THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLV OCTOBER 1966 NUMBER 8



# Information Rate of a Coaxial Cable with Various Modulation Systems

By J. R. PIERCE

(Manuscript received May 16, 1966)

*In contrast to inherently broadband media, such as radio, $TE_{01}$ waveguide, or guided coherent light, the attenuation of a coaxial cable increases rapidly with frequency. Thus, while for broadband media broadband transmission schemes (FM or PCM, for example) decrease the power required for a given channel capacity, they would seem to be ill-suited to coaxial cable.*

*Idealized comparisons are made among digital systems which transmit pulses of various numbers of amplitudes or levels. These show multilevel digital pulse transmission or analog transmission to have greater channel capacity (in the sense of information theory) than digital pulse transmission. Practical difficulties or cost of instrumentation may, in particular instances, dictate the use of single-sideband frequency-division multiplex for efficient voice transmission or binary pulse transmission for efficient digital transmission. Multilevel pulse transmission is a possible alternative if problems of instrumentation can be overcome.*

## I. INTRODUCTION

In the very early days of information theory, it was proposed that broadband signals might be sent over a narrow-band medium by using more power. Most media (such as radio) are inherently broadband, and it has turned out that for broadband media the advantage lies in the other direction. Broadband modulation systems, such as FM or PCM transmitted by means of binary pulses, increase the signal-to-noise ratio

for a given power and help to guard against distortion and interference. Indeed, there are strong arguments for the advantage of broadband modulation systems for any broadband medium, including radio, $TE_{01}$ waveguide and guided optical transmission. With the rising importance of digital transmission, there are of course very strong arguments for digital forms of modulation, such as binary pulse transmission.

Coaxial cable (and other transmission lines) are unique in that the attenuation arises extremely rapidly with increasing frequency. Qualitatively, this suggests that broadband modulation systems may be unsuited to coaxial cable. What do the numbers show?

The purpose of this paper is to illustrate strong effects, not to make exhaustive comparisons or optimizations or to take the practical matters of details of circuit use and limitations of circuit art into account. To this end, a simple, particular case will be considered — a system using standard $\frac{3}{8}$-inch coaxial cable, with a repeater spacing of two miles. Over most of the useful frequency range the received power $P_2$ will be related to the transmitted power $P_1$ by*

$$P_2 = P_1 \exp\left[-(f/0.30 \times 10^6)^{\frac{1}{2}}\right].$$

Here $f$ is the frequency in hertz. An average transmitter output power of a tenth of a watt will be assumed, and a repeater noise power density of $1.67 \times 10^{-19}$ watts/hertz, corresponding to a receiver noise temperature of 12,100°K and a noise figure of 16.2 dB. The calculations would equally apply for a tenth the average power and a tenth the noise.

## II. COMPARISONS FOR PERFECT INSTRUMENTATION

In this section we will compare channel capacities, in the sense of information theory, for various signal spectra and, in the case of digital transmission, for various encodings. It is assumed that there is no degradation due to imperfect amplification, imperfect regeneration, imperfect equalization or imperfect timing. The pulse rate of digital pulse systems is taken as $2B$, where $B$ is a sharply limited bandwidth.

As a standard of comparison we will use the channel capacity for the best possible frequency distribution of transmitter power density. This results in a frequency distribution of signal-to-noise ratio which seems unsuited to any useful analog signal, multiplex voice or video. Further, we do not know a practical digital encoding which will realize or even closely approximate this ideal channel capacity.

* This assumes that the loss is due to skin resistance in the conductors. If this is so, there is an unavoidable nonlinear phase lag of $(\frac{1}{2})$ $(f/0.30 \times 10^6)$ $\frac{1}{2}$ radians, which amounts to 13 radians or 740 degrees at 200 MHz.

We will also consider the rate or channel capacity for binary and multilevel digital pulse transmission with regeneration and for analog transmission with a flat signal-to-noise ratio. Details are given in Appendices A through D.

In Fig. 1, rate or channel capacity in megabits/second is plotted against bandwidth $B$ in megahertz. The cross is the optimal channel capacity; the dashed line indicates this rate (1581 megabits/second).

The optimal power density is nearly constant over the band, so that the received power is concentrated at low frequencies for which the cable attenuation is low. If we make the transmitted power density increase with frequency so that the received power density and the signal-to-noise ratio at the receiver are constant, the transmitter power is mostly used at high frequencies where the attenuation of the cable is high. This causes a degradation of performance.

The upper solid curve of Fig. 1 applies to this case of constant signal-to-noise at the receiver. This curve shows the channel capacity, subject to this restriction of signal, as a function of bandwidth $B$. The channel capacity is given by the formula

$$R = B \log_2 (1 + (S/N)). \tag{1}$$

The maximum capacity is about 989 megabits/second at a bandwidth of 80 MHz. This is lower than the 1581 megabits/second for optimal transmitter power distribution, because the power has been concentrated at high frequencies where the attenuation of the cable is large.

We cannot transmit a digitalized signal with the rate given by the upper curve of Fig. 1 because we don't know any practical means of encoding which will give a bit rate very close to the channel capacity. In

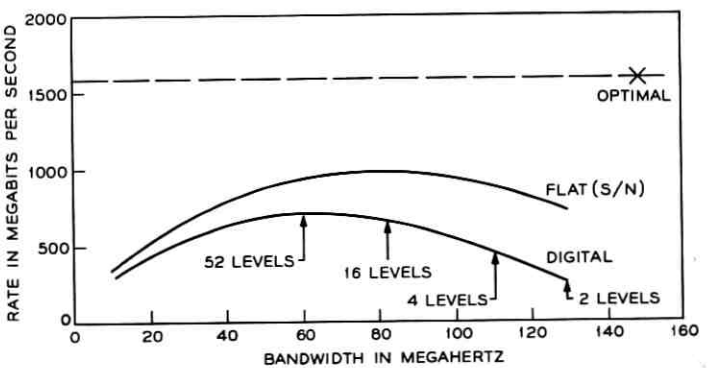



Fig. 1 — Information rate vs frequency for various forms of transmission.

practice, we can transmit digital pulses — either binary or multilevel. The lower curve of Fig. 1 is for pulse transmission with an error rate of one error per repeater in $10^{12}$ Nyquist intervals (one error in $10^8$ for 10,000 repeaters). The optimum rate is about 701 megabits/second at a band width of 60 megahertz. This would call for 52 levels, an impractically large number. The performance for 2, 4, and 16 levels is indicated on the curve.

One can do a little better with pulse transmission by equalizing differently (see Appendix E).

One can compare these bit rates with the channel capacity for analog transmission in an unquantized transmission system. This is done in Appendix D. Equalization for flat signal-to-noise is assumed. A 4000-mile system with 2000 repeaters is assumed, so that the signal-to-noise ratio is only 1/2000 that for a single two-mile link. The bandwidth is taken as 20 MHz — about that of the L4 system, which transmits voice channels by single-sideband frequency-division multiplex.[1] The channel capacity is found to be 286 megabits/second.

It is of some interest to ask what the ideal bit rate would be for digital pulse transmission over such a 4000-mile analog system, with one error in $10^8$. This is also computed in Appendix D; the bit rate is found to be 216 megabits/second and the optimum number of levels 42, which is of course impractically large.

These various results are displayed in Table I. We should remember that there is little received power in the upper part of the "148-megahertz" band of the optimal system.

We see that an analog system with a 20-megahertz bandwidth has a somewhat greater channel capacity than a binary digital system, but a smaller channel capacity than a digital system with four or more levels. As we might expect, the ideal bit rate for multilevel quantized transmission over the analog system is less than the ideal bit rate for binary digital transmission with regeneration. However, the difference is small.

TABLE I

| System | Number of Levels | Bandwidth (MHz) | Rate or Channel Capacity (Mb/s) |
|---|---|---|---|
| Optimal | — | 148 | 1581 |
| Digital | 2 | 129 | 258 |
| Digital | 4 | 111 | 444 |
| Digital | 16 | 82 | 656 |
| Digital | 52 | 60 | 701 |
| Analog | — | 20 | 286 |
| Digital on Analog | 42 | 20 | 216 |

## III. SOME PRACTICAL CONSIDERATIONS

All the comparisons in Section II are made in terms of average power. This is chiefly because Shannon's simple formula for channel capacity[2]

$$R = B \log_2 (1 + (S/N))$$

holds only for average signal power (and additive Gaussian noise). In practice, the limitation on transmitter power is more likely to be a limitation on peak power than a limitation on average power. I do not believe that a comparison based on peak power would give results substantially different from those of Section II.

The comparisons of Section II are made assuming perfect instrumentation. An actual analog system will be inferior to an ideal system chiefly because of nonlinearity. An actual digital system can be inferior to an ideal system because of imperfect equalization in amplitude and phase, imperfect level control, imperfect timing, and imperfect regeneration.

In present practice, well-instrumented analog systems (such as L4) come closer to ideal performance than well-instrumented digital systems. There are good reasons for this. In digital transmission, equalization, level control, and recovery of timing are not easy. They are usually imperfect and sometimes substantially impair performance. Moreover, the complexity of a regenerative repeater increases as the number of levels is increased. Thus, in practice the comparison of digital and an analog system will be less favorable to digital than the comparison for ideal instrumentation, which is given in Table I.

Nonetheless, comparisons of various coaxial cable systems are not easy. For a given repeater spacing an analog system appears to be somewhat better than a binary digital system for speech transmission, but if we have to transmit digital signals over it the analog system will be considerably inferior to a binary digital system for this purpose. A multilevel digital system might be very considerably superior to either an analog system or a binary digital system for either speech or data transmission.

D. G. Holloway[3] and E. D. Sunde (in unpublished work[4]) have pointed out the advantages of multilevel digital transmission.

## IV. ACKNOWLEDGMENT

The author wishes to acknowledge the thoughtful comments of Mr. R. A. Kelley, which have led to considerable clarification and improvement in this paper.

APPENDIX A

*Optimum Regenerative System*

Raisbeck[5] found the optimal power distribution and channel capacity (in the sense of information theory[2]) for a channel for which the average output power density $p'(f)$ is related to the average input power density $p(f)$ by

$$p'(f) = p(f) \exp [-(f/f_0)^{\frac{1}{2}}] \tag{2}$$

and in which the white noise power density at the output is $N$. He found that

$$\begin{aligned} p(f) &= k - N \exp [(f/f_0)^{\frac{1}{2}}], \qquad f \leqq B \\ p(f) &= 0, \qquad\qquad\qquad\qquad f \geqq B. \end{aligned} \tag{3}$$

He defines the parameter $u$ as

$$u = k/N. \tag{4}$$

The total power $P_o$, the channel capacity $C$, which is the maximum possible value of the bit rate $R$ for the power and the medium, and the bandwidth utilized in transmission, $B$, are given by

$$P_o = Nf_0(u \ln^2 u - 2u \ln u + 2u - 2) \tag{5}$$

$$C = f_0(\tfrac{1}{3} \log_2 e) \ln^3 u \tag{6}$$

$$B = f_0 \ln^2 u. \tag{7}$$

The transmitter power density $p(f)$ in watts per cycle is

$$p(f) = N\{u - \exp [(f/f_0)^{\frac{1}{2}}]\}. \tag{8}$$

This is nearly constant over most of the band, and falls rapidly to zero at the top of the band.

We will assume

$$\begin{aligned} P_o &= 0.1 \text{ watt} \\ f_0 &= 0.30 \times 10^6 \\ N &= 1.67 \times 10^{-19} \\ P_o/Nf_0 &= 2 \times 10^{12}. \end{aligned}$$

The value of $N$ chosen corresponds to a noise temperature of 12,100 degrees Kelvin, or to a noise figure of about 16.2 dB.

For these figures,

$$u = 4.435 \times 10^9$$

$$C = 1581 \times 10^6 \text{ bits/second}$$

$$B = 148.0 \times 10^6.$$

APPENDIX B

*Constant Signal-to-Noise Ratio at the Receiver*

We assume that for a transmitted power $P_o$ of frequency $f$ the received power $P_1$ is

$$P_1 = P_o \exp [-(f/f_0)^{\frac{1}{2}}]. \tag{9}$$

Suppose that we deliberately make the received power density constant with frequency. To do this we must make the transmitted power density $p_o(f)$

$$p_o(f) = \frac{P_o \exp [(f/f_0)^{\frac{1}{2}}]}{2f_0\{\exp [(B/f_0)^{\frac{1}{2}}][(B/f_0)^{\frac{1}{2}} - 1] + 1\}}. \tag{10}$$

Here $P_o$ is the total transmitted power and $B$ is the highest frequency at which power is transmitted — the bandwidth.

If the receiver noise power density has a constant value $N$, the signal-to-noise ratio $(S/N)$ at the receiver will be

$$(S/N) = \frac{P_o}{2Nf_0\{\exp [(B/f_0)^{\frac{1}{2}}][B/f_0)^{\frac{1}{2}} - 1] + 1\}}. \tag{11}$$

APPENDIX C

*The Penalty for Digital Pulse Transmission*

The bit rates computed in Appendices A and B are the limiting rates for the specified average power and noise densities. To approach them closely in digital transmission would require elaborate, error-correcting encoding. Suppose that instead of this we simply transmit digital pulses, with a signal-to-noise ratio great enough to insure a very low error rate, and without resorting to error correction.

