this distribution includes the long-term equipment variations in the reference values as well as the short-term scintillations. Note that the channels with the minimum variations in average value from Table II are those with most of their "90 percent hours" in a single level. These are 4-9, 4-11, 6-14, and 6-17. It is assumed that the variations on these channels are due *only* to scintillation and that this effect can be represented by a probability distribution which is normal in dB. The σ of this distribution can be found from the percent values given in Table III with the results shown in Table IV. The agreement between these channels is excellent. The conclusion is that the scintillation effect over a 68-day period is universal with a σ of 0.6 dB superimposed on an equipment variation of ± 0.5 dB. The rms variation in reference value is then ± 0.8 dB.

4.1 *Channel Calibration*

The data on reference values were combined with the MIDAS calibration curve to calibrate the 15 RF channels in dB. First, all the 6- and 4-

TABLE III—SUMMARY OVER ENTIRE 68 DAYS
(Data for 129 Hours on 30 Days)

Channel Freq.	Percent of Time in Level					Hours with 90 Percent or More of Time in Level				
	1	2	3	4	5	1	2	3	4	5
West Unity										
4- 7			0.32	78.25	21.43				63	
- 1	18.1	80.50	1.40			12	93			
- 8		17.52	81.34	1.13	0.01		9	88		
- 2		0.01	86.25	13.74				95		
- 9			7.41	91.94	0.65			1	111	
-11			10.36	88.93	0.71				96	
- 6		0.77	32.50	66.73			1	16	58	
6-11	28.57	69.60	1.83			9	47			
-13	34.56	63.68	1.76			31	71			
-14	6.01	92.65	1.34			4	118			
-15	83.95	16.01	0.04			94	6			
-17	2.10	94.96	2.94				113			
-18	17.88	78.96	3.16			4	69	1		
6-UD			34.65	65.27	0.08			18	42	8
6-LD				33.2	66.80				24	66
Paulding 4- 6	4.09	58.04	37.87			2	59	42		

TABLE IV—LONG-TERM STANDARD DEVIATION

Channel	σ in dB
4-9	0.56
4-11	0.59
6-14	0.58
6-17	0.56

GHz channels were simultaneously lined up at their reference level which was specified as 0 dB. By inspection, 29 dB values were chosen over the fading range in order to give minimum ambiguity over the entire set of channels; thus each quantizing step on each channel was not used more than once. In this way all the 6- and 4-GHz channels were simultaneously calibrated; this was done so that an arbitrary subset could be chosen for analysis without having to recalibrate. Figure 5 gives an example of the results of the calibration procedure for channels 4-7, 4-2, and 4-9 for fades greater than 20 dB.

Because the calibration curve is nonlinear, this process requires some judgment. The minor combined effects of the nonlinear calibration

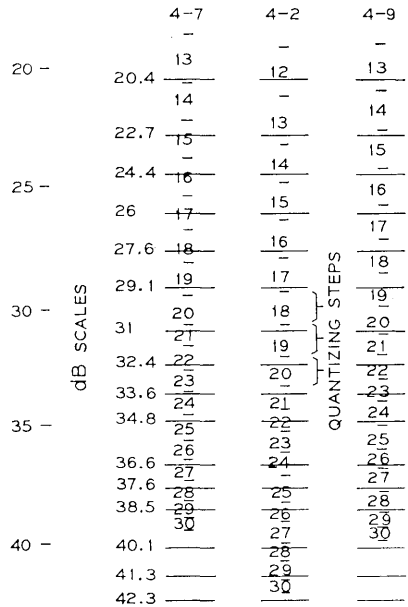


Fig. 5—Calibration example.

curve and differing reference levels for different channels are discussed in conjunction with the single channel outage statistics.

V. SINGLE CHANNEL RESULTS

The raw data were obtained continuously for almost all of the 68 days (5.9×10^6 seconds). Of this total, 5.26×10^6 seconds was used as the data base; the balance was lost mainly because of routine radio maintenance. To condense the data, a criterion was used to select by computer only those time periods which exhibited fading. The start of such a time period was defined by, and included, ten consecutive measurements containing any one channel faded below approximately 10 dB. The end of the time period was defined as that instant for which the next 110 consecutive measurements on any channel did not have a fade exceeding approximately 10 dB.

From the total of 5.26×10^6 seconds, 7.8×10^5 seconds (14.8 percent) were selected for analysis. The average length of the periods selected was sizeable. There were only 96 distinct periods selected; these had an average length of 8.1×10^3 seconds ($2\ \frac{1}{4}$) hours.) Further one-half of the analysis time was in intervals of four hours or longer. Thus any effects due to beginning or ending a time period should be minimal.

The data were processed by computer to determine the total amount of time during which each signal was less than a certain amount. The 4-GHz single-channel fading results are given on Fig. 6 for fades greater than 20 dB. These results and all those to follow are given as a fraction of 5.26×10^6 seconds. It is apparent that these statistics are essentially the same for all the 4-GHz channels and have the Rayleigh slope, that is, 10 dB per decade of probability over the entire range of data points. The solid line on the figure is a least-square fit of a Rayleigh slope line to the data points, most of which are within ± 1 dB as shown by the dashed lines. This scatter is due to both the uncertainties in the reference value and to the nonlinear calibration.

The 4-2 points outside the 2-dB corridor from 22 to 29 dB are due to the nonlinear quantized calibration. From Fig. 5 note that for 4-2 the dB values used lie near the bottom of the quantizing levels up to 31 dB at which point they change to the middle of the quantizing levels. This gives the effect noted on Fig. 6, that is, the data points are shifted to higher fade values for a constant probability. Other anomalies of this type in the single-channel results are explainable in this manner. For these results and for all others described here, the polarization of the signal (s) had no apparent effect.

The 6-GHz signal channel results are given on Fig. 7 for fades greater

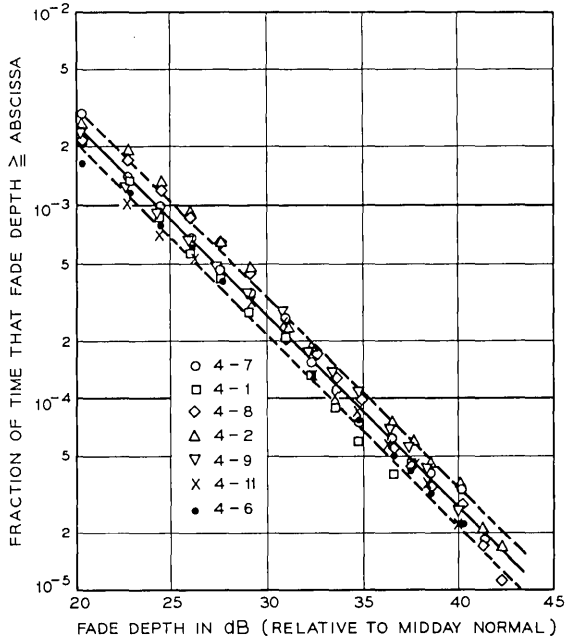


Fig. 6—Fade-depth distribution; 4-GHz channels.

than 20 dB. Again all 6-GHz channels have essentially the same statistics with the solid line being the least-squares fit with a Rayleigh slope. Almost all data points are within ±1 dB of the average above 40 dB except for 6-15 from 37 to 40 dB. This discrepancy is attributable to nonlinear quantizing as discussed for 4 GHz. The increased scatter above 40 dB is thought to be due to decreasing measurement sensitivity.

The single-channel results for the space diversity grouping[10] (the 6-18 signal is received on the horn reflector and two dishes) and for the 4-GHz channel on the Paulding route are given on Fig. 8. The lines are the least-squares fit with a Rayleigh slope.

Figure 9 gives a summary of the single-channel statistics and for comparison, the true Rayleigh curve. The equations of the least-square lines are

West Unity	4:	$P = 0.25\ 10^{-F/10}$
	6:	$P = 0.53\ 10^{-F/10}$
	SD:	$P = 0.43\ 10^{-F/10}$
Paulding	4:	$P = 0.77\ 10^{-F/10}$

where F is the fade depth expressed in dB ($F \geqq 20$ dB). The channel with the most fading was the 4-GHz Paulding followed by 6 GHz, space

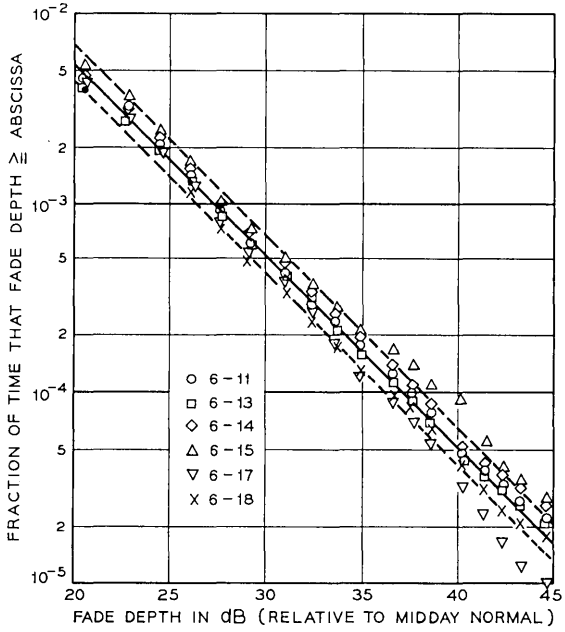


Fig. 7—Fade-depth distribution; 6-GHz channels.

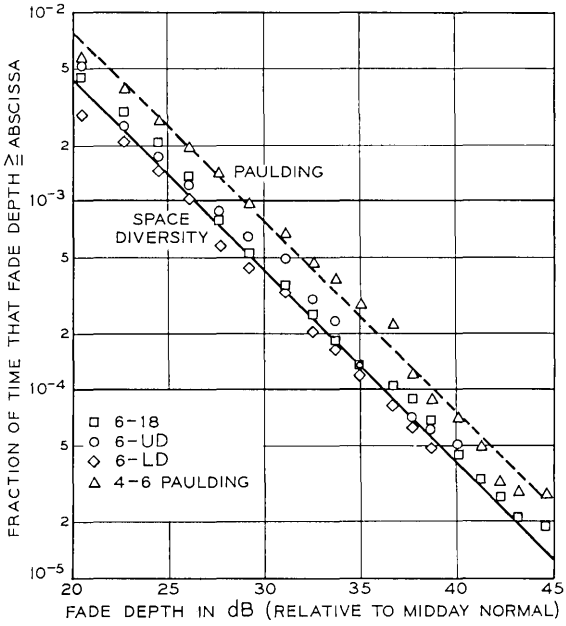


Fig. 8—Fade-depth distribution; space diversity grouping and Paulding.

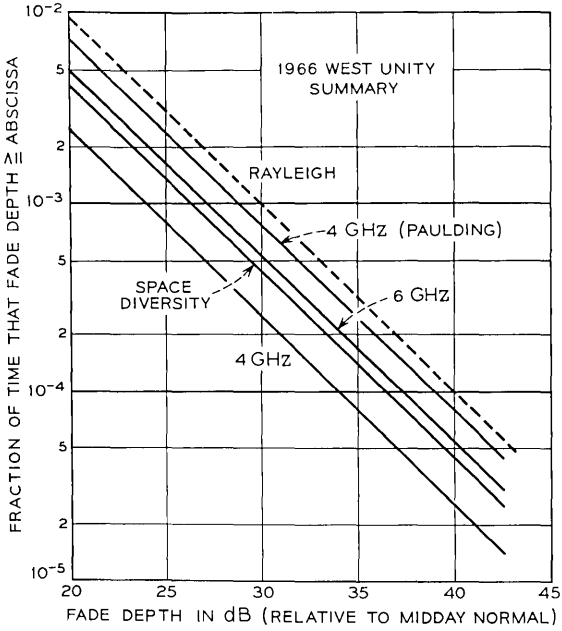


Fig. 9—Summary of fade-depth distributions.

diversity grouping (SD) and 4 GHz. The 4-GHz channels have significantly less fading than either 6-GHz (by 3.3 dB) or 4-GHz Paulding (by 4.9 dB). The 0.9-dB difference between signals on antennas of different height, that is, 6 GHz compared to SD, is thought to be more apparent than real, although it may be a small height effect.

It should be noted that having essentially the same fade distribution for the 6-18 signal as received on both the dishes and horn reflector implies two things. First, the effect of 6-GHz multimoding in the horn reflector, circular waveguide, and combining networks must be negligible because the dishes use dominant mode elliptical waveguide and no combining networks. Secondly the effect of decreased clearance at midpath for the lowest dish is less than 0.9 dB.

One way of explaining the significant differences shown on Fig. 9 is to examine them in terms of the terrain and the radio path lengths. Pearson[7] has given data taken in Britain on the relation between worst-month fading and the terrain as characterized by the path roughness.*

* Path roughness is the standard deviation of terrain height measurements at one-mile intervals on a line between transmitter and receiver with the end points of the path excluded.

Assuming that the 68-day period is equivalent to the British worst-month data, Table V can be compiled from Fig. 9 and Ref. 7.

The 6-GHz British point has been obtained by assuming that the path length is 50 percent longer at 6 GHz than it is at 4 GHz; that is, the path length is cast in terms of wavelengths.

There is good agreement between the British and the West Unity data. Thus the difference in depth of fade for a given percentage of time is apparently directly related to the terrain roughness and to the path length in wavelengths. Of course this is not sufficient evidence to justify the extensive use of these parameters. It has long been known, at least qualitatively, that fading is more severe over smooth terrain or water than on rough paths of comparable frequency, length, and atmospheric conditions.

VI. FREQUENCY DIVERSITY RESULTS

The simultaneous measurements on a number of different frequencies, together with computer processing of the differences in signal level with frequency, have provided much more quantitative information than previously available on the improvements to be expected from the use of frequency diversity. The diversity results specify the total amount of time during which the stronger of two signals was less than a certain amount (this means that both signals simultaneously were less than the given amount).

6.1 6 *GHz*

The results for the 6-GHz pairs for fade depths $\geqq$ 20 dB are given on Figs. 10 through 16. Fifteen pairs were obtained from the six 6-GHz channels and they are grouped according to frequency separations as shown in Table VI.

Four lines are shown on each figure. The uppermost is the nondiversity line which is the average single-channel fade-depth distribution as dis-

TABLE V—PATH ROUGHNESS EFFECTS

			0.1 Percent Fade Depth	
		Roughness	*British*	*West Unity*
Pleasant Lake	4 GHz	16.0 meters	23.5 dB	24.0 dB
	6 GHz	16.0 meters	29.0 dB	27.3 dB
Paulding	4 GHz	8.5 meters	28.0 dB	28.9 dB

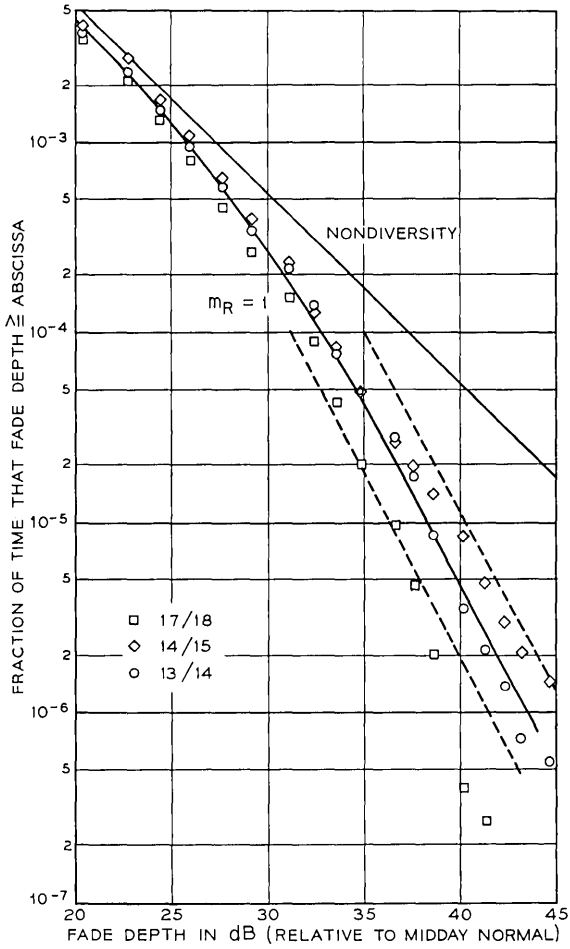


Fig. 10—6-GHz frequency diversity; 30-MHz separation.

cussed previously. The bottom solid line which is tagged with a value of a parameter m_R is a curve fitted to the data. The dashed lines are relative to the fitted line and denote a ± 2-dB corridor which is an estimate of the uncertainties in the data due to nonlinear calibration and reference value determination. The fitted curve is obtained by assuming that the diversity data is jointly Rayleigh distributed with respect to the nondiversity curve. The parameter m_R is related to the amount of correlation between the two components of the distribution. This concept will be discussed in more detail later.

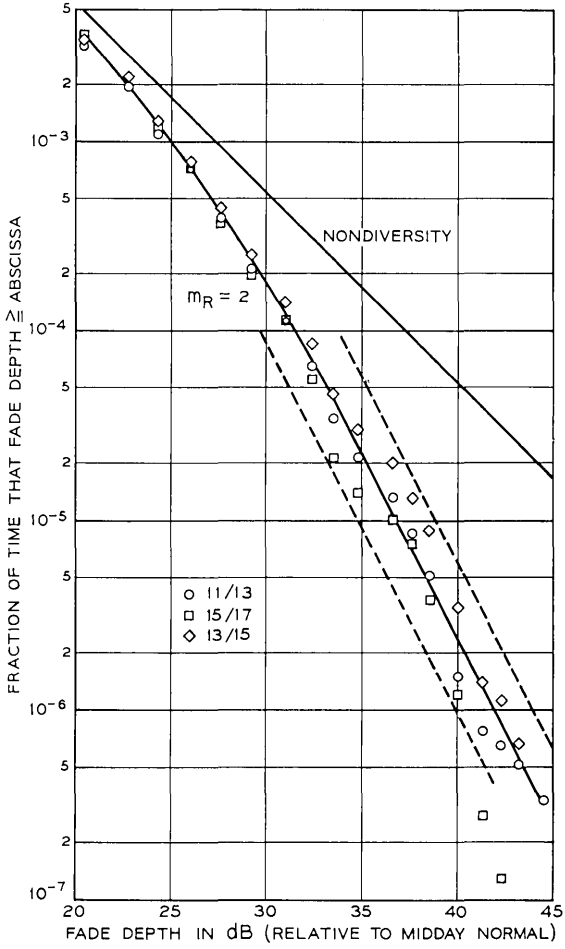


Fig. 11—6-GHz frequency diversity; 60-MHz separation.

Inspection of these results (Figs. 10 through 16) shows that for a fixed frequency separation, the scatter of the data points with respect to the fitted diversity line is small below 30-dB fade depth but increases somewhat for larger fade depths.* However for fade depths of 40 dB or less, all the data points lie within the ±2-dB corridors except for

* On the figures, $10^{-6} = 5.26$ seconds which means that there were few samples at the higher fade depths.

17/18 on Fig. 10. This latter result is an anomaly because all other combinations which include 6–17 or 6–18 are quite consistent within their group. In fact, the consistency of the data points for different pairs having the same frequency separation is remarkable. Also note the excellent agreement between the data and the fitted line for the pair with the maximum frequency spacing (210 MHz).

As the frequency separation increases, it is to be expected that the

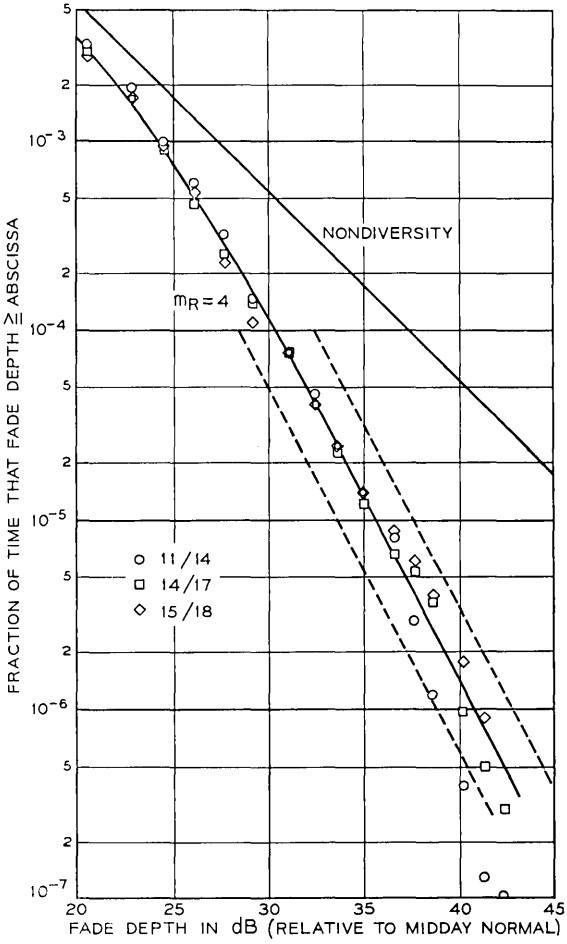


Fig. 12—6-GHz frequency diversity; 90-MHz separation.

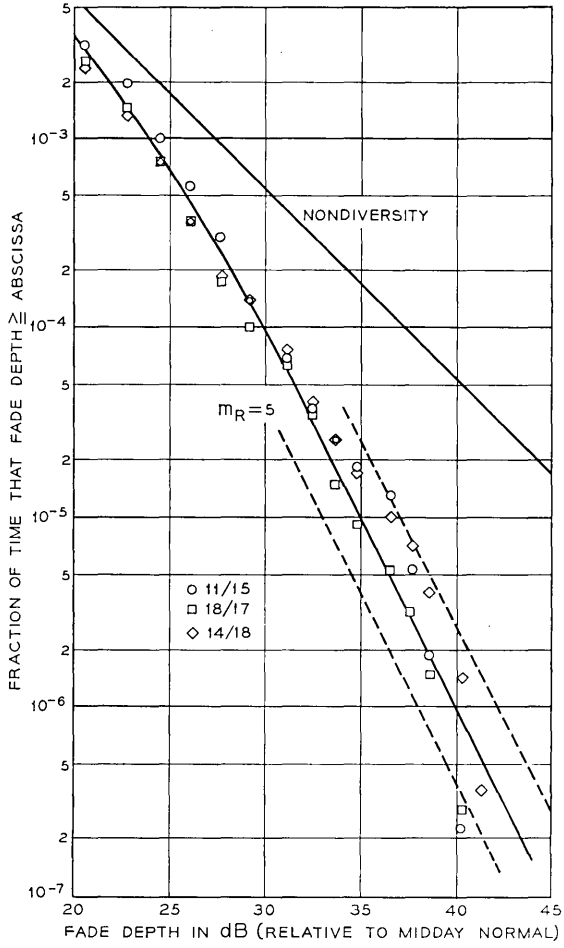


Fig. 13—6-GHz frequency diversity; 120-MHz separation.

diversity performance will improve. This is borne out on Figs. 10 through 16 and is described by increasing values of m_R for increasing frequency separation. The performance of frequency diversity relative to nondiversity will be discussed in a later section.

6.2 4 *GHz*

The results for the 4-GHz frequency diversity pairs are given on Figs. 17 through 24. Twenty-one pairs were obtained from the seven

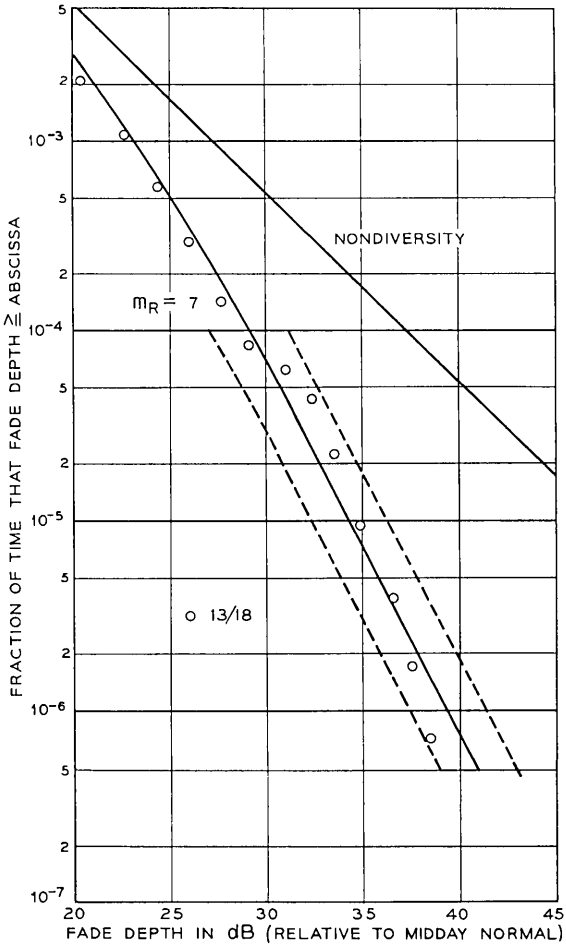


Fig. 14—6-GHz frequency diversity; 150-MHz separation.

different TD-2 channels and they are grouped according to frequency separations as shown in Table VII.

The lines on the figures have exactly the same meaning as in the 6-GHz case discussed in the previous paragraphs.

Inspection of the results shows that the scatter of the points with respect to the fitted diversity line is small for fade depths less than 30 dB except for 7/1 on Fig. 17 which has been ignored as an anomaly.

For greater fade depths, the scatter increases and the data points tend to fall off faster than the fitted line except for Fig. 23 which has a distinct upward bulge. The fast rolloff might result from noise or interference effects in the radio system. Since the 6-GHz results do not exhibit these effects, the MIDAS system and the data reduction procedures are probably not the source of this rolloff since all of the radio channels were treated identically. Further some of the pairs follow the fitted

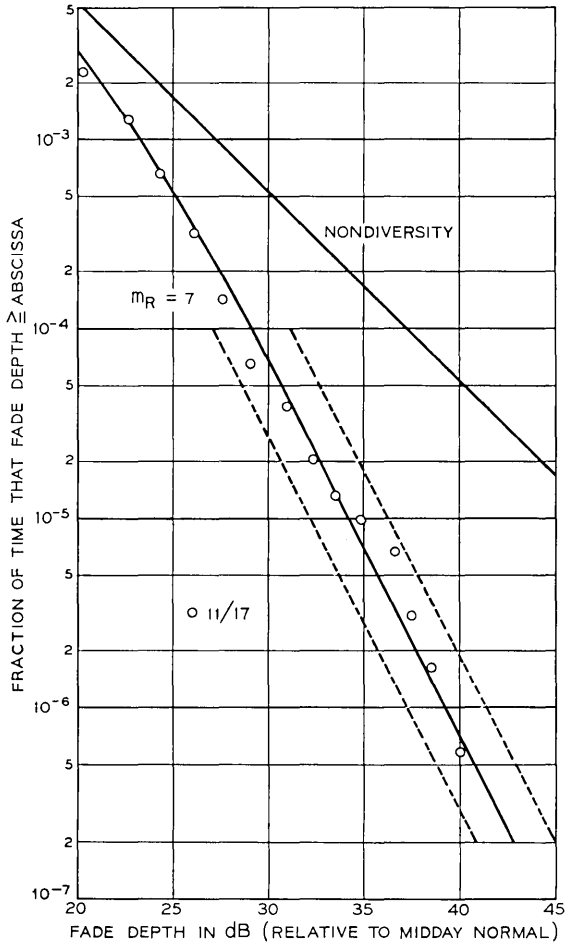


Fig. 15—6-GHz frequency diversity; 180-MHz separation.

line without any rolloff, for example, 4–9/11 on Fig. 21 and 4–7/2 on Fig. 20. The reasons for the anomalies are not explicitly known but it is assumed that they are *not* generated by multipath fading. In any case, the fitted line is a conservative approximation to the data except for Fig. 23.

Just as in the 6-GHz case when the frequency separation increases, the diversity performance improves. This is described by increasing

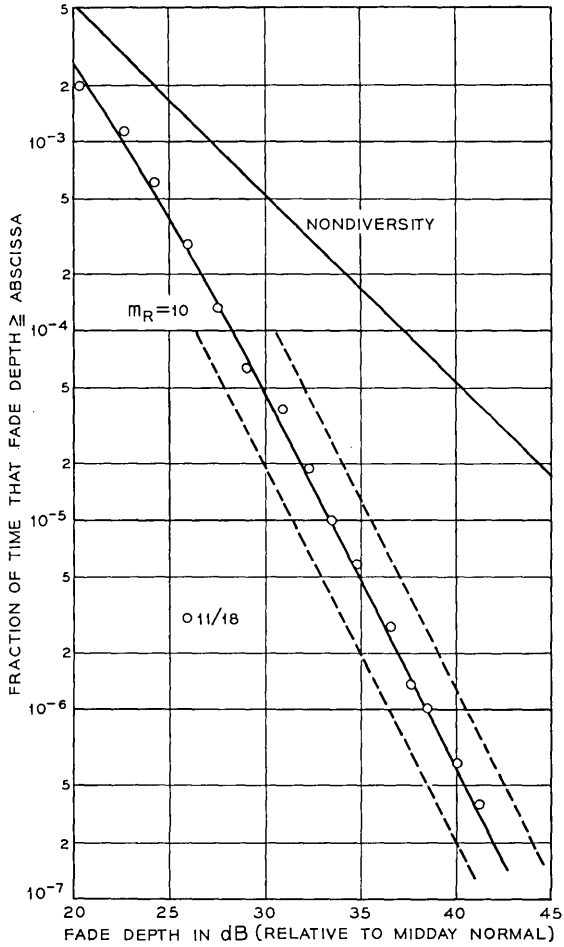


Fig. 16—6-GHz frequency diversity; 210-MHz separation.

TABLE VI—6-GHz FREQUENCY DIVERSITY RESULTS

Figure	Frequency Separation (MHz)	Number of Pairs
10	30*	3
11	60	3
12	90	3
13	120	3
14	150	1
15	180	1
16	210	1

* This is also the nominal bandwidth of the working channel.

values of m_R for increasing frequency separation and will be discussed in a later section.

VII. DESCRIPTION OF SIMULTANEOUS FADING AT DIFFERENT FREQUENCIES

Multipath fading is caused by complicated interference phenomena and it is possible that various descriptions of simultaneous fading are useful. Models for fading can be postulated on two levels. First there is a mathematical (statistical) description of the characteristics of multipath fading. Second, on a more fundamental level, there is the model for the physical process that creates fading and from which the mathematical (statistical) model could be derived. At the present time there is no physical process model which gives results that agree well with the experimental data. On the other hand, a statistical model based on the joint Rayleigh probability distribution has been useful in the description of space diversity, and it is applied here (with considerable success) to frequency diversity. However the physical process model is still the ultimate goal and the experimental data and empirical formulas presented here should aid in attaining this goal.

The following discussion briefly gives the relevant details of the joint Rayleigh distribution as applied to the data. For a Rayleigh variate, the probability that the envelope voltage R_1 of the signal normalized to its rms value has a value less than L is

$$\Pr(R_1 < L) = 1 - \exp(-L^2). \tag{1}$$

Similarly the probability distribution of the envelope voltage R_2 of a second signal normalized to the rms value of the first signal is

$$\Pr(R_2 < L) = 1 - \exp\left(-\frac{L^2}{v^2}\right) \tag{2}$$

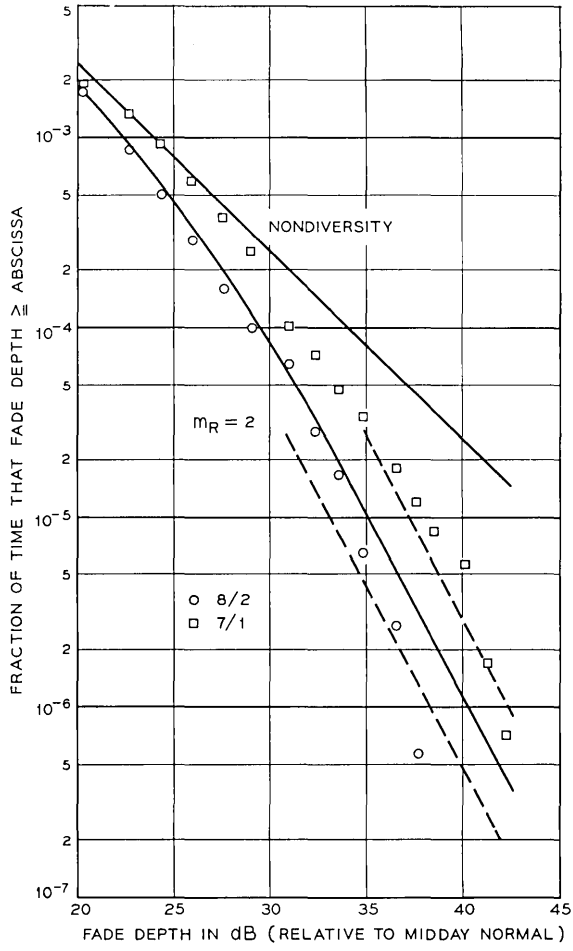


Fig. 17—4-GHz frequency diversity; 20-MHz separation.

where

$$v^2 = \langle R_2^2 \rangle_{av} (\langle R_1^2 \rangle_{av})^{-1}. \tag{3}$$

The joint probability distribution function of the variables R_1 and R_2 is[10]

$$\Pr(R_1 < L, R_2 < L) = \int_0^{L^2/(1-k^2)} dX_1 \int_0^{(L/v)^2/(1-k^2)} dX_2 P(X_1, X_2) \tag{4}$$

with

$$P(X_1, X_2) = (1 - k^2)I_0[2k(X_1X_2)^{1/2}] \exp[-(X_1 + X_2)$$

where k^2 is the correlation coefficient of R_1^2 and R_2^2. For use in this paper m_R has been defined as

$$m_R = 10^3(1 - k^2). \tag{5}$$

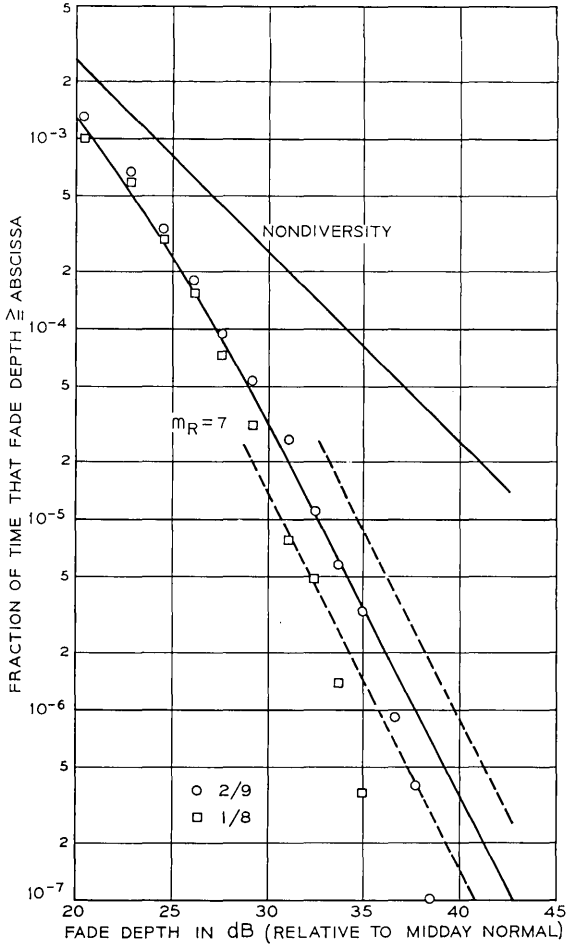


Fig. 18—4-GHz frequency diversity; 60-MHz separation.

Typical computed results are shown in Fig. 25 for $v^2 = 1$. For deep fades, asymptotic forms of equations (2) and (4) are quite useful.

$$\Pr(R_2 < L) \cong L^2 v^{-2} \tag{6}$$

and

$$\Pr(R_1 < L, R_2 < L) \cong (10^3/m_R)(L^4/v^2). \tag{7}$$

The region of validity of equations (6) and (7) depends on v, L, and m_R.

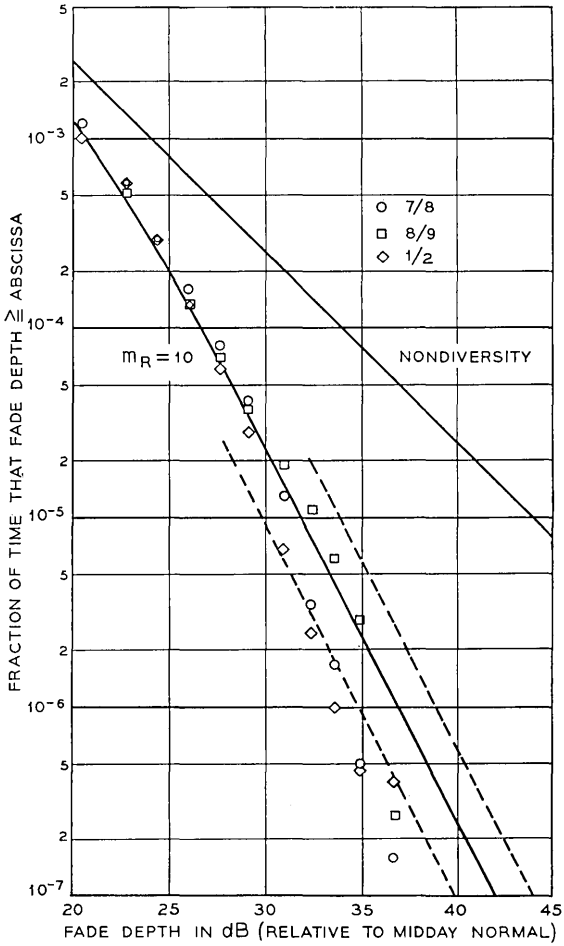


Fig. 19—4-GHz frequency diversity; 80-MHz separation.

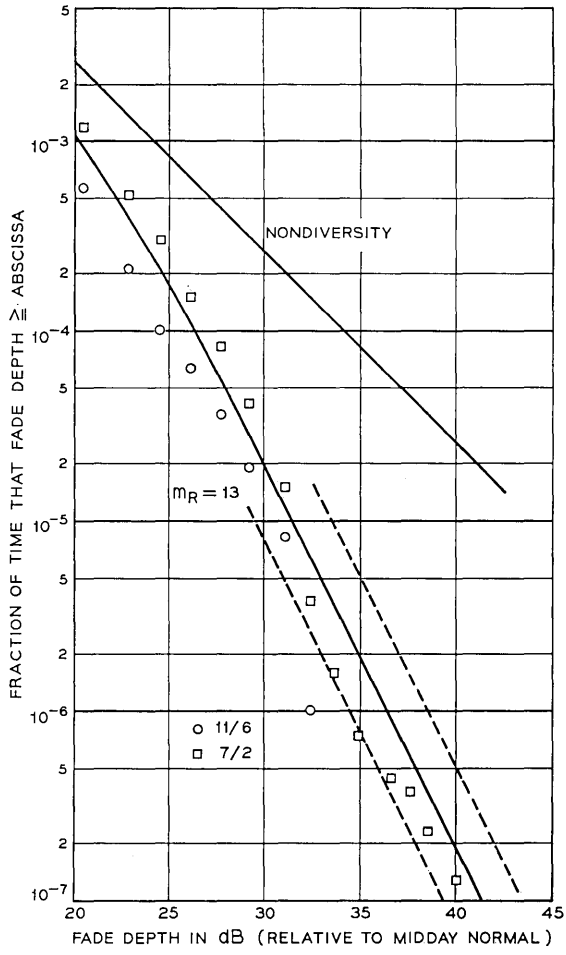


Fig. 20—4-GHz frequency diversity; 100-MHz separation.

For example it is the region in Fig. 25 where the lines are parallel to the $m_R = 10^3$ line.

The joint Rayleigh distribution, calculated from equation (4), was fitted to the diversity data points by overlaying plots of the joint distribution for various values of m_R and choosing the one with the best apparent fit. The results of this are the bottom solid lines on the diversity plots with the value of m_R next to each line. In the fitting, somewhat

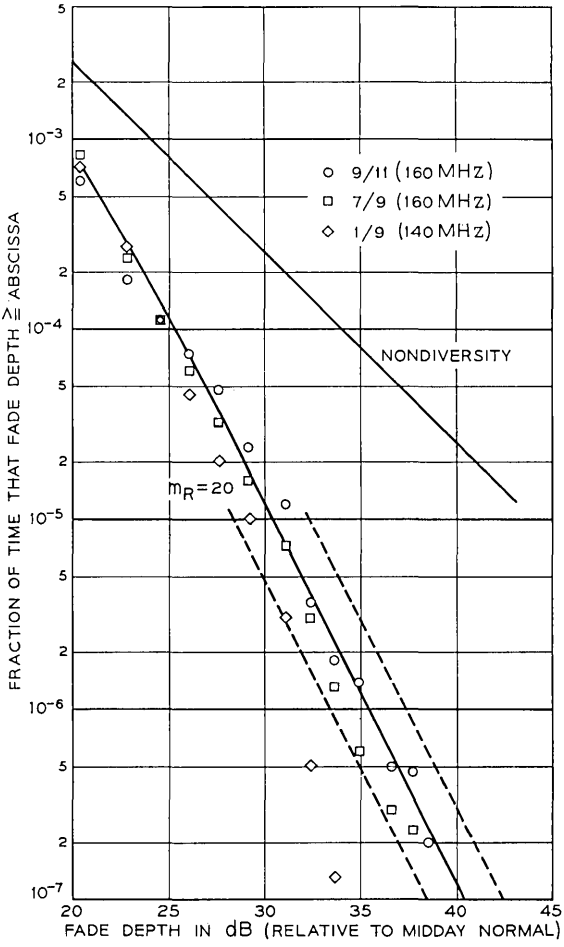


Fig. 21—4-GHz frequency diversity; 140/160-MHz separation.

more weight was given to the values at 30-dB fade rather than at 40 dB because of relative sample size. Also note that the curvature of the joint Rayleigh fits the curvature of the data points for the smaller fade values.