Let us first consider binary transmission in which the pulse voltage is $\pm V/2$, where $V$ is the voltage difference between levels. If $V/2$ is the peak pulse voltage of a sin $x/x$ pulse, the average signal power is $V^2/4$. If the ratio of this average signal power to the average power of Gauss-

ian noise is 50, there will be an error rate of one in $10^{12}$ for one repeater or 1 in $10^8$ for 10,000 repeaters; this seems a reasonable rate.

For multilevel pulse transmission we say that the error rate will be nearly constant if we keep the ratio of the square of the level spacing, $V^2$, to the mean square noise voltage constant. This is nearly true for the number of errors in level when the error rate is low. The corresponding number of errors in the binary stream depends on how multilevel-to-binary encoding is done. For a Gray code an error of one level in the multilevel code will cause an error of only one bit in the corresponding binary code.

In computing the average power in the multilevel case we will assume that all levels are equally likely. Then for the same level spacing $V$ the ratio of $P_n$, the power for $n$ levels to $P_2$, the power for two levels, is

$$(P_n/P_2) = (n^2 - 1)/3. \tag{12}$$

The ratio of average signal power to average noise power, $(S/N)$, will be

$$(S/N) = 50(P_n/P_2). \tag{13}$$

The rate $r$ in bits per Nyquist interval will be

$$r = \log_2 n \text{ bits/Nyquist interval.} \tag{14}$$

The theoretical limiting rate for a flat signal-to-noise ratio is, in bits per Nyquist interval,[2]

$$c = (\tfrac{1}{2}) \log_2 (1 + (S/N)) \text{ bits/Nyquist interval.} \tag{15}$$

In Table II, $(S/N)$, $r$, $c$ and $(c - r)$ are given for several values of $n$.

TABLE II

| $n$ | $(S/N)$ | $r$ | $c$ | $c - r$ |
|---|---|---|---|---|
| 2 | 50 | 1 | 2.83 | 1.83 |
| 3 | 133.3 | 1.59 | 3.54 | 1.95 |
| 4 | 250 | 2 | 3.98 | 1.98 |
| 8 | 1050 | 3 | 5.01 | 2.01 |
| 16 | 4250 | 4 | 6.03 | 2.03 |

For larger values of $n$, $(c - r)$ is 2.03.

Thus, for a flat signal-to-noise ratio, the penalty for using digital pulse instead of optimum encoding is about two bits per Nyquist interval. If the bandwidth is $B$, this means a reduction of rate below optimum of

about

$$4B \text{ bits/second.}$$

The penalty for using digital pulse transmission is a shade less than this for binary, and a shade more for large numbers of levels.

The lowest curve in Fig. 1 shows the rate for digital pulse transmission as a function of frequency. The curve is, of course, meaningful only for integer values of $n$ — it has been drawn as a continuous curve merely for sake of appearance. This digital pulse transmission curve falls below the middle curve (ideal rate for flat signal to noise) because of the digital pulse transmission penalty.

The maximum rate for multilevel digital pulse transmission is about 700 bits/second for a bandwidth of 60 MHz. This requires a number of levels (52) which seems impractically large. Rates and bandwidths for various numbers of levels are given in Table III.

## APPENDIX D

### *Binary Digital Transmission Compared with a 20-Megahertz Channel*

It is of some interest to try to compare binary digital pulse transmission with a 20-MHz analog channel (the approximate bandwidth of L4).[1]

According to (11) of Appendix B, for $(P/Nf_0) = 2 \times 10^{12}$, the signal-to-noise ratio of a 20-megahertz channel is $3.96 \times 10^7$. This is, however, for one two-mile link. If we do not use a regenerative system, noise will accumulate. For a 4000-mile system the noise will be 2000 times as great and the signal-to-noise ratio will be $1.98 \times 10^4$. The corresponding channel capacity will be 286 megabits/second. This is slightly larger than the 258-megabit rate for binary digital transmission for the same value of $(P/Nf_0)$.

The conclusion must be that for the repeater spacing, attenuation, power and noise assumed, for transmission of analog signals, the cost of going to the large bandwidth needed for binary transmission, together

TABLE III

| $n$, number of levels | $B$, bandwidth (MHz) | Mb/s |
|---|---|---|
| 2 | 129 | 258 |
| 4 | 111 | 444 |
| 16 | 82 | 656 |
| 52 | 60 | 701 |

with the digital pulse transmission penalty, a little more than outweighs the accumulation of noise in a nonregenerative system. As an example, for voice transmission, single-sideband frequency-division transmission will give more voice channels than binary digital transmission for the same repeater spacing.

It is of interest to compute the ideal rate at which we could transmit multilevel digital pulses over this analog system with an error of one in $10^8$. For binary transmission and an error rate of one in $10^8$ the required signal-to-noise ratio is 32. The signal-to-noise ratio for the 4000-mile, 20-MHz analog system is $1.98 \times 10^4$, or 620 times as great. According to (12) of Appendix C, this should allow the number of levels $n$ to be $n = 42$, and $\log_2 42 = 5.4$. Hence, for a 20-MHz bandwidth the ideal transmission rate is $(2)(20)(5.4) = 216$ megabits/second.

Transmission of 42 levels is impractical; transmission of 16 levels might be practical, and for 40 million pulses a second this would mean 160 megabits/second.

## APPENDIX E

### *Optimum Power Density for Digital Pulse Transmission*

A channel so equalized as to give a flat signal-to-noise ratio in the received pulse train is not quite optimum for digital pulse transmission. The optimum power distribution is that which will give the greatest signal-to-noise ratio when the received signal is finally equalized to give a flat transmission band of some width $B$. It can be shown that if the ratio of received power $P_1$ to transmitted power $P_o$ is

$$P_1 = P_o \exp\left[-(f/f_0)^{\frac{1}{2}}\right], \tag{16}$$

the optimum transmitter power density $p(f)$ is

$$p(f) = \frac{P_o \exp\left[(f/4f_0)^{\frac{1}{2}}\right]}{8f_o\{\exp\left[(B/4f_0)^{\frac{1}{2}}\right]\left[(B/4f_0)^{\frac{1}{2}} - 1\right] + 1\}}. \tag{17}$$

At the receiver, equalization of the signal for flat overall frequency response will result in a noise density which rises with frequency. The overall signal-to-noise ratio $S/N$ will be

$$(S/N) = \frac{P_o(B/4f_0)}{16Nf_o\{\exp\left[(B/4f_0)^{\frac{1}{2}}\right]\left[(B/4f_0)^{\frac{1}{2}} - 1\right] + 1\}^2}. \tag{18}$$

## REFERENCES

1. Graham, R. S., The L-4 System — A 3600-Channel Coaxial System with Transistor Repeaters, Globecom Conf., June 8, 1965, pp. 419–424.

2. Shannon, C. E., A Mathematical Theory of Communication, B.S.T.J., *27*, July and October, 1948, pp. 379–423 and 623–656.
3. Holloway, D. G., Optimum Coding for Maximum Repeater Spacing, ATE J., *10*, July, 1954, pp. 188–193.
4. Sunde, E. D., Unpublished memorandum, November 19, 1962.
5. Raisbeck, G., Optimal Distribution of Signal Power in a Transmission Link Whose Attenuation is a Function of Frequency, IRE PGIT Trans., *IT-4*, September, 1958, pp. 129–130.
6. Costas, J. P., Coding with Linear Systems, Proc. IRE, *40*, September, 1952, pp. 1101–1103.

# Design of an Electro-Optic Polarization Switch for a High-Capacity High-Speed Digital Light Deflection System

BY S. K. KURTZ

(Manuscript received April 26, 1966)

*Modulator requirements for an active electro-optic polarization switch to operate in a digital light deflector (DLD) are derived. It is shown that a simple capacity-speed product of the form (capacity)$^{\frac{1}{2}}$ × (address rate) ≦ (constant) × (driver power) can be derived for both linear and biased quadratic electro-optic modulator materials. The usefulness of this relation is demonstrated by applying it to a biased quadratic electro-optic material (KTN) and two linear electro-optic materials ($LiNbO_3$ and ZnTe).*

*The results indicate that KTN will operate a DLD at a rate of $10^6$ random addresses/sec and a capacity of $10^6$ addresses with a reactive power of 2.6 watts, a bias voltage of 1200 volts, and a driver voltage of 42 volts, provided,*

*(i) fluctuations in the Curie temperature and ambient operating temperature are held to less than 0.01°C,*

*(ii) some form of ac bias is used to circumvent space charge effects, and*

*(iii) strain and defect-free material meeting these requirements can be grown to a size of at least 1 × 1 × 2 cm.*

*A linear electro-optic material such as ZnTe with a reduced half-wave voltage (unity aspect ratio) in the 2 to 3-kV range (2.5 kV at 6000 Å) will provide 3.6 × $10^6$ addresses at a rate of $10^6$ addresses/sec with a reactive driver power of 10 watts, delivered at a drive voltage of 1250 volts.*

*Experimental results obtained using KTN as a high-speed pulsed light modulator are also presented.*

## I. INTRODUCTION

In this paper we examine the design of a high-speed optical polarization switch utilizing the electro-optic properties of certain crystalline solids. Primary emphasis has been placed on potassium tantalate-niobate, but linear electro-optic materials are also considered. The design

equations are applied to a switch for a $10^6$ addresses/sec digital light deflector (DLD) described by Nelson[1] and Tabor.[2]

In Section II a derivation of the capacity-speed equation is given. Sections III and IV discuss the reactive power limitations due to heating of a KTN modulator, and typical operating characteristics for a high-speed modulator are tabulated. Section V discusses additional limitations on the capacity-speed product of KTN due to composition inhomogeneities and ambient temperature fluctuations. Space charge effects and ac biasing are treated in Section VI. In Section VII the advantages of a rectangular aperture are considered and a comparison is made of the capacity speed product of KTN with those of the linear electro-optic materials ZnTe and $LiNbO_3$. Finally, in Section VIII some experimental results are presented for pulse modulation of light using KTN.

## II. DERIVATION OF THE CAPACITY-SPEED RELATION

A polarization switch in the DLD performs the function of "rotating" the plane of polarization of a light beam rapidly through 90°. The plane of polarization thus selected determines whether the beam traverses a Wollaston prism[2] as an ordinary or extraordinary ray (i.e., determines in which direction it is deflected). This is illustrated in Fig. 1 for one module (deflection unit) of the DLD. It is well known that such a 90° change in the direction of polarization of a light beam is produced by inserting a half-wave plate into a linearly polarized light beam with the preferred axes of the plate at a 45° angle with respect to the direction of polarization of the incoming light beam.

By substituting a crystal whose refractive indices can be varied electro-optically in place of the half-wave plate, we have an electrically variable phase retardation "plate." The desired "rotation" of 90° is achieved by applying an electric field to the crystal of the correct magnitude to produce a half-wave of phase retardation between the ordinary and extraordinary ray. This is illustrated in Fig. 2. It is obvious that for a given aperture A the total length of the DLD must be restricted for some upper limit in order to prevent the optical beam from "walking off" the aperture with consequent loss of intensity in the outermost positions of the beam. This, in turn, places an upper limit on the length of the individual modulation and deflection elements.

As originally described by Nelson[1] the DLD consists of an $X$ deflection bank and an orthogonal $Y$ deflection bank in series. Each bank consists of $n$ modular units of varying length $l_n$, each module containing

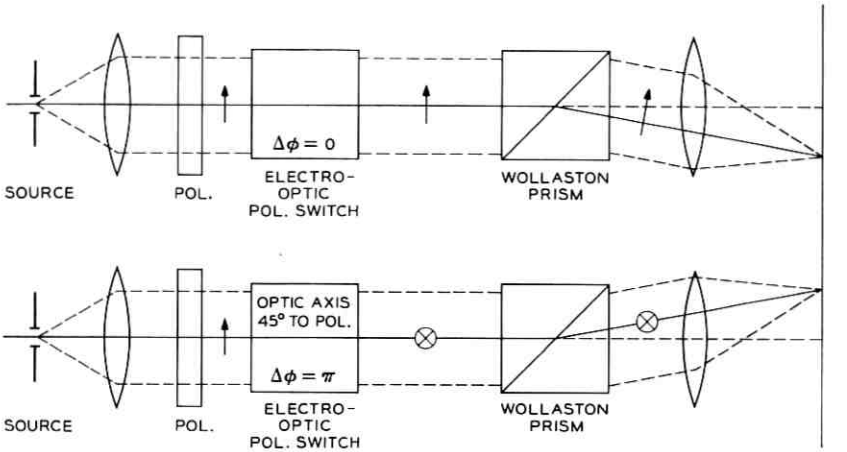



Fig. 1 — Operation of polarization switch in one deflection module of a DLD system.

an active polarization switch (modulator) and a passive birefringent deflector. The deflector thickness is predetermined to give a transverse linear displacement of the beam by distances which increase as multiples of 2, e.g., $2^0t$, $2^1t_o$, $2^2t_o$, $2^3t_o \cdots 2^nt_o$. An improved version of the DLD described by Tabor[2] utilizes Wollaston prisms which give angular rather than linear transverse displacements. The thickness of the prisms is predetermined to give angular displacements $\pm 2^0\theta_o$, $\pm 2^1\theta_o$, $\pm 2^2\theta_o$, $\pm 2^3\theta_o \cdots \pm 2^n\theta_o$, resulting in a total of $2^n \equiv R$ angular positions in each dimension. In order that each of these angular positions be resolvable the basic angular unit of deflection $2\theta_o$ is chosen to be somewhat

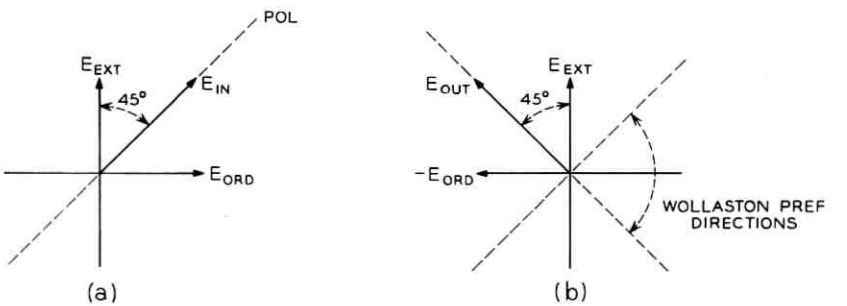



Fig. 2 — Retardation of plane of polarization by half-wave plate.

greater than the diffraction angle $\theta_D$,

$$\theta_D = \frac{\lambda}{d} \tag{1}$$

where $d$ is the width of the aperture.

It is convenient to define $\beta = (2\theta_o/\theta_D) > 1$. Assuming that the DLD consists of $2n$ modules of equal* length $l_m$, we order the modules serially in terms of increasing angular deflection.

$$\theta_{1x}, \theta_{1y}, \theta_{2x}, \theta_{2y}, \cdots, \theta_{nx}, \theta_{ny}. \tag{2}$$

This ordering minimizes walk-off since it puts the largest deflections closest to the exit part of the assembly. For this configuration the maximum cumulative walk-off (displacement at the exit part transverse to the axis of the DLD) is given by,

$$\Delta x = \frac{3}{2\sqrt{2}} \beta l_m \theta_D R \tag{3}$$

$$\Delta y = \frac{\beta l_m \theta_D R}{\sqrt{2}} \tag{4}$$

where the linear capacity $R$ is defined as

$$2^n = R. \tag{5}$$

The deviation of (3) and (4) is given in Appendix A.

If we restrict the loss of intensity in the extreme positions to be less than 20 percent (i.e., $\Delta x, \Delta y \leqq 0.14d$) then from (3) and (4) we obtain the following restriction on the length to aperture ratio for the modulator,

$$\frac{l}{A} \leqq \frac{3}{20\beta\lambda R\left[1 + \left(\frac{l_m - l}{l_m}\right)\right]} \tag{6}$$

where $A = d^2$ and $l \cong l_m$.

The next step in the derivation is to show that for both linear transverse electro-optic materials and biased quadratic electro-optic materials the reactive power is proportional to the cross-sectional area divided by the modulator length.

When an electric field is applied along a crystallographic {100} axis the principal refractive indices of KTN become[3,4]

* While in principal the prism length varies as $2^n$, in practice each prism unit is the same length, being made up of an optically isotropic support section and a thin birefringent section which varies in the prescribed fashion.

$$n_o \cong n - \frac{n^3}{2} g_{12} P_z^2$$
$$n_e \cong n - \frac{n^3}{2} g_{11} P_z^2 \tag{7}$$

hence,

$$\Delta n = n_o - n_e = \frac{n^3}{2} (g_{11} - g_{12}) P_z^2 \tag{8}$$

where $P_z$ is the induced lattice polarization in the {100} direction produced by the electric field. Here $n$ is the isotropic refractive index in zero field and the $g_{ij}$ are the quadratic electro-optic coefficients. The phase retardation can thus be expressed as

$$\Delta\varphi = \frac{2\pi}{\lambda} aP^2 l \tag{9}$$

where $a \equiv n^3/2\ (g_{11} - g_{12})$, $l$ is the length of the KTN crystal in the light direction, and the $z$ subscript on $P$ has been dropped for simplification. From (9) the polarization required to give the first half-wave of phase retardation ($\Delta\varphi = \pi$) is

$$P_\pi = \left(\frac{\lambda}{2al}\right)^{\frac{1}{2}}. \tag{10}$$

The quadratic dependence of phase retardation on lattice polarization leads to a successively closer spacing of half-wave points ($\Delta\varphi = m\pi$, $m$ a positive integer) as shown in Fig. 3. If we define the dielectric permittivity at a bias polarization $P = P_b$ as

$$\epsilon_b = \left(\frac{\partial P}{\partial E}\right)_{P=P_b}, \tag{11}$$

then the incremental voltage $\Delta V_{\pi b}$ which will produce a change in retardation of one half-wave at this bias point $P_b$ can be written

$$\Delta V_{\pi b} \cong \frac{\lambda}{4a}\left(\frac{d}{l}\right)\left(\frac{1}{P_b \epsilon_b}\right). \tag{12}$$

For a material which exhibits a linear transverse electro-optic effect[5,6,7] the induced birefringence can be expressed,

$$\Delta n = n^3 r \frac{V}{d} \tag{13}$$

where $V$ is the voltage applied perpendicular to the light path, $d$ is the

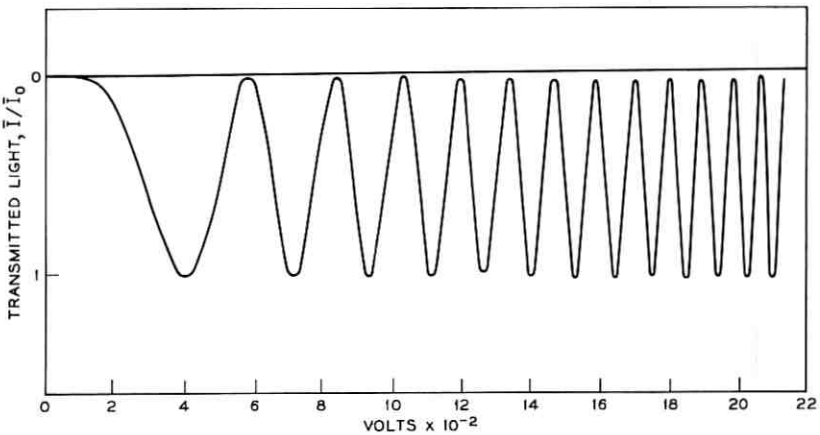



Fig. 3 — Light transmitted by KTN polarization switch as function of applied voltage.

electrode separation, and $r$ is a function of the linear electro-optic coefficient(s) determined by the orientation of the crystallographic axes relative to the electric field and light directions.[4] The half-wave voltage $V_\pi$ and reduced half-wave voltage $v_\pi$ (for unity aspect ratio) can thus be defined from (13) as,

$$\Delta V_\pi = \frac{\lambda}{2n^3 r}\left(\frac{d}{l}\right) = v_\pi\left(\frac{d}{l}\right). \tag{14}$$

The reactive power delivered by the RF driver can be expressed* as

$$\mathcal{P}_r = \tfrac{1}{2}C(\Delta V_\pi)^2\nu_r = \alpha\left(\frac{bd}{l}\right)\nu_r \tag{15}$$

where

$$\alpha = \frac{\epsilon v_\pi^{\,2}}{2}, \tag{16}$$

for linear transverse electro-optic materials, and

$$\alpha = \frac{1}{2\epsilon_b}\left(\frac{\lambda}{4aP_b}\right)^2 \tag{17}$$

for biased quadratic electro-optic materials. An expression for $\alpha$ similar to (16) but valid for biased quadratic electro-optic materials is given in

* This is the case of a pulse train of the form 1,1,1,1,1, $\cdots$. A more complete discussion of the power is given in Appendix D.

Section VII (66). Substitution of (15) ($b = d$) into (6) yields the desired capacity-speed product for a square aperture modulator.

$$R\nu_r \leqq \Lambda \mathcal{P}_r \tag{18}$$

where

$$\Lambda = \frac{3}{20\alpha\beta\lambda\left[1 + \left(\frac{l_m - l}{l_m}\right)\right]}$$

and $\alpha$ is given in (16) and (17) for the linear transverse and biased quadratic cases, respectively, and we have assumed the term $(n + 1)/R$ is small compared to unity (for $n = 10$, $R = 1024$). The total capacity is of course $R^2$ and not $R$. The capacity-speed product is easily generalized to rectangular aperture (see Appendix B) with the results

$$(R_{x'}R_{y'})^{\frac{1}{2}}\nu_r \leqq \Lambda_{x'y'}\mathcal{P}_r \tag{19}$$

where

$$\Lambda_{x'y'} = \frac{1}{5\alpha\sqrt{\beta_x\beta_y}\lambda\left[1 + \frac{l_m - l}{l_m}\right]}\sqrt{\frac{2d}{b}}. \tag{20}$$

Let us consider the implications of the capacity-speed relation for a system with a square aperture. If the dielectric constant and reduced half-wave voltage of a linear electro-optic material are fixed constants, and the address rate $\nu_r$ is also fixed by the application to be made of the DLD, then the total capacity $R_xR_y$ varies directly as the square of power available to drive the modulator. Conversely, if the capacity is fixed, the address rate varies linearly with the available power. The constant of proportionality $\Lambda$ can be calculated for a given electro-optic material and hence the capacity-speed product becomes an important design equation for determining which materials can meet the capacity-speed product required in a specific application of the DLD. It also provides a significant comparison between linear and biased quadratic electro-optic performance. The remaining sections of this paper are concerned with evaluating the optimum capacity-speed product which can be obtained using KTN, and comparing this with the capacity-speed product obtainable using known linear transverse electro-optic materials. It is clear from the form of the capacity-speed product that the question which must be answered in both cases is: What are the limitations on the power with which the modulator can be driven?

### III. DRIVER POWER LIMITATIONS FOR KTN

The purpose of a bias polarization $P_b$ is to reduce the half-wave voltage $V_\pi$ required to produce the 90° rotation of the light polarization. This is evident if we compare the unbiased half-wave voltage,

$$V_\pi = \left[\left(\frac{\lambda}{2a}\right)^{\frac{1}{2}} \frac{1}{\epsilon}\right]\left(\frac{d}{l^{\frac{1}{2}}}\right) \tag{21}$$

with the biased half-wave voltage in (12), and substitute $P_\pi$ from (10) to obtain the following relation between biased and unbiased half-wave voltage,

$$\Delta V_{\pi b} = \frac{1}{2}\left(\frac{P_\pi}{P_b}\right) V_\pi \,. \tag{22}$$

The biased half-wave voltage therefore, decreases as the inverse of the bias polarization. It is helpful conceptually to express the bias polarization $P_b$ in terms of the equivalent number of half-waves of retardation $m$ it produces. Since the phase retardation (see (9)) varies as the square of polarization we can write

$$P_b = \sqrt{m}\, P_\pi \tag{23}$$

hence,

$$\Delta V_{\pi b} \cong \frac{1}{2\sqrt{m}}\, V_\pi \,. \tag{24}$$

The introduction of a bias polarization $P_b$ can thus be used to reduce the drive voltage needed for the switch. Because of saturation effects in the induced polarization it is necessary at this point to differentiate between the low field permittivity $\epsilon$ and the small signal permittivity $\epsilon_b$ about the bias point $P_b$. Saturation behavior of KTN is describable in terms of the Devonshire free energy formalism.[8] Writing the free energy, as,

$$G = \left(\frac{T - T_o}{2\epsilon_o C}\right) P^2 + \frac{\xi}{4} P^4 + \frac{\zeta}{6} P^6 \cdots \tag{25}$$

we obtain,

$$E = \frac{\partial G}{\partial P} = \left(\frac{T - T_o}{\epsilon_o C}\right) P + \xi P^3 + \zeta P^5 \cdots \,. \tag{26}$$

Some useful relations which follow from (25) and (26) are given in Appendix C.