VIII. IMPROVEMENT

The quantity of interest in any diversity scheme is the amount of improvement relative to the nondiversity performance. Here this per-

formance measure is defined as the ratio of fractional outage of the nondiversity signal to that of the diversity signal for a fixed fade depth. Description by this factor (I) is convenient because it avoids detailed description of the many schemes that are used to process the two signals. The best of these switching or combining schemes will provide performance equal to or somewhat better than that described by the fade reduction factor.

The fraction of the total time that a nondiversity signal is faded

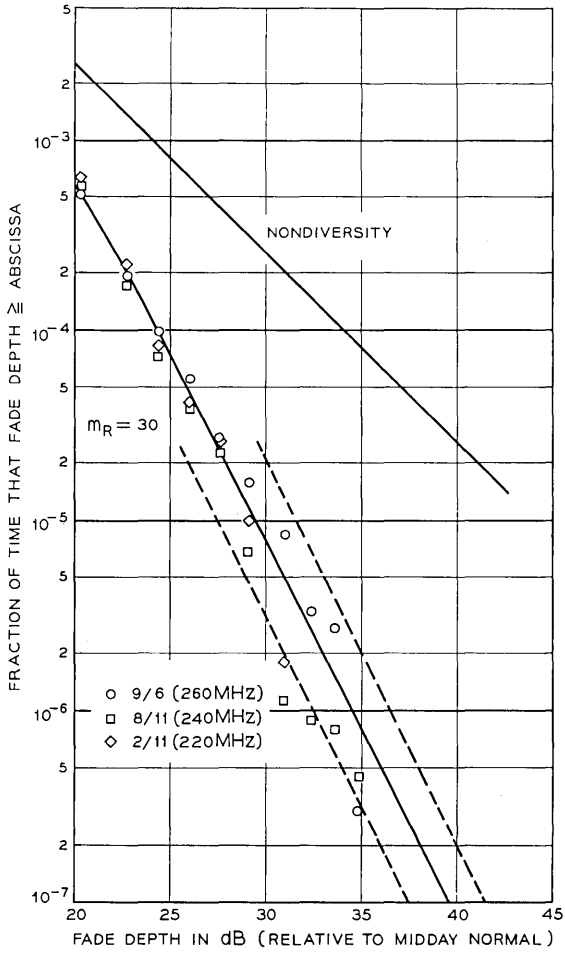


Fig. 22—4-GHz frequency diversity; 220/240/260-MHz separation.

depends on frequency, path length, terrain, antenna placement, and climate. The last of these determines the fraction of the total time that fading conditions exist on a given path. The periods used in analysis were those for which fading conditions were in existence. Any change in the total time of such fading periods would have no effect on the statistics since they pertain to the fading phenomena and not to the length of time (assuming an adequate number of samples are available). However, the statistics have been normalized by adding in the remaining

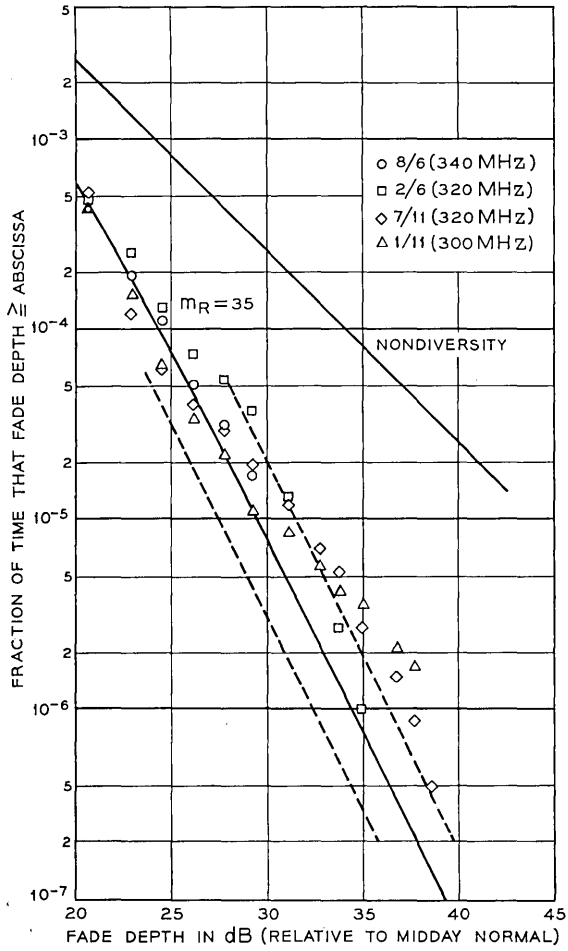


Fig. 23—4-GHz frequency diversity; 300/320/340-MHz separation.

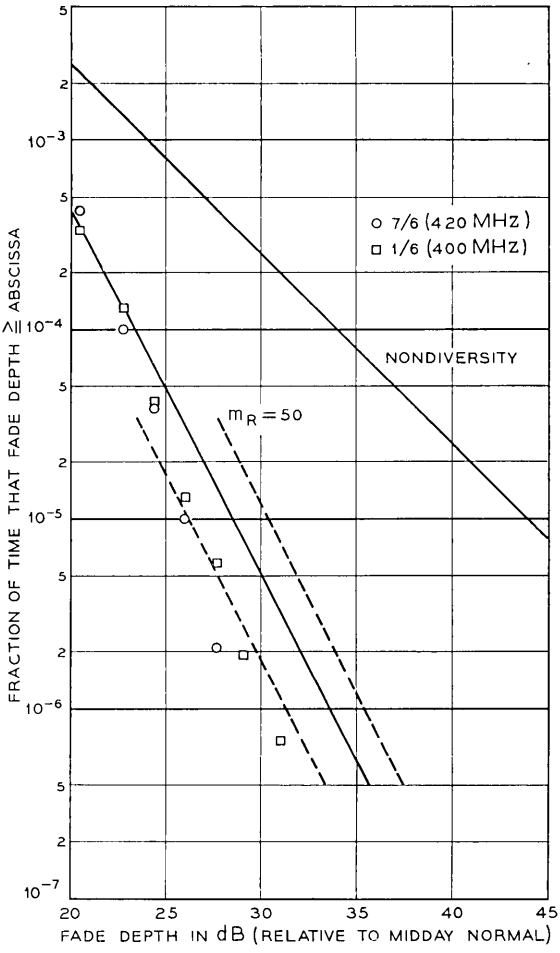


Fig. 24—4-GHz frequency diversity; 400/420-MHz separation.

or nonfading time. The effect of this or any like change in the amount of the nonfading time is a uniform shift in the nondiversity and diversity curves without changing their shape or their ratio; that is, the fractional time scale is multiplied by a constant. This last fact has been heavily utilized in the analysis where this ratio has been called the improvement factor (I). Note that the improvement factor does not depend on how often fading conditions exist but rather upon what happens within these selective fading periods.

TABLE VII—4-GHz FREQUENCY DIVERSITY RESULTS

Figure	Frequency Separation (MHz)	Number of Pairs
17	20*	2
18	60	2
19	80	3
20	100	2
21	140	1
	160	2
22	220	1
	240	1
	260	1
23	300	1
	320	2
	340	1
24	400	1
	420	1

* This is also the nominal bandwidth of the working channel.

Referring to the asymptotic forms for the joint Rayleigh model, equations (6) and (7), the asymptotic form of the improvement factor (I) for Rayleigh fading can be stated as

$$I = \frac{\Pr(R_1 < L)}{\Pr(R_1 < L, R_2 < L)} = \frac{(m_R/10^3)}{\Pr(R_1 < L)} \tag{8}$$

where, for the time being, it is assumed that both signals have the same rms value (that is, $v^2 = 1$).

The experimental improvement factors were obtained from the ratio between the fitted diversity line and the nondiversity lines for the 6-GHz and 4-GHz frequency pairs at a 40-dB fade depth. The values are plotted on Fig. 26 versus the parameter $\Delta f/f$. Here f is taken as 3950 MHz for the 4-GHz band and 6175 MHz for the 6-GHz band and Δf is the average frequency separation for a grouping on a single figure, for example, $\Delta f = 240$ MHz for Fig. 22. If the ± 2-dB uncertainty were included, the points plotted on Fig. 26 would change to vertical lines between 1.58 and 1/1.58 of the average value shown. Even with this large range of uncertainty, it appears that the improvement and $\Delta f/f$ are linearly related as shown by the lines on the figure. The equations of the lines are

$$\left.\begin{aligned} &4\text{ GHz: } I = \frac{1}{2}\left(\frac{\Delta f}{f}\right)L^{-2} \\ &6\text{ GHz: } I = \frac{1}{4}\left(\frac{\Delta f}{f}\right)L^{-2} \end{aligned}\right\} \text{ for } I \geqq 10, \quad \text{good accuracy;} \tag{9}$$

$$\text{for } 1 \leqq I \leqq 10, \text{ less accurate but conservative;}$$

where $F = -20 \log L$ is the fade depth in dB. This is the asymptotic form of the formulas including the variation with fade depth as shown in equation (8).

Using equation (8) as a guide, it is conjectured that the experimental improvement can be separated into two parts which contain respectively the nondiversity fading and the frequency diversity effect, that is,

$$I = \frac{m/10^3}{P(L)} \tag{10}$$

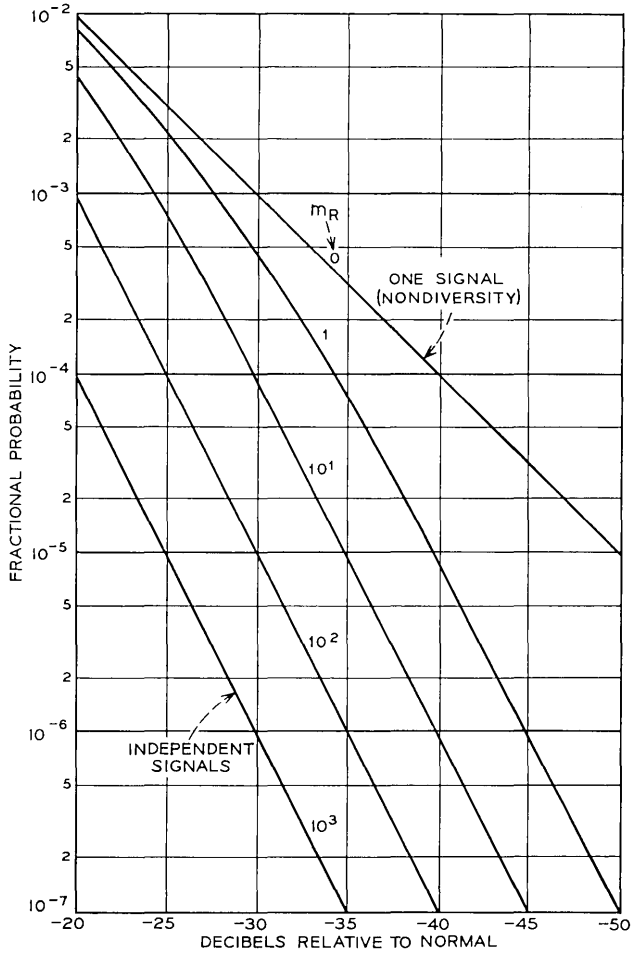


Fig. 25—Probability that both signals are simultaneously less than a given amount (Joint Rayleigh Distribution).

where $P(L)$ is the measured probability that a nondiversity fade exceeds $-20 \log L$ dB and m is a frequency diversity parameter. Of course both these quantities are functions of frequency, path geometry, terrain, and antenna placement.

Consider first the variation of m with $\Delta f/f$ and secondly the difference in improvement between the 4-GHz and 6-GHz bands.

The nondiversity results $[P(L)]$ can be written as (see Fig. 9)

$$\begin{aligned} &6 \text{ GHz:} \quad P_6 = (.53)L^2, \\ &4 \text{ GHz:} \quad P_4 = (.25)L^2, \end{aligned} \tag{11}$$

where $F = -20 \log L$ is the fade depth in dB. Then from equations (9) and (10)

$$\begin{aligned} 6 \text{ GHz:} \quad m_6 &= (10^3)\left(\frac{1}{4}\right)\left(\frac{\Delta f}{f}\right)L^{-2}(.53)L^2, \\ &= 132.5\,\frac{\Delta f}{f}\,; \end{aligned} \tag{12}$$

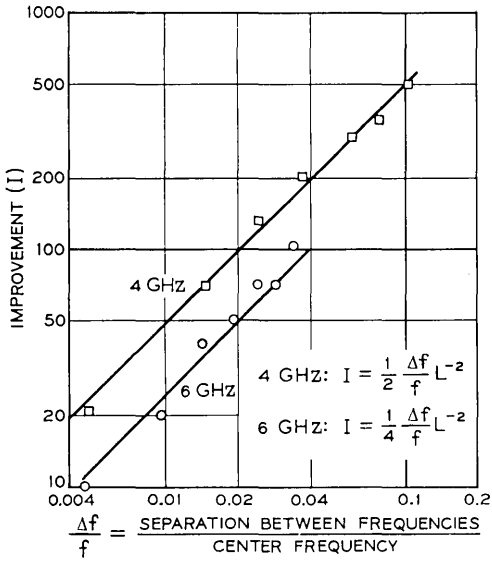


Fig. 26—1966 West Unity in-band frequency diversity improvement ratio at 40-dB Fade Depth ($L = 0.01$).

$$\begin{aligned} 4 \text{ GHz:} \quad m_4 &= (10^3)\left(\frac{1}{2}\right)\left(\frac{\Delta f}{f}\right)L^{-2}(.25)L^2, \\ &= 125\,\frac{\Delta f}{f}\,. \end{aligned} \tag{13}$$

The difference between m_6 and m_4 is small. Thus as desired m depends primarily upon normalized frequency spacing and not upon either the nondiversity fading distribution or the radio frequency band (4 vs 6 GHz). For further use it is assumed that $m = 130\ \Delta f/f$.

Using equations (9) and (10) again and forming a ratio gives

$$\frac{I_4}{I_6} = \frac{P_6}{P_4} = 2.1 \tag{14}$$

which agrees very closely with the experimental ratio of 2 shown on Fig. 26. Thus equation (10) correctly predicts the relative improvement between the 6- and 4-GHz bands. Further this relative improvement depends upon the nondiversity fading results and not upon the normalized frequency spacing.

To recapitulate, the asymptotic value of improvement of an in-band frequency diversity pair relative to the nondiversity signal at a fade depth of $-20 \log L$ dB can be stated for the experimental data as

$$I = \frac{0.13\,\dfrac{\Delta f}{f}}{P(L)} \tag{15}$$

where $P(L)$ is the probability that the nondiversity signal fades below the given depth. In this formula, I is not affected by the relative amount of time that fading conditions do or do not exist. However both the numerator and denominator in equation (15) would change by the same multiplicative constant when the ratio of nonfading to fading time changes. Thus the terms $P(L)$ and $0.13\ \Delta f/f$ individually apply only to the experimental path but their ratio is more generally useful.

This ratio (I) characterizes frequency diversity during multipath fading periods. Although I was obtained from experimental data on one path, it should pertain to other paths of about the same length but having different terrain and climate. The terrain and climate play a major role in determining the fraction of time that multipath fading conditions exist but they probably will have only a secondary effect on the relation between a nondiversity signal and a diversity signal within a multipath fading period.

IX. CROSSBAND FREQUENCY DIVERSITY

Results were also obtained for a subset of the 4-GHz and 6-GHz channels where the diversity pair consists of one channel from each group. The channels used for analysis were 4-2, 4-1, 4-7, 4-6 and 6-11, 6-15, 6-18. The results are given in Figs. 27 through 30. The groupings for each figure are for one of the 4-GHz channels in diversity with each of the 6-GHz channels. As before, there are several curves on each figure. The two uppermost are the average nondiversity results for each band with the 6 GHz being 3.3 dB poorer than the 4 GHz for a fixed probability.

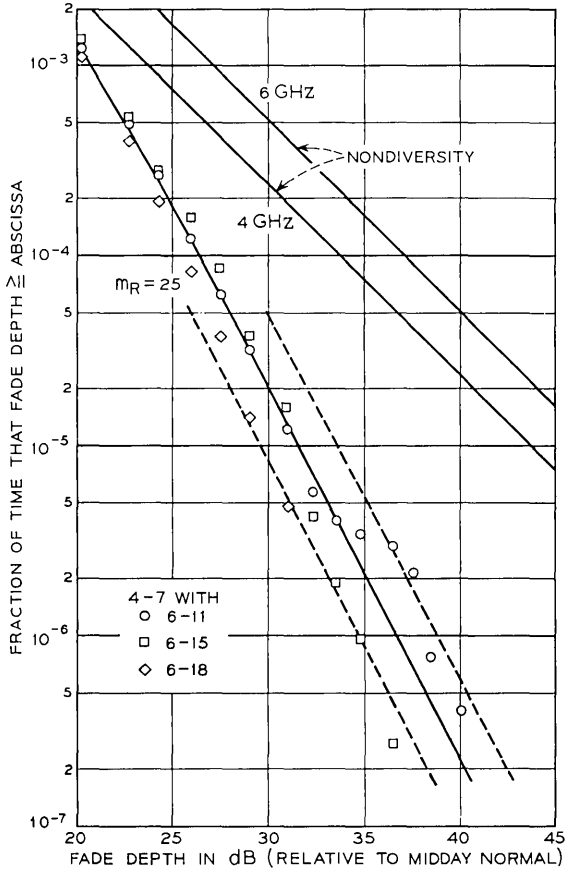


Fig. 27—4/6-GHz crossband frequency diversity.

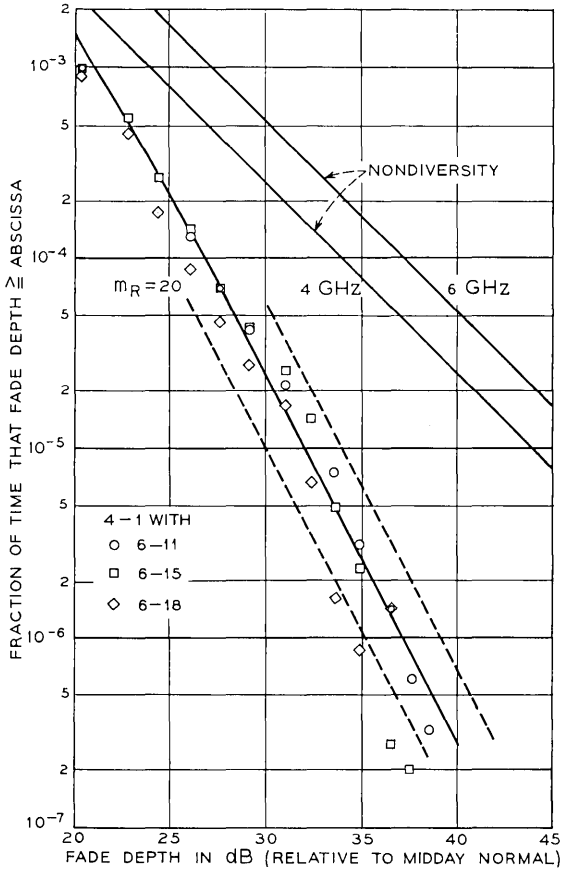


Fig. 28—4/6-GHz crossband frequency diversity.

Again the joint Rayleigh distribution was fitted to the data by overlaying plots of the joint distribution for various values of m_R. In this case the rms values are unequal by an amount

$$-10 \log v^2 = 3.3 \text{ dB} \tag{16}$$

or

$$v^2 = 0.47.$$

The asymptotic form of the improvement factor I between the diversity curve and the top nondiversity curve (6 GHz) is given as in equation (8)

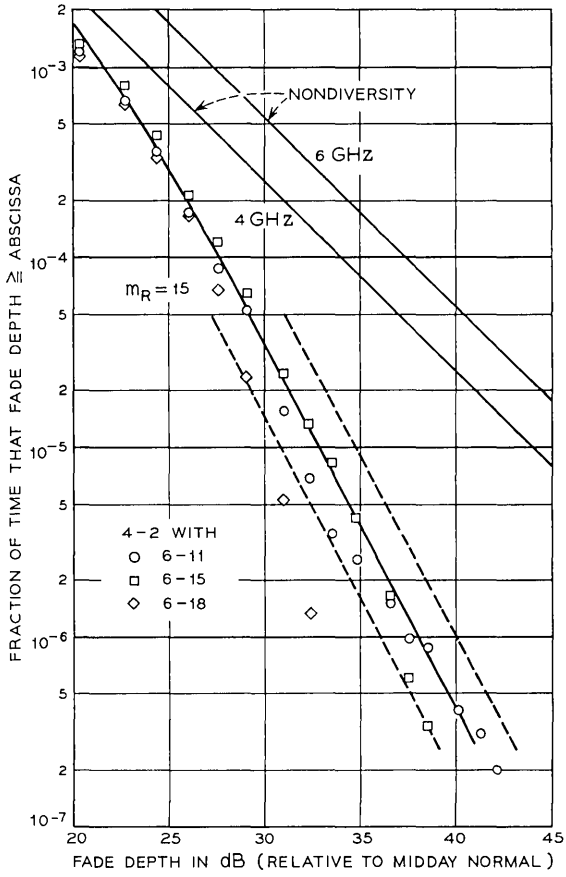


Fig. 29—4/6-GHz crossband frequency diversity.

by

$$I_{\max} = \frac{m_R/10^3}{\Pr\,(R_1 < L)}. \tag{17}$$

This corresponds to the improvement obtained if a 4-GHz channel were used to protect a 6-GHz channel.

The asymptotic ratio between the bottom nondiversity curve (4 GHz) and the diversity curve is then

$$I_{\min} = v^2 \frac{m_R/10^3}{\Pr\,(R_1 < L)} = v^2 I_{\max} \tag{18}$$

which corresponds to the improvement obtained if a 6-GHz channel were used to protect a 4-GHz channel. In these formulas, m_R and Pr $(R_1 < L)$ are Rayleigh quantities with m_R related to the correlation coefficient and $-20 \log L$ equal to the fade depth in dB exceeded by the envelope voltage R_1.

Inspection of the results shows that the points have more scatter than the 6-GHz in-band diversity data and just about the same scatter as the 4-GHz in-band diversity data, that is, the fitted line is a good representation of the data from 20 to 30 dB with increasing divergence for greater fade depths.

As to quantitative interpretation, the results do not appear to be as

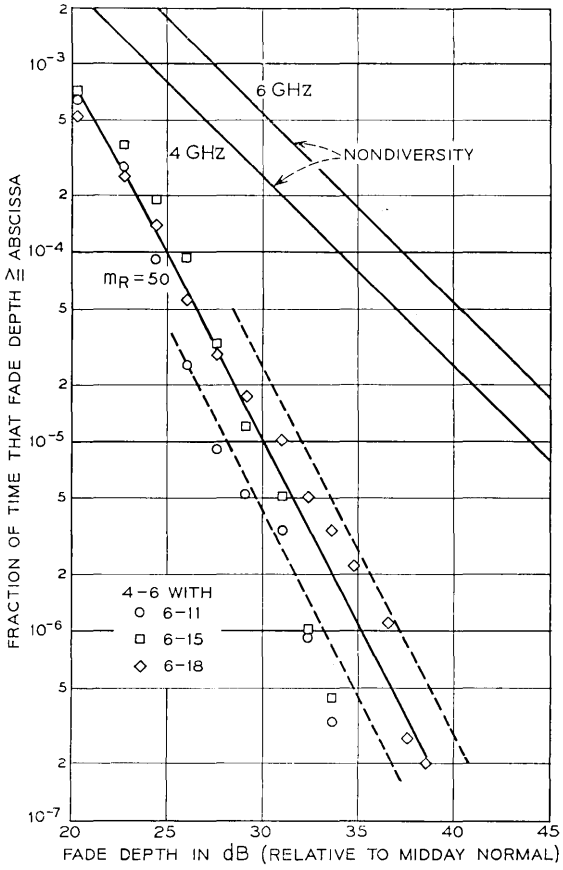


Fig. 30—4/6-GHz crossband frequency diversity.

yielding to analysis as the in-band diversity. This may result because the frequency spacings are a significant percentage of the average center frequency, for example, spacings of 1775 MHz (4-6/6-11) to 2405 MHz (4-7/6-18). The improvement values obtained from the fitted lines on the figures at 40-dB fade depth are presented in Table VIII.

The results do not show a consistent behavior as a function of frequency separation. 4-7, which has the largest frequency separation, shows slightly more improvement than 4-1 or 4-2 but less than 4-6 which has the smallest frequency separation. However, 4-2, 4-1, 4-7 are tightly bunched in frequency whereas 4-6 is about 300 MHz closer to the 6-GHz band.

In any case, these results are comparable to the in-band diversity results, that is, the improvement from crossband diversity was not significantly better than in-band diversity of two percent or more separation. Thus there may be a saturation effect which will appear for frequency separations above say 10 percent. There is neither enough data nor a theory to prove or disprove such speculation.

X. CROSS ROUTE DIVERSITY

Diversity results were obtained for various 4-GHz and 6-GHz channels on the Pleasant Lake hop in diversity with the single 4-GHz channel measured on the Paulding hop. The previous data strongly implies that it may be very misleading to rely on the results for a single channel. However, this data is included for completeness. To review: the Paulding data is for a *different* path but for the same time periods. One would therefore expect the diversity performance to be very good since the signals from the pair of paths should be reasonably independent. However this did not appear to be the case.

The data are shown in Figs. 31 and 32 in the groupings presented in Table IX. The lines on the figure have exactly the same meaning as the corresponding ones in the crossband section. In this case the 6-GHz fit

TABLE VIII—CROSSBAND IMPROVEMENT VALUES

	I_{max}	I_{min}*
Fig. 27	250	125
Fig. 28	200	100
Fig. 29	150	75
Fig. 30	500	250

* $I_{min} = v^2 I_{max}$ from equation (16) with $v = 0.47$ but 0.5 has been used in this table for convenience.

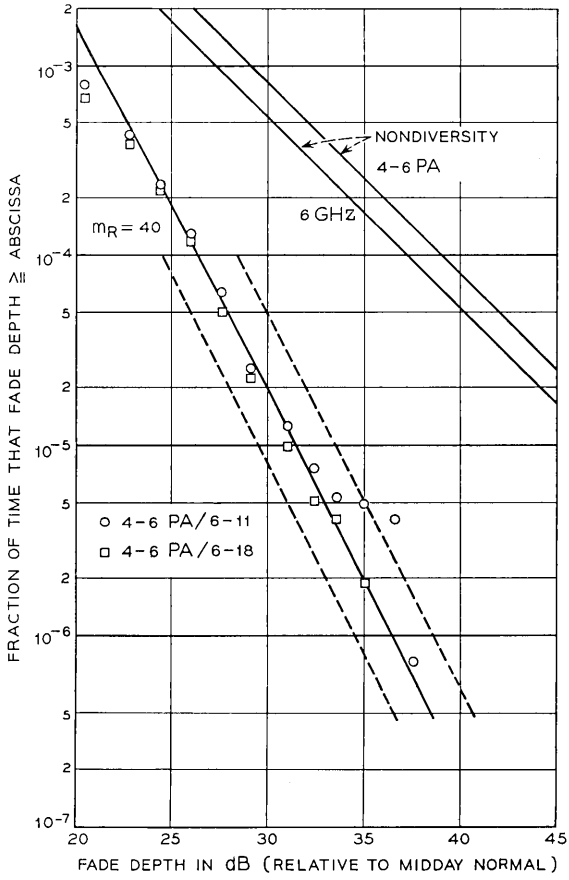


Fig. 31—4/6-GHz cross route diversity.

is good but the 4-GHz fit is poor below 30 dB in that the data has an upward bulge. There is no explanation available for this anomaly.

In any case, the improvement obtained when the two channels in the diversity pair are on different hops is not significantly better than in-band diversity (see Fig. 26). This is surprising and raises questions about the correlation between fading on adjacent hops, for example, the maximum possible diversity improvement may be limited to values less than that expected from independent fading.

To repeat, this is based on a single channel and as such the data base

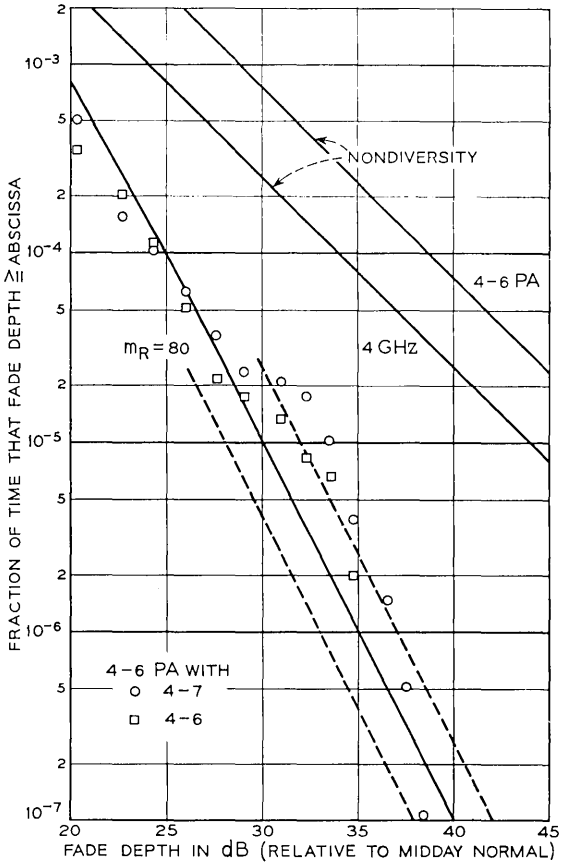


Fig. 32—4-GHz cross route diversity.

TABLE IX—CROSS-ROUTE RESULTS

Figure	Diversity Pairs	I_{max}	I_{min}*
31	6-11 with 4-6 PA 6-18 with 4-6 PA	400	300
32	4- 7 with 4-6 PA 4- 6 with 4-6 PA	800	250

* $I_{min} = v^2 I_{max}$
4/4-PA, $v^2 = 0.322$ (used 0.31)
6/4-PA, $v^2 = 0.715$ (used 0.75)

is simply not sufficient to draw any profound conclusions about cross-route effects.

XI. COMPARISON OF FREQUENCY AND SPACE DIVERSITY IMPROVEMENT

The empirical results for space diversity given in Vigants[10] can be compared with those obtained here for frequency diversity. The improvement factor for space diversity is

$$I_{SD} = \frac{s^2}{2.75D\lambda} 10^{F/10} \tag{19}$$

where

s is vertical separation between equal antennas in feet,
D is path length in feet,
λ is wavelength in feet, and
F is fade depth in dB.

Using $D = 28.5$ miles and equation (9) gives the various diversity improvement factors as presented in Table X.

These are plotted in Fig. 33 for a fade depth of 40 dB. Several points are immediate. First the improvement increases with frequency for space diversity and decreases with frequency for frequency diversity; that is, space diversity becomes relatively more effective as the operating frequency increases. The maximum improvement for frequency diversity is 100 for the maximum allowable spacing of 4 percent in the standard 6-GHz frequency plans. Space diversity of 26 feet will give this improvement. Since this spacing is reasonable, it can be said that space diversity is "better" than frequency diversity at 6 GHz. At 4 GHz, the corresponding values are $I = 625$ for 12.5 percent and 79′ spacings. In this case, frequency and space diversity are comparable in performance.

These comparisons have been made only for one-for-one space and frequency diversity on a single hop; additional data and studies are needed to clarify our understanding.

TABLE X—DIVERSITY IMPROVEMENT FACTORS

	4 GHz	*6 GHz*
Frequency	$0.5(\Delta f/f)10^{F/10}$	$0.25(\Delta f/f)10^{F/10}$
Space	$(s^2/10^5)10^{F/10}$	$1.5(s^2/10^5)10^{F/10}$

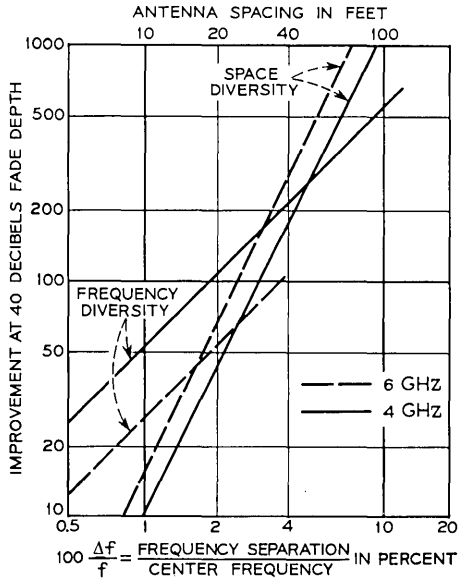


Fig. 33—Comparison of space and frequency diversity at a 40-dB fade depth.

XII. ACKNOWLEDGMENTS

The author is indebted to colleagues at Bell Telephone Laboratories for the experimental data and for many discussions. In particular the MIDAS is the creation of G. A. Zimmerman. The tabulated data used here was extracted from the raw data through the skilled labors of C. H. Menzel. There were many fruitful discussions with A. Vigants, and the interest and support of K. Bullington were invaluable.

REFERENCES

1. Kerr, D. E., *The Propagation of Short Radio Waves*, MIT Radiation Lab. series No. 13, New York: McGraw-Hill, 1951.
2. Crawford, A. B., and Jakes, W. C., "Selective Fading of Microwaves," B.S.T.J., *31*, No. 1 (January 1952), pp. 68–90.
3. De Lange, O. E., Propagation Studies at Microwave Frequencies by Means of Very Short Pulses," B.S.T.J., *31*, No. 1 (January 1952), pp. 91–103.
4. Sharpless, W. M., "Measurement of the Angle of Arrival of Microwaves," Proc. I.R.E., *34*, No. 11 (November 1946), pp. 837–845.
5. Crawford, A. B., and Sharpless, W. M., "Further Observations of the Angle of Arrival of Microwaves," Proc. I.R.E., *34*, No. 11 (November 1946), pp. 845–848.
6. Yonezawa, S., and Tanaka, N., *Microwave Communication*, Tokyo: Maruzen Co., Ltd., 1965, pp. 25–60.

7. Pearson, K. W., "Method for the Prediction of the Fading Performance of a Multisection Microwave Link," Proc. IEEE, *112*, No. 7 (July 1965), pp. 1291–1300.
8. Beckmann P., and Spizzichino, A., *The Scattering of Electromagnetic Waves from Rough Surfaces*, New York: Pergamon Press, 1963, pp. 355–367.
9. Kaylor, R. L., "A Statistical Study of Selective Fading of Super High Frequency Radio Signals," B.S.T.J., *32*, No. 5 (September 1953), pp. 1187–1202.
10. Vigants, A., "Space Diversity Performance as a Function of Antenna Separation," IEEE Trans. Comm. Tech., *COM-16*, No. 6 (December 1968), pp. 831–836.

Computed Transmission Through Rain at Microwave and Visible Frequencies

By DAVID E. SETZER

(Manuscript received May 5, 1970)

In this paper we present tables which contain the Mie scattering coefficient, absorption coefficient, extinction coefficient, equivalent medium index of refraction and phase delay for rains conforming to the Laws and Parsons drop-size distribution. These transmission characteristics have been calculated for microwave frequencies of interest in common carrier radio relay systems, 300 to 1.43 GHz, that is, 0.1 to 21.0 cm, at rain rates from 0.25 to 150.0 mm/hr. We also include the extinction coefficients for the visible wavelength 0.6328 μ.

The microwave tables were generated by using a Mie scattering computer program similar to that designed and previously reported by Deirmendjian. The calculations at 0.6328 μ were made separately by employing the usual assumptions for droplets with very large circumference to wavelength ratios.

I. INTRODUCTION

The Mie extinction properties are of basic importance to those interested in developing an understanding of the influence of rainfall on open air communication systems. In this connection we have generated a rather extensive set of tables of extinction properties of rain. The tables have been used within Bell Laboratories to study a variety of transmission problems, examples of which are the investigation of satellite ground station interference by Gusler and Hogg (1970),* the study of microwave transit time variations by Gray (1970), Pierce's (1969) investigation of the problems associated with the synchronization of digital networks and Setzer's (1969) study of the extinction properties of atmospheric aerosols.[1-4] A set of tables with similar results was published by Medhurst (1965); however, his presentation only includes

* The attenuation constants used by Gusler and Hogg were based on empirical data. The calculated values presented in this paper were used for comparison purposes only.

total attenuation.[5] Our tables include the Mie scattering coefficient, absorption coefficient, extinction coefficient, the van de Hulst equivalent medium index of refraction and the van de Hulst phase delay for rains conforming to the Laws and Parsons drop-size distribution. These transmission characteristics have been calculated for incident microwave wavelengths of 0.1, 0.2, 0.3, 0.5, 1.0, 1.62, 1.88, 2.73, 5.0, 7.5, 10.0, 15.0 and 21.0 cm (corresponding to 300, 150, 100, 60, 30, 18.5, 16, 11, 6, 4, 3, 2 and 1.43 GHz) at rain rates of 0.25, 1.25, 2.5, 5.0, 12.5, 50.0, 100.0 and 150.0 mm/hr. Also included are the extinction coefficients for the visible wavelength 0.6328 μ at the above rain rates.

The calculations in the microwave region were performed on a GE 635 computer using a scattering program similar to that previously presented by Deirmendgian (1963).[6] Since the raindrop circumference-to-wavelength ratio ($\pi d/\lambda$), that is, size parameter, for the visible wavelength, is outside the range of validity of the computer program, approximate characteristics were calculated for 0.6328 μ. The usual assumptions for spheres with very large parameters were employed.

The indices of refraction used in this report and shown in Table I are for a rain temperature of 20°C. They were obtained by cross checking many of the standard optical and microwave references and are thought to be reliable.

II. DROPLET SIZE DISTRIBUTION

All computations in this paper are based on the assumption that raindrops are spherical and the distribution of rain is as was measured by Laws and Parsons and quoted by Kerr (1951).[7] The Laws and Parsons distribution is presented in Table II as the percentage of total water volume within specific size ranges. In order to use the computer program, it is necessary to express the distribution in terms of the number of droplets per unit volume within specific size ranges. If the droplets are assumed to fall at the terminal velocity V_0, that is, up and down drafts are neglected, then the conversion is

$$D(d_{i+1}, d_1) \approx R_f \cdot P(d_{i+1}, d_i)/[V(\bar{d})V_0(\bar{d})], \tag{1}$$

where $D(d_{i+1}, d_1)$ represents the size distribution in units of droplets per unit volume in the droplet diameter range d_{i+1} to d_i. Henceforth, the diameter range d_{i+1} to d_i will be called Δd_i. R_f is the total rainfall rate which is typically specified in mm/hr; $P(d_{i+1}, d_i)$ is the volume percentage rainfall in the diameter range Δd_i as measured by Laws and Parsons; $\bar{d}$ is the average diameter in the range Δd_i; and $V(\bar{d})$

is the volume of a sphere of diameter $\bar{d}$. The terminal velocities of raindrops $V_0(\bar{d})$ are presented in Table III.[7]

For an example of the function $D(d_{i+1}, d_i)$ resulting from the use of equation (1), refer to Fig. 1.

III. TRANSMISSION PARAMETERS FOR MICROWAVE FREQUENCIES

The Mie coefficients and the equivalent index of refraction of the rain medium are defined by van de Hulst (1957).[8] For a detailed description of these parameters, please refer to his work. Essentially, the scattering coefficient β_{scat} (λ) and the absorption coefficient β_{abs} (λ) are measures of the total energy scattered and absorbed by a unit volume of rainfall. In the simple case of a single scattering aerosol the ratio of intensity of the transmitted beam $I_T(\lambda)$ to that of the incident beam $I_0(\lambda)$ is

$$I_T(\lambda)/I_0(\lambda) = \exp[-\beta_{ext}(\lambda) \cdot l], \tag{2}$$

where l is the length of the propagation path through the rain and the extinction coefficient $\beta_{ext}(\lambda)$ is

$$\beta_{ext}(\lambda) = \beta_{scat}(\lambda) + \beta_{abs}(\lambda). \tag{3}$$

A plane-parallel medium containing many scattering particles can be represented by a slab of homogeneous material having a complex refractive index $\tilde{m}$. Carefully note that this sort of representation can

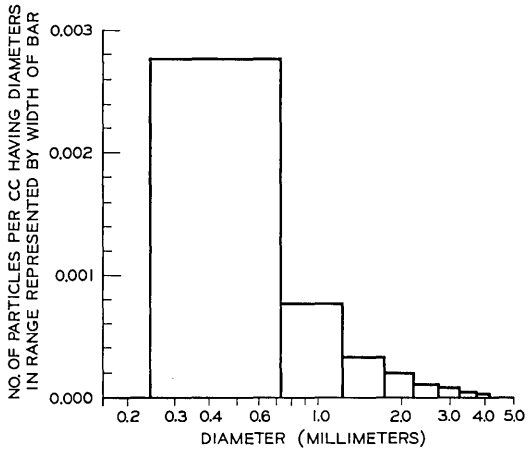


Fig. 1—Laws and Parsons drop-size distribution for 150 mm/hr rain.

be designed, by selecting the appropriate value of $\tilde{m}$, to preserve the input-output relationships but all detail of the scattering process within the medium is lost. According to the van de Hulst (1957) definition of $\tilde{m}$, the amplitude and phase of the incident wave are changed by the slab in the proportion

$$\exp[-(2\pi l/\lambda)\mathrm{Im}(1 - \tilde{m})] \cdot \exp[-(2\pi li/\lambda)\mathrm{Re}(\tilde{m} - 1)], \qquad (4)$$

where the first term is recognized as defining the amplitude ratio and the second the phase.[8] The values of β_{ext} , β_{abs} , β_{scat} , $\tilde{m}$ and the phase angle described above have been calculated for the specified microwave frequencies. The results appear in Tables IV through XVI. The reader is advised to use special care when attempting to apply the van de Hulst phase angle and medium index $\tilde{m}$. It is recommended that van de Hulst's derivation be studied carefully so that the meaning and limitations of these functions are well understood. For example, light reflected from the slab cannot be derived by using the refractive index $\tilde{m}$, but should be computed by means of the actual scattering functions.