A plot of (26) illustrating saturation effects along with some experimentally measured points is shown in Fig. 4. In Fig. 5 we have plotted

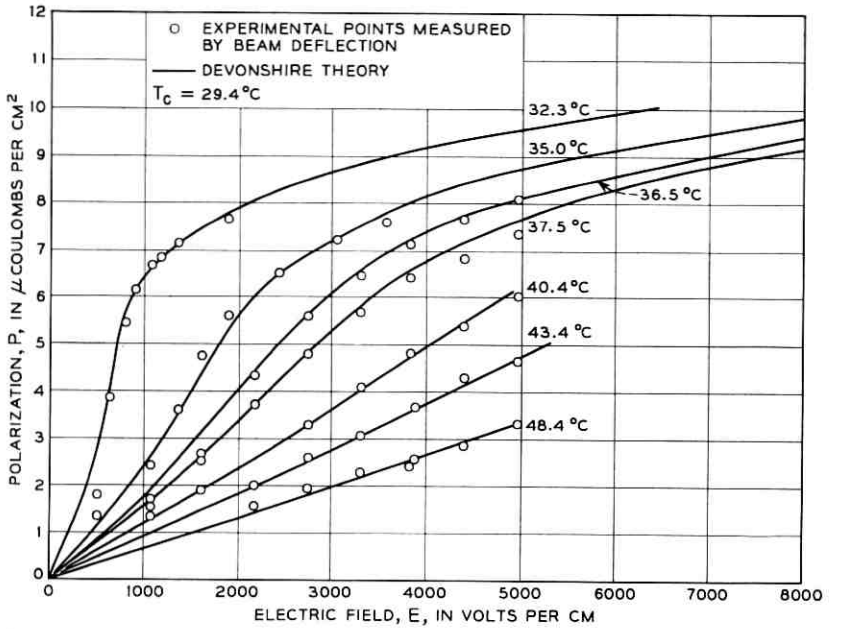



Fig. 4 — Saturation effects in the induced polarization of KTN as a function of temperature.

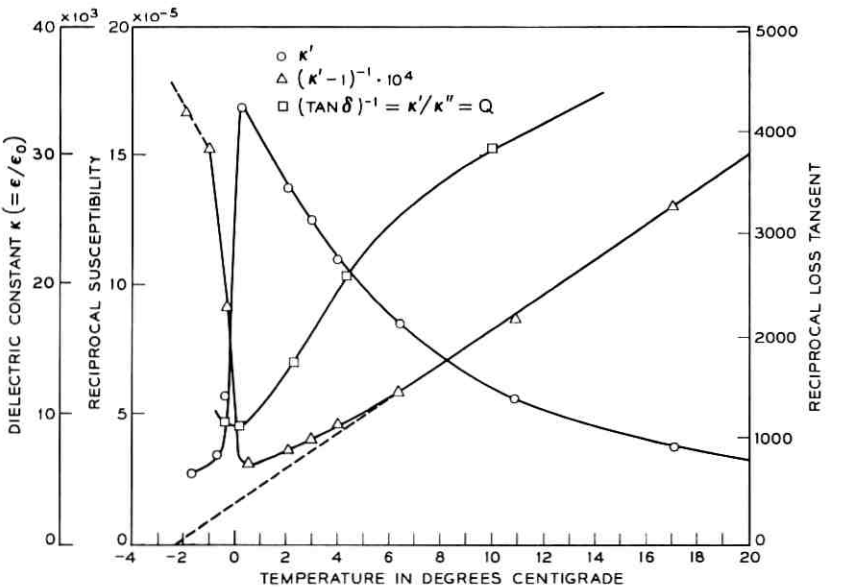



Fig. 5 — Dielectric behavior of KTN as a function of temperature.

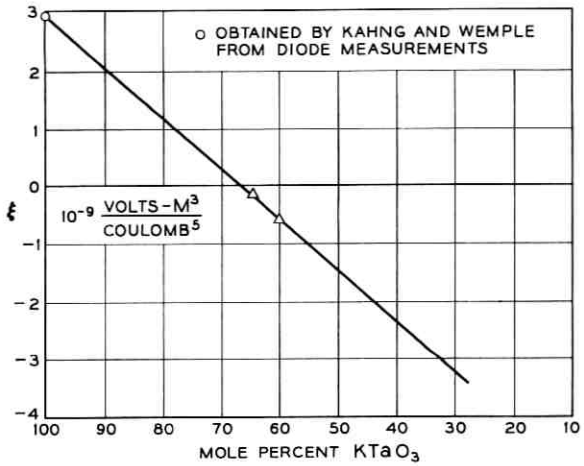



Fig. 6 — Compositional dependence of Devonshire nonlinear parameter $\xi$.

a typical low field dielectric constant $\kappa \equiv \epsilon/\epsilon_o$ versus temperature curve. Fig. 6 shows a plot of the saturation parameter $\xi$ as a function of composition for KTN. Fig. 7 shows an analogous plot of the phase transition temperature $T_o$ for KTN. Using the information given above we can continue the discussion of biasing and derive conditions for optimizing the bias polarization.

From (12) and the small signal permittivity $\epsilon_b$

$$\epsilon_b = \frac{\epsilon}{1 + 3\epsilon\xi P_b^2 + 5\epsilon\zeta P_b^4 \cdots} \tag{27}$$

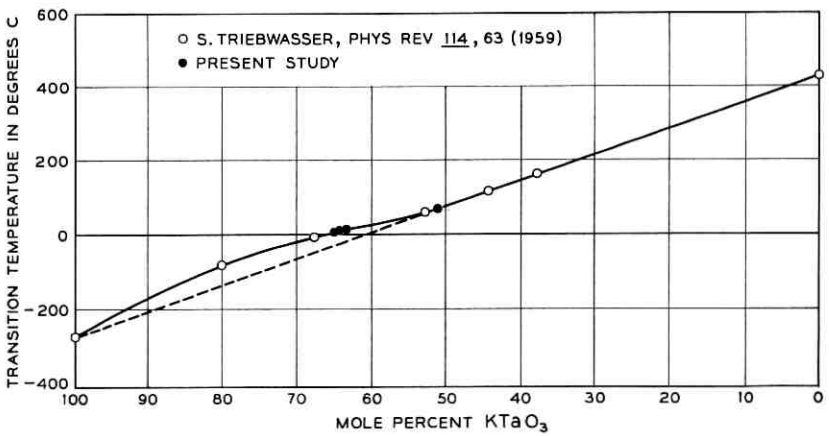



Fig. 7 — Compositional dependence of phase transition temperature in $KTa_xNb_{1-x}O_3$ system.

we see that as the bias polarization is increased, the incremental half-wave voltage has a minimum at,

$$P_b = \left(\sqrt{\frac{2}{15}}\right) P_o \left[1 + \sqrt{1 + \frac{5}{4}\left(\frac{T - T_o}{T_c - T_o}\right)}\right]^{\frac{1}{2}} \tag{28}$$

where $P_o$ is the spontaneous polarization at $T_c$ (see (77)) and $T_c - T_o$ is the difference between the transition temperature and Curie temperature (see (78)). This minimum in incremental half-wave voltage is not necessarily the desired optimum bias point, since it does not correspond to minimum reactive power. The average power dissipated in the sample $\mathcal{P}_d$, assuming it is driven by a square wave of repetitive frequency $\frac{1}{2}\nu_r$, zero-to-peak amplitude $\Delta V_{\pi b}$ and rise time $\sim (1/5\nu_r)$ is

$$\mathcal{P}_d \cong \gamma\left(\frac{\pi C_b (\Delta V_{\pi b})^2 \nu_r}{Q}\right) \tag{29}$$

where $\gamma$ is approximately 1.2. This expression is derived in Appendix C. Each half cycle of the drive voltage is capable of rotating the plane of polarization by 90°. This is done so that there will be no dc component in the drive signal. The necessity of this restriction is discussed in Section VI.

The reactive power defined in the manner of (15) is given by

$$\mathcal{P}_r = \left(\frac{Q}{2\gamma\pi}\right) \mathcal{P}_d \tag{30}$$

Substitution of this result in (18) gives a capacity-speed product,

$$(R_x R_y)^{\frac{1}{2}} \nu_r \leqq \left(\frac{Q}{2\gamma\pi} \Lambda\right) \mathcal{P}_d \tag{31}$$

where $\Lambda$ is given by,

$$\Lambda = \frac{3}{20\alpha\beta\lambda\left[1 + \frac{l_m - l}{l}\right]} \tag{32}$$

and

$$\alpha = \frac{1}{2\epsilon_b}\left(\frac{\lambda}{4aP_b}\right)^2. \tag{33}$$

If the upper limit on the dissipated power $\mathcal{P}_d$ is independent of bias polarization then the capacity-speed product has its maximum as a function of bias polarization at

$$P_b = \left(\frac{1}{5\epsilon\zeta}\right)^{\frac{1}{4}}. \tag{34}$$

The reason we have derived an expression for the capacity-speed product in terms of the dissipated power is that the primary limitation on the reactive driver power for a KTN polarization modulator can be directly related to heating caused by power dissipation within the modulator. Since we are dealing with pulse modulation the drive signal contains frequency components substantially higher than the pulse repetition rate $\nu_r$. The Fourier series expansion of a square wave contains all odd harmonics of the fundamental ($\nu_r/2$). In order to obtain an adequate rise time, the system driver plus modulator should encompass as many harmonics as possible (see (92)). If the Fourier series is terminated on the third or fifth harmonic the waveform will be that shown in Fig. 8. This places a restriction on the driver impedance if a flat response is desired. If we assume that the 3-dB power point occurs at a frequency $\nu_u$ (which we take to be an odd integer multiple of $\nu_r$) then the generator impedance is given by

$$R_g = \frac{1}{2\pi\nu_u C_b}. \tag{35}$$

Up to this point we have not set any upper limit on the power $\mathcal{P}_d$ dissipated within the modulator crystal. The heating caused by this dissipated power is not negligible. Even in the presence of large heat sinks the finite thermal conductivity of the modulator crystal gives rise to thermal gradients which affect the device performance because we are

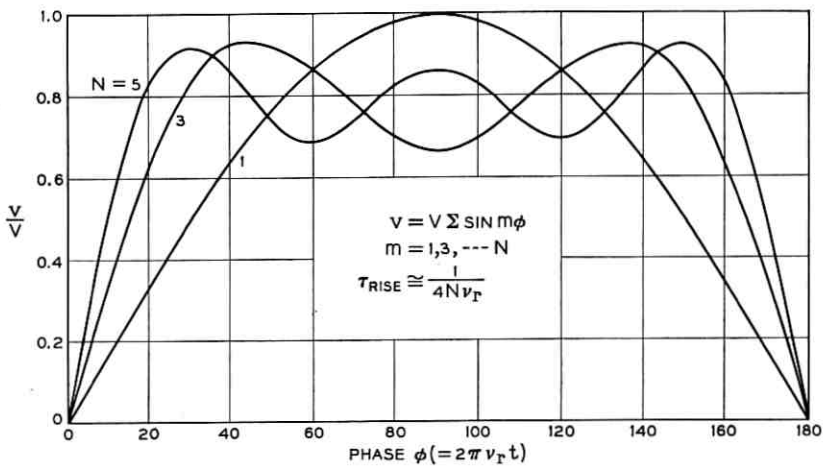



Fig. 8 — Waveshapes of a "square" wave with varying harmonic content.

operating close to the Curie point. These thermal gradients form the basic limitation on the intensity ratio which the modulator can maintain between the two senses of polarization accepted by the Wollaston prisms. Present estimates[9] of the minimum required intensity ratio for the DLD are around 20 dB.* This extinction ratio thus becomes an important design parameter which we shall now introduce into the analysis.

From (9) we can express the phase retardation $\Delta\varphi$ for two light rays traveling along paths of slightly different temperature $T$ and $T + dT$,

$$\Delta\varphi(T + dT) = \Delta\varphi(T) + \frac{\partial\Delta\varphi}{\partial T}\, dT \tag{36}$$

where

$$\frac{\partial\Delta\varphi}{\partial T} = \frac{4\pi a l P_b^{\,2}}{\lambda(T - T_o)}.$$

For light polarized at 45° with respect to principal axes of the modulator the transmission functions for the two orthogonal polarization states of the Wollaston deflectors are,

$$\begin{aligned} I_\perp &= I_o \sin^2\left(\frac{\Delta\varphi}{2}\right) \\ I_\parallel &= I_o \cos^2\left(\frac{\Delta\varphi}{2}\right). \end{aligned} \tag{37}$$

If we define $\Delta$ as

$$\Delta^{\frac{1}{3}} \equiv \frac{2\pi a l P_b^{\,2} dT}{\lambda(T - T_o)} \tag{38}$$

and take $\Delta\varphi(T) = m\pi$ where $m$ is an even integer then it is readily seen that the extinction ratio has changed from

$$\frac{I_\perp(T)}{I_\parallel(T)} = 0$$

to

$$\frac{I_\perp(T + \Delta T)}{I_\parallel(T + \Delta T)} \cong \Delta \qquad \text{where} \qquad \Delta \ll 1. \tag{39}$$

In order to find $\Delta$ as a function of position across the aperture of the modulator, we must solve the heat transfer equation. Assuming that

* A nonlinear optical absorber might reduce this to 10 dB.

the power $\mathcal{P}_d$ given in (29) is dissipated uniformly in the modulator crystal the heat transfer equation can be written,

$$\underline{\nabla} \cdot k \underline{\nabla} T = -\frac{\mathcal{P}_d}{\mathcal{U}} \tag{40}$$

where $\mathcal{U} = bdl$ is the volume of the modulator crystal, $k$ is the thermal conductivity and $T = T(xyz)$ is the temperature function. If cooling is provided primarily at two opposite faces of the modulator (40) reduces to,

$$\frac{\partial^2 T}{\partial x^2} = -\frac{\mathcal{P}_d}{k\mathcal{U}} \tag{41}$$

where $x$ is directed along the cooling surface normal. The solution to (41) is

$$T = T_s + \frac{\mathcal{P}_d}{2k\mathcal{U}}\left(\frac{d^2}{4} - x^2\right) \tag{42}$$

where $T_s$ is the surface temperature and $x$ varies from $-d/2$ to $+d/2$ if cooling is at the electroded faces. Putting this result in (38) and integrating over the aperture $bd$ we find

$$\mathcal{P}_d = \frac{\sqrt{30}\,\lambda k (T_s - T_o)(\bar{\Delta})^{\frac{1}{2}}}{\pi a P_b^2}\left(\frac{b}{d}\right) \tag{43}$$

where $\bar{\Delta}$ is the average extinction ratio over the aperture of the modulator. Substitution of this result in (31) gives the capacity-speed product,

$$(R_x R_y)^{\frac{1}{2}} \nu_r \leqq \left\{\frac{12\sqrt{30}\,aQk(\bar{\Delta})^{\frac{1}{2}}(T_s - T_o)\epsilon_b}{5\pi^2\gamma\beta\lambda^2\left[1 + \left(\frac{l_m - l}{l_m}\right)\right]}\right\}\frac{b}{d}. \tag{44}$$

This expression is independent of bias polarization and to first approximation independent of temperature. The latter statement rests on the condition that at the operating point $(T_s - T_o, P_b)$ dielectric saturation is negligible, (i.e., $\epsilon_b \cong \epsilon = C/(T_s - T_o)$ where $C$ is the Curie constant). For unity aspect ratio $(b/d)$ the capacity speed product is thus primarily determined by material parameters such as thermal conductivity $k$, electrical quality factor $Q$, Curie constant $C$, etc. The only adjustable parameters are the extinction ratio $(\bar{\Delta})$, wavelength $\lambda$, and resolution limit $\beta$. In many instances these will be determined by the choice of memory plane in a particular application. The short wavelength limit of the modulator above is determined by the width of the forbidden energy gap (in KTN 3.45 eV), which restricts use of KTN to wave-

lengths longer than 4000 Å. The advantage of using a rectangular aperture is also evident from (44). For a specific operating speed the total capacity increases linearly with the aspect ratio. Even a modest 3:1 ratio give a factor of 6 improvement in capacity. The optical image of the source is of course distended by roughly the same ratio ($b/d$) but for many applications this might not be a limitation.

An expression for the generator impedance $R_g$ can be derived if we let $\nu_u = \frac{5}{2}\nu_r$ in (35) and substitute (12), (29), and (43),

$$R_g = \frac{\pi\gamma\lambda}{80\sqrt{30}\, aQk(\bar{\Delta})^{\frac{1}{2}}\epsilon_b^2(T_s - T_o)}\left(\frac{d}{b}\right)\left(\frac{d}{l}\right)^2. \tag{45}$$

## IV. Application of the Design Equations to a $10^6$ Addresses/Sec DLD System Utilizing KTN Modulators

In order to obtain a better idea of the implications of the various relations derived in the preceding sections it is necessary at this point to substitute some of the physical constants. Taking the values of the constants listed in Table I, we can evaluate the capacity-speed product from (44),

$$(R_xR_y)^{\frac{1}{2}}\nu_r \leqq \frac{7{,}950(\bar{\Delta})^{\frac{1}{2}}}{1 + \left(\frac{l_m - l}{l_m}\right)}\left(\frac{b}{d}\right) \text{ addresses MHz.} \tag{46}$$

The choice of $\beta = 4$ is based on an extrapolated improvement of 2 in the value of 8 obtained by Tabor[9] in a DLD using passive modulators.

Taking an extinction ratio of 20 dB ($\bar{\Delta} = 0.01$) and a speed of $10^{-6}$ sec/address we obtain from (46) the maximum capacity,

$$(R_xR_y)^{\frac{1}{2}} = \frac{795}{1 + \left(\frac{l_m - l}{l_m}\right)}\left(\frac{b}{d}\right) \text{ addresses.} \tag{47}$$

It is therefore, advantageous to make $(l_m - l)/l_m$ as small as possible.

The Wollaston prisms, plus support sections, plus clearance (i.e.,

Table I

| | |
|---|---|
| Thermal conductivity $k$ | 50 mW/cm °C |
| Electrical quality factor $Q$ | 1000 (at 1 MHz) |
| Electro-optic parameter $a$ | 1.13 $m^4$/coulomb$^2$ (at 5000 A) |
| Curie constant | 1.4 $\times 10^5$ °K$^{-1}$ |
| Light wavelength | 5000 Å |
| Resolution factor $\beta$ | 4 |

$l_m - l$) can be conservatively set at a lower limit of 2 mm. Let us first consider the case of a square aperture (i.e., $b = d$). For this case, (6) becomes

$$A \geqq \frac{4}{3} l_m 10^{-3} R(\text{cm}^2) \tag{48}$$

where

$$R_x = R_y = R.$$

Taking $l_m = 1$ cm ($l = 0.8$ cm) we arrive at the modulator and DLD characteristics shown in Table II. Increasing $l_m$ beyond 1 cm (e.g., to 2 cm) would only increase the capacity by 20 percent. The operation of a $10^6$ bit/sec DLD system with a square aperture using KTN modulators is thus limited to a capacity of about $0.5 \times 10^6$ addresses. However, several factors which have been neglected serve to further limit this capacity. These factors are discussed in the next two sections, as is the rectangular aperture which enables the capacity to be increased to $10^6$ addresses.

## V. Effects of Compositional Inhomogeneities and Ambient Temperature Fluctuations on KTN Modulator Performance

Compositional inhomogeneities occur during the growth of KTN crystals[10] which give rise to fluctuations in the Curie temperature throughout the crystal. The exact nature of the inhomogeneities and their elimination is beyond the scope of this paper. The relevant point in this discussion is that Curie temperature variations do exist and should be included in the modulator analysis. Examination of (38) shows that we can extend the interpretation of $dT$ as

$$dT = d(T - T_o) = dT(x,y) + dT_s - dT_o(x,y). \tag{49}$$

Table II — Performance Characteristics of DLD using KTN Modulator

| | |
|---|---|
| Modulator dimensions | $b = d = 0.92$ cm<br>$l = 0.8$ cm |
| DLD capacity | $0.4 \times 10^6$ addresses |
| DLD speed | $10^{-6}$ seconds/address |
| Dissipated power | 9.7 mW |
| Reactive power | 1.3 W |
| Generator impedance | 65 Ω |
| Bias polarization | 2 $\mu$ coulombs/cm$^2$ |
| Bias voltage | 1500 V |
| Driver voltage | 51 V |
| Number of half-waves bias | 15 |
| Capacitance of modulator | 1000 pF |

In the previous treatment of Sections III and IV we neglected $dT_s$ and $dT_o$. From (42) we can calculate $dT(x = 0)$ for a dissipated power of 10 mW, a volume of 0.65 cm$^3$ and $d = 0.92$ cm corresponding to the modulator discussed in Section IV, and find

$$dT(x = 0) = 0.03°\text{C}. \tag{50}$$

The previous analysis is thus valid for $dT_s$ and $dT_o$ much less than 0.03°C. This corresponds to temperature regulation in the thousandths of a degree region which is within capabilities of present technology but requires some sophistication and cost.

Curie temperature variations occurring during growth are presently[11] in the range 1°C to 10°C for samples several mm's on a side. To hold variations of 0.01°C requires a control of the solid-solution to within roughly 20 ppm of the 65/35 mixture. Let us for the sake of discussion see what effect a constant variation $\delta \equiv \Delta T_s - \Delta T_o$ of 0.01°C would have on the derivation of (43). The result of this calculation is the following equation,

$$\mathcal{P}_d = \frac{2k\mathcal{U}}{d^2}\left[-\frac{5}{2}\delta + \frac{1}{2}\left\{-95\delta^2 + 120\bar{\Delta}\left(\frac{\lambda(T_s - T_o)}{2\pi a P_b{}^2 l}\right)^2\right\}^{\frac{1}{2}}\right]. \tag{51}$$

For $\delta \to 0$, $\mathcal{P}_d$ of course reduces to the expression given in (43). For $\delta \neq 0$ the negative sign in the radical, combined with the requirement that $\mathcal{P}_d$ be a real positive quantity, indicates that there is a lower limit to the quantity $[(\lambda(T_s - T_o)/2\pi a P_b{}^2 l]$. The previous analysis did not place any limit on $T_s - T_o$ and $P_b$ ; and $l_m$ could be made larger if $A$ was increased. The additional restriction coming from (51) leads to a modified capacity-speed product,

$$(R_x R_y)^{\frac{1}{4}}\nu_r \leqq \frac{24a^2 Qk\epsilon_b(P_b{}^2 l)\left[-\frac{5}{2}\delta + \frac{1}{2}\left\{-95\delta^2 + 120\left(\frac{\lambda(T_s - T_o)}{2\pi a}\right)^2\left(\frac{1}{P_b{}^2 l}\right)^2\bar{\Delta}\right\}^{\frac{1}{2}}\right]\left(\frac{b}{d}\right)}{5\pi\gamma\beta\lambda^3\left[1 + \left(\frac{l_m - l}{l}\right)\right]}. \tag{52}$$

Remembering that $l_m - l/l_m < 1$ it can be shown that the capacity-speed product has a maximum, as a function of $P_b{}^2 l$ at

$$P_b{}^2 l = \sqrt{\frac{5}{19}}\,\frac{(\bar{\Delta})^{\frac{1}{2}}\lambda(T_s - T_o)}{2\pi a\delta} \tag{53}$$

and goes to zero at,

$$P_b^2 l = \frac{(\bar{\Delta})^{\frac{1}{2}} \lambda (T_s - T_o)}{2\pi a \delta} . \tag{54}$$

Putting the values $\lambda = 5000$ Å, $(T_s - T_o) = 10°\text{C}$, $a = 1.13$ m$^4$/coulomb$^2$, and $\delta = 0.01°\text{C}$ into (53) and (54) we find $R\nu_r$ has a maximum at

$$P_b^2 l = 3.6 \frac{(\mu \text{ coulomb})^2}{\text{cm}^3} \tag{55}$$

and zero at

$$P_b^2 l = 7 \frac{(\mu \text{ coulomb})^2}{\text{cm}^3} . \tag{56}$$

In the earlier calculation of Section III where $\delta = 0$ we arrived at the values $P_b = 2$ $\mu$ coulombs/cm$^2$ and $l = 8$ mm giving $P_b^2 l = 3.2$ which is only slightly smaller than the optimum above. Using the parameters in Table I and taking $T_s - T_o = 10°\text{C}$ we can write (52) as

$$R\nu_r \leqq \frac{2010 P_b^2 l \left[ -\frac{5}{2}\delta + \frac{1}{2}\left\{ -95\delta^2 + \frac{61\Delta}{(P_b^2 l)^2} \right\}^{\frac{1}{2}} \right]}{\left[ 1 + \left( \frac{l_m - l}{l_m} \right) \right]} \cdot \frac{b}{d} \text{ address-MHz.} \tag{57}$$

Putting $P_b^2 l = 3.6$ $\mu$ coulomb 2/cm$^3$, $\bar{\Delta} = 0.01$ and $\delta = 0.01°\text{C}$, $\nu_r = 1$ MHz, and $l = 8$ mm into (57) we find the maximum capacity is $0.18 \times 10^6$ addresses. This is 35 percent lower linear capacity (418) than the maximum linear capacity (635) obtained when $\delta$ was assumed to be negligible. It can in fact be shown that if $P_b^2 l$ is chosen to satisfy (53) then the capacity-speed products in (44) and (52) are related by a constant multiplier,

$$(R\nu_r)_{\delta \neq 0} = (R\nu_r)_{\delta=0} \left[ \left( 1 - \frac{95}{456} \right)^{\frac{1}{2}} - \frac{5}{\sqrt{456}} \right] = 0.656 (R\nu_r)_{\delta=0} . \tag{58}$$

Thus, the primary effect of introducing compositional nonuniformities and ambient temperature fluctuations has been to decrease the capacity-speed product by 35 percent, and also to prescribe an optimum value for $P_b^2 l$ given by (53). It is interesting to note that for $l > 2$ mm the value of $P_b$ obtained in this fashion is substantially less than the value required to minimize the power or drive voltage (see (28) and (34)). Since $P_b^2 l$ from (53) is proportional to $1/\delta$ an increase in $\delta$ beyond 0.01°C would reduce $P_b^2 l$ to less than 3.6 $\mu$ coulombs$^2$/cm$^3$. A substantial

increase in $\delta$ beyond 0.01°C is not desirable, however, since $l$ must be greater than 2 mm to prevent the denominator in the capacity-speed product from becoming large, and secondly $P_b$ cannot be reduced much below 2.0 $\mu$ coulombs/cm$^2$ as the drive voltages become excessive. Thus, it is clear that close temperature regulation to several hundredths of a degree, and precise control of chemical composition to tens of ppm are both essential. In the next section, an ac biasing scheme is discussed which relaxes these requirements.