Also, it should be noted that although $\tilde{m}$ is calculated, $(\tilde{m} - 1)$ is used to determine the phase angle. Since $\tilde{m}$ is very close to one, cancellation of the leading terms reduces the significant places in the numerical value of the phase angle to one or two at most. Consequently the values given in the phase change column of Tables IV through XVI exhibit noticeable discontinuous jumps.

IV. TRANSMISSION PARAMETERS FOR 0.6328 μ

The Mie coefficients β_i are defined as

$$\beta_i(\lambda) = \int_0^\infty \gamma_i(\lambda, r)n(r)\, dr, \qquad i = 1, 2, 3, \qquad (5)$$

where r is the droplet radius; $n(r)$ is the continuous size distribution, and $\gamma_i(\lambda, r)$, $i = 1, 2, 3$ are the extinction, scattering and absorption cross sections, respectively for droplets of radius r. The smallest ratio of raindrop circumference to wavelength for the combination of a Laws and Parsons rain and 0.6328 μ is approximately 1500. For most purposes, the laws of geometric optics can be applied in such cases and therefore

$$\gamma_{\mathrm{ext}}(\lambda, r) \approx 2\pi r^2. \qquad (6)$$

Also, since the index of refraction of water at 0.6328 μ is a real number, 1.33, the absorption coefficient will be zero. It follows from equations (3), (5) and (6) that

$$\beta_{\mathrm{ext}}(0.6328\mu) = \beta_{\mathrm{scat}}(0.6328\mu), \qquad (7)$$

$$\approx 2\pi \int_0^\infty r^2 n(r)\, dr. \tag{8}$$

This expression and the Laws and Parsons distribution were used to generate Table XVII. In this connection the Laws and Parsons distribution $D(d_{i+1}, d_i)$ was used to approximate the continuous function $n(r)$.

V. GRAPHICAL REPRESENTATION

For the purpose of illustration, a graph of extinction coefficients versus total water content and rain rate is included (see Fig. 2). Not all wavelengths are represented because some of the curves are too closely grouped in the neighborhood of those shown. Those that were excluded, were excluded for reasons of clarity only. One point of some interest is the location of the attenuation curve for 0.6328 μ in Fig. 2. Note that it represents a reversal of the trend exhibited as wavelength

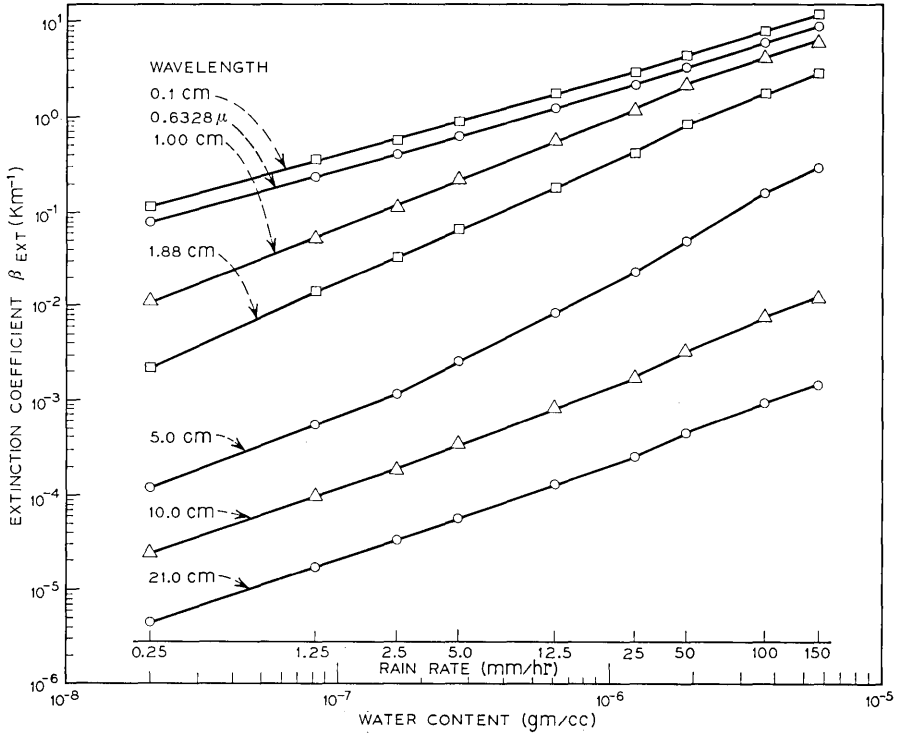


Fig. 2—Rainfall water content and rain rate versus extinction coefficient β_{ext}.

TABLE I—INDEX OF REFRACTION OF WATER AT 20°C

Wavelength (cm)	Index of Refraction
0.10	2.587 - 0.937(i)
0.20	3.039 - 1.575(i)
0.30	3.505 - 2.007(i)
0.50	4.364 - 2.521(i)
1.00	5.900 - 2.900(i)
1.62	7.001 - 2.544(i)
1.88	7.500 - 2.500(i)
2.73	8.070 - 1.990(i)
5.00	8.670 - 1.202(i)
7.50	8.770 - 0.915(i)
10.00	8.871 - 0.628(i)
15.00	8.916 - 0.422(i)
21.00	9.000 - 0.275(i)
0.6328 μ	1.33 - 0.0(i)

TABLE II— LAWS AND PARSONS DROP-SIZE DISTRIBUTIONS FOR VARIOUS PRECIPITATION RATES

Drop Diameter (cm)	Rain Rate (mm/hour)								
	0.25	1.25	2.5	5	12.5	25	50	100	150
	Percent of Total Volume								
0.05	28.0	10.9	7.3	4.7	2.6	1.7	1.2	1.0	1.0
0.1	50.1	37.1	27.8	20.3	11.5	7.6	5.4	4.6	4.1
0.15	18.2	31.3	32.8	31.0	24.5	18.4	12.5	8.8	7.6
0.2	3.0	13.5	19.0	22.2	25.4	23.9	19.9	13.9	11.7
0.25	0.7	4.9	7.9	11.8	17.3	19.9	20.9	17.1	13.9
0.3		1.5	3.3	5.7	10.1	12.8	15.6	18.4	17.7
0.35		0.6	1.1	2.5	4.3	8.2	10.9	15.0	16.1
0.4		0.2	0.6	1.0	2.3	3.5	6.7	9.0	11.9
0.45			0.2	0.5	1.2	2.1	3.3	5.8	7.7
0.5				0.3	0.6	1.1	1.8	3.0	3.6
0.55					0.2	0.5	1.1	1.7	2.2
0.6						0.3	0.5	1.0	1.2
0.65							0.2	0.7	1.0
0.7									0.3

TABLE III—RAINDROP TERMINAL VELOCITY

Radius, cm	Velocity, m/sec
0.025	2.1
0.05	3.9
0.075	5.3
0.10	6.4
0.125	7.3
0.15	7.9
0.175	8.35
0.2	8.70
0.225	9.0
0.25	9.2
0.275	9.35
0.30	9.5
0.325	9.6

TABLE IV—MIE EXTINCTION PARAMETERS AT 0.1 CM WAVELENGTH (300 GHz), H_2O INDEX OF REFRACTION 2.587—0.937i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.05390	0.05878	0.1127	0.0×10^{-6}	0.1051×10^{-7}	0.0
1.25	0.1705	0.1723	0.3428	0.0×10^{-6}	0.3051×10^{-7}	0.0
2.5	0.2760	0.2693	0.5452	0.0×10^{-6}	0.4763×10^{-7}	0.0
5.0	0.4550	0.4306	0.8856	0.0×10^{-6}	0.7630×10^{-7}	0.0
12.5	0.8913	0.8133	1.705	0.0×10^{-6}	1.451×10^{-7}	0.0
25.0	1.452	1.284	2.736	0.0×10^{-6}	2.305×10^{-7}	0.0
50.0	2.270	1.914	4.187	0.0×10^{-6}	3.471×10^{-7}	0.0
100.0	3.993	3.354	7.347	0.0×10^{-6}	6.109×10^{-7}	0.0
150.0	5.636	4.730	10.37	0.0×10^{-6}	8.636×10^{-7}	0.0

TABLE V—MIE EXTINCTION PARAMETERS AT 0.2 CM WAVELENGTH (150 GHz), H_2O INDEX OF REFRACTION 3.039—1.575i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.05581	0.05445	0.1103	0.0×10^{-6}	0.2276×10^{-7}	0.0
1.25	0.1828	0.1685	0.3514	0.0×10^{-6}	0.6657×10^{-7}	0.0
2.5	0.2991	0.2686	0.5677	0.0×10^{-6}	1.042×10^{-7}	0.0
5.0	0.4965	0.4349	0.9314	0.0×10^{-6}	1.671×10^{-7}	0.0
12.5	0.9766	0.8283	1.805	0.0×10^{-6}	3.176×10^{-7}	0.0
25.0	1.596	1.315	2.911	0.0×10^{-6}	5.044×10^{-7}	0.0
50.0	2.512	1.988	4.500	0.0×10^{-6}	7.592×10^{-7}	0.0
100.0	4.406	3.449	7.856	0.1×10^{-6}	13.33×10^{-7}	18.0
150.0	6.212	4.846	11.06	0.1×10^{-6}	18.83×10^{-7}	18.0

TABLE VI—MIE EXTINCTION PARAMETERS AT 0.3 CM WAVELENGTH (100 GHz), H_2O INDEX OF REFRACTION 3.505—2.007i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.04252	0.04991	0.09243	0.0×10^{-6}	0.2886×10^{-7}	0.0
1.25	0.1544	0.1586	0.3130	0.0×10^{-6}	0.8927×10^{-7}	0.0
2.5	0.2634	0.2555	0.5189	0.1×10^{-6}	1.432×10^{-7}	12.0
5.0	0.4520	0.4175	0.8695	0.1×10^{-6}	2.342×10^{-7}	12.0
12.5	0.9211	0.8026	1.723	0.1×10^{-6}	4.542×10^{-7}	12.0
25.0	1.542	1.283	2.825	0.2×10^{-6}	7.320×10^{-7}	24.0
50.0	2.502	1.957	4.459	0.2×10^{-6}	11.27×10^{-7}	24.0
100.0	4.382	3.389	7.770	0.4×10^{-6}	19.72×10^{-7}	48.0
150.0	6.153	4.751	10.90	0.6×10^{-6}	27.76×10^{-7}	72.0

TABLE VII—MIE EXTINCTION PARAMETERS AT 0.5 CM WAVELENGTH (60 GHz), H_2O INDEX OF REFRACTION 4.364—2.521i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.01638	0.02856	0.04493	0.0×10^{-6}	0.2281×10^{-7}	0.0
1.25	0.08590	0.1085	0.1945	0.1×10^{-6}	0.9094×10^{-7}	7.2
2.5	0.1667	0.1876	0.3544	0.1×10^{-6}	1.608×10^{-7}	7.2
5.0	0.3157	0.3236	0.6393	0.2×10^{-6}	2.832×10^{-7}	14.4
12.5	0.7145	0.6574	1.372	0.4×10^{-6}	5.924×10^{-7}	28.8
25.0	1.279	1.089	2.368	0.6×10^{-6}	10.06×10^{-7}	43.3
50.0	2.233	1.743	3.977	0.8×10^{-6}	16.59×10^{-7}	57.6
100.0	3.934	2.999	6.933	1.3×10^{-6}	28.89×10^{-7}	93.6
150.0	5.505	4.165	9.670	1.8×10^{-6}	41.65×10^{-7}	129.6

TABLE VIII—MIE EXTINCTION PARAMETERS AT 1.0 CM WAVELENGTH (30 GHz), H_2O INDEX OF REFRACTION 5.9—2.9i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.001459	0.009006	0.01046	0.0×10^{-6}	0.01046×10^{-7}	0.0
1.25	0.01303	0.04392	0.05695	0.1×10^{-6}	0.5248×10^{-7}	3.6
2.5	0.03112	0.08465	0.1158	0.2×10^{-6}	1.038×10^{-7}	7.2
5.0	0.07387	0.1617	0.2355	0.4×10^{-6}	2.066×10^{-7}	14.4
12.5	0.2210	0.3751	0.5961	0.8×10^{-6}	5.106×10^{-7}	28.8
25.0	0.4939	0.6890	1.183	1.4×10^{-6}	9.978×10^{-7}	50.5
50.0	1.071	1.234	2.305	2.1×10^{-6}	19.16×10^{-7}	75.6
100.0	2.184	2.207	4.397	3.6×10^{-6}	36.26×10^{-7}	130.0
150.0	3.280	3.106	6.386	4.9×10^{-6}	52.58×10^{-7}	177.0

TABLE IX—MIE EXTINCTION PARAMETERS AT 1.62 CM WAVELENGTH (18.5 GHz), H_2O INDEX OF REFRACTION 7.001—2.544i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m}-1)$	$Im(1-\tilde{m})$	
0.25	0.0001932	0.002970	0.003162	0.0×10^{-6}	0.05064×10^{-7}	0.0
1.25	0.002003	0.01814	0.02015	0.1×10^{-6}	0.2982×10^{-7}	2.2
2.5	0.005166	0.03855	0.04372	0.3×10^{-6}	0.6316×10^{-7}	6.6
5.0	0.01373	0.08067	0.09440	0.5×10^{-6}	1.336×10^{-7}	11.1
12.5	0.04672	0.2093	0.2560	1.0×10^{-6}	3.544×10^{-7}	22.2
25.0	0.1198	0.4172	0.5370	1.7×10^{-6}	7.320×10^{-7}	37.8
50.0	0.3051	0.8111	1.116	2.9×10^{-6}	15.01×10^{-7}	64.5
100.0	0.7365	1.525	2.262	5.2×10^{-6}	30.19×10^{-7}	115.5
150.0	1.209	2.210	3.420	7.4×10^{-6}	45.49×10^{-7}	164.4

TABLE X—MIE EXTINCTION PARAMETERS AT 1.88 CM WAVELENGTH (16 GHz), H_2O INDEX OF REFRACTION 7.5—2.5i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.0001035	0.002018	0.002121	0.0×10^{-6}	0.03934×10^{-7}	0.0
1.25	0.001083	0.01299	0.01407	0.1×10^{-6}	0.2410×10^{-7}	1.91
2.5	0.002821	0.02836	0.03118	0.3×10^{-6}	0.5216×10^{-7}	5.74
5.0	0.007659	0.06135	0.06901	0.5×10^{-6}	1.131×10^{-7}	9.56
12.5	0.02667	0.1661	0.1928	1.0×10^{-6}	3.094×10^{-7}	19.1
25.0	0.07037	0.3422	0.4126	1.8×10^{-6}	6.524×10^{-7}	34.4
50.0	0.1853	0.6849	0.8702	3.1×10^{-6}	13.58×10^{-7}	59.2
100.0	0.4643	1.308	1.773	5.5×10^{-6}	27.44×10^{-7}	105.5
150.0	0.7701	1.907	2.678	7.9×10^{-6}	41.31×10^{-7}	151.0

TABLE XI—MIE EXTINCTION PARAMETERS AT 2.73 WAVELENGTH (11 GHz), H_2O INDEX OF REFRACTION 8.07—1.99i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.00002176	0.0006630	0.0006630	0.0×10^{-6}	0.01786×10^{-7}	0.0
1.25	0.0002257	0.004325	0.004550	0.1×10^{-6}	0.1123×10^{-7}	1.3
2.5	0.0005954	0.009876	0.01047	0.3×10^{-6}	0.2525×10^{-7}	4.0
5.0	0.001664	0.02377	0.02543	0.5×10^{-6}	0.6008×10^{-7}	6.6
12.5	0.005980	0.07349	0.07947	1.1×10^{-6}	1.844×10^{-7}	14.5
25.0	0.01652	0.1725	0.1890	1.9×10^{-6}	4.330×10^{-7}	25.0
50.0	0.04603	0.3936	0.4396	3.4×10^{-6}	9.956×10^{-7}	44.8
100.0	0.1227	0.8482	0.9710	6.3×10^{-6}	21.80×10^{-7}	83.0
150.0	0.2057	1.303	1.508	8.9×10^{-6}	33.72×10^{-7}	118.7

TABLE XII—MIE EXTINCTION PARAMETERS AT 5.0 CM WAVELENGTH (6 GHz), H_2O INDEX OF REFRACTION 8.670—1.202i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.000001855	0.0001138	0.0001156	0.0×10^{-6}	0.05728×10^{-8}	0.0
1.25	0.00001769	0.0005516	0.0005692	0.1×10^{-6}	0.2604×10^{-8}	0.7
2.5	0.00004546	0.001138	0.001183	0.3×10^{-6}	0.5262×10^{-8}	2.2
5.0	0.0001254	0.002589	0.002714	0.5×10^{-6}	1.177×10^{-8}	3.6
12.5	0.0004493	0.007932	0.008380	1.0×10^{-6}	3.554×10^{-8}	7.2
25.0	0.001290	0.02076	0.02205	1.9×10^{-6}	9.200×10^{-8}	13.7
50.0	0.003760	0.05599	0.05975	3.5×10^{-6}	24.63×10^{-8}	25.2
100.0	0.01065	0.1509	0.1615	6.8×10^{-6}	66.19×10^{-8}	49.0
150.0	0.01777	0.2542	0.2720	10.1×10^{-6}	111.4×10^{-8}	80.0

TABLE XIII—MIE EXTINCTION PARAMETERS AT 7.5 CM WAVELENGTH (4 GHz), H_2O INDEX OF REFRACTION 8.77—0.915i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.0000003639	0.00004853	0.00004889	0.0×10^{-6}	0.03631×10^{-8}	0.0
1.25	0.000003433	0.0002078	0.0002112	0.1×10^{-6}	0.1456×10^{-8}	0.5
2.5	0.000008733	0.0003957	0.0004044	0.3×10^{-6}	0.2712×10^{-8}	1.4
5.0	0.00002337	0.0007859	0.0008092	0.4×10^{-6}	0.5307×10^{-8}	1.9
12.5	0.00008091	0.002002	0.002083	1.0×10^{-6}	1.335×10^{-8}	4.8
25.0	0.0002185	0.004256	0.004474	1.8×10^{-6}	2.823×10^{-8}	8.6
50.0	0.0006030	0.009508	0.01011	3.3×10^{-6}	6.285×10^{-8}	15.8
100.0	0.001629	0.02251	0.02414	6.4×10^{-6}	14.90×10^{-8}	30.7
150.0	0.002705	0.03549	0.03820	9.5×10^{-6}	23.54×10^{-8}	45.6

TABLE XIV—MIE EXTINCTION PARAMETERS AT 10.0 CM WAVELENGTH (3 GHz), H_2O INDEX OF REFRACTION 8.871—0.6280i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m} - 1)$	$Im(1 - \tilde{m})$	
0.25	0.0000001149	0.00002309	0.00002320	0.0×10^{-6}	0.02296×10^{-8}	0.0
1.25	0.000001081	0.00009474	0.00009582	0.1×10^{-6}	0.08815×10^{-8}	0.36
2.5	0.000002746	0.0001758	0.0001786	0.2×10^{-6}	0.1599×10^{-8}	0.72
5.0	0.000007326	0.0003369	0.0003443	0.4×10^{-6}	0.3017×10^{-8}	1.44
12.5	0.00002528	0.0008147	0.0008400	1.0×10^{-6}	0.7200×10^{-8}	3.6
25.0	0.00006790	0.001627	0.001696	1.8×10^{-6}	1.430×10^{-8}	6.5
50.0	0.0001857	0.003336	0.003521	3.3×10^{-6}	2.924×10^{-8}	11.9
100.0	0.0004946	0.007142	0.007637	6.2×10^{-6}	6.302×10^{-8}	22.3
150.0	0.0008220	0.01099	0.01181	9.2×10^{-6}	9.729×10^{-8}	33.2

TABLE XV—MIE EXTINCTION PARAMETERS AT 15.0 CM WAVELENGTH (2 GHz), H_2O INDEX OF REFRACTION 8.916—0.4220i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m}-1)$	$Im(1-\tilde{m})$	
0.25	0.002267×10^{-5}	0.9807×10^{-5}	0.9830×10^{-5}	0.0×10^{-6}	0.01459×10^{-8}	0.00
1.25	0.02129×10^{-5}	3.894×10^{-5}	3.916×10^{-5}	0.1×10^{-6}	0.05408×10^{-8}	0.24
2.5	0.05403×10^{-5}	7.087×10^{-5}	7.141×10^{-5}	0.2×10^{-6}	0.09604×10^{-8}	0.48
5.0	0.1439×10^{-5}	13.22×10^{-5}	13.37×10^{-5}	0.4×10^{-6}	0.1760×10^{-8}	0.96
12.5	0.4957×10^{-5}	30.73×10^{-5}	31.23×10^{-5}	1.0×10^{-6}	0.4022×10^{-8}	2.40
25.0	1.328×10^{-5}	58.70×10^{-5}	60.03×10^{-5}	1.8×10^{-6}	0.7612×10^{-8}	4.32
50.0	3.622×10^{-5}	113.2×10^{-5}	116.8×10^{-5}	3.2×10^{-6}	1.456×10^{-8}	7.68
100.0	9.608×10^{-5}	227.4×10^{-5}	237.0×10^{-5}	6.1×10^{-6}	2.941×10^{-8}	14.64
150.0	15.96×10^{-5}	341.7×10^{-5}	357.6×10^{-5}	9.0×10^{-6}	4.429×10^{-8}	21.60

TABLE XVI—MIE EXTINCTION PARAMETERS AT 21.0 CM WAVELENGTH (1.43 GHz), H_2O INDEX OF REFRACTION 9.00—0.275i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$	Medium Index of Refraction $\tilde{m}$		Phase Change $\left(\frac{deg}{km}\right)$
				$Re(\tilde{m}-1)$	$Im(1-\tilde{m})$	
0.25	0.0005909×10^{-5}	0.4531×10^{-5}	0.4537×10^{-5}	0.0×10^{-6}	0.009192×10^{-8}	0.0
1.25	0.005543×10^{-5}	1.746×10^{-5}	1.752×10^{-5}	0.1×10^{-6}	0.03347×10^{-8}	0.17
2.5	0.01406×10^{-5}	3.137×10^{-5}	3.151×10^{-5}	0.2×10^{-6}	0.05885×10^{-8}	0.34
5.0	0.03743×10^{-5}	5.774×10^{-5}	5.811×10^{-5}	0.4×10^{-6}	0.1065×10^{-8}	0.68
12.5	0.1288×10^{-5}	13.17×10^{-5}	13.30×10^{-5}	1.0×10^{-6}	0.2390×10^{-8}	1.71
25.0	0.3449×10^{-5}	24.66×10^{-5}	25.01×10^{-5}	1.8×10^{-6}	0.4431×10^{-8}	3.09
50.0	0.9399×10^{-5}	46.29×10^{-5}	47.23×10^{-5}	3.2×10^{-6}	0.8237×10^{-8}	5.50
100.0	2.491×10^{-5}	90.75×10^{-5}	93.24×10^{-5}	6.1×10^{-6}	1.619×10^{-8}	10.5
150.0	4.138×10^{-5}	134.8×10^{-5}	139.0×10^{-5}	8.9×10^{-6}	2.408×10^{-8}	15.3

TABLE XVII—MIE EXTINCTION PARAMETERS AT 0.6328 μ WAVELENGTH, H_2O INDEX OF REFRACTION 1.33—0.0i, FOR LAWS AND PARSONS RAIN

Rain Rate (mm/hr)	Scattering Coef. $(km)^{-1}$	Absorption Coef. $(km)^{-1}$	Extinction Coef. $(km)^{-1}$
0.25	0.08093	0.00	0.08093
1.25	0.2482	0.00	0.2482
2.5	0.3977	0.00	0.3977
5.0	0.6519	0.00	0.6519
12.5	1.273	0.00	1.273
25.0	2.069	0.00	2.069
50.0	3.221	0.00	3.221
100.0	5.689	0.00	5.689
150.0	8.046	0.00	8.046

decreased from 21 to 0.1 cm. This phenomenon is also illustrated in Fig. 26 of a paper previously presented in this journal by Chu and Hogg (1968).[9] It serves to warn the reader that he should be very careful when applying the common rules of thumb relating wavelength and attenuation.

VI. ACKNOWLEDGMENTS

Mr. D. S. Drumheller helped write the computer program and Mr. A. J. D'Alessio helped to organize the results.

REFERENCES

1. Gusler, L. T., and Hogg, D. C., "Some Calculations on the Interference Between Satellite Communications and Terrestrial Radio-Relay Systems Due to Scattering by Rain," B.S.T.J., *49*, No. 7 (September 1970), pp. 1491–1511.
2. Gray, D. A., "Transit Time Variations in Line of Sight Troposphere Propagation Paths," B.S.T.J., *49*, No. 6 (July–August 1970), pp. 1059–68.
3. Pierce, J. R., "Sychronizing Digital Networks," B.S.T.J., *48*, No. 3 (March 1969), pp. 615–636.
4. Setzer, D. E., "Comparison of Measured and Predicted Aerosol Scattering Functions," Appl. Opt., *8*, No. 3 (May 1969), pp. 905–911.
5. Medhurst, R. G., "Rainfall Attenuation of Centimeter Waves: Comparison of Theory and Measurement," IEEE Trans. on Antennas and Propagation, *13*, No. 4 (July 1965), pp. 550–564.
6. Deirmendjian, D., *Electromagnetic Scattering*, M. Kerker, editor, New York: MacMillan Company, 1963, p. 171.
7. Kerr, D. E., *Propagation of Short Radio Waves*, New York: Dover Publications, Inc., p. 671.
8. van de Hulst, H. C., *Light Scattering by Small Particles*, New York: John Wiley Sons, Inc.,1957, pp. 31, 114, 129.
9. Chu, T. S., and Hogg, D. C., "Effects of Precipitation on Propagation at 0.63, 3.5 and 10.6 Microns," B.S.T.J., *47*, No. 5 (May–June 1968), pp. 723–759.

A Linear Phase Modulator for Large Baseband Bandwidths

By C. L. RUTHROFF and W. F. BODTMANN

(Manuscript received June 3, 1970)

A linear phase modulator with a stable carrier frequency would be a useful component in radio systems—especially in coherent phase-shift-keyed PCM systems with baud rates of the order of 100 megabauds per second.

The Armstrong modulator appears adequate for these applications; the circuit functions required for its realization are well understood and amenable to the techniques of integrated circuitry.

In this paper, an analysis of the signal and distortion properties of the Armstrong circuit and variations of it are presented and applied to three system applications: as a replacement modulator for existing low-index analog systems; for multilevel coherent phase-shift-keyed PCM systems; and for frequency-division frequency-modulation multiplex systems which are of interest in radio trunk systems.

I. INTRODUCTION

A linear phase modulator with a stable carrier frequency would be a useful component for the following three applications.

(*i*) As a replacement modulator for the reflex Klystron in an otherwise all solid-state repeater of the TL System.[1]

(*ii*) For frequency-division frequency-modulation multiplex systems with baseband bandwidths of the order of 100 MHz.[2]

(*iii*) For multi-level coherent phase-shift-keyed PCM systems with baud rates of the order of 100 megabauds per second.[3]

The modulator described in this paper appears adequate for these applications. It is based upon the original Armstrong circuit which is well suited to large baseband bandwidths and is reasonably linear for low modulation indexes.[4] An important feature of this method of modulation is that the carrier frequency can be stable with respect

to ambient effects since it can be derived from a temperature-stabilized quartz crystal oscillator. The baseband bandwidths which may be achieved are those for which low-index double sideband amplitude modulators can be built.

An analysis of distortion is presented for the types of baseband signals used in the three applications discussed above, and a circuit is described in which the phase deviation can be increased to any desired value.

II. CIRCUIT DESCRIPTION

The Armstrong modulator is illustrated in Fig. 1. The baseband signal is modulated in a double-sideband suppressed-carrier amplitude modulator with a sufficiently low index of modulation to ensure suitable linearity. At the modulator output another carrier, 90° out of phase with the first, is added to the sidebands. The residual AM is removed by the limiter whose output is a low-index phase-modulated signal. The phase distortion can be made arbitrarily small by choice of the carrier to sideband power ratio at the limiter input; the result is a nearly linear, low-index phase-modulated signal.

Let the baseband signal be

$$e = v(t), \quad \text{with} \quad |v(t)| \leqq 1. \tag{1}$$

The output of the double-sideband suppressed-carrier amplitude modulator is

$$e_a = mv(t) \cos \omega_0 t \tag{2}$$

where $m \leqq 1$ is the index of modulation.

A quadrature carrier is added to e_a in approximately the correct phase to obtain

$$e_p = \sin (\omega_0 t + \epsilon) + mv(t) \cos \omega_0 t. \tag{3}$$

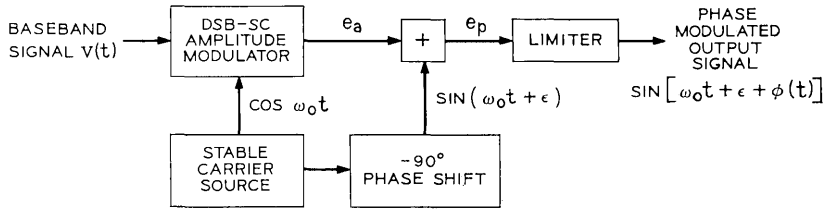


Fig. 1—Armstrong phase modulator.

$$e_p = \sqrt{1 + 2mv(t) \sin \epsilon + m^2v^2(t)} \cdot \sin \left[\omega_0 t + \epsilon + \tan^{-1} \frac{mv(t) \cos \epsilon}{1 + mv(t) \sin \epsilon} \right], \quad (4)$$

where ϵ is small and represents any error in carrier phase.

If this signal is passed through a perfect limiter the envelope becomes constant, leaving an angle modulated signal whose phase modulation is

$$\varphi(t) = \tan^{-1} \frac{mv(t) \cos \epsilon}{1 + mv(t) \sin \epsilon}. \quad (5)$$

When the nonlinear distortion is small, the controlling distortions will be second and third order so terms in the expansion of equation (5) beyond the third will be omitted and (5) becomes

$$\varphi(t) \approx mv(t) \cos \epsilon - m^2v(t)^2 \sin \epsilon \cos \epsilon + m^3v(t)^3 \sin^2 \epsilon \cos \epsilon - \frac{m^3}{3} v(t)^3 \cos^3 \epsilon. \quad (6)$$

Ideally, $\epsilon = 0$ and the first term in equation (6) is the desired modulating signal; the second and third terms will be zero and the last term is the third-order distortion. When $\epsilon \neq 0$, second-order distortion occurs and the desired output signal amplitude is reduced by the factor cos ϵ.

It can be seen from equation (6) that the distortion can be made as small as desired by the proper choice of m, which is proportional to the phase deviation. In order to determine suitable values of m, $v(t)$ must be specified; we shall consider three signals of interest, corresponding to the three applications listed in Section I.

2.1 *Case I*

The signal $v(t)$ is gaussian noise uniformly distributed in a bandwidth extending from $O - W$ Hz.

For nonlinearities of the type described in equation (6) the desired results can be computed by well-known methods.[5]

$$\frac{S_0(f)}{S_2(f)} = \frac{1}{2m^2\sigma^2 \sin^2 \epsilon \left(1 - \frac{|f|}{2W}\right)}, \qquad 0 \leqq |f| \leqq W, \quad (7)$$

$$\frac{S_0(f)}{S_3(f)} = \frac{2}{m^4\sigma^4 \cos^4 \epsilon \left[1 - \frac{1}{3}\left(\frac{f}{W}\right)^2\right]}, \qquad 0 \leqq |f| \leqq W, \quad (8)$$

where,

$S_0(f) = m^2 \cos^2 \epsilon(\sigma^2/2W)$, with $-W \leqq f \leqq W$, is the spectral density of the phase of the output signal,

$S_2(f)$, $S_3(f)$ are the spectral densities of the second- and third-order distortion terms, respectively.

σ^2 is the mean square value of $v(t)$, that is, the power in $v(t)$, and $m\sigma$ is the rms phase deviation.

2.2 *Case II*

$$v(t) = \sum_{n=1}^{N} Q \cos (np + q_n)t. \tag{9}$$

The baseband signal, $v(t)$, is a frequency-division frequency-modulated multiplex signal. Each term in equation (9) is an FM carrier with its own frequency modulation q_n. Bennett has derived the number and types of modulation products produced by the last three terms of equation (6) for $v(t)$ as in equation (9).[6] The second-order term of largest amplitude has the form

$$e_2 = m^2Q^2 \sin \epsilon \cos \epsilon \cos [(m \pm n)p + (q_m \pm q_n)]t. \tag{10}$$

Similarly, the controlling third-order product has the form

$$e_3 = \frac{m^3}{2} Q^3 \cos^3 \epsilon \cos [(l \pm m \pm n)p + (q_l \pm q_m \pm q_n)]t. \tag{11}$$

The total power in the signal of equation (9) is

$$\sigma^2 = N \frac{Q^2}{2} \tag{12}$$

where N is the number of terms in equation (9). From equation (6) the output phase modulation for an individual channel is

$$e_1 = mQ \cos \epsilon \cos (np + q_n)t. \tag{13}$$

The ratios of signal-to-distortion power for single modulation products are,

$$\frac{|e_1|^2}{|e_2|^2} = \left[\frac{1}{mQ \sin \epsilon}\right]^2, \tag{14}$$

$$\frac{|e_1|^2}{|e_3|^2} = \left[\frac{2}{m^2Q^2 \cos^2 \epsilon}\right]^2. \tag{15}$$

In order to determine the total signal-to-distortion power ratios it is

necessary to compute the number of products falling in the kth channel, $1 \leqq k \leqq N$. Assuming power addition for these products the total signal to distortion ratios become

$$\frac{S}{D_2} = \frac{1}{2m^2\sigma^2 \sin^2 \epsilon}\left(\frac{N}{N_2}\right) \tag{16}$$

$$\frac{S}{D_3} = \frac{2}{m^4\sigma^4 \cos^4 \epsilon}\left(\frac{N^2}{2N_3}\right) \tag{17}$$

where

N is the total number of channels, i.e., the number of terms in equation (9),

N_2 is the equivalent number of $m \pm n$ type products and includes other second-order products weighted in accordance with their contribution to the distortion power. It is a function of k and N, and

N_3 is the equivalent number of $l \pm m \pm n$ type products and includes other third-order products weighted in accordance with their contribution to the distortion power. It is a function of k and N.

Expressions (16) and (17) for the signal consisting of N sine waves are much like expressions (7) and (8) for the case of the noise-like signal. It has been shown by Bennett that the sum of randomly phased sine waves of equation (9) behave like noise as N increases without bound and if the power and bandwidth are finite.[7] It is of interest to see in the present context how fast expressions (16) and (17) approach (7) and (8) as N increases; this is shown in Figs. 2 and 3. It is evident from the figures that the signal-to-distortion ratios are not a strong function of the number of channels, the ratios changing a maximum of 2 dB while the number of channels goes from 10 to infinity.

For a more detailed look at the behavior of the distortion products, the number of the various types of products falling in the kth channel for the 500-channel case are shown in Figs. 4 and 5.

2.3 *Case III*

In this case the baseband signal is a sequence of pulses which phase modulate a carrier in the format of a phase-shift-keyed system. A 4-level polar baseband signal is written

$$v(t) = V_0 \sum_{n=-\infty}^{\infty} k_n p(t - nT), \tag{18}$$

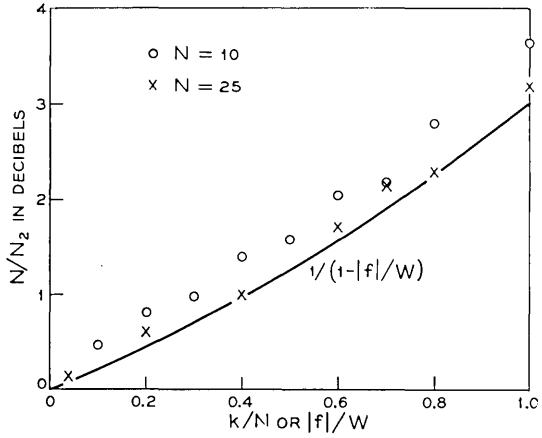


Fig. 2—The effect of the number of channels on the ratio of signal-to-second-order distortion.

where,

$p(t)$ is the pulse shape,

T is the time interval between adjacent pulses, and

$$k_n = \pm 1, \pm 3.$$

In a 4-level PSK system, a maximum peak deviation of $\pm 3\pi/4$ radians

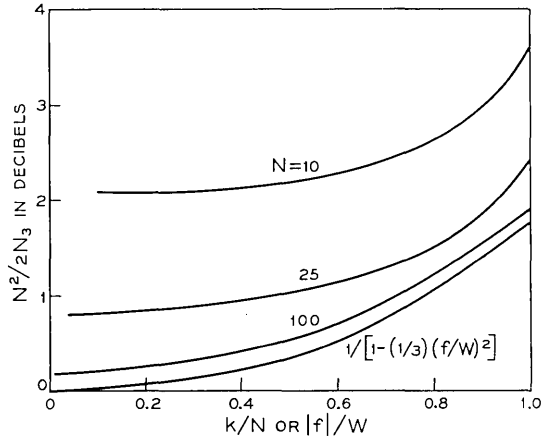


Fig. 3—The effect of the number of channels on the ratio of signal-to-third-order distortion.

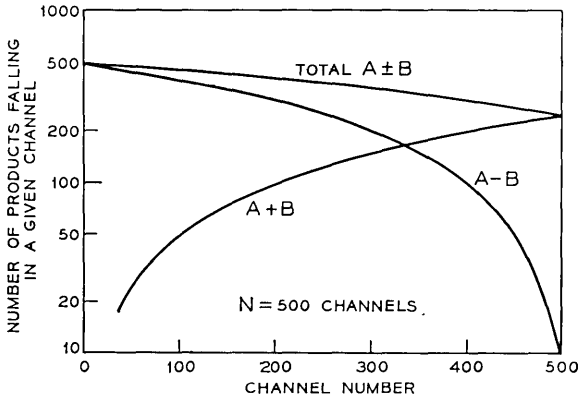


Fig. 4—Number of second-order distortion products.

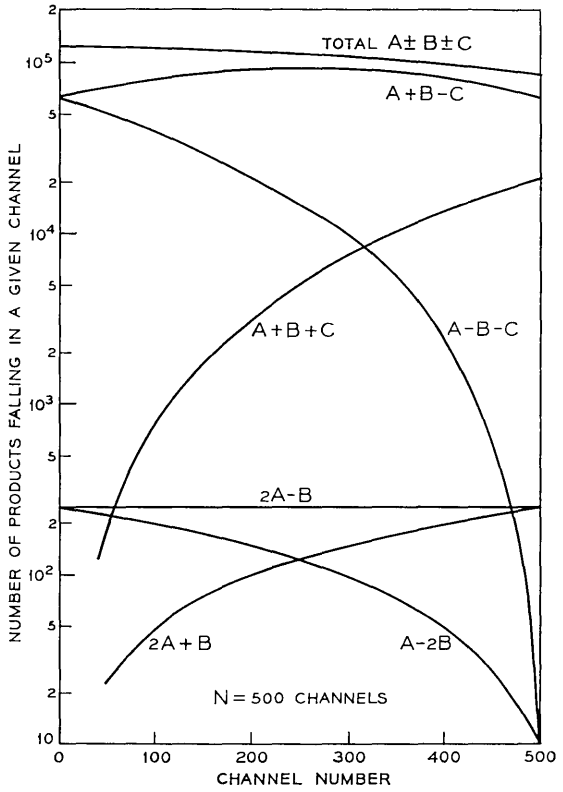


Fig. 5—Number of third-order distortion products.

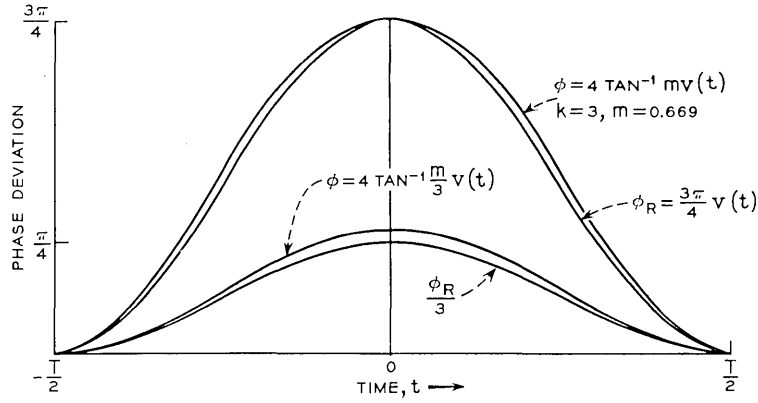


Fig. 6—Modulator input and output pulses.

is required. Deviations of this magnitude may be obtained by multiplying the output of the modulator in a resistive multiplier circuit.

As an example, suppose the modulator output is multiplied by four. The peak deviation required in the modulator is then $3\pi/16$ radians. Raised cosine input pulses, $v(t)$, and the corresponding phase deviations in the output of the modulator are shown in Fig. 6 for this case. The output pulses were computed from equation (5) for $\epsilon = 0$. The value of m was chosen to result in a peak deviation of $3\pi/16$ radians for the pulse corresponding to $k_n = 3$. For this example, $m = \tan 3\pi/16$, and

$$v(t) = \frac{k_n}{6}\left[1 + \cos\frac{2\pi t}{T}\right], \qquad -\frac{T}{2} \leqq t \leqq \frac{T}{2}.$$

In Fig. 6, the phase deviation, φ, is shown for pulses having $k_n = +1$, $+3$. Some pulse compression is present in the larger pulse and the parameter m has been chosen for the correct peak deviation. For the smaller pulse the peak deviation is seen to be too large by about five degrees. If uncorrected, this error would cause the system performance to be degraded a few tenths of a dB.[8] The peak deviation can be corrected by a gain adjustment in the circuits in which the smaller pulses are generated.[3]

III. MODIFIED ARMSTRONG MODULATORS

There may be applications in which it is desirable that the output carrier frequency equal the frequency of the source carrier. The circuit

of Fig. 7 will accomplish this purpose while minimizing the degradation due to tones generated in the final mixer. The carrier frequencies of any high-order products of the two input signals which fall into the output band will be exactly at the carrier frequency of the output signal and result in minimum degradation.

If the times $(N - 1)$ frequency multiplier is replaced by a times M multiplier the flexibility in the choice of output carrier frequency is increased while the feature described above is retained. In either case the frequency multipliers should be resistive rather than reactive.

Finally, in the balanced modulator illustrated in Fig. 8 the phase deviation is doubled for a specified ratio of signal-to-third-order distortion.

IV. CONCLUDING REMARKS

The Armstrong modulator has three attractive features.

(*i*) The carrier frequency can be derived from a frequency stabilized oscillator. For example, a single source can be used in both modulators used to derive two cross-polarized channels for a short hop radio system or a satellite radio system. The identical carrier frequencies serve to minimize the effect of co-channel interference due to cross-polarization coupling.

(*ii*) The functions required to realize the modulator—limiting, mixing, and multiplication—are amenable to circuit integration.

(*iii*) The modulator is suitable for very large baseband bandwidths, particularly high-speed pulse sequences for PSK-PCM systems.

A short hop radio system has been described recently which has about the same communication capacity for either large index analog phase modulation or digital PSK-PCM.[9] In a system designed for either type of operation, it is convenient to do the digital processing at the inter-

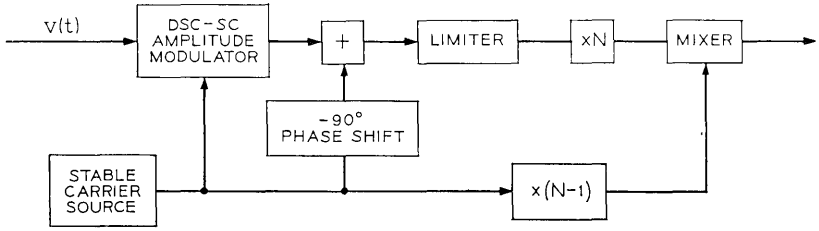


Fig. 7—A modulator with output frequency equal to frequency of stable carrier source.

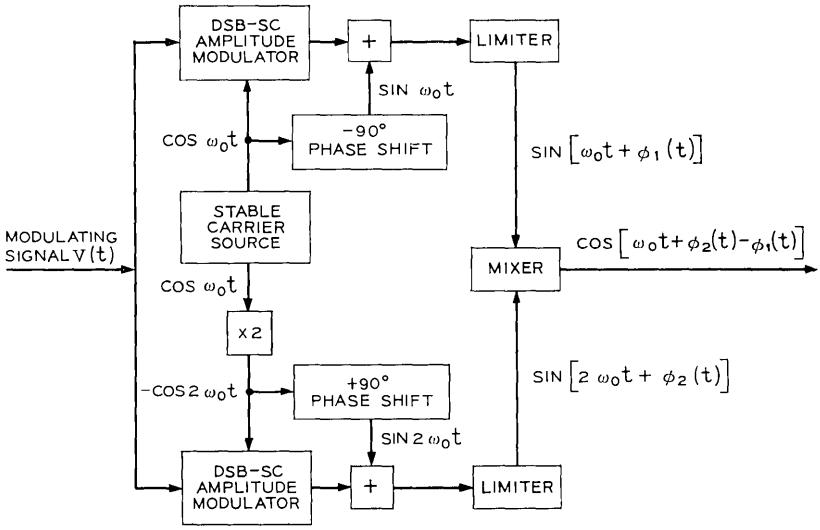


Fig. 8—Balanced Armstrong modulator.

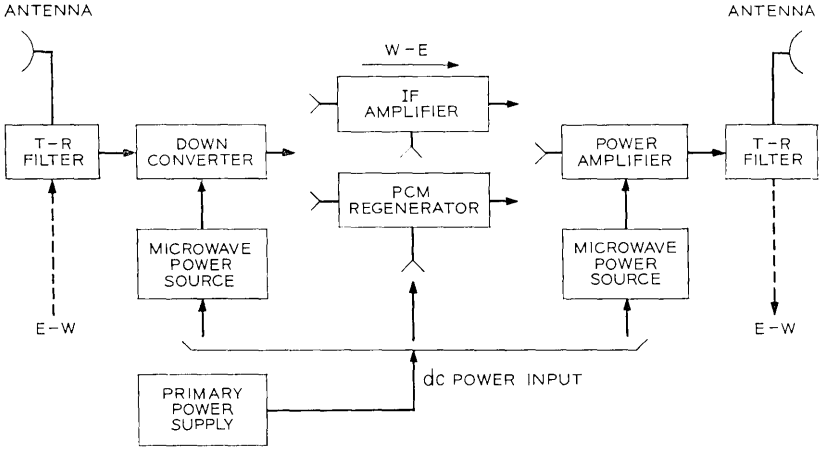


Fig. 9—Repeater of configuration for analog phase modulation or digital CPSK-PCM modulation.

mediate frequency; if PSK-PCM is to be used, the IF amplifier can be replaced by a digital regenerative repeater and no other changes need be made (See Fig. 9).

A digital regenerative repeater has been described which is appropriate for this application; it requires a phase modulator with requirements which are met by the configuration of Figure 7: that incidental AM be small, that the frequency be stable, that the linearity be adequate for multi-level operation, and that the power consumption be small.[3]

REFERENCES

1. Hathaway, S. D., Sagaser, D. D., and Ward, J. A., "The TL Radio Relay System," B.S.T.J., *42*, No. 5 (September 1963), pp. 2297–2353.
2. Bodtmann, W. F., "Phase Locked Frequency Modulated Multiplex," unpublished work.
3. Ruthroff, C. L., and Bodtmann, W. F., "A Digital Repeater for the Short Hop Microwave Radio System," unpublished work.
4. Armstrong, E. H., "A Method of Reducing Disturbances in Radio Signalling by a System of Frequency Modulation," Proc. IRE, *24*, No. 5 (May 1936), pp. 689–740.
5. Davenport, W. B., and Root, W. L., *Random Signals and Noise*, New York: McGraw-Hill, 1958, pp. 277–311.
6. Bennett, W. R., "Cross-Modulation Requirements on Multichannel Amplifiers Below Overload," B.S.T.J., *19*, No. 4 (October 1940), pp. 587–610.
7. Bennett, W. R., "Distribution of the Sum of Randomly Phased Components," Quarterly of Applied Math., *5*, No. 1 (January 1948), pp. 385–393.
8. Prabhu, V. K., "Error-Rate Considerations for Digital Phase-Modulation Systems," IEEE Trans. on Communication Technology, *COM-17*, No. 1 (February 1969), pp. 33–42.
9. Tillotson, L. C., "Use of Frequencies above 10 GHz for Common Carrier Applications," B.S.T.J., *48*, No. 6 (July–August 1969), pp. 1563–1576.

Eventual Stability for Lipschitz Functional Differential Systems

By GERALD A. SHANHOLT

(Manuscript received April 3, 1970)

In this paper it is established that for Lipschitz functional differential systems, the eventual uniform asymptotic stability of the origin is preserved under absolutely diminishing perturbations.

I. INTRODUCTION AND NOTATION

In two recent papers, A. Strauss and J. A. Yorke have investigated "eventual" stability properties for systems of ordinary differential equations.[1,2] In particular, they have shown that for Lipschitz systems, diminishing perturbations preserve eventual uniform asymptotic stability.[1] It is the purpose of this paper to extend a somewhat weaker form of this result to functional differential systems. Namely, it will be shown that for Lipschitz functional differential systems, the eventual uniform asymptotic stability of the origin is preserved under absolutely diminishing perturbations.

The following notation will be used in this paper: E^n is the space of n-vectors, and for x in E^n, $|x|$ denotes any vector norm. For a given number $\tau > 0$, C denotes the linear space of continuous functions mapping the interval $[-\tau, 0]$ into E^n, and for ϕ in C, $\|\phi\| = \sup |\phi(\theta)|$, $-\tau \leqq \theta \leqq 0$. For $H > 0$, C_H denotes the set of ϕ in C for which $\|\phi\| < H$. For any continuous function $x(u)$ whose domain is $-\tau \leqq u \leqq a$, $a \geqq 0$, and whose range is in E^n, and any fixed t, $0 \leqq t \leqq a$, the symbol x_t will denote the function $x_t(\theta) = x(t + \theta)$, $-\tau \leqq \theta \leqq 0$; that is, x_t is in C, and is that segment of the function $x(u)$ defined by letting u range in the interval $t - \tau \leqq u \leqq t$.

Let $F(t, \phi)$ be a function defined on $D_H = [0, \infty) \times C_H$ into E^n, and let $\dot{x}(t)$ denote the right hand derivative of $x(u)$ at $u = t$. Consider the functional differential system

$$\dot{x}(t) = F(t, x_t). \tag{1}$$

Let (s, ϕ) be in D_H. A function $x(s, \phi)(t)$ is said to be a solution of equation (1) with initial function ϕ at $t = s$ if there exists a number $b > 0$ such that

(*i*) for $t \in [s, s + b)$, $x_t(s, \phi)$ is defined and in C_H;
(*ii*) $x_s(s, \phi) = \phi$; and
(*iii*) $x(s, \phi)(t)$ satisfies equation (1) for $s \leqq t < s + b$.

$x(s, \phi)(t)$ is unique if every other solution with the same initial function ϕ at $t = s$ agrees with $x(s, \phi)(t)$ in their common domain of definition.

If F is continuous on D_H, then for every (s, ϕ) in D_H there is at least one solution of equation (1) with initial function ϕ at $t = s$.[3] If, moreover, F is Lipschitzian in ϕ, that is, there is a constant L such that for every ϕ_1, ϕ_2 in C_H

$$| F(t, \phi_1) - F(t, \phi_2) | \leqq L ||\phi_1 - \phi_2 || \tag{2}$$

for $t \geqq 0$, then there is only one such solution. Generally, under such assumptions, one can only expect solutions to exist over a finite interval.

II. PRELIMINARIES

We now define the stability concepts to be used herein. These definitions are stated for equation (1) in which it is assumed that for some H, $0 < H \leqq \infty$, F is continuous and Lipschitzian on D_H.

Definition 1: The origin is *eventually uniformly stable* (EvUS) if for every $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ and $\alpha = \alpha(\epsilon) \geqq 0$ such that $|| x_t(s, \phi) || < \epsilon$ for all $|| \phi || < \delta$ and $t \geqq s \geqq \alpha$. It is *uniformly stable* (US) if one can choose $\alpha(\epsilon) \equiv 0$.

Definition 2: The origin is *eventually uniformly attracting* (EvUA) if there exists constants $\eta > 0$ and $\beta \geqq 0$, and if for every $\epsilon > 0$ there exists a $T = T(\epsilon) > 0$ such that $|| x_t(s, \phi) || < \epsilon$ for $|| \phi || < \eta$, $s \geqq \beta$, and $t \geqq s + T$. It is *uniformly attracting* (UA) if one can choose $\beta = 0$.

Definition 3: The origin is *eventually uniform-asymptotically stable* (EvUAS) if it is both EvUS and EvUA. It is *uniform-asymptotically stable* (UAS) if it is both US and UA.

The above definitions show that EvUS, EvUA, and EvUAS are weaker stability concepts than their respective Lyapunov counterparts: US, UA, and UAS. Also, it should be noted that in these definitions we do not require that the zero function be a solution of equation (1). When the origin is US, this implies that the zero function is a unique

solution of equation (1) for any $s \geqq 0$. Thus, we see that EvUS (EvUAS) is a natural generalization of US (UAS) in which it is not assumed that the zero function is a solution. Finally, it is important to note that UA does not imply that the zero function is a solution (Ref. 1, example 2.8).

Definition 4: Let $V(t, \phi)$ be a function defined for (t, ϕ) in D_H. The derivative of V along solutions of equation (1) will be denoted by $\dot{V}_{(1)}[t, x_t(s, \phi)]$ and is defined to be

$$\dot{V}_{(1)}[t, x_t(s, \phi)] = \limsup_{h \to 0^+} \frac{1}{h} \{V[t + h, x_{t+h}(s, \phi)] - V[t, x_t(s, \phi)]\}.$$

If F is continuous and Lipschitzian, and if the origin is EvUAS, then the existence of a Lyapunov type comparison function can be established. By following D. Wexler[4] and A. Halanay[5] one can prove the following theorem.

Theorem 1: Let F be continuous and Lipschitzian on D_H, and let the origin be EvUAS. Then there exists a number K, $0 < K < H$, and a function $V(t, \phi)$ with the properties: (i) there exists functions $a(r)$, $b(r)$ continuous, positive, and monotone increasing for $r > 0$, with $a(0) = b(0) = 0$, such that for m in $(0, K]$

$$a(\| \phi \|) \leqq V(t, \phi) \leqq b(\| \phi \|)$$

for $m \leqq \| \phi \| \leqq K$, $t \geqq d(m)$, where $d(r)$ is a continuous, nonnegative, and nonincreasing function for $r > 0$; (ii) there exists a function $c(r)$ continuous, positive, and monotone-increasing for $r > 0$, with $c(0) = 0$ such that

$$\dot{V}_{(1)}[t, x_t(s, \phi)] \leqq -c[\| x_t(s, \phi) \|]$$

for $\| \phi \| \leqq K$, $t \geqq s \geqq d(K)$; and (iii) for $0 < r \leqq \| \phi_i \| \leqq K$, $t \geqq d(K)$

$$| V(t, \phi_1) - V(t, \phi_2) | \leqq M(r) \| \phi_1 - \phi_2 \|,$$

where $M(r)$ is continuous and monotone-decreasing on $(0, K]$.

III. PERTURBED EQUATION

We now prove a theorem which shows that the EvUAS of the origin of the nominal equation

$$\dot{y}(t) = F(t, y_t) \tag{N}$$

is preserved for the perturbed equation

$$\dot{x}(t) = F(t, x_t) + G(t, x_t) \tag{P}$$

when F and G satisfy certain conditions. In particular, $G(t, \phi)$ is required to be *absolutely diminishing*, that is, for every m in $(0, H)$, there exists a $\gamma_m \geqq 0$ and a function $g_m(t)$ continuous on $[\gamma_m, \infty)$ such that for $m \leqq \| \phi \| < H$, $t \geqq \gamma_m$

$$|G(t, \phi)| \leqq g_m(t) \quad \text{and} \quad I_m(t) \triangleq \int_t^{t+1} g_m(s)\, ds \to 0 \quad \text{as} \quad t \to \infty .$$

Theorem 2: Suppose that F and G are continuous and Lipschitzian on D_H, that G is absolutely diminishing, and that the origin is EvUAS for equation (N). Then the origin is EvUAS also for equation (P).

Proof: Define $J_m(t) = \sup [I_m(s): t - 1 \leqq s < \infty]$ for $t \geqq 1$. Since $I_m(t) \to 0$ as $t \to \infty$, this implies $J_m(t) \to 0$ monotonically as $t \to \infty$.

Let $0 < \epsilon \leqq K$, choose $\| \phi \| < \delta(\epsilon) = b^{-1}[a(\epsilon)/2]$, and pick $\theta = \theta(\epsilon) \geqq 0$ and such that

$$2LM(\delta)J_\delta(t) < \min [a(\epsilon), c(\delta)] \tag{3}$$

for $t \geqq \theta$, where L is the Lipschitz constant associated with F. Then for $t \geqq s \geqq \alpha(\epsilon) = \max [1, \theta(\epsilon), d(\delta)]$, $\| x_t(s, \phi) \| < \epsilon$. Suppose not, that is, for some $t \geqq s$, $\| x_t(s, \phi) \| = \epsilon$. Let q be the first t-value greater than s for which $\| x_q(s, \phi) \| = \epsilon$, and let p be the last t-value less than q for which $\| x_p(s, \phi) \| = \delta$. Then

$$\delta \leqq \| x_t(s, \phi) \| \leqq \epsilon, \qquad p \leqq t \leqq q. \tag{4}$$

For t in an interval on which $x(s, \phi)(t)$ exists, we evaluate

$$\begin{aligned}
\dot{V}_{(P)}[t, x_t(s, \phi)] &\leqq \dot{V}_{(N)}[t, x_t(s, \phi)] \\
&\quad + \limsup_{h \to 0^+} \frac{1}{h} \left(V\{t + h, x_{t+h}[t, x_t(s, \phi)]\} \right. \\
&\qquad \left. - V\{t + h, y_{t+h}[t, x_t(s, \phi)]\} \right) \\
&\leqq -c[\|x_t(s, \phi)\|] \\
&\quad + \limsup_{h \to 0^+} \frac{M}{h} \{ \|x_{t+h}[t, x_t(s, \phi)] \\
&\qquad - y_{t+h}[t, x_t(s, \phi)]\| \}
\end{aligned}$$

where the function V is as described in Theorem 1. By assuming—with no loss of generality—that $L > 1$, we obtain[5] from the above inequality

$$\dot{V}_{(P)}[t, x_t(s, \phi)] \leqq -c[\, \|x_t(s, \phi)\|] + LM \mid G[t, x_t(s, \phi)] \mid .$$

Employing the absolute diminishing character of G and equation (4), we obtain by integrating the above from p to q

$$a(\epsilon) \leqq b(\delta) - (q - p)c(\delta) + LM \int_p^q g_\delta(t)\, dt. \tag{5}$$

Using the easily shown fact that

$$\int_u^t g_m(s)\, ds \leqq \int_{u-1}^t I_m(s)\, ds, \qquad t \geqq u \geqq 1,$$

and equations (3) and (5), we see that

$$a(\epsilon) \leqq b(\delta) - (q - p)c(\delta) + LM(q - p + 1)J_\delta(p)$$
$$< b(\delta) + a(\epsilon)/2 = a(\epsilon).$$

Hence, we arrive at a contradiction which shows that the origin is EvUS.

Let $\eta = \delta(K)$, $\beta = \alpha(K)$, and

$$T(\epsilon) = \alpha(\epsilon) + 2[LMJ_\delta(1) + b(K)]/c(\delta). \tag{6}$$

Consider $s \geqq \beta$ and $\| \phi \| < \eta$. Thus, $x(s, \phi)(t)$ exists for all $t \geqq s$. Moreover, since the origin is EvUS, to prove EvUA it is sufficient to show the existence of a u, $s + \alpha \leqq u \leqq s + T$, such that $\| x_u(s, \phi) \| < \delta(\epsilon)$. Assume the contrary, that is,

$$\delta \leqq \| x_t(s, \phi) \| \leqq K, \; s + \alpha \leqq t \leqq s + T.$$

Employing the same procedure as above, we arrive at the estimate

$$a(\delta) < b(K) - (T - \alpha)c(\delta) + ML(T - \alpha + 1)J_\delta(s + \alpha).$$

Using the monotonicity of J_δ and equations (3) and (6), we compute

$$a(\delta) < b(K) - \frac{c(\delta)}{2}(T - \alpha) + MLJ_\delta(1) = 0.$$

This contradiction then completes the proof of this theorem.

REFERENCES

1. Strauss, A., and Yorke, J. A., "Perturbing Uniform Asymptotically Stable Nonlinear Systems," J. Differential Equations, *6*, No. 3 (November 1969), pp. 452-483.
2. Strauss, A., and Yorke, J. A., "Perturbing Uniformly Stable Linear Systems With and Without Attraction," SIAM J. Appl. Math., *17*, No. 4 (July 1969), pp. 725–738.
3. Oguztoreli, N. N., *Time-Lag Control Systems*, New York: Academic Press, 1966, pp. 20–30.
4. Wexler, D., "Note on the Eventual Stability," *Revue Roumaine de Mathematiques Pures et Appliquees, 11*, No. 7 (1966), pp. 819–824.
5. Halanay, A., *Differential Equations: Stability, Oscillations, Time Lags*, New York: Academic Press, 1966, pp. 340–349.

Information Theory and Approximation of Bandlimited Functions

By DAVID JAGERMAN

(Manuscript received April 15, 1970)

For bandlimited functions, simultaneous approximation of a function and several of its derivatives is considered. Concomitant entropy estimates are obtained. A feasible algorithm for the transmission of information is discussed. This algorithm has been applied to the design of a class of PCM systems.[1]

I. INTRODUCTION

It is the purpose of this paper to discuss both the best approximation of sets of bandlimited functions under Sobolev norms and the concomitant information-theoretic estimates. The Sobolev norms are useful when it is desired to approximate simultaneously the function and some of its derivatives. This requires an amount of information beyond that for approximating only the function. Section II gives the necessary background definitions of width, entropy, and capacity; theorems providing representations of bandlimited functions, as well as a form of Mitjagin's inequality relating approximability to entropy, are proved. The distinction between capacity and entropy is comparable to that between communication and storage, since capacity refers to the number of distinguishable functions transmitted from a signal source while entropy measures a bit requirement for the reproduction of a function to within a specified accuracy. A constructive approach to communication requirements implies an explicit means of representing any function of the signal source by numbers with a uniformly bounded number of digits. The procedure or algorithm used is usually obtained from an infinite series representation with subsequent truncation and quantization. Pulse code modulation systems provide examples of this procedure. Section II gives a precise definition, while Section III presents an explicit construction of a feasible algorithm. This algorithm has been applied to the design of a class of PCM systems.[1]

Sections III and IV contain the theorems and proofs which provide upper bounds on widths and entropies. Section III discusses signal sources with finite instantaneous power. Section IV considers signal sources in which the total energy is finite.

II. PRELIMINARIES

Let A be a subset of a Banach space X; it is desired to approximate A, that is, uniformly all elements of A by means of n-dimensional subspaces X_n of X. The deviation $E_{X_n}(A)$ of A from X_n is defined by

$$E_{X_n}(A) = \sup_{f \in A} \inf_{g \in X_n} || f - g ||. \tag{1}$$

The deviation provides information on how well A may be uniformly approximated by elements of the given space X_n ; however, another choice of X_n might provide a smaller deviation. Accordingly the nth width, $d_n^X(A)$, of A relative to the space X is defined by[2]

$$d_n^X(A) = \inf_{X_n \subset X} E_{X_n}(A). \tag{2}$$

If the infimum is attained, then a corresponding X_n is called an extremal subspace. The following properties are immediate.

$$0 \leqq d_{n+1}^X(A) \leqq d_n^X(A), \qquad n \geqq 0, \tag{3}$$

$$d_0^X(A) = \sup_{X \in A} || x ||, \tag{4}$$

$$B \subset A \Rightarrow d_n^X(B) \leqq d_n^X(A). \tag{5}$$

If X has finite dimension m, then $d_n^X(A) = 0$ for $n \geqq m$.

A set of sets whose diameters do not exceed $2\epsilon(\epsilon > 0)$ and whose union contains A is called an ϵ-covering of A. A finite set $S \subset X$ such that for $f \in A$ there is a $g \in S$ with $|| f - g || \leqq \epsilon$ is called an ϵ-net of A. Clearly $d_n^X(A) \leqq \epsilon$ for a set A possessing an ϵ-net of n elements. If A is totally bounded then $\lim_{n \to \infty} d_n^X(A) = 0$. To see this, choose a covering of A consisting of n ϵ-balls, then their centers constitute an ϵ-net of A.

Let $N_\epsilon(A)$ (presumed finite) be the number of sets in a minimal ϵ-covering of A; then the *absolute ϵ-entropy*, $H_\epsilon(A)$, of A is defined by

$$H_\epsilon(A) = \log N_\epsilon(A) \tag{6}$$

in which the logarithm is taken to base two.[2-4]

Let $N_\epsilon^X(A)$ be the number of elements in a minimal ϵ-net $S \subset X$ of A; then the relative ϵ-entropy, $H_\epsilon^X(A)$, is defined by

$$H_\epsilon^X(A) = \log N_\epsilon^X(A) \tag{7}$$

in which the logarithm is taken to base two.[2-4] For A totally bounded, let $x_1, \cdots, x_n$ be the elements of an ϵ-net, and let $B_j (1 \leqq j \leqq n)$ be a ball of radius ϵ about x_j; then the sets $U_j = B_j \cap A$ constitute an ϵ-covering of A; hence

$$H_\epsilon(A) \leqq H_\epsilon^X(A). \tag{8}$$

The minimum number of binary digits, d, of an integer expressed in radix two needed to identify uniquely every element in a minimal ϵ-covering of A satisfies

$$[H_\epsilon(A)] \leqq d \leqq [H_\epsilon(A)] + 1 \tag{9}$$

in which $[x]$ designates the *integral part* of x, that is, the unique integer satisfying $x - 1 < [x] \leqq x$. Thus $H_\epsilon(A)$ may serve as an absolute measure of efficiency for processes designed for the storage and transmission of information.

Let a set ω of n real numbers be chosen, and also a mapping from A onto $\Omega_p = \omega \times \cdots \times \omega$ (p times); that is,

$$x \in A \rightarrow \alpha = (\alpha_1, \cdots, \alpha_p) \in \Omega_p, \alpha_1, \cdots, \alpha_p \in \omega.$$

Let the algorithm Γ define a one-to-one and onto mapping of Ω_p to an ϵ-net S of A in which $\Gamma(\alpha) \in S$ approximates $x \in A$ to within ϵ; then the volume $V(\Gamma)$ is defined by

$$V(\Gamma) = p \log n. \tag{10}$$

In view of expression (8), one has

$$V(\Gamma) \geqq H_\epsilon^X(A) \geqq H_\epsilon(A). \tag{11}$$

Thus the greater $V(\Gamma)$ is, the less efficient is the algorithm Γ compared to the absolute standard $H_\epsilon(A)$.

If $D \subset A$ has the property that

$$f \neq g, \qquad f, g \in D \Rightarrow \| f - g \| > \epsilon, \tag{12}$$

then D is called ϵ-distinguishable. Let $M_\epsilon(A)$ be the number of elements (presumed finite) in a maximal ϵ-distinguishable subset of A, then the *ϵ-capacity*, $C_\epsilon(A)$ is defined by

$$C_\epsilon(A) = \log M_\epsilon(A), \tag{13}$$

the logarithm being again taken to base two.[3] For a transmission system, $C_\epsilon(A)$ measures the number of distinguishable signals of the source or of the processed signal at the output of the receiver depending on the identification of A. The following inequalities hold between

ϵ-capacity and ϵ-entropy:

$$C_{2\epsilon}(A) \leqq H_\epsilon(A) \leqq C_\epsilon(A). \tag{14}$$

To show this, consider the inequality on the right. Let D be a maximal ϵ-distinguishable subset of A; then ϵ-balls about each element of D constitute an ϵ-covering of A for, otherwise, there would be an $x \in A$ not covered and hence more than ϵ away from every element of D. This would contradict the maximality of D. For the inequality on the left, let D be a 2ϵ-distinguishable subset of A, then the number of elements of D cannot exceed the number of covering sets of diameter 2ϵ or less in an ϵ-covering of A for, otherwise, there would be at least two elements of D in the same covering set. This would contradict the 2ϵ-distinguishability of D.

It is possible to bound $H_\epsilon^X(A)$ above and below in terms of $d_n^X(A)$ (refer to Ref. 2 where Mitjagin's inequalities are given). An improved form of Mitjagin's upper bound is proved below.

Theorem 1: *Let A be a totally bounded subset of a real, normed, vector space X. Let the nth widths relative to X be $d_n^X(A)$, and let*

$$N = \max_n [n : d_{n-1}^X(A) \geqq (1 - \alpha)\epsilon]$$

with α an arbitrary number satisfying $0 < \alpha < 1$; then

$$H_\epsilon^X(A) \leqq N \log \left(\frac{2d_0^X}{\alpha\epsilon} + \frac{2 - \alpha}{\alpha}\right).$$

Proof: Let E_N be an N-dimensional subspace of X for which $E_{E_N}(A) < (1 - \alpha)\epsilon$, then $\forall x \in A \; \exists y \in E_N \ni \| x - y \| < (1 - \alpha)\epsilon$. Let A_N be the set of all such y for every $x \in A$. An $\alpha\epsilon$ net of A_N is an ϵ-net of A; hence $H_\epsilon^X(A) \leqq H_{\alpha\epsilon}^X(A_N) \leqq C_{\alpha\epsilon}(A_N)$. Let $y_1, \cdots, y_M$ be an $\alpha\epsilon$-distinguishable subset of A_N, and let $B_k \subset E_N$ be balls with centers y_k and radius $\frac{1}{2}\alpha\epsilon$, then they are disjoint and are all contained in the ball B with center the origin and radius $d_0^X + (1 - \frac{1}{2}\alpha)\epsilon$. Let λ_N be the volume element in E_N; then $\lambda_N M(\frac{1}{2}\alpha\epsilon)^N \leqq \lambda_N[d_0^X + (1 - \frac{1}{2}\alpha)\epsilon]^N$. The inequality of the theorem follows on taking logorithms.

The class of functions to be studied consists of the space B_σ defined by the conditions that $f(t) \in B_\sigma$ be analytically continuable into the complex plane as an entire function of exponential order one and type σ, and that it be bounded on the whole real axis $-\infty < t < \infty$. The following inequality is valid for B_σ :[5]

$$\sup_{-\infty<t<\infty} | \dot{f}(t) | \leqq \sigma \sup_{-\infty<t<\infty} | f(t) |. \tag{15}$$

Important subspaces of the space B_σ are the space C_σ defined by

$$f \in C_\sigma \Rightarrow A(\eta) = o(e^{\sigma|\eta|}) \tag{16}$$

in which

$$A(\eta) = \sup_{-\infty<\xi<\infty} |f(\xi + i\eta)|, \qquad \xi, \eta \text{ real}, \tag{17}$$

and the space W_σ defined by

$$f \in W_\sigma \Rightarrow f \in L^2(-\infty, \infty). \tag{18}$$

Several representations for B_σ exist;[6] however, the following representations are needed for the present investigation. Let

$$\phi(t, \sigma) = \frac{\sin \sigma t}{\sigma t}, \tag{19}$$

$$\phi_i(t, \sigma) = \phi(t - jh, \sigma), \qquad h = \pi/\sigma, \tag{20}$$

then one has the following:

Theorem 2:

$$f(t) \in C_\sigma \Leftrightarrow f(t) = \sum_{j=-\infty}^{\infty} f(jh)\phi_j(t, \sigma)$$

for all complex t. The series converges uniformly in every closed, bounded region.

Proof: Consider the integral

$$I_N = \frac{1}{2\pi i} \int_{C_N} \frac{f(\zeta)}{(\zeta - t) \sin \sigma\zeta} d\zeta, \qquad \zeta = \xi + i\eta, \tag{21}$$

taken over a square C_N with corners at $(N + \frac{1}{2})(\pm 1 \pm i)h$, and N so large that t is in the interior of the region bounded by C_N. The theorem is clearly true when $t = kh$ (k integral); it will hence be assumed $t \neq kh$ for any integral k. The index $N \geqq 0$ is an integer. Evaluation of I_N by use of residues yields

$$f(t) = \sum_{j=-N}^{N} f(jh)\phi_j(t, \sigma) + I_N \sin \sigma t; \tag{22}$$

thus, to prove the implication to the right, it is sufficient to show $I_N \to 0$, $N \to \infty$. Let $I_N^{(1)}$ be the integral (21) extended over that part of C_N given by $\xi = (N + \frac{1}{2})h$; then

$$|I_N^{(1)}| \leqq \frac{1}{2\pi} \int_{-(N+\frac{1}{2})h}^{(N+\frac{1}{2})h} \frac{A(\eta)}{|(N + \frac{1}{2})h + i\eta - t| \, |\sin(\pi(N + \frac{1}{2}) + i\sigma\eta)|} d\eta. \tag{23}$$

Since

$$| (N + \tfrac{1}{2})h + i\eta - t | \geqq (N + \tfrac{1}{2})h - | t | ,$$

$$| \sin (\pi(N + \tfrac{1}{2}) + i\sigma\eta) | = \cosh \sigma\eta \geqq \tfrac{1}{2}e^{\sigma|\eta|},$$

one has

$$| I_N^{(1)} | \leqq \frac{1}{\pi} \frac{1}{(N + \tfrac{1}{2})h - | t |} \int_{-(N+\frac{1}{2})h}^{(N+\frac{1}{2})h} e^{-\sigma|\eta|} A(\eta)\, d\eta. \tag{25}$$

Writing equation (25) in the form

$$| I_N^{(1)} | \leqq \frac{2}{\pi} \frac{(2N + 1)h}{(2N + 1)h - 2 | t |} \frac{1}{(2N + 1)h} \int_{-(N+\frac{1}{2})h}^{(N+\frac{1}{2})h} e^{-\sigma|\eta|} A(\eta)\, d\eta, \tag{26}$$

using equation (16) and the following lemma[7]

$$f(\eta) \to 0, \qquad | \eta | \to \infty \Rightarrow \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(\eta)\, d\eta = 0, \tag{27}$$

shows that $I_N^{(1)} \to 0$ uniformly in t. The same conclusion applies to the integral extended over $\xi = -(N + \frac{1}{2})h$.

Let $I_N^{(2)}$ be the integral (21) extended over $\eta = (N + \frac{1}{2})h$; then

$$| I_N^{(2)} | \leqq \frac{1}{2\pi} \cdot \int_{-(N+\frac{1}{2})h}^{(N+\frac{1}{2})h} \frac{f(\xi + i(N + \frac{1}{2})h)}{| \xi + i(N + \frac{1}{2})h - t | \, | \sin (\sigma\xi + i\pi(N + \frac{1}{2})) |} d\xi. \tag{28}$$

Since

$$| \xi + i(N + \tfrac{1}{2})h - t | \geqq (N + \tfrac{1}{2})h - | t |, \tag{29}$$

$$| \sin (\sigma\xi + i\pi(N + \tfrac{1}{2})) | \geqq \frac{1 - e^{-\pi}}{2} e^{\pi(N+\frac{1}{2})},$$

one has

$$| I_N^{(2)} | \leqq \frac{2}{\pi(1 - e^{-\pi})} \frac{(2N + 1)h}{(2N + 1)h - 2 | t |} e^{-\pi(N+\frac{1}{2})} A((N + \tfrac{1}{2})h). \tag{30}$$

In view of equation (16), $I_N^{(2)} \to 0$ uniformly in t. The same conclusion applies to the integral extended over $\eta = -(N + \frac{1}{2})h$. For the implication to the left, one may observe that $\phi_i(t, \sigma) \in C_\sigma$, and that the series converges uniformly.

The series of Theorem 2, which is clearly interpolatory, is called the *cardinal series*.[8]

For $f(t) \in L^2(-\infty, \infty)$, the Fourier transform relations are given by

$$F(u) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{-iut} f(t)\, dt, \tag{31}$$

$$f(t) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} e^{iut} F(u)\, du. \tag{32}$$

The Fourier transform of $\phi_j(t, \sigma)$ is

$$\begin{aligned} \Phi_j(u, \sigma) &= \frac{1}{\sigma} \left(\frac{\pi}{2}\right)^{\frac{1}{2}} e^{-iujh}, \qquad | u | < \sigma; \\ &= 0, \qquad | u | > \sigma. \end{aligned} \tag{33}$$

The Parseval relation now shows that the sequence $\phi_j(t, \sigma)$, $-\infty < j < \infty$ is orthogonal over $(-\infty, \infty)$ with respect to unit weight; thus,

$$\begin{aligned} \int_{-\infty}^{\infty} \phi_j(t, \sigma)\phi_k(t, \sigma)\, dt &= 0, \qquad j \neq k; \\ &= h, \qquad j = k. \end{aligned} \tag{34}$$

The following theorem may now be stated for $f \in W_\sigma$.

Theorem 3: $f \in W_\sigma$

$$\Rightarrow \int_{-\infty}^{\infty} | f(t) |^2\, dt = h \sum_{j=-\infty}^{\infty} | f(jh) |^2,$$

$$f(t) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\sigma}^{\sigma} e^{iut} F(u)\, du,$$

$$F(u) = \frac{h}{(2\pi)^{\frac{1}{2}}} \sum_{j=-\infty}^{\infty} f(jh) e^{-iujh}, \qquad | u | < \sigma.$$

Proof: The Paley–Wiener theorem[9] shows that $f \in W_\sigma$ has the representation given in Theorem 3; hence, by the Cauchy–Schwartz inequality

$$| f(\xi + i\eta) | \leqq \left\{ \frac{\sinh \sigma\eta}{2\pi\eta} \int_{-\sigma}^{\sigma} | F(u) |^2\, du \right\}^{\frac{1}{2}} = o(e^{\sigma|\eta|}). \tag{35}$$

Equation (35) shows that $W_\sigma \subset C_\sigma$; thus, by Theorem 2, f is in the closure of the system $\phi_j(t, \sigma)$, $-\infty < j < \infty$. The Parseval relation now follows from equation (34). To establish the formula for $F(u)$, it is only necessary to show

$$\int_{-\sigma}^{\sigma} e^{iut} \sum_{j=M}^{N} f(jh) e^{-iujh}\, du \to 0, \qquad M, N \to \infty, \quad M, N \to -\infty, \tag{36}$$

because each term is the Fourier transform of the corresponding ϕ_j term

of the cardinal series. One has

$$\left| \int_{-\sigma}^{\sigma} e^{iut} \sum_{j=M}^{N} f(jh)e^{-iujh}\, du \right|^2 \leqq 2\sigma \int_{-\sigma}^{\sigma} \left| \sum_{j=M}^{N} f(jh)e^{-iujh} \right|^2 du, \tag{37}$$

$$\left| \int_{-\sigma}^{\sigma} e^{iut} \sum_{j=M}^{N} f(jh)e^{-iujh}\, du \right|^2 \leqq 4\sigma^2 \sum_{j=M}^{N} |\, f(jh)\,|^2 \to 0. \tag{38}$$

The limit zero is obtained as a consequence of the Parseval relation of Theorem 3.

To obtain a representation for the class B_σ,[10] let

$$\theta(t) = \phi\left(t, \frac{\delta\sigma}{(1-\delta)m}\right)^m \phi\left(t, \frac{\sigma}{1-\delta}\right) \tag{39}$$

$$m > 0 \text{ integral}, \qquad 0 < \delta < 1,$$

$$\theta_j(t) = \theta(t - jh), \qquad h = \frac{\pi}{\sigma}(1-\delta); \tag{40}$$

then one has

Theorem 4: $f \in B_\sigma$

$$\Rightarrow f(t) = \sum_{j=-\infty}^{\infty} f(jh)\theta_j(t).$$

The series converges absolutely and uniformly in every closed, bounded region.

Proof: The function

$$f(t)\left\{\frac{\sin\frac{\delta\sigma}{(1-\delta)m}(s-t)}{\frac{\delta\sigma}{(1-\delta)m}(s-t)}\right\}^m \tag{41}$$

belongs to $W_{\sigma/(1-\delta)}$ for each positive integer m and arbitrary s, hence the cardinal series applied to this function yields the expansion

$$f(t)\left\{\frac{\sin\frac{\delta\sigma}{(1-\delta)m}(s-t)}{\frac{\delta\sigma}{(1-\delta)m}(s-t)}\right\}^m = \sum_{j=-\infty}^{\infty} f(jh)\left\{\frac{\sin\frac{\delta\sigma}{(1-\delta)m}(s-jh)}{\frac{\delta\sigma}{(1-\delta)m}(s-jh)}\right\}^m \phi_j\left(t, \frac{\sigma}{1-\delta}\right). \tag{42}$$

Let $s = t$; then the required representation is obtained. The absolute convergence follows from the boundedness of $| f(jh) |$ and

$$| \theta_j(t) | = 0(|j|^{-m-1}). \tag{43}$$

Approximation will be studied in the uniform norm and the following Sobolev norm

$$|| f ||_s = \left\{ \int_{-T/2}^{T/2} (| f(t) |^2 + \mu_1 | \dot{f}(t) |^2 + \cdots + \mu_s | f^{(s)}(t) |^2) \, dt \right\}^{\frac{1}{2}} \tag{44}$$

in which $\mu_1, \cdots, \mu_s$ are positive numbers. For the space B_σ, the symbol $B^T_{\sigma,s}$ will be used for the corresponding normed space. The symbol $B^T_{\sigma,s}(M)$ will be used for the subset defined by $| f(t) | \leqq M$, $-\infty < t < \infty$. For the space W_σ, the corresponding normed space will be denoted by $W^T_{\sigma,s}$, and $W^T_{\sigma,s}(B)$ for the subset in which

$$\left\{ \int_{-\infty}^{\infty} | f(t) |^2 \, dt \right\}^{\frac{1}{2}} \leqq B. \tag{45}$$

III. THEORETICAL INVESTIGATION OF B_σ

Let B^T_σ designate the vector space B_σ normed by

$$|| f ||_u = \max_{-T/2 \leqq t \leqq T/2} | f(t) |, \tag{46}$$

and let $B^T_\sigma(M)$ be the subset of B^T_σ satisfying

$$| f(t) | \leqq M, \qquad -\infty < t < \infty. \tag{47}$$

The completion of B^T_σ is the space C^T of functions continuous over $[-T/2, T/2]$ and normed by equation (46).