Table III lists the new operating characteristics for a square aperture system which result from the modified design equations derived in this section.

## VI. SPACE CHARGE EFFECTS AND AC BIASING

In this section, we shall consider the adverse effects of finite electrical conductivity in the presence of a dc bias, and discuss the ac biasing scheme proposed by Warter[12] for overcoming these effects. We shall also consider the advantages of a rectangular aperture and the problems associated with finding other materials than KTN for use as the active electro-optic medium.

The static electrical conductivity of KTN in the vicinity of 300°K falls in the range $10^{-11}$ to $10^{-12}$ mhos/cm. This conduction has been demonstrated to be extrinsic and due to holes having a very low trap controlled mobility of $10^{-6}$ cm$^2$/V sec. The filled acceptor level density is less than $10^{13}$/cm$^3$ and is peaked around 0.6 to 0.8 eV above the valence band.

This small but finite conductivity gives rise to several types[13] of nonuniform electric polarization distribution within the sample when a dc electric field is applied. The type of nonuniformity and the associated time constant depend on the nature of the electrical contact (e.g., ohmic

TABLE III — PERFORMANCE CHARACTERISTICS OF DLD USING KTN MODULATOR IF $\delta$ = 0.01°C

| | |
|---|---|
| Modulator dimensions | $b = d = 0.75$ cm<br>$l = 0.8$ cm |
| DLD capacity | $0.18 \times 10^6$ addresses |
| DLD capacity-speed product | $0.42 \times 10^9$ sec$^{-1}$ |
| Dissipated power | 6.4 mW |
| Reactive power | 0.86 W |
| Generator impedance | 43 Ω |
| Bias polarization | 2 $\mu$ coulombs/cm$^2$ |
| DC bias voltage | 1200 V |
| Driver voltage | 42 V |
| Number of half-waves bias | 15 |
| Capacitance of modulator | 1000 pF |

or blocking). In the case of blocking contacts[14] two phenomena occur with different time constants. The first effect is the build-up of a space charge in the vicinity of nonuniform conductivity and/or nonuniform dielectric constant. It can be shown from Maxwells equations that the time constant for this build-up is approximately given by the dielectric relaxation time,

$$\tau_r = \rho\kappa\epsilon_o \tag{59}$$

where $\rho$ is the average resistivity and $\kappa$ the average dielectric constant. For a dielectric constant of $10^4$ and a resistivity of $10^{12}$ Ω-cm, $\tau_r$ is of the order of $10^3$ seconds. The second effect is the formation of a depletion layer[14] in the vicinity of the blocking (positive for $p$-type conduction) electrode. The time constant for depletion layer formation is given by,

$$\tau_{sh} = \frac{d(\infty)L}{2\mu_{eff}V} \tag{60}$$

where $d(\infty)$ is the final depletion layer width

$$d(\infty) = \sqrt{\frac{2\epsilon V}{N_T e}}$$

and $L$ is the electrode separation, $N_T$ is the density of filled trapping levels, and $\mu_{eff}$ is the trap-controlled or "effective" mobility. For KTN having a resistivity of $10^{12}$ ohm-cm this depletion layer buildup time is between $10^3$ and $10^4$ seconds.

In the case of ohmic contacts[14,15] space charge buildup can occur due to nonuniformities in the conductivity, as in the case of blocking contacts, or a space charge may also occur which is associated with operation in range of space charge limited current flow. It can be shown that the transition from ohmic current flow to space charge limited current flow occurs when

$$V > \frac{8}{9}\left(\frac{N_T e L^2}{\epsilon}\right). \tag{61}$$

The large dielectric constant of KTN causes this transition to occur at much lower current levels than in other materials.

The nonuniformities in polarization caused by each of these effects exceeds that which the system can tolerate by several orders of magnitude. While the possibility of a direct solution to the dc bias problem cannot be ruled out, two alternatives exist which eliminate the need for a dc bias. The first of these involves operation of the modulator in the ferroelectric region[16,17] with the spontaneous polarization acting as the

bias polarization. The previous analysis of Sections II to V is applicable to this case except that the dielectric constant no longer obeys a Curie-Weiss law but varies in the fashion described by (82) in Appendix C. The temperature dependence of the spontaneous polarization is also different, which changes the derivation of (43).

A comparison of (80) and (81) shows that the dielectric constant drops discontinuously to 1/4 of its peak value at the phase transition. For KTN with a transition temperature of 10°C this means a drop from 36,000 to 9,000. Thereafter, it drops to below 2,000 within a temperature range of less than 10°C. If one modifies the analysis of the preceding sections to take these facts into account, one finds that ferroelectric biasing further reduces the maximum capacity by a factor of at least 4. The further assumption is made here that large samples will remain single domain when operated within 5 to 10°C of the Curie point.

A second biasing scheme which does not appear to have this liability has been proposed by P. J. Warter, Jr.[12] The basic idea in Warter's scheme is to use two modulator crystals in series (1 and 2). An ac bias source provides separate current drives in quadrature to the two sections, as indicated in Fig. 9 for one of the two modulators,

$$\begin{aligned} i_1 &= i_o \sin \omega_o t \\ i_2 &= i_o \cos \omega_o t. \end{aligned} \tag{62}$$

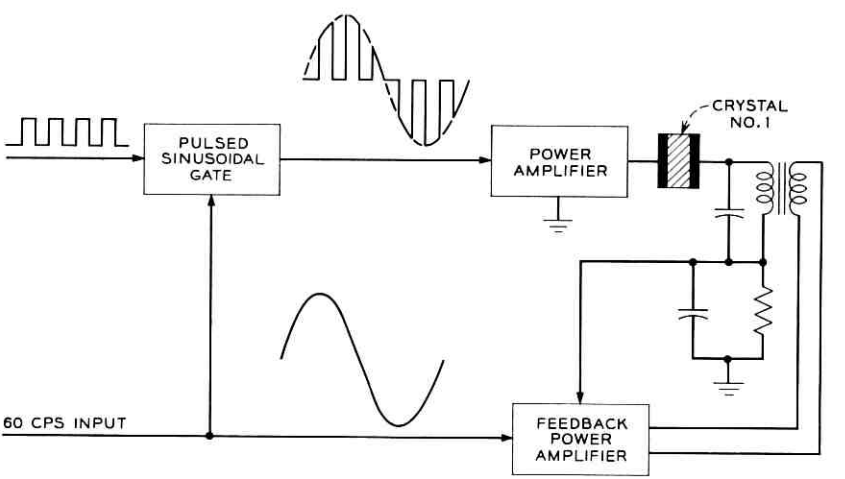



Fig. 9 — Block diagram of driver circuit for modulator 1 in Warter's ac bias scheme.

The charge on the electrodes of the two samples produces an electric polarization in the sections which varies as

$$\begin{aligned} P_1 &= P_o \cos \omega_o t \\ P_2 &= P_o \sin \omega_o t \end{aligned} \tag{63}$$

where $P_o = i_o/\omega_o A$, $A$ being the electrode area. The phase retardation through the two sections is the sum of retardations in each section and hence,

$$\Delta\varphi = \frac{2\pi}{\lambda} a(P_1^2 + P_2^2)\ell = \frac{2\pi a}{\lambda} P_o^2 \ell \tag{64}$$

with $P_o$ being the bias polarization. Since the bias signal on each section contains no dc component the previously discussed conduction phenomenon does not occur if the contacts are electrically blocking and the period $\tau_o = 2\pi/\omega_o$ is short compared to the dielectric relaxation period $\tau_r$.

In addition the use of a current drive rather than a voltage drive insures a constant polarization along a path normal to the electrodes ($x$ direction) even if the temperature changes slightly. This also holds for nonuniform temperature variations along the same path within the sample. The electric fields adjust internally for regions of varying dielectric constant such as to maintain uniform polarization along the $x$ direction. This relaxes the stringent requirements on ambient temperature control and chemical homogeneity. A detailed analysis of the limitations of the ac bias is needed before any reliable statements can be made as whether it will allow a significant increase in the capacity speed products calculated in the previous sections. Such an analysis depends critically on distortions in the drive signal from a pure sinusoidal behavior which are not known at present. It should be noted that each modulator in this scheme must be the same length $l$ as a single modulator in the previous analysis, giving a reduction of 4 (see (20)) in the capacity $R^2$ if the cross-sectional area $A$ is held constant.

Another adverse optical effect which has been observed in dc biased KTN polarization switches occurs when the diameter of the optical beam (of several mW power) is reduced to around 0.2 mm. Under these conditions a severe distortion of the optical transmission function from its expected form (see (37) and Fig. 3) was observed as shown in Fig. 10. If the light was switched on rapidly (in $<0.1$ sec.) the initial transmission function was that shown in Fig. 3, but went over into that shown in Fig. 10 after several seconds of illumination. If the light beam was moved rapidly to a different region the same sequence was observed. A return

to the original spot gave the pattern in Fig. 10 without any buildup time, indicating that the cause of the refractive index distortion was still present, and had not decayed after many seconds. Insertion of 30 dB of optical attenuation in the input beam was found to completely eliminate the occurrence of distortion. Further measurements[18] suggest that the distortion is being produced by an internal bias field acting in opposition to the applied field but nonuniformly distributed in the immediate vicinity of the affected area. Since no distortion was observed in the absence of a dc bias field one is tempted to postulate that the intense light beam is generating some sort of charged centers or charge carriers which drift under the influence of the external field and are then trapped near the edge of beam. Chen[18] and Boyd[19] have made further optical measurements of the distortion of the refractive index ellipsoid of KTN in the vicinity of beam which indicate that the effect being described here may be related to "optical damage" effects observed recently in $LiNbO_3$ and $LiTaO_3$.

Since no satisfactory explanation of either effect is available at the present time we shall conclude this discussion by noting that this problem is not one of concern for the DLD polarization switch since optical levels will probably be somewhat less than 0.1 W/cm$^2$, and for KTN the ac bias scheme of Waters would circumvent the problem even at high light levels.

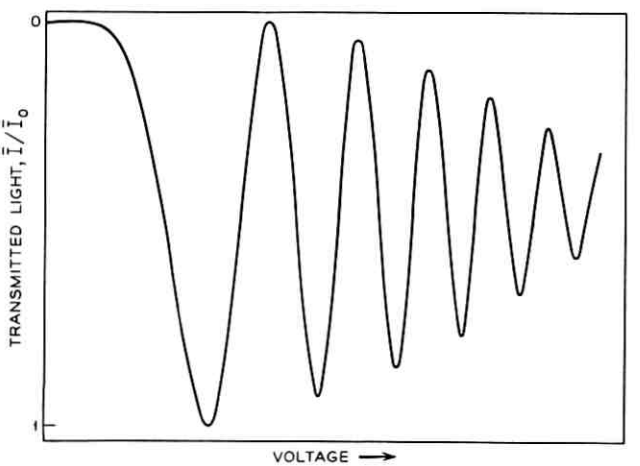



Fig. 10 — Light transmitted by optically "damaged" region of KTN polarization switch.

## VII. RECTANGULAR APERTURE AND COMPARISON OF LINEAR ELECTRO-OPTIC MATERIALS WITH KTN

### 7.1 *Rectangular Aperture*

It was shown in Section II that the capacity speed product depended primarily on the amount of power one can use to drive the modulator. In Section III we showed that this power for KTN is limited by the extinction requirements of the DLD. Holding the extinction constant it was shown that one could increase this power hence the capacity-speed product by using a rectangular aperture (i.e., $b/d > 1$). This conclusion was unchanged when Curie temperature variations and ambient temperature fluctuations were included (see (52)). Let us consider the improvement which a modest aspect ratio of 3:1 ($= b/d$) can make in the capacity-speed product of the KTN modulator described in Section V. Table IV lists the operating characteristics of a KTN nodulator with $b/d = 3$. The capacity can therefore, be brought up to the megabit region with only a modest 3/1 aspect ratio, while maintaining reasonable driver requirements, well within the capabilities of transistor circuitry.

### 7.2 *Comparison of Linear Electro-optic Materials with KTN*

In Section II it was shown that the capacity-speed product for different electro-optic modulator materials (e.g., $A$ and $B$) being driven with identical reactive powers differed only in the factor (see (20)).