Let

$$c = \frac{\sigma T}{2}, \qquad \delta_n = 1 - \left(\frac{2c}{\pi n}\right)^{\frac{1}{2}}, \qquad n > \frac{2c}{\pi}, \tag{48}$$

$$m = \left[\frac{\pi \delta^2_{n-1}}{2e} (n-1) \right], \qquad m \geqq 1;$$

then the following theorem provides a bound on the nth width, $d^{C^T}_n(B^T_\sigma(M))$, of $B^T_\sigma(M)$ relative to the space C^T.

Theorem 5: $d^{C^T}_n(B^T_\sigma(M) \leqq (2M/\pi m)e^{-m}$.

Proof: The series representation of Theorem 4 will be used. The function

$$g(t) = \sum_{|j| \leqq N} f(jh)\theta_j(t) \tag{49}$$

establishes an approximation to $f(t)$ whose error is given by

$$f(t) - g(t) = \sum_{|j|>N} f(jh)\theta_j(t). \tag{50}$$

From equations (39) and (40), one has

$$|\theta_j(t)| \leqq \frac{1}{\pi}\left(\frac{m}{\pi\delta}\right)^m \frac{1}{\left(|j| - \frac{T}{2h}\right)^{m+1}}, \quad |j| > \frac{T}{2h}, \quad |t| \leqq \frac{T}{2}. \tag{51}$$

Define the function $\rho(x)$ by

$$\begin{aligned} \rho(x) &= \tfrac{1}{2} - x \qquad 0 \leqq x < 1, \\ &= \rho(x+1) \quad \text{for all} \quad x, \end{aligned} \tag{52}$$

then the Sonin (Euler–Maclaurin) summation formula[11] is

$$\sum_{a<j\leq b} W(j) = \int_a^b W(x)\,dx + \rho(x)W(x)\Big]_a^b - \int_a^b \rho(x)W'(x)\,dx \tag{53}$$

in which $a < b$ are arbitrary numbers. Use of equation (53) with

$$W(x) = \frac{1}{\pi}\left(\frac{m}{\pi\delta}\right)^m \frac{1}{\left(x - \frac{T}{2h}\right)^{m+1}}, \qquad x > T/h, \tag{54}$$

$$a = N + \tfrac{1}{2}, \qquad b = \infty$$

yields

$$\sum_{|j|>N} |\theta_j(t)| \leqq \frac{2}{\pi m}\left[\frac{m}{\pi\delta\left(N + \frac{1}{2} - \frac{T}{2h}\right)}\right]^m. \tag{55}$$

Let

$$m = \left[\frac{\pi\delta}{e}\left(N + \frac{1}{2} - \frac{T}{2h}\right)\right] \geqq 1; \tag{56}$$

then

$$\sum_{|j|>N} |\theta_j(t)| \leqq \frac{2}{\pi m} e^{-m}. \tag{57}$$

Thus, from equation (50), one obtains

$$\|f - g\|_u \leqq \frac{2M}{\pi m} e^{-m}, \tag{58}$$

and hence

$$d_{2N+1}^{C\,T}\left(B_\sigma^T(M)\right) \leqq \frac{2M}{\pi m}\, e^{-m}. \tag{59}$$

For n odd, one has

$$d_n^{C\,T}\left(B_\sigma^T(M)\right) \leqq \frac{2M}{\pi} \frac{\exp\left\{-\left[\frac{\pi\delta}{2e}\left(n-\frac{T}{h}\right)\right]\right\}}{\left[\frac{\pi\delta}{2e}\left(n-\frac{T}{h}\right)\right]}\,; \tag{60}$$

while if n is even, one has $d_n^{C\,T} \leqq d_{n-1}^{C\,T}$; hence, in all cases

$$d_n^{C\,T} \leqq \frac{2M}{\pi} \frac{\exp\left\{-\left[\frac{\pi\delta}{2e}\left(n-1-\frac{T}{h}\right)\right]\right\}}{\left[\frac{\pi\delta}{2e}\left(n-1-\frac{T}{h}\right)\right]}. \tag{61}$$

The fractional guardband δ is now chosen as in equation (48) from which the inequality of the theorem follows.

When n is large, a more accurate estimate of $d_n^{C\,T}$ may be obtained by using a polynomial approximation to B_σ^T. Let

$$f(t) = f\left(\frac{T}{2}\,x\right) = g(x), \tag{62}$$

and let $L(x)$ be the Lagrange interpolation polynomial established for $g(x)$ on the zeros of $T_n(x)$, the nth Tchebysheff polynomial of first kind, over $[-1, 1]$; then the standard error formula for Lagrange interpolation[11] yields

$$\max_{-1\leqq x\leqq 1} \mid g(x) - L(x) \mid \leqq \frac{1}{n!\,2^{n-1}} \max_{-1\leqq x\leqq 1} \mid g^{(n)}(x) \mid. \tag{63}$$

Bernstein's inequality (15) and equation (62) now yield

$$\left|\left| f(t) - L\left(\frac{2}{T}\,t\right) \right|\right|_u \leqq \frac{2M}{n!}\left(\frac{c}{2}\right)^n; \tag{64}$$

hence one has

Theorem 6:

$$d_n^{C\,T}\left(B_\sigma^T(M)\right) \leqq \frac{2M}{n!}\left(\frac{c}{2}\right)^n, \qquad n \geqq 0.$$

Let H_*^T be the space of functions $f(t)$ possessing derivatives up to

order s satisfying $f, \dot{f}, \cdots, f^{(s)} \in L^2(-T/2, T/2)$ and normed by equation (44); then Theorem 7 provides an estimate of the nth width of $B^T_{\sigma,s}(M)$ relative to H^T_s.

Theorem 7: Let

$$\Gamma = \left\{ \sum_{r=0}^{s-1} \frac{\mu_r}{T^{2r}(s-r-1)!^2\,(2s-2r-1)(s-r)} + \frac{2}{T^{2s}}\,\mu_s \right\}^{\frac{1}{2}},$$

in which $\mu_0 = 1$ *and the sum is considered zero when* $s = 0$, *then*

$$d^{H_s T}_{n+s}(B^T_{\sigma,s}(M)) \leqq \frac{M\Gamma\, T^{\frac{1}{2}}(2c)^{n+s}}{n!\binom{2n}{n}(2n+1)^{\frac{1}{2}}}.$$

Proof: For the function $g(x)$ of equation (62), the identity

$$g(x) = P(x) + \int_{-1}^{x} \frac{(x-u)^{s-1}}{(s-1)!}\, g^{(s)}(u)\, du, \tag{65}$$

(in which $P(x)$ is a polynomial of degree not exceeding $s-1$), will be used to obtain a polynomial approximation to $g(x)$ in the Sobolev norm (44). Let $L(x)$ be the Lagrange interpolation polynomial for $g^{(s)}(x)$ formed with n nodal points on $[-1, 1]$ and $\omega(x)$ the corresponding fundamental poynomial; then one has

$$g^{(s)}(x) = L(x) + \frac{1}{n!}\, g^{(n+s)}(\xi)\omega(x), \qquad \xi \in [-1, 1]. \tag{66}$$

The polynomial $I(x)$ defined by

$$I(x) = P(x) + \int_{-1}^{x} \frac{(x-u)^{s-1}}{(s-1)!}\, L(u)\, du \tag{67}$$

will be used to approximate $g(x)$ in the Sobolev norm; its degree does not exceed $n + s - 1$. Let

$$| g^{(j)}(x) | \leqq M_j, \qquad | x | \leqq 1; \tag{68}$$

then, from equation (66), one has

$$| g^{(r)}(x) - I^{(r)}(x) | \leqq \frac{M_{n+s}}{n!} \int_{-1}^{x} \frac{(x-u)^{s-r-1}}{(s-r-1)!}\, | \omega(u) |\, du, \qquad 0 \leqq r < s, \tag{69}$$

$$| g^{(s)}(x) - I^{(s)}(x) | \leqq \frac{M_{n+s}}{n!}\, | \omega(x) |. \tag{70}$$

The norm (44) for the interval $[-1, 1]$ may be written

$$|| g - I ||_s^2 = \int_{-1}^{1} \sum_{r=0}^{s} \nu_r \mid g^{(r)}(x) - I^{(r)}(x) \mid^2 dx, \tag{71}$$

in which $\nu_0 , \cdots , \nu_s \geqq 0$; the ν_r and μ_r are related through the change of variable $t = (T/2)x$. Using equations (69) and (70), one has

$$|| g - I ||_s^2 \leqq \frac{M_{n+s}^2}{n!^2} \cdot \int_{-1}^{1} \left\{ \sum_{r=0}^{s-1} \nu_r \left(\int_{-1}^{x} \frac{(x-u)^{s-r-1}}{(s-r-1)!} \mid \omega(u) \mid du \right)^2 + \nu_s \omega(x)^2 \right\} dx. \tag{72}$$

Define the function $k(u, v)$ by

$$k(u, v) = \int_{\max(u,v)}^{1} (x-u)^{s-r-1}(x-v)^{s-r-1} \, dx; \tag{73}$$

then equation (72) may be written:

$$|| g - I ||_s^2 \leqq \frac{M_{n+s}^2}{n!^2} \left\{ \sum_{r=0}^{s-1} \frac{\nu_r}{(s-r-1)!^2} \cdot \int_{-1}^{1} \int_{-1}^{1} k(u, v) \mid \omega(u)\omega(v) \mid du \, dv + \nu_s \int_{-1}^{1} \omega(x)^2 \, dx \right\}. \tag{74}$$

The Cauchy–Schwartz inequality shows that

$$k(u, v) \leqq \frac{(1-u)^{s-r-\frac{1}{2}}(1-v)^{s-r-\frac{1}{2}}}{2s-2r-1}; \tag{75}$$

hence,

$$|| g - I ||_s^2 \leqq \frac{M_{n+s}^2}{n!^2} \left\{ \sum_{r=0}^{s-1} \frac{\nu_r}{(s-r-1)!^2 (2s-2r-1)} \cdot \left(\int_{-1}^{1} (1-u)^{s-r-\frac{1}{2}} \mid \omega(u) \mid du \right)^2 + \nu_s \int_{-1}^{1} \omega(x)^2 \, dx \right\}. \tag{76}$$

Further application of the Cauchy–Schwartz inequality yields

$$|| g - I ||_s^2 \leqq \frac{M_{n+s}^2}{n!^2} \cdot \left\{ \sum_{r=0}^{s-1} \frac{\nu_r 2^{2s-2r}}{2(s-r-1)!^2 (2s-2r-1)(s-r)} + \nu_s \right\} \int_{-1}^{1} \omega(x)^2 \, dx. \tag{77}$$

A good choice for $\omega(x)$ is

$$\omega(x) = \frac{2^n}{\binom{2n}{n}} P_n(x), \tag{78}$$

where $P_n(x)$ is the nth Legendre polynomial. The coefficient of $P_n(x)$ in equation (78) makes $\omega(x)$ monic. Since

$$\int_{-1}^{1} P_n(x)^2 \, dx = \frac{2}{2n+1}, \tag{79}$$

one obtains

$$\| g - I \|_s \leqq \frac{2^n M_{n+s}}{n! \binom{2n}{n}} \left(\frac{2}{2n+1}\right)^{\frac{1}{2}} \cdot \left\{ \sum_{r=0}^{s-1} \frac{\nu_r 2^{2s-2r}}{2(s-r-1)!^2 (2s-2r-1)(s-r)} + \nu_s \right\}^{\frac{1}{2}}. \tag{80}$$

The Bernstein inequality (15) shows that

$$M_{n+s} \leqq M c^{n+s}; \tag{81}$$

hence

$$\| g - I \|_s \leqq \frac{M 2^n c^{n+s}}{n! \binom{2n}{n}} \left(\frac{2}{2n+1}\right)^{\frac{1}{2}} \cdot \left\{ \sum_{r=0}^{s-1} \frac{\nu_r 2^{2s-2r}}{2(s-r-1)!^2 (2s-2r-1)(s-r)} + \nu_s \right\}^{\frac{1}{2}}. \tag{82}$$

Finally the change of variable $t = (T/2)x$ and the replacement of ν_r by the original μ_r yield the result of the theorem.

The results of Theorems 5, 6, and 7, may be translated into estimates of entropy by use of the Mitjagin inequality of Theorem 1. The estimates so obtained will apply only to the subset of $B_\sigma(M)$ for which $f(t)$ is real. Doubling the bounds will provide estimates for complex valued $f(t)$.

Theorem 8: *Let* $0 < \alpha < 1$, $(1-\alpha)\epsilon < (2M/\pi e)$, $f(t)$ *real,*

$$m = \left\lfloor \ln \frac{2M}{\pi(1-\alpha)\epsilon} \cdot \frac{1 + \ln \frac{2M}{\pi(1-\alpha)\epsilon}}{1 + \ln \frac{2M}{\pi(1-\alpha)\epsilon} + \ln \ln \frac{2M}{\pi(1-\alpha)\epsilon}} \right\rfloor;$$

then

$$H_\epsilon(B_\sigma^T(M)) \leqq \left\{2 + \left[\frac{2c}{\pi}\left(1 + \left(\frac{e}{c}(m+1)\right)^{\frac{1}{2}}\right)^2\right]\right\} \log\left(\frac{2M}{\alpha\epsilon} + \frac{2-\alpha}{\alpha}\right).$$

Proof: According to Theorems 1 and 5, one must solve the inequality

$$\frac{2M}{\pi m} e^{-m} \geqq (1-\alpha)\epsilon \tag{83}$$

for the largest integer m; thus

$$me^m \leqq \frac{2M}{\pi(1-\alpha)\epsilon}. \tag{84}$$

Consider the function

$$F(x) = \delta - x - \ln x, \qquad \delta > 1. \tag{85}$$

One has

$$F'(x) = -1 - \frac{1}{x}; \tag{86}$$

hence, by the mean value theorem,

$$F(\delta - h) = -\ln \delta + h\left(1 + \frac{1}{\xi}\right), \qquad \delta - h < \xi < \delta. \tag{87}$$

Let

$$F(\delta - h) = 0; \tag{88}$$

then, since h is positive,

$$0 < -\ln \delta + h\left(1 + \frac{1}{\delta - h}\right), \tag{89}$$

$$0 < -\delta \ln \delta + h(1 + \delta + \ln \delta) - h^2; \tag{90}$$

thus

$$h > \frac{\delta \ln \delta}{1 + \delta + \ln \delta}, \tag{91}$$

and

$$\delta - h < \delta \frac{1 + \delta}{1 + \delta + \ln \delta}. \tag{92}$$

The inequality

$$x + \ln x \leqq \delta \tag{93}$$

is thus satisfied by

$$x < \delta \frac{1 + \delta}{1 + \delta + \ln \delta}; \tag{94}$$

hence, setting

$$\delta = \ln \frac{2M}{\pi(1 - \alpha)\epsilon} \tag{95}$$

and taking cognizance of the integral character of m, one has

$$m \leqq \left[\ln \frac{2M}{\pi(1 - \alpha)\epsilon} \cdot \frac{1 + \ln \frac{2M}{\pi(1 - \alpha)\epsilon}}{1 + \ln \frac{2M}{\pi(1 - \alpha)\epsilon} + \ln \ln \frac{2M}{\pi(1 - \alpha)\epsilon}} \right] \tag{96}$$

provided

$$(1 - \alpha)\epsilon < \frac{2M}{\pi e}. \tag{97}$$

For the computation of d_{n-1}, one has from equation (48)

$$m = \left[\frac{\pi}{2e} \delta_{n-2}^2 (n - 2) \right]. \tag{98}$$

Hence

$$\frac{\pi}{2e} \delta_{n-2}^2 (n - 2) < m + 1, \tag{99}$$

$$\left\{ 1 - \left(\frac{2c}{\pi(n - 2)} \right)^{\frac{1}{2}} \right\}^2 (n - 2) < \frac{2e}{\pi} (m + 1). \tag{100}$$

Let

$$n = 2 + \frac{2c}{\pi} \nu; \tag{101}$$

then

$$\nu \left(1 - \frac{1}{\nu^{\frac{1}{2}}} \right)^2 < \frac{e}{c} (m + 1); \tag{102}$$

accordingly

$$\nu < \left\{ 1 + \left(\frac{e}{c} (m + 1) \right)^{\frac{1}{2}} \right\}^2. \tag{103}$$

Thus n satisfies

$$n \leqq 2 + \left[\frac{2c}{\pi}\left\{1 + \left(\frac{e}{c}(m+1)\right)^{\frac{1}{2}}\right\}^2\right]; \tag{104}$$

the theorem now follows from equations (96), (97), (104), and Theorem 1.

When ϵ is small, a more accurate estimate of entropy may be obtained by use of Theorem 6 in place of Theorem 5. Accordingly one has

Theorem 9: *Let* $0 < \alpha < 1$, $\eta = (2M/(1-\alpha)\epsilon(\pi ec)^{\frac{1}{2}})$,

$$\eta \geqq \max\left(\frac{1}{\epsilon c}, e^{c/2}\right), \qquad f(t) \quad real;$$

then

$$H_\epsilon(B_\sigma^T(M)) \leqq \left\{1 + \left[\frac{2\ln\eta + \frac{1}{2} - \frac{1}{2}\ln\left(\frac{2}{ec}\ln\eta\right)}{\ln\left(\frac{2}{ec}\ln\eta\right) + 1 + \frac{1}{2\ln\eta}}\right]\right\}\log\left(\frac{2M}{\alpha\epsilon} + \frac{2-\alpha}{\alpha}\right).$$

Proof: According to Theorem 6, one may consider

$$\frac{2M}{n!}\left(\frac{c}{2}\right)^n \geqq (1-\alpha)\epsilon. \tag{105}$$

Stirling's formula provides the inequality

$$n! > n^n e^{-n}(2\pi n)^{\frac{1}{2}}, \tag{106}$$

and hence one may consider

$$\frac{2M}{(2\pi n)^{\frac{1}{2}}}\left(\frac{ec}{2n}\right)^n \geqq (1-\alpha)\epsilon. \tag{107}$$

Let

$$n = \frac{ec}{2}x, \qquad \eta = \frac{2M}{(1-\alpha)\epsilon(\pi ec)^{\frac{1}{2}}}; \tag{108}$$

then equation (107) becomes

$$x^{x+1/ec} \leqq \eta^{2/ec}. \tag{109}$$

Consider the function

$$F(x) = \delta - (x+a)\ln x; \tag{110}$$

then, by the mean value theorem

$$F(\delta - h) = \delta - (\delta + a) \ln \delta + h\left(\ln \delta + 1 + \frac{a}{\delta}\right) - \frac{h^2}{2}\left(\frac{1}{\xi} - \frac{a}{\xi^2}\right), \qquad \delta - h < \xi < \delta. \tag{111}$$

Let

$$x = \delta - h \geqq a, \qquad F(\delta - h) = 0; \tag{112}$$

then

$$0 \leqq \delta - (\delta + a) \ln \delta + h \ln \left(\delta + 1 + \frac{a}{\delta}\right). \tag{113}$$

Let

$$\delta \geqq 1/e; \tag{114}$$

then

$$h \geqq \frac{(\delta + a) \ln \delta - \delta}{\ln \delta + 1 + \dfrac{a}{\delta}}, \tag{115}$$

and hence

$$x \leqq \frac{2\delta + a - a \ln \delta}{\ln \delta + 1 + \dfrac{a}{\delta}}. \tag{116}$$

Thus, in terms of n and η, one has

$$n \leqq \left\lceil \frac{2 \ln \eta + \frac{1}{2} - \frac{1}{2} \ln \left(\dfrac{2}{ec} \ln \eta\right)}{\ln \left(\dfrac{2}{ec} \ln \eta\right) + 1 + \dfrac{1}{2 \ln \eta}} \right\rceil. \tag{117}$$

The lower bound on η in the theorem assures the satisfaction of the conditions on x and δ in equations (112) and (114). Use of Theorem 1 now provides the inequality of the theorem.

Theorem 10 provides an entropy estimate deduced from the width result of Theorem 7.

Theorem 10: $0 < \alpha < 1$, $\eta = (M\Gamma(2c)^s/(1 - \alpha)\epsilon(e\sigma)^{\frac{1}{2}})$

$$\eta \geqq \max\left(\frac{1}{\epsilon c}, e^{c/2}\right), \qquad \gamma = (1 + \mu_1\sigma^2 + \cdots + \mu_s\sigma^{2s})^{\frac{1}{2}}, \quad f(t) \text{ real};$$

then

$$H_\epsilon(B^T_{\sigma,s}(M)) \leqq \left\{ s + 1 + \left[\frac{2 \ln \eta + \frac{1}{2} - \frac{1}{2} \ln \left(\frac{2}{ec} \ln \eta \right)}{\ln \left(\frac{2}{ec} \ln \eta \right) + 1 + \frac{1}{2 \ln \eta}} \right] \right\} \log \left(\frac{2M\gamma T^{\frac{1}{2}}}{\alpha\epsilon} + \frac{2 - \alpha}{\alpha} \right).$$

Proof: The investigation parallels that of Theorem 9. A difference occurs in the estimation of d_0 . The Bernstein inequality of (15) shows that

$$d_0 \leqq M\gamma T^{\frac{1}{2}}; \tag{118}$$

hence the estimate of the theorem.

The case $m = 1$ of the representation given in Theorem 4 may be used to obtain an explicit ϵ-net for $B_\sigma(M)$, and hence to provide a constructive algorithm for the transmission of information from such a source. The representation for $f(t) \in B_\sigma(M)$ takes the form

$$f(t) = \sum_{j=-\infty}^{\infty} f(jh) \frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh) \sin \frac{\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh) \quad \frac{\sigma}{1-\delta}(t - jh)},$$

$$\sigma h = \pi(1 - \delta). \tag{119}$$

In order to proceed, it is necessary to estimate the quantity $A(\delta)$ given by

$$A(\delta) = \sup_{-\infty<t<\infty} \sum_{j=-\infty}^{\infty} \left| \frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh) \sin \frac{\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh) \quad \frac{\sigma}{1-\delta}(t - jh)} \right|. \tag{120}$$

Theorem 11: $A(\delta) \leqq 1/\delta^{\frac{1}{2}}$ for $0 < \delta < 1$.

Proof: The Cauchy–Schwartz inequality yields

$$A(\delta)^2 \leqq \sup_{-\infty<t<\infty} \sum_{j=-\infty}^{\infty} \left[\frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh)} \right]^2 \cdot \sup_{-\infty<t<\infty} \sum_{j=-\infty}^{\infty} \left[\frac{\sin \frac{\sigma}{1-\delta}(t - jh)}{\frac{\sigma}{1-\delta}(t - jh)} \right]^2. \tag{121}$$

From the Parseval relation of Theorem 3, one has

$$\sum_{j=-\infty}^{\infty} \left[\frac{\sin \frac{\sigma}{1-\delta}(t-jh)}{\frac{\sigma}{1-\delta}(t-jh)} \right]^2 = \frac{1}{h} \int_{-\infty}^{\infty} \left[\frac{\sin \frac{\sigma}{1-\delta}(t-s)}{\frac{\sigma}{1-\delta}(t-s)} \right]^2 ds = 1, \quad (122)$$

$$\sum_{j=-\infty}^{\infty} \left[\frac{\sin \frac{\delta\sigma}{1-\delta}(t-jh)}{\frac{\delta\sigma}{1-\delta}(t-jh)} \right]^2 = \frac{1}{h} \int_{-\infty}^{\infty} \left[\frac{\sin \frac{\delta\sigma}{1-\delta}(t-s)}{\frac{\delta\sigma}{1-\delta}(t-s)} \right]^2 ds = \frac{1}{\delta}. \quad (123)$$

The theorem is established.

Let

$$S = \sup_{-\infty<j<\infty} |(fjh)|; \quad (124)$$

then a corollary to Theorem 11 is

Corollary:

$$\sup_{-\infty<t<\infty} |f(t)| \leqq S/\delta^{\frac{1}{2}}.$$

Proof: From equations (119) and (120), one has

$$\sup_{-\infty<t<\infty} |f(t)| \leqq SA(\delta). \quad (125)$$

The result follows from Theorem 11.

The function

$$g(t) = \sum_{|j| \leqq N} f(jh) \frac{\sin \frac{\delta\sigma}{1-\delta}(t-jh)}{\frac{\delta\sigma}{1-\delta}(t-jh)} \frac{\sin \frac{\sigma}{1-\delta}(t-jh)}{\frac{\sigma}{1-\delta}(t-jh)} \quad (126)$$

constitutes an approximation to $f(t)$. The error may be assessed by application of equation (51) for $m = 1$, and Sonin's formula (53); thus

$$||f - g||_u \leqq \frac{M}{\pi^2\delta} \left(\left(N + \frac{1}{2} - \frac{T}{2h} \right)^{-1} + \left(N + \frac{1}{2} + \frac{T}{2h} \right)^{-1} \right). \quad (127)$$

For $0 < \alpha < 1$, let

$$N = \left[\frac{M}{(1-\alpha)\epsilon\pi^2\delta} \left(1 + \left\{ 1 + \left(\frac{c(1-\alpha)\epsilon\,\delta\pi}{(1-\delta)M} \right)^2 \right\}^{\frac{1}{2}} \right) - \frac{1}{2} \right] + 1; \quad (128)$$

then direct verification establishes

$$||f - g||_u < (1-\alpha)\epsilon. \quad (129)$$

It may be observed that for large c, one has

$$N \cong \frac{c}{\pi(1-\delta)} = \frac{T}{2h}\,; \tag{130}$$

that is, N is approximately the number of nodal points jh in $(-T/2, T/2)$.

Let

$$\beta_i(f) = \left[\frac{A(\delta)f(jh)}{2\alpha\epsilon}\right] \tag{131}$$

and

$$\beta(f) = (\beta_{-N}(f), \cdots, \beta_N(f)); \tag{132}$$

then the set U_β is defined to consist of all $f(t)$ generating the same vector $\beta = \beta(f)$. It will now be shown that the diameter of U_β does not exceed 2ϵ. Let $f_1(t), f_2(t)$ ε U_β ; then

$$|f_1(jh) - f_2(jh)| \leqq \frac{2\alpha\epsilon}{A(\delta)}. \tag{133}$$

One has

$$f_1(t) - f_2(t) = \sum_{j=-\infty}^{\infty} (f_1(jh) - f_2(jh)) \cdot \frac{\sin\frac{\delta\sigma}{1-\delta}(t-jh)}{\frac{\delta\sigma}{1-\delta}(t-jh)} \frac{\sin\frac{\sigma}{1-\delta}(t-jh)}{\frac{\sigma}{1-\delta}(t-jh)}\,; \tag{134}$$

and hence, by equation (129),

$$|f_1(t) - f_2(t)| \leqq \sum_{|j|\leq N} |f_1(jh) - f_2(jh)| \cdot \left|\frac{\sin\frac{\delta\sigma}{1-\delta}(t-jh)}{\frac{\delta\sigma}{1-\delta}(t-jh)} \frac{\sin\frac{\sigma}{1-\delta}(t-jh)}{\frac{\sigma}{1-\delta}(t-jh)}\right| + 2(1-\alpha)\epsilon \tag{135}$$

in which N is chosen as in equation (128). From equation (133), one has

$$|f_1(t) - f_2(t)| \leqq \frac{2\alpha\epsilon}{A(\delta)} \cdot \sum_{|j|\leq N} \left|\frac{\sin\frac{\delta\sigma}{1-\delta}(t-jh)}{\frac{\delta\sigma}{1-\delta}(t-jh)} \frac{\sin\frac{\sigma}{1-\delta}(t-jh)}{\frac{\sigma}{1-\delta}(t-jh)}\right| + 2(1-\alpha)\epsilon. \tag{136}$$

Use of equation (120) shows that

$$\| f_1 - f_2 \|_u \leqq 2\epsilon. \tag{137}$$

The sets U_β are centerable with respect to themselves; that is, there exists an element $g(t) \in U_\beta$ whose distance from any other element of U_β does not exceed ϵ. Consider the function $g(t)$ defined by

$$g(t) = \frac{2\alpha\epsilon}{A(\delta)} \sum_{|j| \leqq N} (\beta_j(f) + \tfrac{1}{2}) \frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh)} \frac{\sin \frac{\sigma}{1-\delta}(t - jh)}{\frac{\sigma}{1-\delta}(t - jh)}. \tag{138}$$

One has

$$| f(jh) - g(jh) | \leqq \frac{\alpha\epsilon}{A(\delta)}, \qquad | j | \leqq N, \tag{139}$$

and

$$\begin{aligned} f(t) - g(t) &= \sum_{|j| \leqq N} (f(jh) - g(jh)) \frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh)} \frac{\sin \frac{\sigma}{1-\delta}(t - jh)}{\frac{\sigma}{1-\delta}(t - jh)} \\ &\quad + \sum_{|j| > N} f(jh) \frac{\sin \frac{\delta\sigma}{1-\delta}(t - jh)}{\frac{\delta\sigma}{1-\delta}(t - jh)} \frac{\sin \frac{\sigma}{1-\delta}(t - jh)}{\frac{\sigma}{1-\delta}(t - jh)}; \end{aligned} \tag{140}$$

hence, by equations (120), (129), and (139)

$$\| f - g \|_u \leqq \epsilon. \tag{141}$$

The required constructive algorithm, Γ, is thus given by the mapping $f \to g$ in equations (131) and (138).

Theorem 12: $V(\Gamma) = (2N + 1) \log \{[A(\delta)M/2\alpha\epsilon] - [-A(\delta)M/2\alpha\epsilon] + 1\}$, *in which* N *is given in equation* (128).

Proof: It is necessary to enumerate the number of distinct $g(t)$ which are generated by $\Gamma(B_\sigma(M))$. Since

$$\frac{A(\delta) | f(jh) |}{2\alpha\epsilon} \leqq \frac{A(\delta)M}{2\alpha\epsilon}, \tag{142}$$

the number of distinct values of $\beta_i(f)$ is

$$\left[\frac{A(\delta)M}{2\alpha\epsilon}\right] - \left[-\frac{A(\delta)M}{2\alpha\epsilon}\right] + 1, \tag{143}$$

and hence the number of distinct vectors $\beta(f)$ is

$$\left\{\left[\frac{A(\delta)M}{2\alpha\epsilon}\right] - \left[-\frac{A(\delta)M}{2\alpha\epsilon}\right] + 1\right\}^{2N+1}. \tag{144}$$

The theorem follows from equation (144).

Corollary 1: $V(\Gamma) \leqq (2N + 1) \log ([M/2\alpha\epsilon\delta^{\frac{1}{2}}] - [-M/2\alpha\epsilon\delta^{\frac{1}{2}}] + 1)$.
Proof: Theorem 11.

Corollary 2. $V(\Gamma) \leqq (2N + 1) \log (M/\alpha\epsilon(\delta)^{\frac{1}{2}} + 2)$.
Proof: Corollary 1 and the inequalities

$$\left[\frac{M}{2\alpha\epsilon(\delta)^{\frac{1}{2}}}\right] \leqq \frac{M}{2\alpha\epsilon(\delta)^{\frac{1}{2}}}, \tag{145}$$

$$-\left[-\frac{M}{2\alpha\epsilon(\delta)^{\frac{1}{2}}}\right] < \frac{M}{2\alpha\epsilon(\delta)^{\frac{1}{2}}} + 1.$$

IV. THEORETICAL INVESTIGATION OF W_σ

Using Theorem 3 for $f, g \in W_\sigma$, the Sobolev inner product

$$(f, g)_s = \int_{-T/2}^{T/2} (f\bar{g} + \mu_1 \dot{f}\dot{\bar{g}} + \cdots + \mu_s f^{(s)}\overline{g^{(s)}})\, dt \tag{146}$$

takes the form

$$(f, g)_s = \int_{-\sigma}^{\sigma}\int_{-\sigma}^{\sigma} \frac{\sin \frac{T}{2}(u - v)}{\pi(u - v)} \cdot (1 + \mu_1 uv + \cdots + \mu_s u^s v^s) F(u)\bar{G}(v)\, du\, dv, \tag{147}$$

in which $F(u)$, $G(u)$ are the Fourier transforms of f, g respectively. The corresponding positive definite quadratic form Q is

$$Q = \|f\|_s^2 = \int_{-\sigma}^{\sigma}\int_{-\sigma}^{\sigma} \frac{\sin \frac{T}{2}(u - v)}{\pi(u - v)} \cdot (1 + \mu_1 uv + \cdots + \mu_s u^s v^s) F(u)\bar{F}(v)\, du\, dv, \tag{148}$$

and an operator K generating Q is given by

$$KF = \int_{-\sigma}^{\sigma} \frac{\sin \frac{T}{2}(u - v)}{\pi(u - v)} \cdot (1 + \mu_1 uv + \cdots + \mu_s u^s v^s) F(v)\, dv, \qquad |u| \leqq \sigma; \tag{149}$$

thus

$$Q = \int_{-\sigma}^{\sigma} \bar{F} K F\, du. \tag{150}$$

The equation defining the eigenvalues and eigenfunctions of K is

$$K\Phi_k = \lambda_k \Phi_k, \qquad k \geqq 0, \tag{151}$$

in which the ordering $\lambda_0 \geqq \lambda_1 \geqq \lambda_2 \geqq \cdots$ is used. It follows from the Hilbert–Schmidt theory[12] that the eigenvalues are denumerable and of finite multiplicity and the eigenfunctions form an orthonormal set which, from the positive definite character of K, is complete in $L^2(-\sigma, \sigma)$.

Let

$$\varphi_k(t) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\sigma}^{\sigma} e^{iut} \Phi_k(u)\, du; \tag{152}$$

then the Parseval relation for Fourier transforms shows that the sequence $\phi_0(t), \phi_1(t), \phi_2(t), \cdots$ is orthonormal over $(-\infty, \infty)$; further, from equations (147), (150), and (151), one has

$$(\phi_j, \phi_k)_s = \int_{-\sigma}^{\sigma} \bar{\Phi}_k K \Phi_j\, du = \lambda_j \int_{-\sigma}^{\sigma} \Phi_j \bar{\Phi}_k\, du = 0 \quad j \neq k$$
$$= \lambda_j \quad j = k. \tag{153}$$

Thus the sequence $\{\phi_k(t)\}_0^\infty$ forms an orthogonal system with respect to the Sobolev inner product (146). The system $\{\phi_k(t)\}_0^\infty$ is also complete in $W_{\sigma,s}^T$ as a consequence of the completeness of the system $\{\Phi_k(u)\}_0^\infty$ in $L^2(-\sigma, \sigma)$.

Define the n-dimensional subspace $X_n \subset W_\sigma$ by

$$X_n = X_n(\phi_0, \cdots, \phi_{n-1}) \tag{154}$$

then Theorem 13 provides the nth width of $W_{\sigma,s}^T(B)$, relative to H_s^T, in terms of the eigenvalues of K.

Theorem 13: $d_n^{H_s{}^T}(W_{\sigma,s}^T(B)) = B\lambda_n^{\frac{1}{2}}$.

Proof: Let $f(t) \in W_{\sigma,s}(B)$; then

$$f(t) = \sum_{k=0}^{\infty} a_k \phi_k(t). \tag{155}$$

Let

$$g(t) = \sum_{k=0}^{n-1} a_k \phi_k(t) \, \varepsilon \, X_n \, ; \tag{156}$$

then the orthogonality of the $\phi_k(t)$, (153), yields

$$|| f - g ||_s^2 = \sum_{k=n}^{\infty} | a_k |^2 \lambda_k \, . \tag{157}$$

Thus

$$\inf_{g \varepsilon X_n} || f - g ||_s^2 \leqq \sum_{k=n}^{\infty} | a_k |^2 \lambda_k \, . \tag{158}$$

From the monotonicity of the λ_k , one has

$$\inf_{g \varepsilon X_n} || f - g ||_s^2 \leqq \lambda_n \sum_{k=n}^{\infty} | a_k |^2 ; \tag{159}$$

however, the orthonormality of the $\phi_k(t)$ over $(-\infty, \infty)$ shows that

$$\int_{-\infty}^{\infty} | f(t) |^2 \, dt = \sum_{k=0}^{\infty} | a_k |^2 \leqq B^2, \tag{160}$$

and hence from equation (159)

$$E_{X_n}(W_{\sigma,s}^T(B)) = \sup_{f \varepsilon W_{\sigma,s}(B)} \inf_{g \varepsilon X_n} || f - g ||_s \leqq B \lambda_n^{\frac{1}{2}}. \tag{161}$$

Thus

$$d_n(W_{\sigma,s}^T(B)) \leqq B \lambda_n^{\frac{1}{2}}. \tag{162}$$

Consider the ball U_{n+1} defined by

$$g(t) = \sum_{k=0}^{n} a_k \phi_k(t), \qquad || g ||_s \leqq B \lambda_n^{\frac{1}{2}}; \tag{163}$$

then, by a theorem on balls in a finite dimensional subspace of a Banach space,[2]

$$d_n(U_{n+1}) = B \lambda_n^{\frac{1}{2}}. \tag{164}$$

Thus the theorem will be established if it is shown that the ball U_{n+1} defined in equation (163) is contained in $W_{\sigma,s}(B)$. It is only necessary to verify that

$$\int_{-\infty}^{\infty} | g(t) |^2 \, dt = \sum_{k=0}^{n} | a_k |^2 \leqq B^2. \tag{165}$$

One has from

$$|| g ||_s^2 = \sum_{k=0}^{n} | a_k |^2 \lambda_k \leqq B^2 \lambda_n \tag{166}$$

that

$$\sum_{k=0}^{n} | a_k |^2 \leqq \sum_{k=0}^{n} | a_k |^2 \frac{\lambda_k}{\lambda_n} \leqq B^2, \tag{167}$$

and hence the theorem is proved.

Use of the series representation of Theorem 4 permits one to estimate $d_n^{H_0{}^T}(W_{\sigma,0}^T(B))$. The quantities m and δ_n are as in equation (48); additionally, the corresponding interval h_n is defined by $h_n = \pi(1 - \delta_n)/\sigma$.

Theorem 14:

$$d_n^{H_0{}^T}(W_{\sigma,0}^T(B)) \leqq \frac{2}{\pi} \frac{B}{h_{n-1}^{\frac{1}{2}}} \frac{e^{-m}}{\left((2m+1)\left(n - 1 - \frac{T}{h_{n-1}}\right)\right)^{\frac{1}{2}}}.$$

Proof: Let

$$f(t) = \sum_{j=-\infty}^{\infty} f(jh)\theta_j(t), \tag{168}$$

and

$$g(t) = \sum_{|j| \leqq N} f(jh)\theta_j(t); \tag{169}$$

then

$$| f(t) - g(t) | \leqq \sum_{|j|>N} | f(jh) | | \theta_j(t) |. \tag{170}$$

Since, by Parseval's relation of Theorem 3

$$\int_{-\infty}^{\infty} | f(t) |^2 dt = h \sum_{j=-\infty}^{\infty} | f(jh) |^2 \leqq B^2, \tag{171}$$

Schwartz's inequality applied to equation (170) yields

$$| f(t) - g(t) |^2 \leqq \frac{B^2}{h} \sum_{|j|>N} \theta_j(t)^2. \tag{172}$$

One has, from equation (51)

$$|| f - g ||_U^2 \leqq \frac{2B^2}{\pi^2 h} \left(\frac{m}{\pi\delta}\right)^m \sum_{j>N} \frac{1}{\left(j - \frac{T}{2h}\right)^{2m+2}}, \qquad N > \frac{T}{2h}. \tag{173}$$

One may use Sonin's formula, equation (53), to effect the summation in equation (173); thus

$$|| f - g ||_v \leqq \frac{B}{\pi} \frac{\sqrt{2}}{\left(h(2m+1)\left(N + \frac{1}{2} - \frac{T}{2h}\right)\right)^{\frac{1}{2}}} \left[\frac{m}{\pi\delta\left(N + \frac{1}{2} - \frac{T}{2h}\right)}\right]^m. \tag{174}$$

The choice

$$m = \left[\frac{\pi\delta}{e}\left(N + \frac{1}{2} - \frac{T}{2h}\right)\right] \tag{175}$$

leads to

$$|| f - g ||_v \leqq \frac{B}{\pi} \frac{\sqrt{2}\, e^{-m}}{\left(h(2m+1)\left(N + \frac{1}{2} - \frac{T}{2h}\right)\right)^{\frac{1}{2}}}. \tag{176}$$

Thus equation (176) shows that

$$d_{2N+1}^{H_0{}^T}(W_{\sigma,0}^T(B)) \leqq \frac{B}{\pi}\left(\frac{2}{h}\right)^{\frac{1}{2}} \frac{e^{-m}}{\left((2m+1)\left(N + \frac{1}{2} - \frac{T}{2h}\right)\right)^{\frac{1}{2}}}; \tag{177}$$

and, hence, for n odd

$$d_n^{H_0{}^T}(W_{\sigma,0}^T(B)) \leqq \frac{2}{\pi}\frac{B}{h^{\frac{1}{2}}} \frac{e^{-m}}{\left((2m+1)\left(n - \frac{T}{h}\right)\right)^{\frac{1}{2}}}. \tag{178}$$

For n even, one has

$$d_n^{H_0{}^T} W_{\sigma,0}^T(B) \leqq \frac{2}{\pi}\frac{B}{h^{\frac{1}{2}}} \frac{e^{-m}}{\left((2m+1)\left(n - 1 - \frac{T}{h}\right)\right)^{\frac{1}{2}}}; \tag{179}$$

thus equation (179) applies in all cases. The fractional guardband is now chosen as in equation (48), and the inequality of the theorem follows.

Theorem 13 permits an immediate corollary to be obtained from Theorem 14.

Corollary: *For* $s = 0$, *one has*

$$\lambda_n \leqq \frac{4}{\pi^2 h_{n-1}} \frac{e^{-2m}}{(2m+1)\left(n - 1 - \frac{T}{h_{n-1}}\right)}.$$

As was done in Theorem 7, polynomial approximation will be used to estimate $d_n^{H^{T,s}}(W_{\sigma,s}^T(B))$. The estimate is given in Theorem 15.

Theorem 15:

$$d_{n+s}^{H^{T,s}}(W_{\sigma,s}^T(B)) \leqq B\Gamma\left(\frac{2c}{\pi}\right)^{\frac{1}{2}} \frac{(2c)^{n+s}}{n!\binom{2n}{n}((2n+1)(2n+2s+1))^{\frac{1}{2}}}.$$

Proof: The estimate will be obtained from equation (80). In order to estimate M_{n+s}, consider

$$f(t) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\sigma}^{\sigma} e^{iut} F(u)\, du, \tag{180}$$

from which one has

$$g(x) = \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\sigma}^{\sigma} e^{iu(T/2)x} F(u)\, du. \tag{181}$$

Accordingly

$$g^{(r)}(x) = \left(\frac{T}{2}\right)^r \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\sigma}^{\sigma} e^{iu(T/2)x} (iu)^r F(u)\, du. \tag{182}$$

By use of the Schwartz inequality, one obtains

$$| g^{(r)}(x) |^2 \leqq \left(\frac{T}{2}\right)^{2r} \frac{1}{2\pi} \int_{-\sigma}^{\sigma} u^{2r}\, du \int_{-\sigma}^{\sigma} | F(u) |^2\, du. \tag{183}$$

The Parseval relation for Fourier transforms

$$\int_{-\infty}^{\infty} | f(t) |^2\, dt = \int_{-\sigma}^{\sigma} | F(u) |^2\, du \leqq B^2 \tag{184}$$

and equation (184) now yields

$$| g^{(r)}(x) |^2 \leqq B^2 \frac{\sigma}{\pi} \frac{c^{2r}}{2r+1}. \tag{185}$$

Thus

$$M_{n+s} \leqq B\left(\frac{\sigma}{\pi}\right)^{\frac{1}{2}} \frac{c^{n+s}}{(2n+2s+1)^{\frac{1}{2}}}. \tag{186}$$

The remainder of the analysis is the same as in Theorem 7.

Theorem 13 again permits an immediate corollary to be obtained from Theorem 15.

Corollary:

$$\lambda_{n+s} \leqq \frac{2c}{\pi}\,\Gamma^2\,\frac{(2c)^{2n+2s}}{n!^2\binom{2n}{n}^2(2n+1)(2n+2s+1)}.$$

Theorems 14 and 15 lead to corresponding estimates of entropy through use of Theorem 1.

Theorem 16: Let

$$n \geqq 2 + \left(1 + \left(\frac{2c}{\pi}\right)^{\frac{1}{2}}\right)^2, \qquad 0 < \alpha < 1, \quad f(t) \quad \textit{real},$$

and

$$m = \left[\ln\left(\frac{2B}{\pi\sqrt{3}\,(1-\alpha)\epsilon}\left(\frac{2c}{\pi}\right)^{\frac{1}{4}}\right)\right];$$

then

$$H_\epsilon(W^T_{\sigma,0}(B)) \leqq \left\{2 + \left[\frac{2c}{\pi}\left(1 + \left(\frac{e}{c}(m+1)\right)^{\frac{1}{2}}\right)^2\right]\right\}\log\left(\frac{2B\lambda_0^{\frac{1}{2}}}{\alpha\epsilon} + \frac{2-\alpha}{\alpha}\right).$$

Proof: From Theorem 14, one has

$$d_{n-1}(W^T_{\sigma,0}(B)) \leqq \frac{2}{\pi}B\left(\frac{1}{h_{n-2}}\right)^{\frac{1}{2}}\frac{e^{-m}}{\left((2m+1)\left(n-2-\frac{T}{h_{n-2}}\right)\right)^{\frac{1}{2}}}. \tag{187}$$

From equation (48), one has

$$2m + 1 \geqq 3, \qquad \frac{T}{h_{n-2}} = \left(\frac{2c}{\pi}(n-2)\right)^{\frac{1}{2}}; \tag{188}$$

hence

$$d_{n-1}(W^T_{\sigma,0}(B)) \leqq \frac{2}{\pi\sqrt{3}}B\left(\frac{2c}{\pi}\right)^{\frac{1}{4}}e^{-m}\frac{1}{\left((n-2)^{\frac{1}{2}} - \left(\frac{2c}{\pi}\right)^{\frac{1}{2}}\right)^{\frac{1}{2}}}. \tag{189}$$

Since

$$\frac{1}{\left((n-2)^{\frac{1}{2}} - \left(\frac{2c}{\pi}\right)^{\frac{1}{2}}\right)^{\frac{1}{2}}} \leqq 1 \quad \text{for} \quad n \geqq 2 + \left(1 + \left(\frac{2c}{\pi}\right)^{\frac{1}{2}}\right)^2, \tag{190}$$

$d_{n-1}(W^T_{\sigma,0}(B))$ obeys the inequality

$$d_{n-1}(W^T_{\sigma,0}(B)) \leqq \frac{2}{\pi\sqrt{3}} B\left(\frac{2c}{\pi}\right)^{\frac{1}{4}} e^{-m}. \tag{191}$$

According to Theorem 1, one may consider

$$\frac{2}{\pi\sqrt{3}} B\left(\frac{2c}{\pi}\right)^{\frac{1}{4}} e^{-m} \geqq (1 - \alpha)\epsilon; \tag{192}$$

and hence

$$m = \left[\ln\left(\frac{2B}{\pi\sqrt{3}\,(1-\alpha)\epsilon}\left(\frac{2c}{\pi}\right)^{\frac{1}{4}}\right)\right]. \tag{193}$$

The remaining analysis is the same as that of Theorem 8. The inequality of the theorem now follows.

Theorem 17: *Let* $0 < \alpha < 1$, $\eta = \Gamma B(2c)^s/(1-\alpha)\epsilon e(\pi c)^{\frac{1}{2}}$,

$$\eta \geqq \max\left(\frac{4}{e^2c^2}, e^{c/2}\right), \qquad f(t) \quad real,$$

then

$$H_\epsilon(W^T_{\sigma,s}(B)) \leqq \left\{s + 1 + \left[\frac{2\ln\eta + 1 - \ln\left(\frac{2}{ec}\ln\eta\right)}{\ln\left(\frac{2}{ec}\ln\eta\right) + 1 + \frac{1}{\ln\eta}}\right]\right\} \log\left(\frac{2B\lambda_0^{\frac{1}{2}}}{\alpha\epsilon} + \frac{2-\alpha}{\alpha}\right).$$

Proof: The proof parallels that of Theorem 9.

It may be useful to observe

$$\lambda_0 \leqq \frac{2c}{\pi}\gamma^2. \tag{194}$$

IV. ACKNOWLEDGMENTS

I should like to express my appreciation for the careful review of the manuscript given by the referee in which he brought to my attention an important error. I should also like to thank Dr. H. Heffes for his thorough reading of the paper and for his suggestions.

REFERENCES

1. Heffes, H., Horing, S., and Jagerman, D., "On the Design and Analysis of a Class of PCM Systems," to be published in the March 1971 B.S.T.J.
2. Lorentz, G. G., *Approximation of Functions*, New York: Holt, Rinehart and Winston, 1966.

3. Kolmogorov, A. N., and Tikhomirov, W. M., *Arbeiten Zur Informationstheorie III*, Mathematische Forschungsberichte, X, Berlin: VEB Deutscher Verlag der Wissenschaften, 1960.
4. Vitushkin, A. G., *Theory of the Transmission and Processing of Information*, New York: Pergaman Press, 1961.
5. Achieser, N. I., *Theory of Approximation*, New York: Frederick Ungar 1956.
6. Timan, A. F., *Theory of Approximation of Functions of a Real Variable*, New York: Macmillan, 1963, Chapter IV.
7. Hardy, G. H., *Divergent Series*, London: Clarendon Press, 1949, p. 50.
8. Whittaker, J. M., *Interpolatory Function Theory*, London: Cambridge University Press, 1935.
9. Paley, R. E. A. C., and Wiener, N., "Fourier Transforms in the Complex Domain," American Mathematical Society Colloquium Publications, Vol. XIX, 1934.
10. Helms, H. D., and Thomas, J. B., "Truncation Error of Sampling-Theorem Expansions," Proc. of the I.R.E., *50*, No. 1 (February 1962), pp. 179–182.
11. Fort, Tomlinson, *Finite Differences and Difference Equations in the Real Domain*, England: Oxford University Press, 1948.
12. Tricomi, F. G., *Integral Equations*, New York: Interscience, 1965.

A Satellite System for Avoiding Serial Sun-Transit Outages and Eclipses

By C. W. LUNDGREN

(Manuscript received March 20, 1970)

The motions of satellites phased in particular, slightly inclined orbits are timed so that different satellites are north and south of the equator when sun-caused outages occur in geostationary equatorial systems.

I. INTRODUCTION

Communication satellite systems experience predictable service interruptions involving the sun. A sun-transit outage occurs when the pointing angles from a receiving earth terminal to a satellite and to the sun so nearly coincide that the additional noise power presented by the sun renders transmission unusable.[1] When a satellite passes through the earth's shadow, its solar primary power is interrupted and its sunlight-dependent heat balance is upset.

A geostationary system serving a common coverage region may include several satellites spaced less than 10° (175 mrad) in the synchronous equatorial orbit. Figure 1 illustrates the timing of sun transits and eclipses occurring in rapid series for three geostationary satellites during one day at the spring equinox, observed from an earth terminal located on the equator at longitude 0°W. One sun transit near noon and one eclipse 12 hours later are observed for each satellite served by this terminal. Eclipses of closely spaced satellites may occur at the same time, and sun transits of different satellites may also occur simultaneously within a large coverage region.

Daily sun transits of all geostationary satellites serving an earth terminal occur during one week in the spring and again in the fall. Service interruptions can last five minutes or more per satellite. Affected outage regions are large and move so rapidly that terrestrial restoration is unattractive.

Conversely, a minimum of one working and one spare geostationary satellite are required for restoration independent of terrestrial facilities.

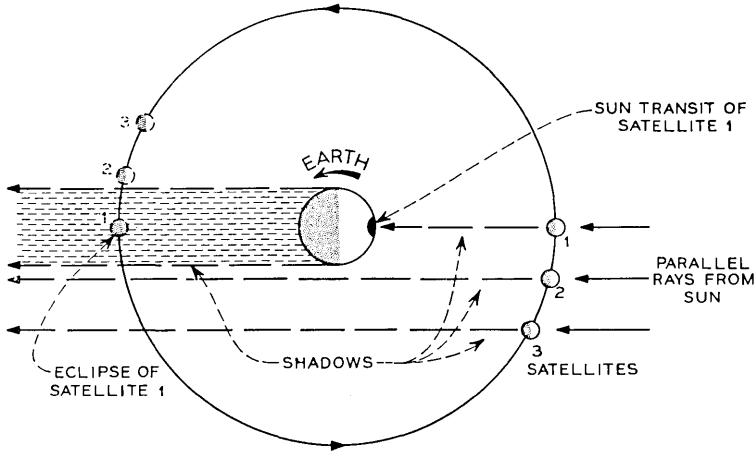


Fig. 1—Sun transits for earth terminal on equator, and eclipses of geostationary satellites at equinox.

Such redundancy is also required for adequate protection against satellite failure, since satellite replenishment intervals are prohibitively large.

A fully redundant geostationary system incorporates duplicate transmissions to working and spare satellites and duplicate reception from these satellites continuously at all earth terminals. Partially redundant systems depend upon redirection of earth antenna beams to spare satellites.*

Rapid, highly coordinated switching between geostationary satellites is required at all earth terminals to restore serial sun-transit outages. Numerous residual transmission "hits" result from such switching. Also, the orbit spacing must be sufficiently large to prevent simultaneous mutual outages of the different satellites at different locations within the coverage region to avoid additional switching complexity. A spacing as large as 8° (140 mrad) is necessary to prevent mutual sun-transit outages within the contiguous United States.†

Alternatively, serial sun transits are avoided by phasing the satellites in particular, slightly inclined orbits with motions timed so that one satellite is north of the equator and the other is south during both the spring and fall outage events. Only one switch of reception between

* If the earth terminals are equipped with duplicate antennas, transmitters, and receivers, the capacity of both satellites can be utilized except during outage periods.
† The 48 continental states, excluding Alaska and Hawaii.

the separated satellites is required per sun-transit season. The exact timing (hour) is unimportant and may be different for the convenience of each earth terminal. Except for these two switches, all earth terminals throughout the coverage region are afforded uninterrupted reception throughout the year. Mutual sun transits within the same coverage region are also avoided by this satellite diversity, and the large orbit spacing discussed above for geostationary satellites is unnecessary.‡

II. SUN TRANSITS AND ECLIPSES

Sun transits and eclipses of geostationary satellites occur during the spring and fall seasons. The exact dates of the former depend primarily upon the latitude of the receiving earth terminal.

2.1 *Sun Transits of Geostationary Satellites*

The geometry and duration associated with a sun transit are controlled by (*i*) the off-axis gain of a properly pointed earth antenna, (*ii*) the receiving system noise temperature, (*iii*) the solar noise power profile, and (*iv*) the minimum acceptable signal-to-noise ratio.

In Fig. 2 the sun's rays are assumed to be parallel; refraction cor-

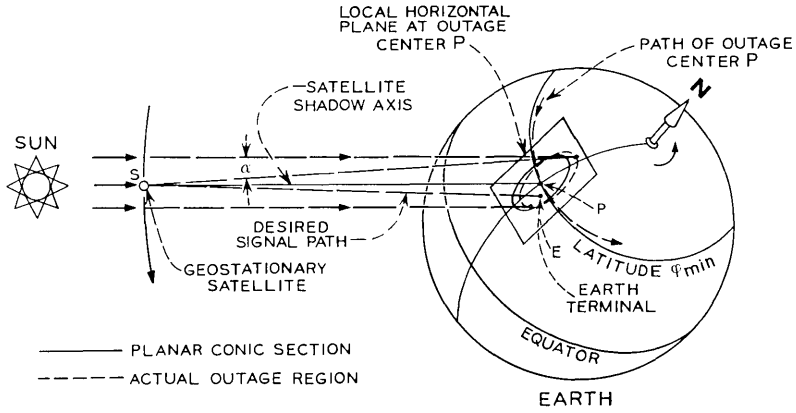


Fig. 2—Planar approximation of geography affected by a sun-transit outage.

rections are neglected, assuming a sufficiently large angle of incidence to the atmosphere for the desired ray **SE**. The affected outage region is defined approximately by the locus of all points on the illuminated earth's surface for which earth antennas aimed at satellite S also point

‡ See Sections 3.3 and A.5.

a prescribed minimum angular distance $\alpha°$ away from the sun's center.

An estimate of the geography involved is provided by the elliptical intersection of a cone of angular radius $\alpha°$, symmetrical about satellite-shadow axis **SP** with its apex at S, and the horizontal plane at P. It is elongated north-south in the figure.

The sun is assumed to be a uniform disk source of thermal noise about 0.5° in diameter.* Shapes and magnitudes of the solar noise power profile vary strongly with time and radio frequency. Edge brightening at the lower microwave frequencies approaches a factor of two, and comparable variations of total flux with time are common.[2,3]

A minimum solar noise temperature for the mean quiet sun (total flux averaged over the disk) is about 25,000 K for a single polarization, inferred from measurements at a wavelength of 10.3 cm.[2-4] This is approximately the minimum temperature presented to a sun-pointed ideal antenna at 4 GHz whose beamwidth is less than 0.5°.

Convolution of an appropriate solar noise profile with a known earth antenna gain pattern provides an estimate of increased noise versus angular displacement of the sun center from the main beam axis. Estimates for the minimum displacement permitting acceptable reception at 4 GHz range from about 0.6° (10 mrad) for very large earth antennas (30 m) to greater than 1° (18 mrad) for small antennas (8m).[1,5] Corresponding minor axes of outage regions range from 800 to 1300 km. Major axes occurring along satellite-earth longitudes are equal to the minor axes at the equator and approach 1.5 times the latter at high latitudes.

Because of synchronism between earth rotation and satellite revolution, each outage region appears to move. One at 41° north latitude traverses the contiguous United States from west to east in approximately one-half hour at noon of the time zone at the satellite's longitude (see Appendix A).

Figure 3 illustrates the path of an outage region. Each path is tangent to the latitude intercept of the center of the satellite's shadow at apparent noon at the satellite's longitude. For all other longitudes in the Northern Hemisphere, the path lies slightly to the north of this latitude.

Hence, in very late February or early March, short daily outages affect earth terminals situated near the United States–Canadian border. Two to three days later these terminals experience maximum outages lasting five minutes or more, depending upon transmission parameters and permissible signal-to-noise ratios. Outages at these terminals end

* The optical disk has a diameter of about 29 minutes of arc, in geocentered angular measure.

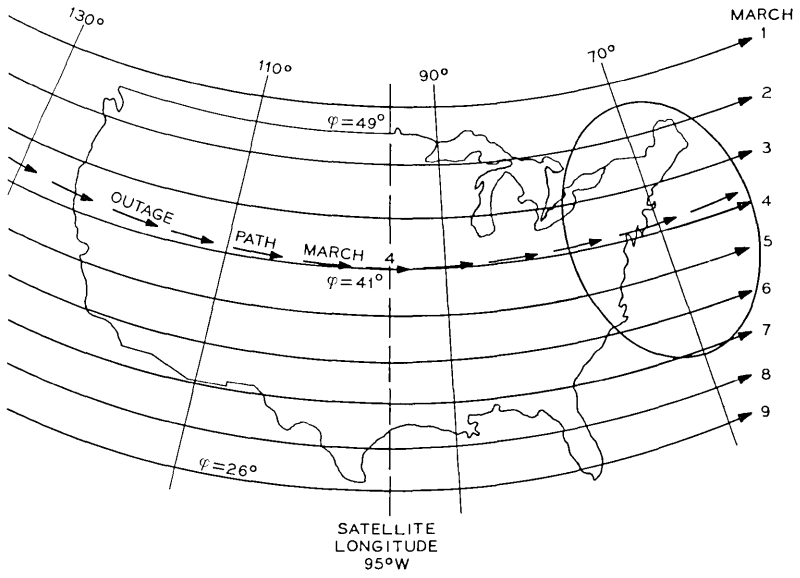


Fig. 3—Approximate paths of sun-transit outages for geostationary satellite.

after an additional two to three days, the outage paths progressing southward at a rate of about 3° latitude per day. All outages affecting United States earth terminals above north latitude 26° cease prior to mid-March.

Conversely, in the fall the daily outage paths progress from south to north, affecting southern United States terminals about October 1 and ending in the north about mid-October.

In Fig. 3, based on parameters adopted in Appendix A, a given earth terminal is affected about six days, twice yearly, while the contiguous United States experiences outages throughout a 14-day period, again twice yearly. If a multiple-feed antenna or a rapid-slewing antenna is employed to switch reception at an earth terminal from an affected satellite to another 6.8° (120 mrad) westward in the geostationary orbit, transmission from the latter satellite is interrupted only 30 minutes later.

2.2 *Eclipses of Geostationary Satellites*

Eclipses of geostationary satellites can be expected for a total of about 90 evenings per year in the spring and fall. Concurrent eclipses occur for geostationary satellites spaced less than 17.6° (310 mrad).

Eclipses occur near apparent midnight of the time zone at each satellite's longitude, beginning in late February or early March and ending by mid-April. Fall events begin about September 1 and end about mid-October. Eclipses lasting about 70 minutes occur on the dates of the spring and fall equinoxes; those lasting longer than one hour occur about 50 days per satellite per year.

Communication satellites are provided with batteries to prevent circuit outages and to maintain antenna pointing, attitude control, station keeping, telemetry, and command capabilities during eclipses. However, concomitant voltage and temperature fluctuations, loss of the solar reference for antenna pointing, and related ground command activities may contribute to an increased likelihood of satelite failure or a reduction in transmission capacity during eclipses.

III. DIVERSITY SYNCHRONOUS SATELLITE SYSTEM

A minimum arrangement of two slightly inclined, circular synchronous orbits with deliberate phasing of one working and one spare satellite in their respective orbits is suggested for providing space diversity during outage periods. The specific orbit parameters and satellite phasing are chosen so that they may remain unchanged throughout the year. Thus satellite station-keeping fuel expenditures are comparable to geostationary values. The parameters are also chosen so that only one noncritical handover of reception between satellites is required per sun-transit season.

3.1 *Basic Satellite Phasing in Specific Inclined Orbits*

Figure 4 illustrates the relationship between a "figure 8" pattern traced out by a synchronous satellite and the magnitude of its orbit inclination. Recent descriptions of such patterns are given by Rowe and Penzias,[6] treating the efficient use of orbit longitude.

Figure 5 illustrates the satellite phasing and timing of motions required for a two-satellite diversity system. The time reference selected for describing these motions is initial time t_0 mean solar hours, marking the advent of 12 o'clock noon (apparent, or sun time) on the date of the vernal equinox at average θ of mean longitudes θ_1 and θ_2 degrees west for satellites S_1 and S_2, respectively ($\theta = \langle \theta_1 + \theta_2 \rangle_{av}$). For satellites sharing radio frequency bands, a minimum orbit spacing between interfering satellites is generally specified consistent with resolving powers of the earth antennas. Accordingly, a minimum satellite spacing x degrees is assumed between mean longitudes θ_1 and θ_2.

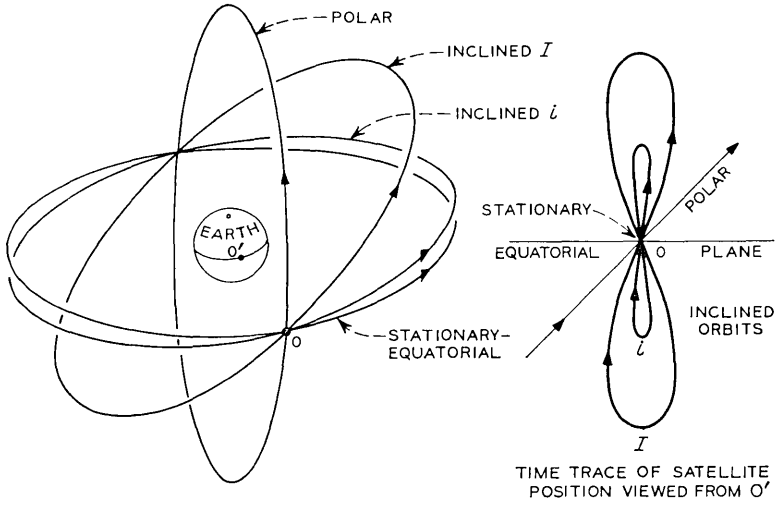


Fig. 4—Earth synchronous orbits and figure 8 patterns.

In terms of θ, chosen for service to a particular geographical region, the mean orbit longitudes shown in Fig. 5 are

$$\theta_1 = (\theta - x/2), \qquad \theta_2 = (\theta + x/2) \quad \text{degrees west.} \tag{1}$$

Dimensions of 8 patterns allowing adequate diversity between properly phased satellites are determined in Appendix B. Peak satellite displacements from the equatorial plane (geostationary orbit in Fig. 5) coincide in the spring and fall with sun transits of each satellite's mean longitude meridian.

For example, in Fig. 5(a) satellite S_1 is northernmost in its 8 pattern prior to apparent noon at average longitude θ. To an observer located at earth longitude θ_1, this coincides with alignment of the sun behind the 8 pattern for satellite S_1.

At apparent noon at longitude θ, satellite S_1 in Fig. 5(b) moves very slowly toward the geostationary orbit, while S_2 is approaching the southernmost point in its 8 pattern. The sun is located midway between the 8 patterns.

Shortly after apparent noon at longitude θ, the sun aligns behind the 8 pattern for satellite S_2, as observed from earth longitude θ_2. At this time, satellite S_2 reaches its peak excursion, while satellite S_1 moves more rapidly towards the geostationary orbit [Fig. 5(c)]. Tick marks

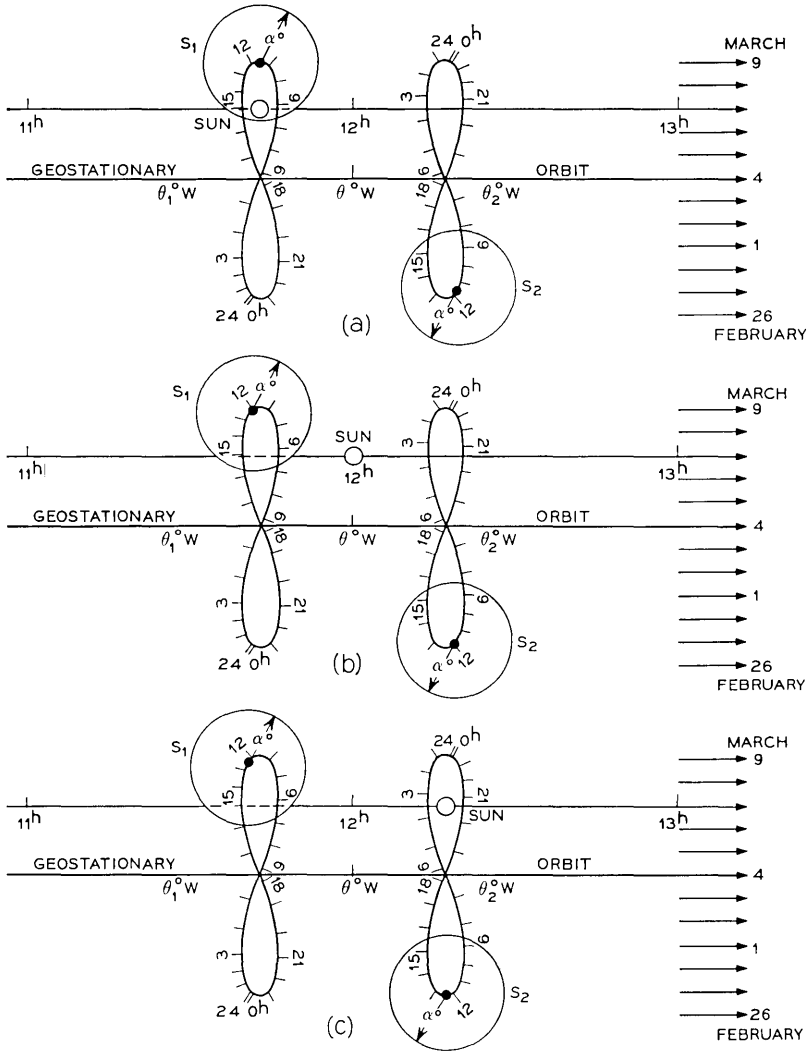


Fig. 5—Phased satellite motions.

on the 8 patterns are labeled according to each satellite's location at times referred to longitude θ.

Paths of the sun on consecutive days during the spring sun-transit season are also indicated. Note that these daily paths progress from south to north in accordance with a decreasing southern declination of the sun's rays at this time of year (cf., Fig. 3).

Circles of radius $\alpha°$ centered at each satellite define the minimum pointing angle to the sun for earth antennas directed at the satellites. Hence, reception from satellite S_1 is interrupted when the sun is within the circle for S_1. Tick marks give positions of the sun along its path, again at times referred to longitude θ.

In Fig. 5, uninterrupted reception from satellite S_1 is assumed throughout the late fall and winter, until March 7. At any convenient time between March 1 and March 7, an earth terminal observing these motions redirects its reception from satellite S_1 to satellite S_2. This allows uninterrupted reception from S_2 until the fall sun-transit season, during which this noncritical procedure is reversed.

Note that the 8 patterns in Fig. 5 are larger than required by a single earth terminal. The dimensions determined in Appendix B are sufficient to prevent serial sun transits throughout the entire latitude range of the coverage region, so that only one outage region from either satellite may traverse any part of the coverage region on any day. This simplifies switching between satellites in restoration schemes involving large numbers of working satellites and a minimum of one spare satellite.* However, for the basic scheme involving duplicate transmission via equal numbers of working and spare satellites, the dimensions of the 8 patterns may be reduced until the outage circles ($\alpha°$) are almost tangent to the geostationary orbit. Redirection of the earth antenna appropriate for Fig. 5 is required on or about March 4 for such reduced 8 patterns.

Note also that the satellites spend most of the time near the extremes of the 8 patterns, providing near-maximum diversity separation for several hours near noon. This tolerance to timing errors is particularly useful since the apparent alignments of the sun in Fig. 5 and the timing of transit events are somewhat different for observers at different locations within the coverage region. Allowances are made in Appendix B in the computation of required diversity separation for both latitude and longitude ranges of the coverage region, assuming that uninterrupted reception from the unaffected satellite is required continuously at all earth terminals throughout the coverage region.

The diversity performance is made nearly independent of arbitrary satellite spacing x by phasing each satellite so that its maximum latitude excursion occurs at sun transit of its mean longitude meridian.

Tick marks in Fig. 6 illustrate a daily progression of satellite positions at apparent noon at longitude θ throughout the year. This regular shift is observed in the *ideal* case at the earth terminals because such

* Discussed in Section 3.3.

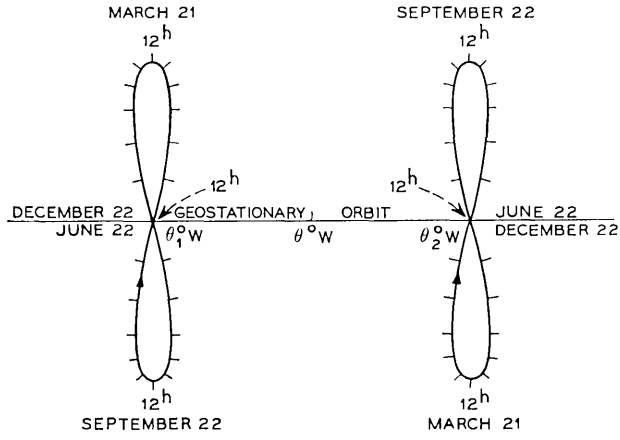


Fig. 6—Shift of daily satellite positions.

inclined synchronous orbits tend to maintain fixed orientations in space as the earth revolves around the sun (about 1° per day), illustrated in Fig. 7, by virtue of conservation of orbit angular momenta $\mathbf{m}_1$ and $\mathbf{m}_2$. Orbit perturbations, or departures from the above ideal motions, are approximately the same as those for geostationary orbits and are corrected by firing small station-keeping rocket motors at intervals throughout the lifetime of the satellites.

Specification of orbit stabilization with respect to the fixed stars is necessary to obtain properly timed satellite diversity automatically throughout the year; the precision required for diversity is needed only during outage seasons.

Hence, the daily period of satellite motions in their figure 8 patterns is less than 24 hours of civil time (mean solar hours). The actual sidereal period is $23^h\ 56^m\ 04^s.09054$ in mean solar time measure.

The daily shift of positions is utilized, by the deliberate orbit orientations and satellite phasing in the orbits, so that the apparent positions of satellites S_1 and S_2 are reversed automatically in time for diversity reception again during the fall outage season (see Fig. 7). Positions are also reversed daily, providing diversity for satellite eclipses near midnight, assuming sufficiently large orbit inclinations.

Conversely, lesser accumulated shifts must also be considered in computing the minimum diversity separation for sun transits for coverage regions located far from the equator, since sun transits occur either before, or after actual equinoxes (see Figs. 5 and 6, and Appen-

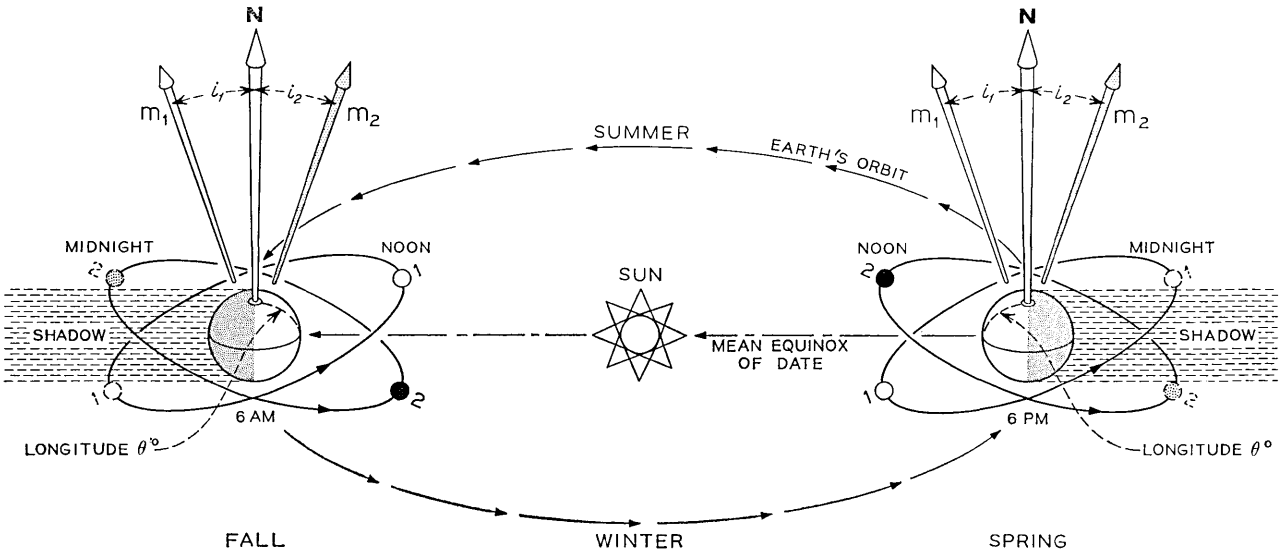


Fig. 7—Satellite space diversity with respect to the sun and seasons.

dix B). Sun transits are observed in the Northern Hemisphere prior to the vernal equinox and again after the autumnal equinox. Offsets of approximately two weeks from the symmetrical case are representative for the contiguous United States. Of course, dates of satellite eclipses are independent of earth latitude and the ideal symmetry is applicable.

3.2 *Orbit Parameters*

Satellite motions and initial conditions are illustrated in Figs. 7 and 8 for two diversity satellites.

3.2.1 *Inclination of Orbits*

The planes of orbits for satellites S_1 and S_2 are tilted slightly with respect to the earth's equatorial plane by inclination angles i_1 and i_2.

For the idealized case of equal inclinations, the minimum required magnitudes range from about 2 degrees for avoiding serial sun transits to about 9 degrees for avoiding serial and concurrent eclipses (see Appendix B).

3.2.2 *Alignment of Inclined Orbit Planes*

Positioning of the figure 8s is accomplished by aligning the orbit planes in slightly offset opposition as shown in Fig. 8. Two plane intersections with the earth's equator result, each forming acute angles $(90 - x/2)$ degrees symmetrically with the mean equinox axis (intersection of planes of the equator and of the earth's orbit around the sun; direction from earth towards the sun at the vernal equinox).

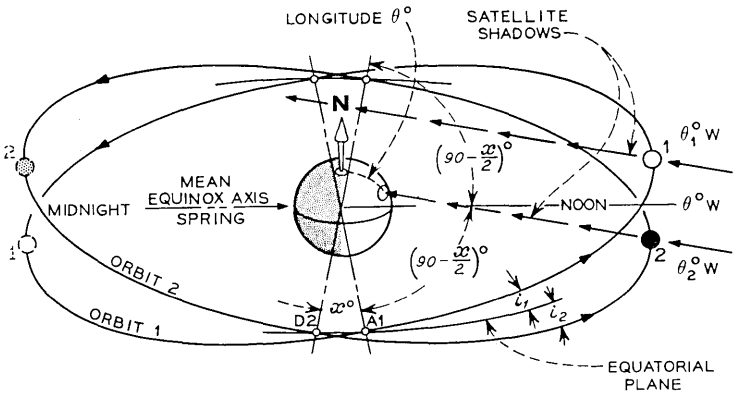


Fig. 8—Synchronous orbits phased for sun diversity.

3.2.3 *Phasing of Satellite S_1*

The time of the ascending node in orbit 1 for satellite S_1, for spacing x degrees is

$$t_1 = t_0 - [(6 + x/30)] \text{ mean solar hours,} \quad (2)$$

so that at time $t = t_0 - x/30$ hours satellite S_1 necessarily assumes its maximum north latitude (upper limit of excursion for left-hand figure 8 pattern in Fig. 5). From Fig. 8, note that the semi-major axis of the 8 in geocentered angular measure is equivalent numerically to orbit inclination i_1.

3.2.4 *Phasing of Satellite S_2*

The descending node in orbit 2 for satellite S_2 is specified by

$$t_2 = t_0 - [(6 - x/30)] \text{ mean solar hours,} \quad (3)$$

for which satellite S_2 assumes its maximum south latitude at time $t = t_0 + x/30$ hours.

3.2.5 *Satellite Motions Related to the Sun and Seasons*

By synchronizing satellite motions and timing with respect to the earth's revolution about the sun as shown in Fig. 7, the required space diversity is obtained during both spring and fall outage seasons.

Satellite motions and timing are specified above in terms of initial conditions at the vernal equinox. Of course, actual satellite launching is not restricted to any season, provided that satellite motions coincide with those for the specified system at the times when sun-caused outages occur in geostationary equatorial systems.

3.3 *Phased Multisatellite Systems*

Two satellites are required for the basic diversity system. The diversity satellites may be placed as desired in orbit longitude consistent with an assumed minimum orbit spacing x.

An obvious system growth is to add uniformly spaced, alternately phased working and spare satellites along the orbit (Fig. 5). Note that one of a diversity pair of spare satellites can restore all working satellites if fast switching may be employed daily at the affected earth terminals. Reception is transferred in sequence between transitted active satellites and the unaffected spare.* The orbit spacing between *second*-adjacent satellites (same phasing) should be sufficient to prevent mutual sun transits of the latter satellites within the coverage region.

* The affected spare is available as an additional working satellite.

For this case, only half of the orbit spacing required by geostationary satellites is required by the diversity satellites (about 4° for avoiding mutual outages within the contiguous United States).

Conversely, for satellites which may be closely spaced ($x \doteq 1°$), efficient use of the orbit may result from judicious incorporation, in a manner consistent with the satellite phasing and timing described above, of orbit loading techniques suggested by Rowe and Penzias.[6] Deliberate relative phases in adjacent 8 patterns may prevent major multiple sun transits of all satellites near the same latitude and minimize daily switching to different unaffected satellites (Section 3.1).

For large orbit inclinations,* (*i*) tracking of satellite and earth antennas is required, (*ii*) reduction in latitude of the coverage region results, (*iii*) transmission at low angles of arrival is more susceptible to atmospheric degradation, and (*iv*) the interference exposure between radio relay and satellite services is increased.

3.4 *Antenna Requirements for Earth Terminals and Satellites*

Only slight geometric departures from the geostationary case are required to obtain diversity for avoiding sun-transit outages; somewhat larger departures are required for avoiding eclipses. Hence, satellite radio transmission parameters appropriate for corresponding geostationary designs are essentially retained.

Earth antennas need follow only slow and very small periodic satellite motions. These motions are accomodated reliably by conventional 24-hour cyclic cam drives (sidereal time measure). Costs and maintenance for such antenna drives are virtually insignificant when compared with those for full automatic tracking. Cyclic drives are appropriate for a large deployment of small earth antennas requiring moderate beam-pointing precision, while costs for full automatic tracking are less significant for a smaller number of large antennas requiring precise beam pointing.

A minimum earth antenna steering requirement accommodating orbit inclinations up to 10° (175 mrad) and satellite longitude drifts from assigned orbit stations of $\pm 10°$, for satellite elevations of 5° or more, is reported by the Communications Satellite Corporation for quasi-stationary satellites.[7] Such earth terminals are compatible with the diversity satellites, since in the ideal application the smaller desired orbit inclinations are also maintained continuously.

The spin axis of a satellite is maintained perpendicular to its orbit plane, in the simplest wheel-mode attitude stabilization. Satellite

---

* For $x = 1°$, $i \doteq 10.7°$ and for $x = 5°$, $i \doteq 24°$, from equation (7) of Ref. 6.

antenna pointing referred to this axis benefits from partial compensation of pointing errors otherwise accompanying departures from the equatorial plane in the inclined orbits.*

IV. CONCLUSIONS

Space diversity is provided automatically at times of sun transits and eclipses by a convenient modification of a geostationary system in which the satellites appear to move in figure 8 patterns. Alternate satellites are oppositely phased, so that when one satellite is north the other is south. Orbit orientations and timing of satellite motions are arranged so that near the spring and fall equinoxes, when geostationary satellites transit the sun, the diversity satellites are at extreme north and south positions, allowing uninterrupted reception from at least one satellite.

The contiguous United States is cleared of serial sun-transit outages if orbit inclinations of about two degrees are employed. Concurrent satellite eclipses are also reduced in frequency and duration, and are avoided by increasing the orbit inclinations to about nine degrees.

Neglecting perturbations common to synchronous orbits including the geostationary orbit, the satellite deployment is steady state. Satellite launching requirements, mean station-keeping precision, and lifetimes are comparable to the geostationary case.

Diversity is provided automatically during both spring and fall outage seasons, requiring two noncritical switches between satellites per year.

Relatively minor modifications of earth terminals and satellites designed for geostationary service are required.

The diversity satellites are positioned as desired in orbit longitude without degrading system performance significantly, consistent with minimum orbit spacings to control interference from neighboring satellites.

Transmission via the unaffected satellite of a diversity pair can be switched in sequence daily to restore all transitted active satellites of a larger system.

One-half the minimum orbit spacing required by geostationary systems to prevent mutual outages of neighboring satellites within large coverage areas is required by the diversity system, since only alternate satellites experience outages on a given day.

Sun-transit outages in satellite circuits can be restored without involving terrestrial facilities.