TABLE IV — PERFORMANCE CHARACTERISTICS OF DLD USING KTN AND RECTANGULAR APERTURE

| | |
|---|---|
| Modulator dimensions | $b = 2.24$ cm<br>$d = 0.75$ cm<br>$l = 0.8$ cm |
| DLD capacity | $1.0 \times 10^6$ addresses |
| $R_x$ | 422 |
| $R_y$ | 3700 |
| DLD capacity-speed product | $1.0 \times 10^9$ sec$^{-1}$ |
| Dissipated power | 19 mW |
| Reactive power | 2.6 W |
| Generator impedance | 15 Ω |
| Bias polarization | 2 $\mu$ coulombs/cm$^2$ |
| dc bias voltage | 1200 V |
| Driver voltage | 42 V |
| Number of half-wave bias | 15 |
| Capacitance of modulator | 3000 pF |
| Dielectric constant | 14,000 |
| Reduced biased half-wave voltage | 45 V |

$$\frac{[(R_xR_y)^{\frac{1}{2}}\nu_r]_A}{[(R_xR_y)^{\frac{1}{2}}\nu_r]_B} = \frac{\alpha_B\left(1 + \frac{l_m - l}{l}\right)_B}{\alpha_A\left(1 + \frac{l_m - l}{l}\right)_A}. \tag{65}$$

Since we assumed that $l_m - l/l_m$ is small, the comparison of different modulator materials reduces to a comparison of their $\alpha$'s. To facilitate this comparison we put (17) for biased quadratic-electro-optic materials in the form,

$$\alpha = \frac{\epsilon_b}{2}(\Delta v_{\pi b})^2 \tag{66}$$

where $\Delta v_{\pi b} = \Delta V_{\pi b}(l/d)$ is the reduced half-wave voltage for the biased material,

$$\Delta v_{\pi b} = \frac{\lambda}{4aP_b\epsilon_b}. \tag{67}$$

For the KTN modulator described in Table IV, $\alpha/\epsilon_o$ has the value $1.4 \times 10^7$ $(\text{V})^2$. For a linear electro-optic material with a dielectric constant of $10\epsilon_o$, a reduced half-wave voltage of 1670 volts would be required to give the same value of $\alpha$ as the biased KTN.

In the past the lowest reported[5] reduced half-wave voltage for a linear transverse effect was 6200 volts for cuprous chloride. Recently both lithium niobate[20] and zinc telluride[6] have been shown to have reduced half-wave voltages of less than 5 kV. Lithium niobate has the disadvantages of a somewhat higher dielectric constant and an appreciable natural birefringence which would tend to limit the angular aperture. Zinc telluride has a relatively small optical band gap ($\sim$2 eV) and would be restricted to use at wavelengths $>6000$ Å. Nevertheless, it is significant that materials with a sufficiently large linear transverse electro-optic effect and low dielectric constant do exist. Table V lists the capacity-speed product and several other operating characteristics of a DLD designed using $LiNbO_3$ and ZnTe. One of the principal advantages is relative insensitivity of these materials to temperature changes. This means that larger reactive powers may be used if one can meet the drive voltage requirements. The rectangular aperture again offers an advantage by allowing a reduction in this drive voltage while maintaining a high capacity. This fact has been used in deriving the numbers in Table V.

TABLE V — COMPARISON OF KTN WITH LINEAR ELECTRO-OPTIC MATERIALS

| | $LiNbO_3$* | ZnTe† | KTN‡ |
|---|---|---|---|
| Modulator dimensions $b$ | 2.64 cm | 2.64 cm | 2.24 cm |
| $d$ | 0.5 cm | 0.5 cm | 0.75 cm |
| $l$ | 1 cm | 1 cm | 0.8 cm |
| Reduced half-wave voltage $v_\pi$ | 400 V | 2500 V | 45 V |
| Dielectric constant | 40 | 10 | 14,000 |
| Capacity-speed product | $1.9 \times 10^8$ sec$^{-1}$ | $1.9 \times 10^9$ sec$^{-1}$ | $1.0 \times 10^9$ sec$^{-1}$ |
| Capacity | $3.6 \times 10^4$ add. | $3.6 \times 10^6$ add. | $1.0 \times 10^6$ add. |
| Speed | $10^{-6}$ sec/add. | $10^{-6}$ sec/add. | $10^{-6}$ sec/add. |
| Driver voltage | 2000 V | 1250 V | 42 V |
| Reactive power 10 W | 10 W | 10 W | 2.6 W |
| $\alpha/\epsilon_o$ | $3.2 \times 10^8$ (V)$^2$ | $3.1 \times 10^7$ (V)$^2$ | $1.4 \times 10^7$ (V)$^2$ |
| DC bias voltage | 0 V | 0 V | 1200 V |
| Capacitance | 19 pF | 4.6 pF | 3000 pF |
| Heating $\Delta T$ | 0.04°C | 0.4°C | 0.02°C |

* $\lambda$ = 5000 Å, $Q$ = 1000.
† $\lambda$ = 6000 Å, $Q \sim 100$.
‡ $\lambda$ = 5000 Å, $Q$ = 1000, dc bias.

## VIII. SOME EXPERIMENTAL RESULTS

Because of the large composition fluctuations in presently available KTN mentioned in Section V the experiments described in this section were carried out using much smaller samples than those needed for a $10^6$ address DLD system. Samples were generally several mm on a side. Pulse experiments were performed using the circuit shown in Fig. 11. At slow sweep speeds, around 1 msec/cm, the expected modulation waveforms (see Fig. 12) were observed with 100 percent modulation occurring when the pulse height equaled the dc incremental half-wave voltage. For faster sweep speeds in the 1 $\mu$sec/cm range, a strong ringing of the

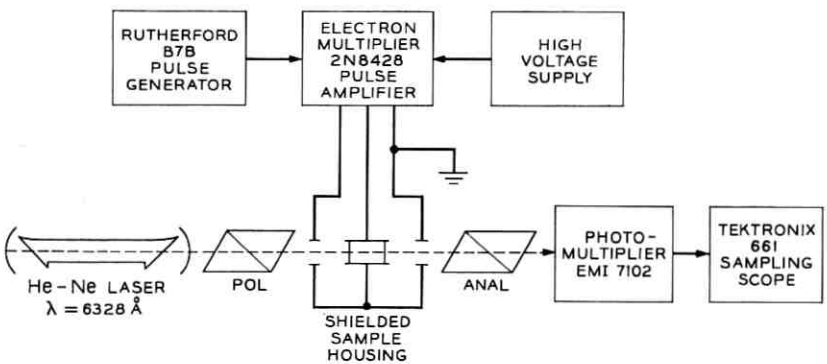



Fig. 11 — Experimental apparatus for unbiased pulse measurements.

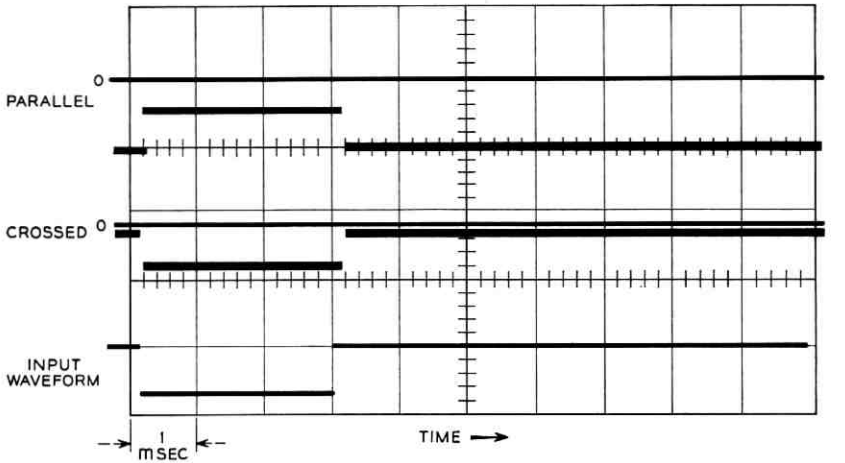



Fig. 12 — Very low frequency light modulation response.

light intensity following the leading and trailing edges of the pulse was observed. The damping of these oscillations was small. Analysis with a Rohde Schwartz receiver showed that a number of frequencies were being superposed and that these corresponded approximately to the low-order mechanical vibration modes[21] of the sample. Typical responses are shown in Figs. 13 and 14. In Figs. 14(a) and 14(b) the pulse width

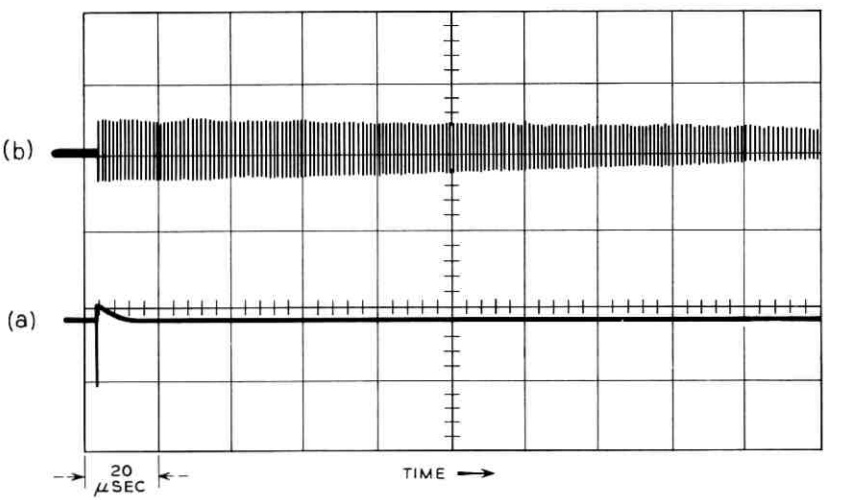



Fig. 13 — Ringing in light modulation at intermediate pulse widths (several μsec).

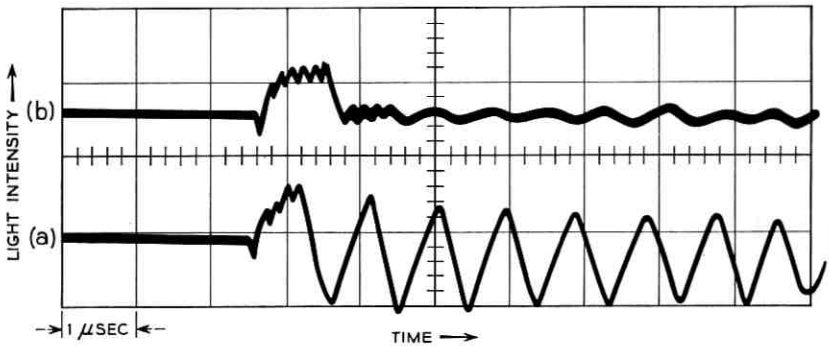



Fig. 14 — Interference effects induced by varying the pulse widths.

has been adjusted to give, respectively, constructive and destructive interference. The dominant frequency excited is that corresponding to the fundamental longitudinal thickness mode. A further study with sine wave excitation revealed that in the unbiased case, resonant excitation of this mode occurred at $\nu_{\text{drive}} = \frac{1}{2}\nu_{\text{fundamental}}$ corresponding to electrostrictive excitation. In the presence of a bias field, excitation occurred at $\nu_{\text{drive}} = \nu_{\text{fundamental}}$ corresponding to piezoelectric excitation. An effective (or induced) piezoelectric coefficient $d_{33}$ can be calculated for a bias polarization $P_b$ from the relation

$$d_{33} = \left[2Q_{11} - Q_{12}\left(\frac{2s_{12}^{D}}{s_{11}^{D} + s_{12}^{D}}\right)\right] P_b \epsilon_b$$

where $s_{ij}$ are the elastic compliances measured at constant electric displacement, and $Q_{ij}$ are the electrostrictive constants.

It was found possible to partially damp these acoustic resonances by two methods. In the first method, the sample and electrodes were imbedded in Armstrong epoxy but the faces through which the light passed were left unobstructed. In the second technique cold-worked aluminum was bonded to the electrodes. An example of the partially damped response is shown in Fig. 15. An interesting feature clearly illustrated in this figure is the initial primary or high-frequency electro-optic response followed by the secondary or elasto-optically induced response having a rise time characteristic of the acoustic travel time across the sample. The clamping effect on the induced birefringence can be estimated from the pulse data to be 25 ± 10 percent for the unbiased crystal. This is in agreement with calculations based on thermodynamic arguments.[3] Both clamping effects and acoustic damping have been neglected in the

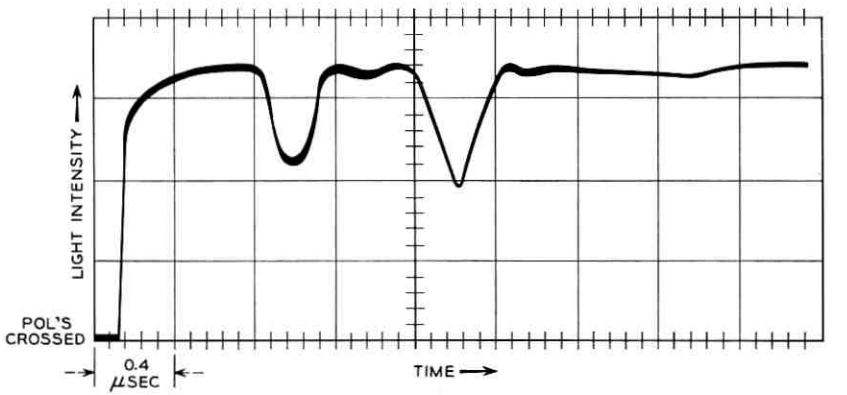



Fig. 15 — Partial damping of acoustic ringing by imbedding in epoxy.

previous sections. Clamping will raise the voltage and power requirements and thus further restrict the capacity. Acoustic damping may be difficult to achieve in practice on large samples without introducing strain. It may also have an adverse effect on the electrical Q in the frequency region near the acoustic resonances. Electrical Q's of several thousand (and up to 10,000) have been measured for KTN up to frequencies of several MHz on undamped samples.