* For $i = 2°$, a peak uncompensated pointing error of 0.3° is representative.

V. ACKNOWLEDGMENTS

The writer wishes to express appreciation to H. W. Evans and to many colleagues who encouraged this presentation, to D. Jarett for many helpful suggestions, and to L. C. Thomas for suggesting a time-correction accounting for earth rotation during outages.

APPENDIX A

Simplified Geometry and Numerical Examples for Geostationary Satellites

A.1 *Minimum-Latitude Circle Tangent to Outage Path*

At the satellite longitude, conjunction of the sun and satellite occurs at apparent noon and the satellite's shadow intercepts a minimum latitude, shown in Fig. 9. On successive days in the spring, the shadow path becomes tangent to a smaller north latitude at the satellite's longitude, and lies slightly to the north of this latitude for all other longitudes.

Figure 9 illustrates the sun's rays on March 4, 1970. From an almanac, the apparent declination of the sun for 0 hours ephemeris time (E.T.) is $-6°\ 40'\ 54''.5$ and on March 5, is $-6°\ 17'\ 49''.0$.[8]

Ephemeris transit of the sun on March 4 is given as $12^h\ 11^m\ 50^s.39$ and the reduction ΔT from universal time (U.T.) to E.T. for the year 1970.5 is approximately 40^s. The ephemeris time corresponding to solar transit at west longitude $\lambda°$ is

$$\text{E.T.} \doteq \text{E.T. (TRANSIT)} + [1.002738]\,\frac{\lambda}{360}\,(24^h)\ \text{hours}, \qquad \lambda < 180°, \tag{4}$$

where the coefficient in brackets is the approximate ratio of the mean solar day to the mean sidereal day. Allowing for a 6-hour time difference from the Greenwich Meridian to the Central Time Zone,

$$\text{C.S.T.} \doteq \text{E.T.} - \Delta T - 6^h\ \text{hours}. \tag{5}$$

Assume a transit of geostationary satellite stationed at $\lambda = 95°$W:

$$\text{C.S.T.} \doteq 12^h\ 11^m.84 + 6^h 20^m.93 - 0^h\ 0^m.67 - 6^h\ 0^m. \tag{6}$$

From equation (4), the ephemeris time of this event is $18^h\ 32^m.77$ on March 4. Interpolating between 0^h on March 4 and 0^h on March 5, the sun's apparent declination is

$$D \doteq -6.682° + \frac{18.55}{24.00}\,(0.385°), \tag{7}$$

$$\doteq -6.38°.$$

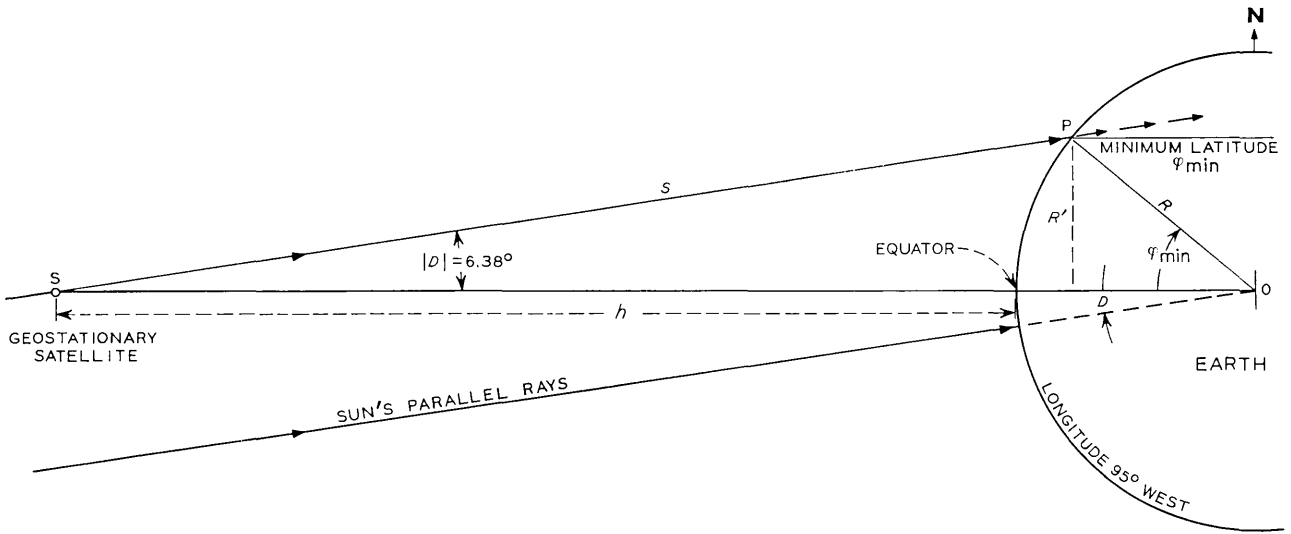


Fig. 9—Determination of minimum outage path latitude φ_{min}.

Note from Fig. 9,

$$\sin \varphi_{\min} = -\frac{S}{R} \sin D, \tag{8}$$

$$S = \frac{h + R(1 - \cos \varphi_{\min})}{\cos D} \text{ km}, \tag{9}$$

where $\varphi_{\min}$ is the north latitude of the satellite's shadow at the time of sun transit.
Then

$$\sin \varphi_{\min} = -\frac{\sin D\left[(1 - \cos \varphi_{\min}) + \frac{h}{R}\right]}{\cos D}, \tag{10}$$

from which it is determined that $\varphi_{\min} \doteq 41.0°$ north latitude, assuming geostationary orbit altitude $h = 35{,}900$ km and mean spherical earth radius $R = 6373$ km.

A.2 *Estimate of Speed with Which Outage Centers Traverse U.S.A.*

Figure 10 shows the contiguous United States represented by a longitude span of 60° centered at the satellite longitude and located at north latitude $\varphi_{\min}(41°)$. Consider a projection of the extreme longitude meridians (i.e., $\pm 30°$ referred to the satellite longitude) parallel to an assumed shadow axis between the span center at B′ and the satellite at B, such that orbit arc intercept $\overset{\frown}{AC}$ is specified.

The geocentric orbit radius is

$$AO = CO = R + h \doteq 42{,}270 \text{ km}. \tag{11}$$

Then the radius of latitude circle $\varphi_{\min}$ is

$$R'' = R \cos \varphi_{\min} \doteq 4810 \text{ km}. \tag{12}$$

The approximate distance measured along latitude circle $\varphi_{\min}$ for this model of the United states is

$$| \overset{\frown}{A'C'} |_{60°} = \frac{2\pi 60}{360} R'' \doteq 5040 \text{ km}. \tag{13}$$

Recognizing equilateral triangle A′OC′, the orbit chord is

$$AC = A'C' = R'' \doteq 4810 \text{ km}. \tag{14}$$

The solution of an oblique triangle with sides a, b, c and opposite angles A, B, C is

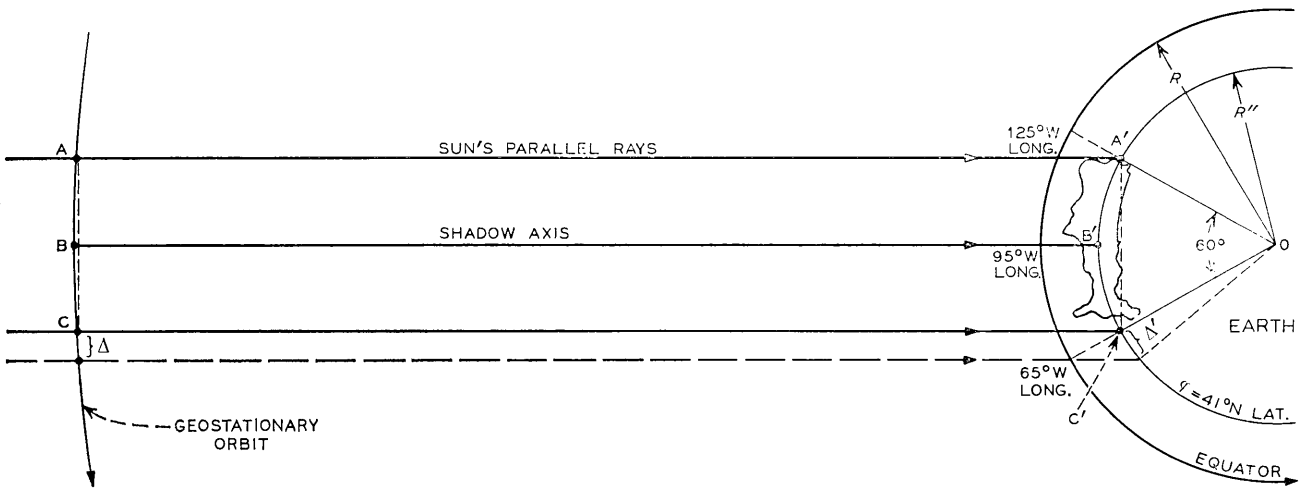


Fig. 10—Geometry describing motion of an outage region across a model of the contiguous United States.

$$\cos A = \frac{-a^2 + b^2 + c^2}{2bc}, \tag{15}$$

so that the orbit arc may be found from

$$\cos (A\text{-}O\text{-}C) = \frac{-(AC)^2 + 2(R + h)^2}{2(R + h)^2}. \tag{16}$$

Then the desired geocentered angle representing the orbit intercept of all parallel sun's rays simultaneously illuminating this model is

$$\overset{\frown}{AC} = \cos^{-1}(A\text{-}O\text{-}C) \doteq 6.5 \text{ degrees}. \tag{17}$$

The time required for a satellite's shadow to traverse a stationary representation of the United States is numerically equivalent to the time for the fractional revolution of a satellite from position A to C:

$$t_{\text{sta}} \doteq 6.5° \times \frac{(24 \times 60)^{\text{m}}}{360°} = 26.0 \text{ min}. \tag{18}$$

However, the actual elapsed time t_t is greater by virtue of earth rotation during this interval. The effective longitude span of the United States is very approximately

$$\overset{\frown}{A'C'} + \overset{\frown}{\Delta'} \doteq 60° + 26^{\text{m}}.0 \times \frac{15°}{60^{\text{m}}} \doteq 66.5°. \tag{19}$$

Accounting for a correspondingly enlarged orbit intercept,

$$t_t \doteq t_{\text{sta}} \times \frac{66.5°}{60°} \doteq 28.8 \text{ min}, \tag{20}$$

so that an outage region traverses the United States from west to east in approximately one-half hour. The exact interval depends primarily upon φ_{min}.

A.3 *Estimation of Size of Outage Region—Example*

A conic figure of revolution about axis **SP** in Fig. 2 defining the affected outage region subtends total angle 2α measured at the satellite. To enable example calculations without specific reference to antenna pattern data, a worst-case minimum angular separation $\alpha = 1°$ between a satellite and the sun center is adopted.*

*The value $\alpha = 1°$ is assumed for a hypothetical 4-GHz satellite system incorporating 55 percent efficient, 30-ft diameter parabolic reflector earth antennas, a receiving system noise temperature of 200 K, and a 3-dB allowable increase in received thermal noise power.

The horizontal plane at the location of the satellite's shadow P at time of transit is shown in edge view in Fig. 11. Slant range $S = SP$ is found from equation (9) to be about 37,470 km. The conic section defined by this plane and the outage cone is elliptical; point P specifies its motions.

The east-west semi-minor axis r is equivalent to the radius of the right circular intersection of the cone and a plane through P normal to satellite-shadow axis **SP**:

$$r = S \sin \alpha$$
$$\doteq 655 \text{ km}. \tag{21}$$

The north-south semi-major axis r' in Fig. 11 is found from a projection of the above circular intersection upon the local horizontal plane at P:

$$r' = \frac{r}{\cos (\varphi_{\min} - D)}$$
$$\doteq 970 \text{ km}. \tag{22}$$

A.4 *Estimate of Outage Duration*

The maximum duration of an outage occurring at an earth terminal located on latitude $\varphi_{\min}$ is approximately that fraction of time t_t [equation (20)] for the satellite's shadow to travel the 1310-km width of the

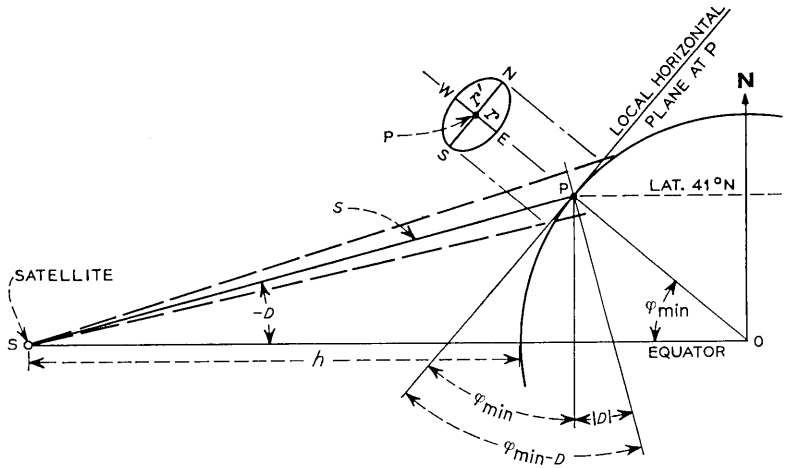


Fig. 11—Determination of outage region at P.

above outage pattern. Allowing for earth rotation, and noting that the 60° longitude span $\overset{\frown}{A'C'}$ at latitude $\varphi_{\min}$ corresponds to chord $A'C'$:

$$t_t \times \frac{2r}{|\overset{\frown}{A'C'}|} < t_d < t_t \times \frac{2r}{A'C'} ; \tag{23}$$

$$7.5 \text{ minutes} < t_d < 7.8 \text{ minutes}.$$

For the satellite stationed at longitude 95° west, the path of its shadow on March 4, 1970, approaches latitude 41° north near Omaha at 12:32 p.m. C.S.T. Taking dimensions of the outage region into account, the West Coast should just begin to experience outages north of Eureka, California, at about 10:16 a.m. Pacific Standard Time, and the last outages, near Boston, should cease about 1:50 p.m. Eastern Standard Time.

A.5 *Geostationary Satellite Spacing and Serial Outages*

Several identical satellites are assumed deployed along the geostationary orbit. Earth terminals are assumed capable of receiving signals from at least one pair of adjacent satellites, either simultaneously or one at a time. The orbit spacing between satellites is assumed to be uniform, but adjustable to alter the timing of serial sun transits. Numerical assumptions made in previous sections are retained for illustration; earth terminals are assumed to be located along the outage path (worst case).

A.5.1 *Case 1–Minimum of 30 Minutes Between Switches at an Earth Terminal*

If each satellite is assumed to possess spare circuit capacity adequate for the restoration of one transitted satellite, it is of interest to estimate the orbit spacing between satellites required for a prescribed outage-free interval between switches at an affected earth terminal. The interval between onsets of serial outages at a given earth terminal for satellites spaced 6.5° in orbit, allowing for earth rotation is about 28.8 minutes (Section A.2). Then, an approximate minimum satellite spacing for a 30-minute clear interval is $(30^{m}/28^{m}.8) \times 6.5° \doteq 6.8°$.

A.5.2 *Case 2–Minimum of 30 Minutes Between Adjacent Outages*

If multiple satellites are deployed without spare capacity and an earth terminal receives simultaneously from adjacent satellites, but does not switch between them, a 30-minute required clear time between outages of the adjacent satellites leads to a greater estimated satellite spacing. The elapsed time for the center of a first (easterly) outage region to depart an affected earth terminal and travel eastward

until reception is regained (a distance equal to the semi-minor dimension; Section A.4) is approximately $7^m.6/2 \doteq 3.8$ minutes. The elapsed time for the center of a second outage region to approach the same earth terminal is also 3.8 minutes, measured from onset of the second outage. The sum of elapsed times and the required 30-minute clear interval is 37.6 minutes. The minimum satellite spacing, scaled from the 28.8-minute interval between arrivals of shadows at the terminal for satellites spaced 6.5° (Section A.2) is approximately $(37^m.6/28^m.8) \times 6.5° \doteq 8.5°$.

A.5.3 *Case 3–Minimum of 30 Minutes Free of United States Outages*

An estimate of the satellite spacing required for a 30-minute clear interval between outages of earth terminals throughout the contiguous United States for the case without switching is desired. A time equivalent of the satellite spacing for a 30-minute clear interval between adjacent outages at a single earth terminal is about 37.6 minutes (Section A.5.2). A satellite spacing of 6.5° is necessary for simultaneous sun transit of a first satellite at the extreme eastern terminal and a second satellite at the extreme western terminal; a time equivalent of this spacing is approximately 28.8 minutes (Section A.2). The sum of these intervals, 66.4 minutes, accounts for transits of all terminals within the assumed 60° longitude span at 41° north latitude. The approximate minimum satellite spacing for a 30-minute clear interval throughout the United States is $(66^m.4/28^m.8) \times 6.5° \doteq 15.0°$.

A.5.4 *Case 4–Minimum of 30 Minutes Free of Outages Throughout One Time Zone*

The time equivalent of spacing for a 30-minute clear interval at a single terminal without switching is 37.6 minutes. The time equivalent of spacing for simultaneous sun transits of adjacent satellites at eastern and western terminals bounding a 15° time zone is approximately $(15°/60°) \times 28^m.8 \doteq 7.2$ minutes. The required interval is about 44.8 minutes, accounting for outage dimensions and all terminals within one time zone. The resulting minimum satellite spacing is approximately $(44^m.8/28^m.8) \times 6.5° \doteq 10.1°$.

APPENDIX B

Estimation of Minimum Required Space Diversity

B.1 *Minimum Orbit Inclinations for a Prescribed Coverage Region*

Figure 12 relates the latitude extremes of a desired coverage region to limits of the sun's apparent declination angle for which sun transits

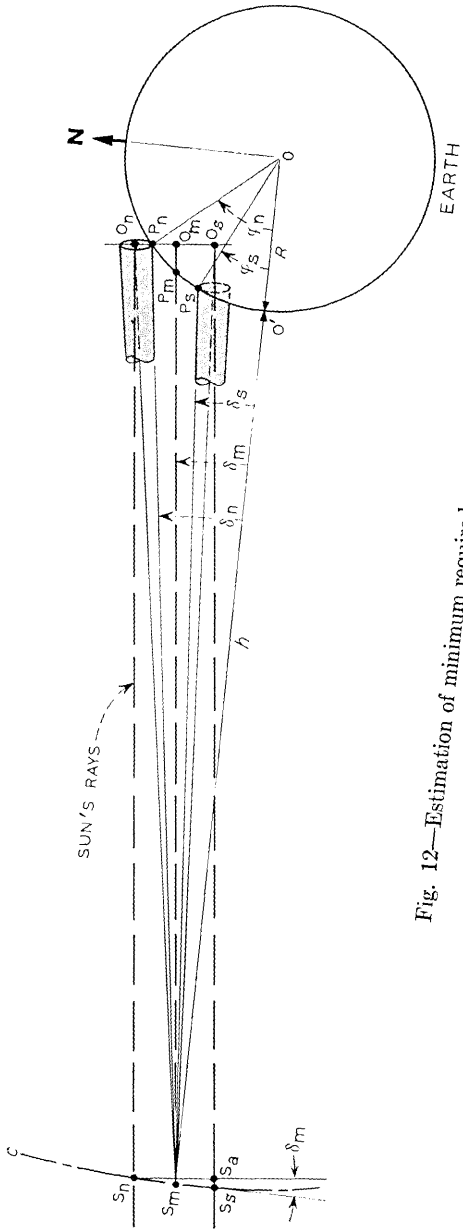


Fig. 12—Estimation of minimum required space diversity.

of diversity satellites can affect transmissions. First, two conic figures of revolution as described in Section A.3 having angular radius α define the northernmost and southernmost outage regions for geostationary satellite S_m. Minimum orbit inclinations for diversity satellites S_n and S_s are estimated by geometric construction. Parallel sun rays are assumed and atmospheric refraction is neglected for all but extreme latitudes in the presence of fairly large angles of incidence.[9]

One approximation implicit in the figure is that the satellites occupy the same mean orbit longitude. This enables a highly simplified geometrical analysis and the uncertainty introduced is shown later to be insignificant.

B.2 *Determination of Minimum Orbit Inclinations*

For geostationary satellite S_m in Fig. 12, apparent declination angle limits δ_n and δ_s are calculated for which the satellite shadows intercept north geographic latitude limits φ_n and φ_s of an assumed coverage region in the Northern Hemisphere between points P_n and P_s respectively. Slant range segment P_nS_m is determined from the solution of oblique triangle P_nOS_m:

$$P_nS_m = [R^2 + (R + h)^2 - 2R(R + h) \cos \varphi_n]^{\frac{1}{2}}, \tag{24}$$

and

$$P_sS_m = [R^2 + (R + h)^2 - 2R(R + h) \cos \varphi_s]^{\frac{1}{2}} \text{ km}. \tag{25}$$

The declination angles corresponding to northern and southern boundaries of the coverage region are

$$\delta_n = \cos^{-1}\left[\frac{-R^2 + (P_nS_m)^2 + (R + h)^2}{2(P_nS_m)(R + h)}\right] \text{degrees } S, \tag{26}$$

and

$$\delta_s = \cos^{-1}\left[\frac{-R^2 + (P_sS_m)^2 + (R + h)^2}{2(P_sS_m)(R + h)}\right] \text{degrees } S, \tag{27}$$

where the units designation S denotes angular displacement south from the celestial equator.

The angle measuring bisector $\mathbf{P_mS_m}$ is denoted by δ_m, where

$$\delta_m = \langle \delta_n + \delta_s \rangle_{av} \quad \text{degrees S}. \tag{28}$$

Synchronous satellites S_n and S_s are shown in Fig. 12 located on great circle C of a geocentered sphere of radius $(R + h)$ whose plane contains the mean geopolar axis and an assumed common satellite

meridian circle. The satellites are also assumed to be symmetrically opposite and equidistant from the equatorial plane. The required distance between parallel sun's rays through satellites S_n and S_s having mean apparent declination δ_m is determined by constructing segment $\mathbf{O_nO_s}$ perpendicular to $\mathbf{S_mP_m}$ through P_n. The base of isosceles triangle $O_nS_mO_s$ represents the required ray separation. Making the approximation

$$O_nS_m \equiv O_sS_m \doteq P_nS_m , \tag{29}$$

and denoting the angle O_n-S_m-O_s by γ,

$$\gamma = \delta_n - \delta_s + 2\alpha \quad \text{degrees.} \tag{30}$$

From the solution of an isosceles triangle,

$$O_nO_s \doteq [2(P_nS_m)^2(1 - \cos \gamma)]^{\frac{1}{2}} \text{ km.} \tag{31}$$

Constructing segment $\mathbf{S_nS_a}$ perpendicular to $\mathbf{S_mP_m}$ through S_n, its length is

$$S_nS_a \equiv O_nO_s \text{ km.} \tag{32}$$

The length of chord S_nS_s between the satellites on circle C is

$$S_nS_s = S_nS_a/\cos \delta_m \text{ km.} \tag{33}$$

The total geocentered arc $\overset{\frown}{S_nS_s}$ on circle C corresponding to chord S_nS_s is found from the solution of isosceles triangle S_nOS_s (not illustrated). Note that

$$OS_n \equiv OS_s \equiv OS_m = (R + h) \text{ km.} \tag{34}$$

Then

$$\cos (\overset{\frown}{S_nS_s}) = \frac{-(S_nS_s)^2 + 2(OS_m)^2}{2(OS_m)^2}. \tag{35}$$

Note from Fig. 12 that equal orbit inclinations i_n and i_s are determined by the minimum geocentered angular displacements of synchronous satellites S_n and S_s from the equatorial plane, necessary for avoiding simultaneous sun-transit outages between latitudes φ_n and φ_s. Hence,

$$i_n = i_s = (\overset{\frown}{S_nS_s})/2 \quad \text{degrees.} \tag{36}$$

While the simplified geometry of Fig. 12 results from an assumption that the satellites' mean longitudes are identical, recall from Section 3.1

that the maximum satellite excursions are made to occur at the instant of zenith transit viewed by an observer at each satellite's longitude. Thus, for earth terminals situated along the longitude meridian of—and receiving from—satellite S_1, the minimum required orbit inclination i_1 is identical to i_n. Very slightly increased inclinations are necessary to accommodate receiving earth terminals far from this longitude.

B.3 *Correction for Longitude Span and Latitude Location of Coverage Region*

The maximum time difference Δt_1 between sun transit of a geostationary satellite centered over the United States and observed along its longitude, and sun transit of the same satellite observed at a longitude displaced by $\pm 30°$, for a minimum latitude of 26°N is about ∓ 0.3 hour, allowing for earth rotation (Fig. 3; Section A.2). The magnitude of accumulated time shift Δt_2 (civil time versus sidereal time) relating positions of satellites at 0^h to 0^h at the vernal equinox, arising from location of the affected coverage region north of the equator, is about 1 hour (Fig. 6). An approximate worst-case adjustment of orbit inclinations providing the required displacement of diversity satellites from the equator at times when sun transits would otherwise be observed is

$$i' \doteq \frac{i_n}{\cos\left[(|\,\Delta t_1\,| + |\,\Delta t_2\,|)(360°/24^h)\right]} \text{ degrees.} \tag{37}$$

B.4 *Illustrative Calculation*

It is assumed that latitude limits φ_n and φ_s for the United States coverage region to be cleared of outages are 49°N and 26°N, respectively. A spherical earth model is assumed with radius $R = 6373$ km. The height of the geostationary orbit h is assumed to be 35,900 km. A conic sun-transit outage figure is assumed (Fig. 12), having a radius in angular measure of $\alpha = 1°$.

Numerical results are obtained using all preceding relationships:

From equations (24) and (25),	$P_nS_m = 38{,}394$ km,
	$P_sS_m = 36{,}652$ km.
From equations (26) and (27),	$\delta_n = 7.200°$,
	$\delta_s = 4.375°$.
From equation (28),	$\delta_m = 5.788°$.
From equation (30),	$\gamma = 4.825°$.
From equation (31),	$O_nO_s = 3{,}232$ km.
From equation (33),	$S_nS_s = 3{,}249$ km.

From equation (34), $OS_m = 42{,}273$ km.

From equation (35), $\cos \overset{\frown}{(S_n S_s)} = 0.997047$.

From equation (36), $i_n = i_s = 2.201°$.

From equation (37), $i' \doteq 2.337°$.

For larger earth terminals, assuming $\alpha = 0.7°$ for 25-m antennas, the corresponding worst-case minimum equal orbit inclinations providing the specified diversity is 2.045 degrees.

B.5 *Minimum Orbit Inclinations for Avoiding Serial Eclipses*

The earth's shadow is assumed to be a circular cylinder with a diameter equal to the mean diameter of the earth. This amounts to neglecting atmospheric refraction and the distinction between the umbra and penumbra shadow regions. For satellites with batteries, the net radiation energy lost per eclipse corresponds to a time integration of the actual solar-array power output. This is nearly the energy loss which would result if the solar source were completely obstructed while the satellite traversed the assumed cylindrical shadow.

An approximate relationship between declination D and the orbit arc eclipsed is illustrated by Figs. 13 and 14. The length of geostationary orbit radius OS is $(R + h)$ km, so that

$$O'S = (R + h) \mid \sin D \mid \text{km}. \tag{38}$$

Angle ψ in Fig. 14 is thus determined:

$$\psi = \sin^{-1}\left(\frac{O'S}{R}\right) \text{degrees}. \tag{39}$$

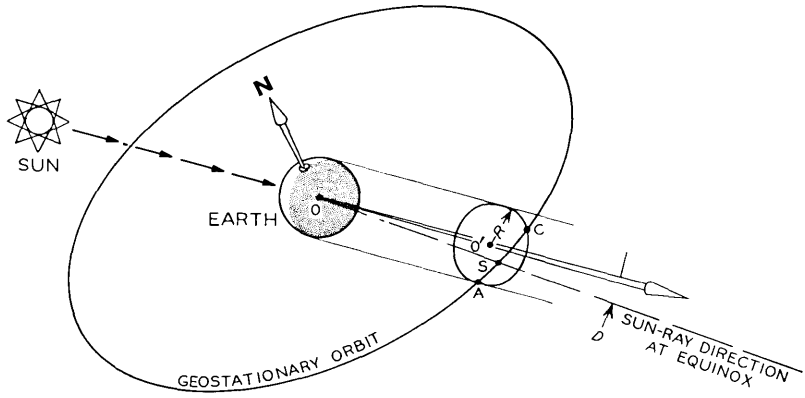


Fig. 13—Simplified geometry describing satellite eclipses.

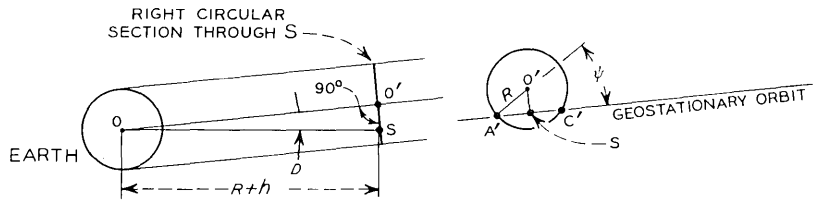


Fig. 14—Projection of points A, C upon right section of earth's shadow through S.

The length of the chord intercept common to both the orbit and the cylindrical earth shadow is determined by a normal projection of the orbit upon a right circular section of the shadow through S. From equations (38) and (39),

$$AC = A'C' \doteq 2R \cos\left\{\sin^{-1}\left[\frac{(R+h)\,|\sin D\,|}{R}\right]\right\} \text{km}. \tag{40}$$

If the fraction in brackets in equation (40) is smaller than unity for a given declination D, an eclipse of the orbiting satellite is indicated. For zero values of apparent declination, the chord AC is simply twice the mean earth radius R.

The corresponding orbit arc $\overset{\frown}{AC}$ is next calculated from the solution of oblique triangles:

$$\overset{\frown}{AC} = \cos^{-1}\left\{\frac{-(AC)^2 + 2R^2}{2R^2}\right\} \text{degrees}. \tag{41}$$

Hence, the minimum space diversity in geocentered angular measure necessary for avoiding serial satellite eclipses is identified numerically with the maximum orbit arc intercept, occurring for $D = 0°$. From equation (41), the maximum resulting geocentered angle, corresponding to one earth diameter, is approximately 17.6°. Then each minimum orbit inclination $i_1 = i_2$ necessary for avoiding serial eclipses in the manner of Section 3.2 is approximately $17.6°/2 = 8.8°$.

Finally, it is of interest to estimate the time required for the satellite to traverse shadow arc $\overset{\frown}{AC}$. The interval Δt_e is numerically equivalent to the resulting arc fraction times the orbit period, corrected for the earth's revolution about the sun:

$$\Delta t_e \doteq [1.002738](24 \times 60)^{\text{m}} \times \frac{\overset{\frown}{AC}}{360} \text{ minutes}. \tag{42}$$

REFERENCES

1. Erler, G., and Schönfeld, N., "Interference on Communication Links Via Satellites Caused by the Sun," NTZ-Commun. Proc. Nachrichtentechnische Gesellschaft VDE *6*, No. 5/6 (1967), pp. 218–223.
2. Hogg, D. C., "Ground-Station Antennas for Space Communication," Chapter 1 of *Advances in Microwaves*, L. Young, editor, New York: Academic Press, 1968.
3. Wrixson, G. T., unpublished work.
4. Giger, A. J., Pardee, S., and Wickliffe, P. R., "The Ground Transmitter and Receiver," B.S.T.J., *42*, No. 4, Part 1 (July 1963), pp. 1063–1107.
5. Lundgren, C. W., unpublished work.
6. Rowe, H. E., and Penzias, A. A., "Efficient Spacing of Synchronous Communication Satellites," B.S.T.J., *47*, No. 10 (December 1968), pp. 2379–2433.
7. Barthle, R. C., and Briskman, R. D., "Trends in Design of Communications Satellite Earth Stations," Microwave J., *10*, No. 11 (October 1967), pp. 26–108.
8. *American Ephemeris and Nautical Almanac for the year* 1970, U. S. Govt. Print. Office, Washington, 1968, pp. 18–33.
9. Lundgren, C. W., and May, A. S., "Radio-Relay Antenna Pointing for Controlled Interference with Geostationary Satellites," B.S.T.J., *48*, No. 10 (December 1969), pp. 3387–3422.

Adaptive Predictive Coding of Speech Signals

By B. S. ATAL and M. R. SCHROEDER

(Manuscript received December 13, 1968)

We describe in this paper a method for efficient encoding of speech signals, based on predictive coding. In this coding method, both the transmitter and the receiver estimate the signal's current value by linear prediction on the previously transmitted signal. The difference between this estimate and the true value of the signal is quantized, coded and transmitted to the receiver. At the receiver, the decoded difference signal is added to the predicted signal to reproduce the input speech signal. Because of the nonstationary nature of the speech signals, an adaptive linear predictor is used, which is readjusted periodically to minimize the mean-square error between the predicted and the true value of the signals.

The predictive coding system was simulated on a digital computer. The predictor parameters, comprising one delay and nine other coefficients related to the signal spectrum, were readjusted every 5 milliseconds. The speech signal was sampled at a rate of 6.67 kHz, and the difference signal was quantized by a two-level quantizer with variable step size. Subjective comparisons with speech from a logarithmic PCM encoder (log-PCM) indicate that the quality of the synthesized speech signal from the predictive coding system is approximately equal to that of log-PCM speech encoded at 6 bits/sample.

Preliminary studies suggest that the binary difference signal and the predictor parameters together can be transmitted at approximately 10 kilobits/second which is several times less than the bit rate required for log-PCM encoding with comparable speech quality.

I. INTRODUCTION

The aim of efficient coding methods[1] is to reduce the channel capacity required to transmit a signal with specified fidelity. To achieve this objective, it is often essential to reduce the redundancy of the transmitted signal. One well-known procedure for reducing signal redundancy

is predictive coding.*[2-5] In predictive coding, redundancy is reduced by subtracting from the signal that part which can be predicted from its past. For many signals, the first-order entropy of the difference signal is much smaller than the first-order entropy of the original signal; thus, the difference signal is better suited to memoryless encoding than the original signal. Predictive coding offers a practical way of coding signals efficiently without requiring large codebook memories.

Many previous speech coding methods[6] have employed schemes which attempt to separate the contributions of the vocal excitation from that of the vocal-tract transmission function. The well-known channel vocoder of Dudley[7] was the first attempt in this direction. Although vocoders can reproduce intelligible speech, there is appreciable loss in naturalness and speech quality. This degradation in speech quality arises from various operations in the vocoding process, which are either inaccurately performed or are based on certain idealized approximations of speech production and perception processes.

The present paper describes a different approach[8,9] to encoding of speech signals, based on predictive coding, which avoids the difficulties encountered in vocoders and vocoder-like devices. Although predictive coding utilizes such well-known characteristics of speech signals as pitch and formant structure, its operation does not rely solely upon a rigid parameterization of the speech signal. That part of the speech signal which cannot be represented in terms of these characteristics is not discarded but suitably encoded and transmitted to the receiver where it is used in the synthesis of a close replica of the original speech waveform.

Previous studies of predictive coding systems for speech signals[10] have been limited to linear predictors with fixed coefficients. However, due to the nonstationary nature of the speech signals, a fixed predictor cannot predict the signal values efficiently at all times. For example, the speech waveform is approximately periodic during voiced portions; thus, a good prediction of the present value of the signal can be based on the value of the signal exactly one period earlier. However, the period of the speech signal varies with time. The predictor, therefore, must change with the changing period of the input speech signal. In the predictive coding system described below, the linear predictor is adaptive; it is readjusted periodically to match the time-varying characteristics of the input speech signal. The parameters of the linear predictor are optimized to obtain an efficient prediction in the sense that

* Another name often used for this kind of encoding is Differential Pulse Code Modulation.

the mean-square error between the predicted value and the true value of the signal is minimum.

II. PREDICTIVE CODING SYSTEM

2.1 *Description*

A block diagram illustrating the principle of predictive coding is shown in Fig. 1. The input signal $s(t)$ is sampled at the Nyquist rate to produce the samples s_n of the signal. The predictor forms an estimate $\hat{s}_n$ of the signal's present value based on the past samples r_{n-1} , r_{n-2} , $\cdots$ of the reconstructed signal at the transmitter. The predicted value $\hat{s}_n$ of the signal is next subtracted from the signal value s_n to form the difference δ_n , which is quantized, encoded, and transmitted to the receiver. At the same time, the transmitted signal is decoded at the transmitter and the signal reconstructed in exactly the same manner as is done at the receiver. The reconstructed signal is then used to predict the next sample of the input signal.

At the receiver, the transmitted signal is decoded and added to the predicted value of the signal to form the samples r'_n of the reconstructed signal. The predictor used at the receiver is identical to one employed at the transmitter. The samples r'_n of the reconstructed signal are finally low-pass filtered to produce the output signal $r'(t)$.

2.2 *Signal-to-Quantizing Noise Ratio*

Consider the predictive coding system shown in Fig. 1. Let P_s be the mean-square value of the input signal samples s_n , P_δ be the mean-square value of the difference signal samples δ_n , P_q be the mean-square value of the quantizing noise in the decoded difference signal δ'_n, and

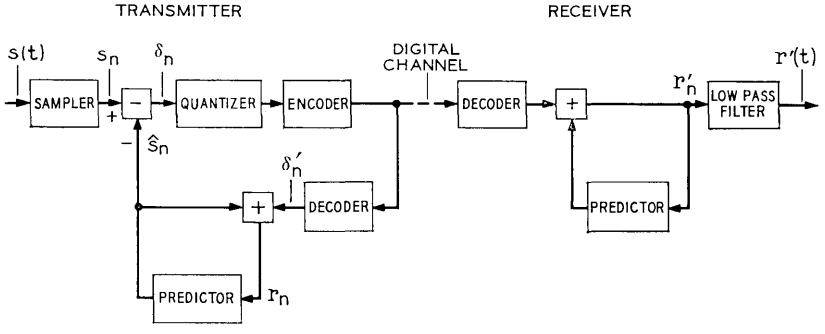


Fig. 1—Block diagram of a predictive coding system.

P_e be the mean-square value of the quantizing noise in the reconstructed signal r_n' . We will now show that, in the absence of digital channel transmission errors, the signal-to-quantizing noise ratio P_s/P_e of the reconstructed signal is given by

$$\frac{P_s}{P_e} = \frac{P_s}{P_\delta} \cdot \frac{P_\delta}{P_q}. \tag{1}$$

In other words, the signal-to-quantizing noise ratio of the reconstructed signal *exceeds* the signal-to-quantizing noise ratio of the decoded difference signal by a factor equal to the ratio of the mean-square value of the input signal to the mean-square value of the difference signal. The predictive coding system is thus superior to a straight PCM system whenever P_s/P_δ is much greater than 1. For a signal such as speech, this is indeed true. The results obtained by computer simulation of the predictive coding system (see Section 3.3) show that P_s/P_δ is about 100 for speech signals. By using predictive coding, one could thus expect improvement of about 20 dB in signal-to-quantizing noise ratio over a PCM system using identical quantizing levels.

To prove equation (1), we will first show that the error between any sample of the reconstructed signal and the corresponding sample of the input signal is identical to the error introduced by the quantizer, the encoder and the decoder.

The error e_n between the sample r_n' of the reconstructed signal and the sample s_n of the input signal is given by

$$e_n = r_n' - s_n . \tag{2}$$

In the absence of digital channel transmission errors, we can replace r_n' in equation (2) by r_n and rewrite equation (2) as

$$e_n = (r_n - \hat{s}_n) - (s_n - \hat{s}_n). \tag{3}$$

It is readily seen in Fig. 1 that

$$r_n = \delta_n' + \hat{s}_n \tag{4}$$

and

$$\delta_n = s_n - \hat{s}_n . \tag{5}$$

On combining equations (3), (4) and (5), one obtains

$$e_n = \delta_n' - \delta_n . \tag{6}$$

The right side of equation (6) represents the error introduced by the quantizer, the encoder, and the decoder. Thus, the error in the nth

sample of the reconstructed signal is identical to the error in the nth sample of the decoded difference signal.

The signal-to-quantizing noise ratio of the reconstructed signal is by definition P_s/P_e and can be written as

$$\frac{P_s}{P_e} = \frac{P_s}{P_\delta}\cdot\frac{P_\delta}{P_e}. \tag{7}$$

Since the mean-square value P_e of the quantizing noise in the reconstructed signal is identical to the mean-square value P_q of the quantizing noise in the decoded difference signal, P_e on the right side of equation (7) can be replaced by P_q, and one obtains

$$\frac{P_s}{P_e} = \frac{P_s}{P_\delta}\cdot\frac{P_\delta}{P_q}. \tag{1}$$

III. APPLICATION OF PREDICTIVE CODING TO SPEECH SIGNALS

3.1 *Linear Prediction of Speech Signals*

Two of the main causes of redundancy in speech are:

(*i*) Quasi-periodicity during voiced segments[6] and,
(*ii*) Lack of flatness of the short-time spectral envelope.[6]

The exact form of the predictor for the speech wave depends on the model used to represent the human speech production process. A reasonable model for the production of voiced speech sounds is obtained by representing them as the output of a discrete linear time-varying filter which is excited by a quasi-periodic pulse train (see Fig. 2). The output of the linear filter at any sampling instant is a linear combination of the past p output samples and the input. The number of past samples p is given by twice the number of resonances (formants) of the vocal tract which are contained in the frequency range of interest. For example, in the case of speech signals band-limited to 3 kHz, it can be

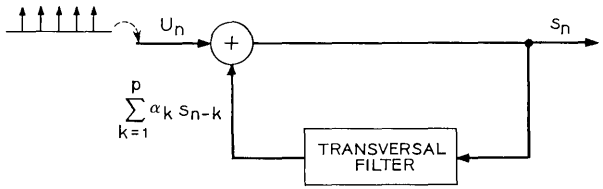


Fig. 2—Model for the production of voiced speech sounds.

assumed that there are typically three to four formants.[6] A suitable value of p is thus 8.

Let s_n and U_n be the amplitudes of the output and input signals (see Fig. 2) at the nth sampling instant. The nth output sample s_n is then given by

$$s_n = \sum_{k=1}^{p} \alpha_k s_{n-k} + U_n , \tag{8}$$

where

$$U_n = \beta U_{n-M} , \tag{9}$$

M is the period of the excitation signal and β takes account of the variation of the amplitude of the input pulse train from one period to the next. For natural speaking conditions, the period of the excitation signal is usually below 15 milliseconds, and, as a first approximation, the effect of time variation of the coefficients α_k from one pitch period to the next can be neglected. Under this assumption, we find

$$s_n - \beta s_{n-M} = \sum_{k=1}^{p} \alpha_k (s_{n-k} - \beta s_{n-k-M}) + U_n - \beta U_{n-M} . \tag{10}$$

Since $U_n = \beta U_{n-M}$, equation (10) reduces to

$$s_n = \beta s_{n-M} + \sum_{k=1}^{p} \alpha_k (s_{n-k} - \beta s_{n-k-M}), \tag{11}$$

which determines completely the structure of the linear predictor.

A block diagram of the predictor as described by equation (11) is shown in Fig. 3. The delay M as well as the parameters $\alpha_1 , \alpha_2 , \cdots , \alpha_p$

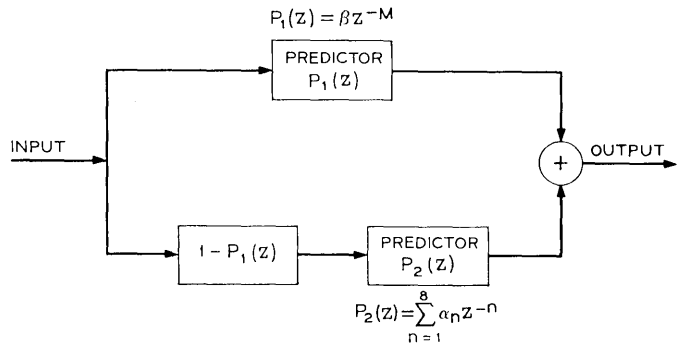


Fig. 3—Block diagram of the predictor for speech signals.

and β are variable and are readjusted periodically to match the characteristics of the input speech signal. Ideally the readjustment of the predictor parameters need be done only when there are significant changes in the characteristics of the speech signal. This implies that the predictor should be readjusted at short intervals during transitions and at long intervals during steady state portions of the speech signal and, consequently, a long buffer storage is needed to ensure transmission of parameters at a uniform rate on the channel. In order to avoid the use of a long buffer storage, the predictor parameters were readjusted at a fixed time interval in our study. This time interval was chosen to be 5 milliseconds to ensure that the prediction be efficient even during rapidly changing segments of the speech wave.

For unvoiced sounds, the quasi-periodic excitation U_n in equation (8) is replaced by a noise-like excitation. Generally speaking, the transfer function of the filter for unvoiced sounds must include poles as well as zeros. However, we find that for all practical purposes it is sufficient to include only the effect of poles. Equation (11), thus, represents the linear predictor for unvoiced sounds too if β is assumed zero.

3.2 *Determination of Predictor Parameters*

The predictor parameters are determined by minimizing the mean-square error between the actual speech sample and its predicted value. The predicted value $\hat{s}_n$ of the nth speech sample is given by

$$\hat{s}_n = \beta s_{n-M} + \sum_{k=1}^{p} \alpha_k (s_{n-k} - \beta s_{n-k-M}). \tag{12}$$

The prediction error sample E_n is then given by

$$\begin{aligned} E_n &= s_n - \hat{s}_n \\ &= (s_n - \beta s_{n-M}) - \sum_{k=1}^{p} \alpha_k (s_{n-k} - \beta s_{n-k-M}). \end{aligned} \tag{13}$$

The mean-square prediction error $\langle E_n^2 \rangle_{\text{av}}$ is given by

$$\langle E_n^2 \rangle_{\text{av}} = \frac{1}{N} \sum_n E_n^2 , \tag{14}$$

where the sum extends over all the samples in the time interval during which the predictor is to be optimum.

The problem of minimizing the mean-square error $\langle E_n^2 \rangle_{\text{av}}$ by suitable selection of the predictor parameters does not admit a straightforward solution due to the presence of the delay parameter M in equation (13).

A sub-optimum solution was obtained by minimizing the total error in two steps. First the parameters β and M are determined such that the error E_1, defined by

$$E_1 = \frac{1}{N} \sum_n (s_n - \beta s_{n-M})^2 = \langle (s_n - \beta s_{n-M})^2 \rangle_{\text{av}} , \tag{15}$$

is minimum. Using these values of β and M, the mean-square error $\langle E_n^2 \rangle_{\text{av}}$ is minimized by a suitable choice of parameters $\alpha_1 , \cdots , \alpha_p$.

To find the values of the parameters β and M which minimize the error E_1 as defined in equation (15), we first set the partial derivative of E_1 with respect to β equal to zero:

$$\begin{aligned} \frac{\partial E_1}{\partial \beta} &= -2\langle (s_n - \beta s_{n-M}) s_{n-M} \rangle_{\text{av}} \\ &= 0, \end{aligned} \tag{16}$$

where the $\langle \ \rangle_{\text{av}}$ indicates the averaging over all the samples in the given 5-millisecond time segment during which the predictor is to be optimum.

On solving for β from equation (16), we obtain

$$\beta = \langle s_n s_{n-M} \rangle_{\text{av}} / \langle s_{n-M}^2 \rangle_{\text{av}} . \tag{17}$$

We next substitute the value of β from equation (17) into equation (15). After rearrangement of terms, we obtain

$$E_1 = \langle s_n^2 \rangle - \langle s_n s_{n-M} \rangle_{\text{av}}^2 / \langle s_{n-M}^2 \rangle_{\text{av}} . \tag{18}$$

Since the first term on the right side of equation (18) does not depend on M, it can be omitted in finding the minimum value of the error. Further, E_1 is minimum if the second term on the right side of equation (18) is maximum. The optimum value of M is thus determined from the location of the maximum of the normalized correlation coefficient ρ given by

$$\rho = \{ \langle s_n s_{n-M} \rangle_{\text{av}} \} / \{ \langle s_n^2 \rangle_{\text{av}} \langle s_{n-M}^2 \rangle_{\text{av}} \}^{\frac{1}{2}}, \qquad M > 0. \tag{19}$$

Next, the predictor parameters $\alpha_1 , \cdots , \alpha_p$ are obtained such that the mean-square error $\langle E_n^2 \rangle_{\text{av}}$ as given in equation (14) with β and M fixed at their optimum values is minimum. Let

$$v_n = s_n - \beta s_{n-M} . \tag{20}$$

The error $\langle E_n^2 \rangle_{\text{av}}$ is then given by

$$\langle E_n^2 \rangle_{\text{av}} = \left\langle \left\{ v_n - \sum_{k=1}^{p} \alpha_k v_{n-k} \right\}^2 \right\rangle_{\text{av}} . \tag{21}$$

The optimum values of the coefficients $\alpha_1 , \cdots , \alpha_p$ which minimize $\langle E_n^2 \rangle_{av}$ are obtained by setting the partial derivatives of $\langle E_n^2 \rangle_{av}$ with respect to $\alpha_1 , \cdots , \alpha_p$ equal to zero. Or,

$$\frac{\partial \langle E_n^2 \rangle_{av}}{\partial \alpha_i} = \left\langle \left(v_n - \sum_{k=1}^{p} \alpha_k v_{n-k} \right) v_{n-i} \right\rangle_{av} ,$$

$$= 0 \quad \text{for} \quad j = 1, 2, \cdots , p. \tag{22}$$

Equation (22) can be rewritten in matrix notation as

$$\Phi \mathbf{a} = \psi, \tag{23}$$

where Φ is a p by p matrix with its (ij)th term φ_{ij} given by

$$\varphi_{ij} = \langle v_{n-i} v_{n-j} \rangle_{av} , \tag{24}$$

$\mathbf{a}$ is a p-dimensional vector whose jth component is a_j and ψ is a p-dimensional vector whose jth component ψ_j is given by

$$\psi_i = \langle v_n v_{n-i} \rangle_{av} . \tag{25}$$

The optimum predictor coefficients $\alpha_1 , \alpha_2 , \cdots , \alpha_p$ are obtained by solving equation (23) for $\mathbf{a}$. For the case when Φ is a nonsingular matrix, the solution of equation (23) presents no difficulty. The vector $\mathbf{a}$ can be obtained by multiplying ψ with the inverse of the matrix Φ. A more efficient computational procedure[11] for solving equation (23), which does not involve matrix inversion, takes advantage of the fact that Φ is a symmetric matrix, and thus can be expressed as the product of a triangular matrix and its transpose. Equation (23) can then be written as three separate matrix equations. These equations involve triangular matrices only and their solutions can be expressed by a set of recursive equations.[11]

A singular Φ matrix implies that one or more of its eigenvalues is zero. The matrix Φ can be modified to become nonsingular by adding a small positive constant to its diagonal elements. Equation (23) is solved again with the matrix Φ replaced by the matrix Φ'. The modified matrix Φ' is symmetric and has the same eigenvectors as the matrix Φ, but its eigenvalues are all positive; thus it is a positive definite symmetric matrix and has a unique inverse Φ'^{-1}.

3.3 *Computer Simulation of the System*

The predictive coding system using adaptive predictors was simulated on a digital computer to determine its effectiveness for coding speech signals. The transmitter and the receiver are illustrated separately in Figs. 4 and 5, respectively. The sampling rate used in this

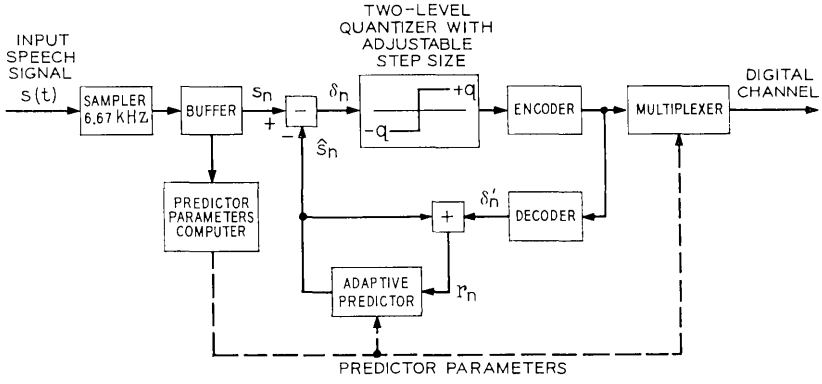


Fig. 4—Transmitter of the predictive coding system.

simulation was 6.67 kHz. Prior to sampling, the input speech signal was filtered with a low-pass filter with 3-dB attenuation at 3.1 kHz and an attenuation of 40 dB or more for frequencies above 3.33 kHz. At the transmitter, the difference δ_n formed by subtracting the predicted value $\hat{s}_n$ from the speech sample s_n was quantized by a *two-level* (1 bit) quantizer with *variable* step size q. The parameter q was readjusted every 5 milliseconds to yield minimum quantization noise power. The parameters of the adaptive predictor were also computed once every 5 milliseconds and sent to the receiver together with the binary difference signal and the step size q of the quantizer. The optimum value of the delay parameter M was obtained by locating the maximum of the correlation coefficient ρ as defined in equation (19) for values of M between 20 and 150. The parameter p was set at 8.

The speech signal was reconstructed at the receiver by a feedback loop containing an adaptive predictor identical to the one used at the transmitter. Here, the predictor too, was reset every 5 milliseconds

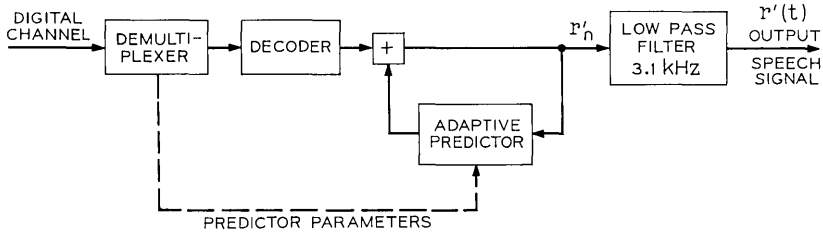


Fig. 5—Receiver of the predictive coding system.

according to the predictor-parameter information received from the transmitter. The reconstructed speech samples were finally smoothed by a 3.1-kHz low-pass filter to form the output speech signal $r'(t)$.

IV. RESULTS OF SUBJECTIVE TESTS

Two different subjective tests were conducted to judge the quality of the reconstructed speech signal produced at the receiver of the predictive coding system. In the first test, trained listeners compared the reconstructed speech signal with speech from a logarithmic PCM (log-PCM) encoder[12] that used the same input signals and a sampling frequency of 6.67 kHz. The compression characteristic employed in a log-PCM encoder is defined by the equation

$$y = \frac{V \log \left[1 + \frac{\mu |x|}{V}\right]}{\log (1 + \mu)} \operatorname{sgn} x, \tag{26}$$

where y represents the output voltage corresponding to an input signal voltage x, μ is a dimensionless parameter which determines the degree of compression and V is the compressor overload voltage.[12] The compressed signal y was quantized at bit rates varying from 5 bits/sample to 7 bits/sample with $\mu = 100$ and $V = 8 \times$ the rms speech signal voltage.† Speech samples from both male and female speakers were used in these tests. The results of the subjective tests indicated that the quality of the reconstructed speech signal was better than that of log-PCM speech with 5 bits/sample but slightly inferior to one with 6 bits/sample. The corresponding measured signal-to-noise ratios for log-PCM speech were 21 dB and 27 dB, respectively.

In the second test, the reconstructed speech signal was compared with the input speech signal contaminated by additive white noise obtained by randomly inverting the polarity of successive Nyquist samples of the input speech signal.[13] This noise is subjectively similar to the distortion introduced by predictive coding and is therefore particularly appropriate for reproducible comparisons. This noise has an added advantage in that its absolute amplitude at any instant of time is proportional to the absolute amplitude of the input speech signal. This proportionality permits the calculation of a precise signal-to-noise ratio (S/N). Based on the results of these tests, the equivalent S/N of the reconstructed speech in the predictive coding system de-

† The integration time for computing the rms value of the speech signal was several seconds and included speech samples from a number of speakers.

scribed above was found to be about 25 dB which is in good agreement with results obtained by the subjective comparison with log-PCM.

V. ADDITIONAL MODIFICATIONS OF THE PREDICTIVE CODING SYSTEM

5.1 *Spectrum of Quantizing Noise and Its Influence on the Subjective Quality of the Reconstructed Speech*

For frequencies above 500 Hz, the frequency spectrum of voiced speech sounds generally falls off with frequency with an average slope between −6 and −12 dB per octave. The spectrum of quantizing noise in the predictive coding system, on the other hand, is approximately uniform. The signal-to-quantizing noise ratio (S/N) of the reconstructed speech, thus, also falls off with frequency. This is illustrated in Fig. 6 where the spectrum of a short segment of the speech signal is compared with the spectrum of the corresponding quantizing noise. As can be seen, the S/N is very poor at high frequencies. Informal listening tests of the reconstructed speech appeared to confirm the above observation. The quality of the reconstructed speech can thus be improved by a suitable shaping of the spectrum of the quantizing noise so that the S/N is more or less uniform over the entire frequency range of the input speech signal. The desired spectral shaping can be achieved by pre-emphasizing the input speech signal at high frequencies by means of a fixed filter whose amplitude versus frequency characteristic rises with frequency above 500 Hz with a

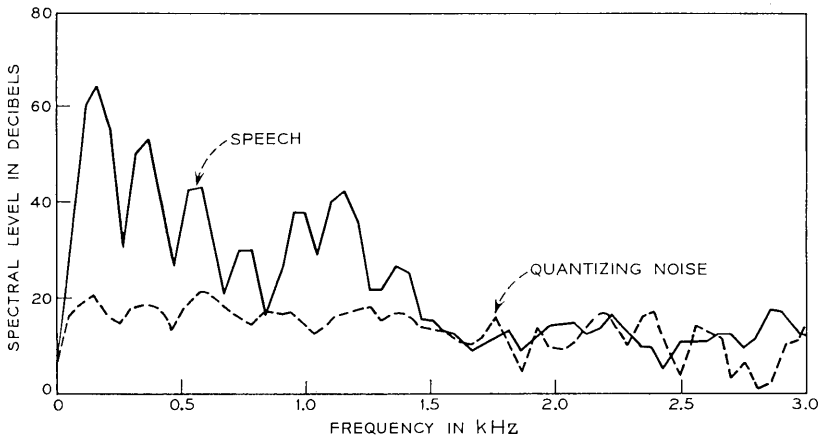


Fig. 6—Spectra of speech and quantizing noise.

slope of 12 dB per octave. The spectral distortion can finally be eliminated by a filter at the output of the receiver whose frequency versus amplitude characteristic is exactly opposite to that of the pre-emphasis filter. The results of computer simulation indicate that the quality of the reconstructed speech in the predictive coding system employing pre-emphasis is considerably better than that of the system without pre-emphasis.

5.2 *Improved Prediction of Voiced Speech*

The redundancy due to the quasi-periodic nature of voiced speech is removed in the predictive coding system described earlier by a predictor $P_1(z)$ consisting of a delay of M samples and an amplifier with gain β as shown in Fig. 3. It is possible to improve the prediction of voiced speech by employing a predictor $P_1(z)$ consisting of two delays and two amplifiers such that

$$P_1(z) = \beta_1 z^{-M} + \beta_2 z^{-2M}. \tag{27}$$

The parameters β_1 and β_2 are calculated by minimizing the mean-square error E_1 defined by

$$E_1 = \langle (s_n - \beta_1 s_{n-M} - \beta_2 s_{n-2M})^2 \rangle_{av} . \tag{28}$$

The modified predictive coding system including pre-emphasis of the input speech signal together with the second-order predictor $P_1(z)$ as given in equation (27) was simulated on the computer. The results of subjective tests similar to those described in Section IV indicated that the quality of the reconstructed speech was somewhat superior to that of log-PCM speech at 6 bits per sample. The equivalent S/N was found to be 30 dB.

VI. QUANTIZATION OF PREDICTOR PARAMETERS

No attempt was made in the study reported here to quantize the predictor parameters. Preliminary calculations were made to estimate the number of bits required to transmit the information to the receiver. Since the predictor parameters (one delay and nine other coefficients) carry the information about the signal spectrum, it should be possible to encode them at a bit rate comparable to one used in conventional formant vocoders. This suggests a bit rate of approximately 10 kilobits per second for transmitting the binary difference signal (6.67 kb/s) and the predictor parameters (3 kb/s). Recent studies by Kelly[14] indicate that it is indeed possible to encode the transmitted information within 9600 b/s without significant loss in speech quality.

VII. CONCLUSIONS

The study reported here shows that predictive coding is a promising approach to digital encoding of speech signals for high-quality transmission at substantial reductions in bit rate. Unlike past speech coding methods based on the vocoder principle, the predictive coding scheme described here attempts to reproduce accurately the speech *waveform*, rather than its spectrum. Listening tests show that there is only slight, often imperceptible, degradation in the quality of the reproduced speech. Although no detailed investigation of the optimum encoding methods of the predictor parameters was made, preliminary studies suggest that the binary difference signal and the predictor parameters together can be transmitted at bit rates of less than 10 kb/s or several times less than the bit rate required for PCM encoding with comparable speech quality.

REFERENCES

1. Davisson, L. D., "The Theoretical Analysis of Data Compression Systems," Proc. IEEE, *56*, No. 2 (February 1968), pp. 176–186.
2. Cutler, C. C., "Differential Quantization of Communication Signals," U. S. Patent 2-605-361, applied for June 29, 1950; issued July 29, 1952.
3. Oliver, B. N., "Efficient Coding," B.S.T.J., *31*, No. 4 (July 1952), pp. 724–750.
4. Elias, P., "Predictive Coding," IRE Trans. Inform. Theor., *IT-1*, No. 1 (March 1955), pp. 16–33.
5. O'Neal, J. B., Jr., "Predictive Quantizing Systems (Differential Pulse Code Modulation) for the Transmission of Television Signals," B.S.T.J., *45*, No. 5 (May–June 1966), pp. 689–721.
6. Schroeder, M. R., "Vocoders: Analysis and Synthesis of Speech," Proc. IEEE, *54*, No. 5 (May 1966), pp. 720–734.
7. Dudley, H., "Remaking Speech," J. Acoust. Soc. Amer., *11*, No. 2 (October 1939), pp. 169–177.
8. Atal, B. S., and Schroeder, M. R., "Predictive Coding of Speech Signals," Proc. 1967 Conf. on Commun. and Processing, November 1967, pp. 360–361.
9. Atal, B. S., and Schroeder, M. R., "Predictive Coding of Speech Signals," 1968 Wescon Technical Papers, August 1968, paper 8/2.
10. McDonald, R. A., "Signal-to-Noise Performance and Idle Channel Performance of Differential Pulse Code Modulation Systems with Particular Applications to Voice Signals," B.S.T.J., *45*, No. 7 (September 1966), pp. 1123–1151.
11. Faddeev, D. K., and Faddeeva, V. N., *Computational Methods of Linear Algebra*, English Translation by R. C. Williams, San Francisco: W. H. Freeman and Company, 1963, pp. 144–147.
12. Smith, B., "Instantaneous Companding of Quantized Signals," B.S.T.J., *36*, No. 3 (May 1957), pp. 653–709.
13. Schroeder, M. R., "Reference Signal for Signal Quality Studies," J. Acoust. Soc. Amer., *44*, No. 6 (December 1968), pp. 1735–1736.
14. Kelly, J. M., unpublished work.

Contributors to This Issue

BISHNU S. ATAL, B.Sc. (Hons.), 1952, University of Lucknow (India); D.I.I.Sc., 1955, Indian Institute of Science, Bangalore, India; Ph.D., 1968, Polytechnic Institute of Brooklyn; Bell Telephone Laboratories, 1961—. Before joining Bell Laboratories, Mr. Atal was a lecturer at the Indian Institute of Science. At Bell Laboratories, he has worked on problems in architectural acoustics, speech, and hearing. Member, Acoustical Society of America.

WILLIAM T. BARNETT, B.S.E.E., 1958, Illinois Institute of Technology; M.E.E., 1960, New York University; Western Electric, 1953–1958; Bell Telephone Laboratories, 1958—. Mr. Barnett has worked on problems related to microwave radio relay systems. Since 1966 he has supervised a group concerned with propagation problems. Member, IEEE.

WILLIAM F. BODTMANN, Monmouth College, 1957-61; Bell Telephone Laboratories, 1941—. Mr. Bodtmann has been engaged in research on long- and short-haul microwave radio systems, frequency feedback receivers, and FM multiplex systems. He is working with communication systems operating at millimeter wavelengths.

C. A. BRACKETT, B.S., 1962, M.S., 1963, and Ph.D., 1968, University of Michigan; Bell Telephone Laboratories, 1968—. Mr. Brackett has been doing research in microwave semiconductor electronics. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi.

P. M. EBERT, B.S., 1958, University of Wisconsin; S.M., 1962, and Sc.D., 1965, Massachusetts Institute of Technology; Bell Telephone Laboratories, 1965—. Mr. Ebert has worked on problems in communications and information theory. Member, IEEE.

R. L. GRAHAM, B.S., 1958, University of Alaska; M.A., Ph.D., 1962, University of California (Berkeley); Bell Telephone Laboratories, 1962—. Mr. Graham's recent interests have been in the fields of combinatorial analysis, graph theory and finite structures and their applications to coding theory and switching theory. He is Head of the Discrete Systems and Control Department. Member, American Mathematical

Society, Mathematical Association of America, SIAM, Sigma Xi, American Association for the Advancement of Science.

R. W. HAMMING, B.S., 1937, University of Chicago; M.S., 1939, University of Nebraska; Ph.D., 1942, University of Illinois; 1942–44, Mathematics Instructor, University of Illinois; 1944–45, Assistant Professor, University of Louisville; 1960–61, Visiting Professor, Stanford University; Bell Telephone Laboratories, 1946—. Since joining Bell Laboratories, Mr. Hamming has specialized in the use of numerical methods for solving problems on large-scale computing machines. He has also been engaged in the design of large-scale computers, and holds a patent for error-detecting and error-correcting codes, and a patent for a remote-controlled system for reducing distortion. He is presently Head of the Computing Science Research Department. Fellow, IEEE; Member, Association for Computing Machinery, American Association for the Advancement of Science, Society for Industrial and Applied Mathematics, Mathematical Association of America.

D. L. JAGERMAN, B.E.E., 1949, Cooper Union; M.S., 1954, and Ph.D., 1962, (mathematics), New York University; Bell Telephone Laboratories, 1964—. Mr. Jagerman has been engaged in mathematical research on numerical quadrature theory, interpolation, mathematical properties of pseudo-random number generators, dynamic programming, and approximation theory. His recent work concerns the theory of widths and entropy with application to the storage and transmission of information. Member, American Mathematical Society, Pi Mu Epsilon.

CARL W. LUNDGREN, E.E., 1957, M.S., 1959, and Ph.D., 1961, University of Cincinnati; U. S. Army Electronics Research and Development Laboratory, 1962–1963; Bell Telephone Laboratories, 1961—. Mr. Lundgren's early work was in electrodynamics and gyro mechanics, resulting in magnetic navigation and spacecraft stabilization techniques. His subsequent interests concerned launch timing for optimum spin-axis orientation and the medium-altitude satellite eclipse environment in support of the TELSTAR® communication satellite experiment. He is studying microwave transmission, interference, and circuit outage problems associated with communication satellite systems. Member, Phi Eta Sigma, Eta Kappa Nu, Tau Beta Pi, Omicron Delta Kappa, IEEE, New York Academy of Sciences.

E. A. J. MARCATILI, Aeronautical Engineer, 1947, and E.E., 1948, University of Cordoba (Argentina); research staff, University of Cordoba, 1947–54; Bell Telephone Laboratories, 1954—. He has been engaged in theory and design of filters in multimode waveguides and in waveguide systems research. More recently he has concentrated on optical transmission media. Fellow, IEEE.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, Dipl. Phys., 1954, Berlin Free University; D.E.E., 1962, Technische Hochschule, Karlsruhe, Germany; Siemens and Halske (Germany), 1954–57; Bell Telephone Laboratories, 1957—. At Siemens and Halske, Mr. Marcuse was engaged in transmission research, studying coaxial cable and circular waveguide transmission. At Bell Telephone Laboratories, he has been engaged in studies of circular electric waveguides and work on gaseous masers. He spent one year (1966–1967) on leave of absence from Bell Telephone Laboratories at the University of Utah where he wrote a book on quantum electronics. He is presently working on the transmission aspect of a light communications system. Member, IEEE, Optical Society of America.

CLYDE L. RUTHROFF, B.S.E.E., 1950, and M.A., 1952 University of Nebraska; Bell Telephone Laboratories, 1952—. Mr Ruthroff has published contributions on the subjects of FM distortion theory, broadband transformers, FM limiters, threshold extension by feedback, and microwave radio systems for satellite and terrestrial use. He is interested in the extension of radio communication into the millimeter and optical wavelengths. Member, A.A.A.S., I.E.E.E., 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.

MANFRED ROBERT SCHROEDER, Diplom Physiker, 1951, and Dr. rer. nat., physics, 1954, University of Göttingen, Germany; Bell Telephone Laboratories, 1954–1969. Mr. Schroeder's work has encompassed fundamental studies of architectural acoustics, electroacoustics, under-

water sound, speech and hearing. Before being appointed Professor of physics and Director of the Drittes Physikalisches Institut, University of Göttingen, he was Director of the Acoustics, Speech and Mechanics Research Laboratory at Bell Labs. Fellow and member of the Executive Council, Acoustical Society of America; senior member, Institute of Electrical and Electronics Engineers; member, German and European Physical Societies.

DAVID E. SETZER, B.S.M.E., 1958, Lehigh University; M.S.E.M., 1960, New York University; Ph.D., Applied Mechanics, 1964, Lehigh University; Instructor of Mechanics, Lehigh University, 1961–1964; Bell Telephone Laboratories, 1958–1961, 1964—. Mr Setzer's early work at Bell Laboratories dealt with the design and construction of the TELSTAR® Earth Stations. Later he engaged in experimental and theoretical studies of transmission through natural aerosols. He is presently working on improving electrical and mechanical properties of communication cables. Member, American Association for the Advancement of Science, Pi Tau Sigma, Tau Beta Pi, Pi Mu Epsilon, Sigma Xi, Newtonian Society.

GERALD A. SHANHOLT, B. S., 1962, and Ph.D., 1968, Polytechnic Institute of Brooklyn; General Electric Co., 1962-1964; U.S. Army active duty, 1967–1969; Bell Telephone Laboratories, 1969—. Mr. Shanholt is engaged in research in stochastic estimation and control. Member, American Mathematical Society.

ALAN N. WILLSON, JR., B.E.E., 1961, Georgia Institute of Technology; M.S.E.E., 1965, and Ph.D., 1967, Syracuse University; International Business Machines Corporation, 1961–1964; Bell Telephone Laboratories, 1967—. Mr. Willson is interested in network and systems theory. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

YU S. YEH, B.S.E.E., 1961, National Taiwan University; M.S.E.E., 1964, and Ph.D., 1966, University of California, Berkeley, Harvard University, 1967; Bell Telephone Laboratories, 1967—. Mr. Yeh is a member of the Radio Transmission Research Department and is doing research work concerning Mobile Radio communication.

B.S.T.J. BRIEF

All Terminal Bubbles Programs Yield the Elementary Symmetric Polynomials

By R. P. KURSHAN

(Manuscript received May 18, 1970)

R. L. Graham has discussed various combinatorial aspects of the behavior of magnetic domains or "bubbles".[1] Representing the initial state of a configuration of n magnetic domains by the n-tuple of indeterminates $B = (X_1, \cdots, X_n)$, he showed that subsequent configurations of magnetic domains obtainable (within the constraints of the problem) correspond exactly to subsequent n-tuples of Boolean expressions in the X_i's* obtainable from B through an application to B of a product of transformations ("commands" in Ref. 1) of the form $T_{ij}(1 \leqq i < j \leqq n)$ where if $P = (P_1, \cdots, P_n)$ is an n-tuple of Boolean expressions in the X_i's, then $T_{ij}(P) = (Q_1, \cdots, Q_n)$,

$$Q_k = \left\{\begin{array}{lll} P_i \bigcup P_j & \text{if} & k = i \\ P_i \bigcap P_j & \text{if} & k = j \\ P_k & \text{otherwise} & \end{array}\right\}, \qquad k = 1, \cdots, n.$$

Furthermore, he showed that

> if $\mathfrak{I}$ is an $\binom{n}{2}$-fold product of such transformations and if T is any other, then $(T \circ \mathfrak{I})(B) = \mathfrak{I}(B)$. (†)

This provides a limitation on the number of distinct n-tuples of the form $\mathfrak{U}(B) = (P_1, \cdots, P_n)$ where $\mathfrak{U}$ is a product of transformations, and hence provides a limitation on the number of distinct P_i's thus obtainable from various $\mathfrak{U}$'s. Graham showed that for $n = 11$, this limitation implies that not all Boolean expressions in the X_i's are realizable as a P_i.

This led to an (as yet unsuccessful) attempt to characterize those expressions which are realizable. The purpose of this note is to observe a fragmentary result in this direction: that if $\mathfrak{I}$ is as above, then $\mathfrak{I}(B) =$

* A Boolean expression in the X_i's is either a term of the form X_i $(1 \leqq i \leqq n)$, a term of the form $P \cup Q$ or a term of the form $P \cap Q$, where both P and Q are Boolean expressions in the X_i's; expressions may be reduced as if the X_i's were sets.

$(S_1, \cdots, S_n)$ where S_i is the elementary symmetric polynomial in $X_1, \cdots, X_n$ of degree i (here interpreting $\cup$ as $+$ and $\cap$ as $\cdot$). The situation will be rephrased in terms of a semiring.

For a fixed n let R be the (Boolean) commutative semiring generated by $X_1, \cdots, X_n$ subject to the relations:

$$\text{for} \quad i = 1, \cdots, n, \quad (1)\ X_i^2 = X_i,$$

$$(2)\ fX_i + f = f \quad \text{for all} \quad f \in R.$$

It follows that $2X_i = X_i (i = 1, \cdots, n)$ and hence, each $f \in R$ is a Boolean polynomial in the indeterminates $X_1, \cdots, X_n$, (that is, the X_i's behave like sets with respect to $+$ and $\cdot$ interpreted as $\cup$ and $\cap$ respectively).

Throughout, if $x \in R^n$ (the set of n-tuples of elements of R), then for $1 \leqq k \leqq n$, x_k will denote the kth coordinate of x, that is, $x = (x_1, \cdots, x_k, \cdots, x_n)$. Let T (or T_n) be the set of transpositions of $\{1, \cdots, n\}$ and for $t \in T$—say $t = (i, j)$, $i < j$—define $t : R^n \to R^n$ by

$$(tf)_k = \begin{cases} f_i + f_j & \text{if} \quad k = i \\ f_i \cdot f_j & \text{if} \quad k = j \\ f_k & \text{otherwise} \end{cases}. \quad \text{Let } B = B_n = (X_1, \cdots, X_n) \in R^n$$

and set $\mathcal{C}_n = \cup_{k=0}^{m} T^k(B)$ where $m = \binom{n}{2}$* and $T^k = \{t_1 t_2 \cdots t_k \mid t_1, t_2, \cdots, t_k \in T\}$. A point $C \in \mathcal{C}_n$ is said to be *terminal* if $t(C) = C$ for all $t \in T$. It is not hard to see that $(S_1, \cdots, S_n)$ is a terminal element of $\mathcal{C}_n$ where $S_i (1 \leqq i \leqq n)$ is the elementary symmetric polynomial in $X_1, \cdots, X_n$ of degree i; in what follows it will be shown that this characterizes the terminal elements of $\mathcal{C}_n$.

The elements of R may be partially ordered by $f \leqq g \Leftrightarrow f + g = g$. For $D \in R^n$, $1 \leqq j \leqq n$, define $D^j \in R^n$ by $D_i^j = D_i(X_1, \cdots, X_{j-1}, 0, X_{j+1}, \cdots, X_n)$, $1 \leqq i \leqq n$.

Lemma 1: C is terminal $\Leftrightarrow C_1 \geqq C_2 \geqq \cdots \geqq C_n$.

Proof: Obvious.

---

* By (†), $\mathcal{C}_n = \cup_{k=0}^{\infty} T^k(B)$; on the other hand $\mathcal{C}_n = \cup_{k=0}^{r} T^k(B) \Rightarrow r \geqq m$: using notation developed below, this can be proved by induction on n as follows. If $n = 1$ it is clear; assuming it is true for a given n, identify $\mathcal{C}_n$ with $\{D^{n+1} \mid D \in \mathcal{C}_{n+1}\} \subset \mathcal{C}_{n+1}$ (see remark following Lemma 3). Using the theorem below and the induction hypothesis, there is a $\mathcal{J}$ such that $\mathcal{J}(B_{n+1}) = (S_1^{n+1}, S_2^{n+1}, \cdots, S_n^{n+1}, X_{n+1})$, and $\mathcal{J}$ is a product of at least $\binom{n}{2}$ transpositions. Let $\mathcal{J}' = (1\ 2)(2\ 3) \cdots (n\ n+1)\mathcal{J}$; then $\mathcal{J}'(B_{n+1}) = (S_1, \cdots, S_{n+1})$, $\mathcal{J}'$ is a product of $\binom{n}{2} + n = \binom{n+1}{2}$ transpositions and if for some $\mathfrak{U}$ $(\mathfrak{U}\mathcal{J})(B_{n+1}) = \mathcal{J}'(B_{n+1})$ then $\mathfrak{U}$ must be a product of at least n transpositions.

Lemma 2: If $f, g \in R$ *are such that* X_i *divides no summand of either, then* $f + X_i h_1 = g + X_i h_2 \Rightarrow f = g$.

Proof: Writing $f + X_i h_1$ as a sum of products of X_m's, both f and g are precisely the sum of those products which are not divisible by X_i.

Lemma 3: If $D \in \mathcal{C}_n$, *then for each* $j = 1, \cdots, n$ *there exists* i *such that* $D_i^j = 0$.

Proof: Assume $D \in \mathcal{C}_n$ and $1 \leqq j \leqq n$. Find $t_1, \cdots, t_r \in T$ such that $tB = D$ where $t = t_r t_{r-1} \cdots t_1$. If $r = 1$, say $t = (\alpha, \beta)$, $\alpha < \beta$; if $j \neq \alpha$ then $D_j^j = 0$ and if $j = \alpha$ then $D_\beta^j = 0$. Now assume the assertion is true whenever $r < u$, and $D = t_u \cdots t_1 B$. Find i such that $(t_{u-1} \cdots t_1 B)_i^i = 0$ and let $t_u = (\alpha, \beta)$, $\alpha < \beta$. As above, if $i \neq \alpha$ then $D_i^i = 0$ and if $i = \alpha$ then $D_\beta^i = 0$. Induction on r completes the proof.

Given $D \in \mathcal{C}_n$, Lemma 3 provides the machinery for associating D^i in a natural way with an element $\tilde{D}^i$ of $\mathcal{C}_{n-1}$: making the initial association $X_i \to X_{i-1}$ in B_n and $i \to i - 1$ in T_n for $i > j$, define $\tilde{D}^i = t_r' \cdots t_1' B_{n-1}$ where if $t_m = (\alpha, \beta), \alpha < \beta$ then

$$t_m' = \begin{Bmatrix} t_m & \text{if} \quad (t_{m-1} \cdots t_1 B_n)_i^i \neq 0 \quad \text{for} \quad i = \alpha, \beta \\ \textit{identity} \text{ otherwise} \end{Bmatrix}$$

for $1 \leqq m \leqq r$. It is clear that $\tilde{D}^i$ represents a collapsing of D at a coordinate i where $D_i^i = 0$ plus a permutation π of the other D_i^i's: $\tilde{D}^i = (D_{\pi(1)}^i, D_{\pi(2)}^i, \cdots) \in R^{n-1}$.

However, the extent of possible permuting is limited by the completeness of the order $\leqq$ on the D_i^i's as is demonstrated in the next two lemmas which apply for $1 \leqq i, j, k \leqq n$.

Lemma 4: $D \in \mathcal{C}_n$, $D_i \leqq D_j \Rightarrow j \leqq i$.

Proof: It suffices to note that an application of a transposition to a member of $\mathcal{C}_n$ preserves the order of the indices.

Lemma 5: $D_i \leqq D_k \Rightarrow D_i^i \leqq D_k^i$.

Proof: Writing $D_i = D_i^i + X_i g$ and $D_k = D_k^i + X_i h$, obtain $D_k^i + X_i h = D_k = D_i + D_k = D_i^i + D_k^i + X_i(g + h)$ which by Lemma 2 implies that $D_k^i = D_i^i + D_k^i$, that is, $D_i^i \leqq D_k^i$.

It follows from Lemmas 1, 3, 4 and 5 that if $C \in \mathcal{C}_n$ is terminal, then $C^i = (\tilde{C}_1^i, \tilde{C}_2^i, \cdots, \tilde{C}_{n-1}^i, 0)$ and $\tilde{C}^i$ is terminal in $\mathcal{C}_{n-1}$ for $1 \leqq j \leqq n$.

Theorem: $C \in \mathcal{C}_n$ is terminal $\Leftrightarrow C_i = S_i$, $(1 \leqq i \leqq n)$.

Proof: $\Leftarrow$. This direction is clear.

$\Rightarrow$. By induction on n—if $n = 1$ then $\mathcal{C} = \{B\}$ and $B = (X_1)$ so the assertion holds. Now assume the assertion holds for $n < k$, and let $C \in \mathcal{C}_k$ be terminal. Then each $\tilde{C}^i$ is terminal in $\mathcal{C}_{k-1}$ and hence by the induction hypothesis each $C_i^i = S_i^i (i = 1, \cdots, k-1; j = 1, \cdots, k)$.

In particular then $C_i \neq X_1 X_2 \cdots X_k$ for $i = 1, \cdots, k-1$. Furthermore, each C_i can be expressed as $C_i = P_1 + \cdots + P_r$ where each P_m is a product of some but not all of the X_i's. It follows for $i < k$ that

$$C_i^i = \sum_{X_i \nmid P_m} P_m, \quad \text{and consequently } C_i = \sum_{i=1}^{k} C_i^i = \sum_{i=1}^{k} S_i^i = S_i.$$

It is left to the reader to show that $C_k = S_k$ and thus complete the induction argument.

REFERENCE

1. Graham, R. L., "A Mathematical Study of a Model of Magnetic Domain Interactions," B.S.T.J., this issue, pp. 1627–1644.

CONTENTS

(Continued from front cover)

A Linear Phase Modulator for Large Baseband Bandwidths — C. L. Ruthroff and W. F. Bodtmann — 1893

Eventual Stability for Lipschitz Functional Differential Systems — G. A. Shanholt — 1905

Information Theory and Approximation of Bandlimited Functions — D. Jagerman — 1911

A Satellite System for Avoiding Serial Sun-Transit Outages and Eclipses — C. W. Lundgren — 1943

Adaptive Predictive Coding of Speech Signals — B. S. Atal and M. R. Schroeder — 1973

Contributors to This Issue — 1987

B.S.T.J. Brief: All Terminal Bubbles Programs Yield the Elementary Symmetric Polynomials — R. P. Kurshan — 1991

Bell System