Pulse response has also been measured in the nano-second range at 100 per cent modulation levels. A typical sampling scope trace is shown in Fig. 16. The modulation voltage pulse was delivered by a Huggins nanosecond pulse generator. The signal was detected in an EMI 7102 photomultiplier terminated in the 50 ohm input impedance of a Tektronix Model 660 Sampling scope. The sample was unbiased with a half-wave voltage about 25 percent larger than the measured dc value.*

## IX. CONCLUSIONS

An analysis of the modulator requirements for a high capacity-high speed digital light deflection system has been carried out. A principal result of this analysis is the derivation of a simple capacity-speed product

$$(R_x R_y)^{\frac{1}{2}} \nu_r \leqq \Lambda P_r$$

where $\Lambda$ is essentially a constant characteristic of a given modulator material, and $P_r$ is the reactive power with which the modulator is

* The triangular shape observed was identical to the input voltage waveform. The triangularity was due to current limitations, the driver being unable to provide the current necessary for a sharp leading and trailing edge.

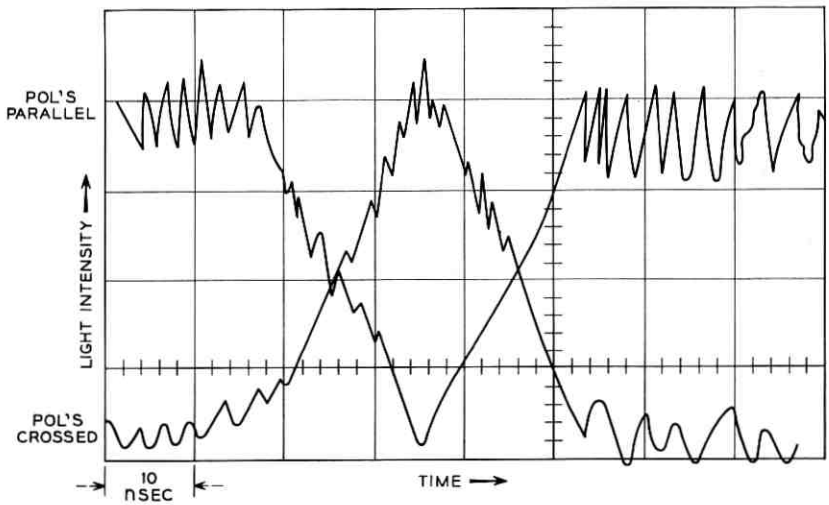



Fig. 16 — Clamped electro-optic pulse modulation.

driven. By examining the limitations on this reactive power one can then compare the performance of different modulator materials in terms of their respective capacity-speed products. For KTN the capacity-speed product is limited to $1.25 \times 10^9$ sec$^{-1}$ by the effects of internal heating, assuming ambient temperature fluctuations are held to less than 0.01°C, and compositional nonuniformities affecting the Curie temperature are less than 20 ppm. Presently available KTN has compositional nonuniformities 100 to 1000 times greater than this. In addition further complications arise from the dc bias in the form of space-charge effects. An ac bias scheme proposed by Warter can be used to eliminate space-charge effects. It is thus clear from the foregoing analysis that the material requirements imposed on the KTN to achieve a capacity-speed product of $10^9$ sec$^{-1}$ (i.e., $10^6$ addresses at a 1 MHz rate) are severe. It is therefore, important to look for other materials where the requirements might not be as severe.

The capacity-speed product is very helpful in such a search since it shows that a linear electro-optic material such as ZnTe with a dielectric constant of 10 and a reduced half-wave voltage of less than 3 kV, has a capacity-speed product of $2 \times 10^9$ sec$^{-1}$ for 10 watts reactive power. The capacity-speed products of all other known linear electro-optic materials are less than $2 \times 10^8$ sec$^{-1}$. On the basis of this study there are at present* only two electro-optic materials which are potentially capable

* Note added in proof: Recent work of Denton, Chen, and Ballman (to be published) on LiTa $O_3$, $x_{11} = 43$, $v_\pi = 2700$ $V$ indicates that this material also has a high-capacity speed product.

of operating in a high-capacity high-speed DLD, namely: KTN and ZnTe.

X. ACKNOWLEDGMENTS

The author would like to express his sincere thanks to K. D. Bowers and S. H. Wemple for many invaluable discussions as well as a critical reading of the manuscript. Appreciation is also extended to L. G. Van Uitert and W. B. Bonner for supplying the KTN samples, to M. C. Huffstuttler Jr., R. Curran and R. L. Barns for their results on compositional nonuniformities in KTN, to W. J. Tabor for his experimental results on the DLD system, and to R. T. Denton and E. H. Turner for helpful discussions on ZnTe.

## APPENDIX A

*Walk-Off Derivation*

We assume modules of constant length $l_m$ consisting of a polarization switch of length $l$ and a Wollaston prism of length $l_m - l$. The length of the Wollaston is assumed to be small compared to $l_m$ so that $l_m \sim l$. Starting from the input the modules are arranged in order of increasing angular deflection, in $x,y$ pairs; e.g., $\pm\theta_{x1'}$, $\pm\theta_{y1'}$, $\pm\theta_{x2'}$, $\pm\theta_{y2'}$, $\cdots$, etc. Consider the transverse displacement of the beam at the output of the $x1',y1'$ modules. In the $x'$ direction the beam is displaced by a distance $l_m\theta_{x1'}$ and in the $y'$ direction it is undisplaced in the present approximation. The beam at this same point projected onto the $x'$-$z$ plane ($z$ is the principal axis of the DLD) makes an angle $\theta_{x1'}$ with respect to the $z$ axis, and projected onto the $y'$-$z$ plane an angle $\theta_{y1'}$ with respect to the $z$ axis. At the output of $x2',y2'$ modules the corresponding displacements and angles are respectively $5l_m\theta_{x1'}\{= l_m\theta_{x1'} + (2l_m)\theta_{x1'} + l_m\theta_{x2'}\}$, $2l_m\theta_{y1'}\{= 0 + (2l_m)\theta_{y1}\}$, $3\theta_{x1}(= \theta_{x1'} + \theta_{x2'})$, and $3\theta_{y1'}(= \theta_{y1'} + \theta_{y2'})$. Table VI lists these displacements and the succeeding ones. In terms of the diffraction $\theta_D = \lambda/d_j$,

$$\Delta x' \cong \frac{3}{2}\beta_x l_m \frac{\lambda}{d_{x'}} 2^n \tag{68}$$

$$\Delta y' \cong \beta_y l_m \frac{\lambda}{d_{y'}} 2^n . \tag{69}$$

## APPENDIX B

*Derivation of Capacity-Speed Product for Rectangular Aperture*

Remembering that the modulator axes $x,y$ (parallel respectively to the $d$ and $b$ dimensions of the modulator) are rotated 45° with respect to

TABLE VI — LINEAR DISPLACEMENT OF BEAM AT OUTPUT OF $i$TH PAIR OF MODULES

| | $\Delta x$ | $\Delta y$ |
|---|---|---|
| $i = 1$ | $l_m\theta_{x1'}$ | 0 |
| 2 | $2l_m\theta_{x1'}(1 + \frac{3}{2})$ | $2l_m\theta_{y1'}$ |
| 3 | $2l_m\theta_{x1'}(1 + 3 + \frac{7}{2})$ | $2l_m\theta_{y1'}(1 + 3)$ |
| 4 | $2l_m\theta_{x1'}(1 + 3 + 7 + \frac{15}{2})$ | $2l_m\theta_{y1'}(1 + 3 + 7)$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $n$ | $2l_m\theta_{x1'}\left[\sum_{j=1}^{n-1}(2^j - 1) + \frac{1}{2}(2^n - 1)\right]$ | $2l_m\theta_{y1'}\sum_{j=1}^{n-1}(2^j - 1)$ |

performing the indicated sums

$$\Delta x = 2l_m\theta_{x1}(\tfrac{3}{2}\,2^n - n - \tfrac{3}{2})$$

$$\Delta y = 2l_m\theta_{y1}(2^n - n - 1).$$

For large $n$ ($n \gg 1$)

$$\Delta x \cong 2l_m\theta_{x1}2^n(\tfrac{3}{2})$$

$$\Delta y \cong 2l_m\theta_{y1}2^n.$$

the deflection axes $x'$,$y'$, and taking $b > d$, we see from Fig. 17(a) that a displacement of the beam $\Delta x'$ (or $\Delta y'$) produces a fractional loss of intensity

$$\frac{\Delta I}{I} = \frac{\Delta A}{A} \cong \frac{1}{\sqrt{2}}\left(\frac{1}{b} + \frac{1}{d}\right)\Delta x'(\text{or } \Delta y')$$

giving

$$\Delta x'(\text{or } \Delta y') \cong \sqrt{2}d\left(\frac{\Delta I}{I}\right). \tag{70}$$

The diffraction angles $\theta_D$ for the $x'$ and $y'$ deflections are given by $\theta_{Dx'} = \sqrt{2}\lambda/b$ and $\theta_{Dy'} = \lambda/\sqrt{2}d$ as shown in Fig. 17(b). Combining these expressions with (68) and (69) we obtain,

$$\Delta x' = \beta_x l_m\left(\frac{\sqrt{2}\lambda}{b}\right)R_{x'} \leqq 0.2\sqrt{2}d \tag{71}$$

$$\Delta y' = \beta_y l_m\left(\frac{\lambda}{\sqrt{2}d}\right)R_{y'} \leqq 0.2\sqrt{2}d \tag{72}$$

for a fractional intensity loss of 20 percent. Combining (71) and (72)

we find

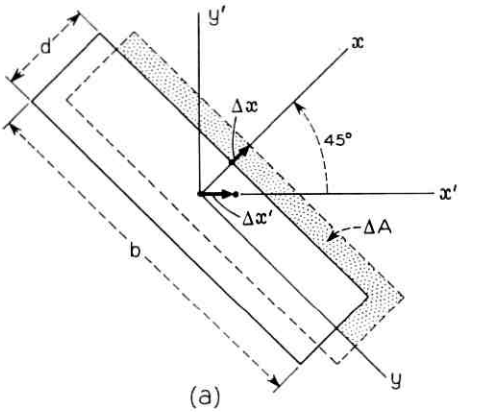
$$(R_{x'}R_{y'})^{\frac{1}{2}} \leqq \frac{0.2}{\sqrt{\beta_x\beta_y}\lambda}\left(\frac{bd}{l_m}\right)\sqrt{\frac{2d}{b}} \tag{73}$$

Substituting for $\mathcal{P}_r$ from (15) we obtain the desired result,

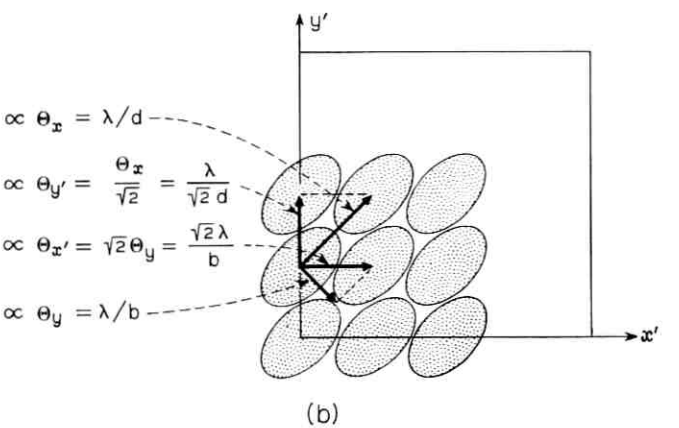
$$(R_{x'}R_{y'})^{\frac{1}{2}}\nu \leqq \Lambda_{x'x'}\mathcal{P}_r \tag{74}$$

where

$$\Lambda_{x'y'} = \frac{1}{5\alpha\sqrt{\beta_x\beta_y}\lambda\left[1+\frac{l_m-l}{l_m}\right]}\sqrt{\frac{2d}{b}} \ . \tag{75}$$




(a)




(b)

Fig. 17 — Rectangular aperture: (a) beam displacement, (b) image plane.

APPENDIX C

*Dielectric Properties of KTN*

From the Devonshire free energy expansion in (25) we can derive expressions for several useful dielectric properties. The spontaneous polarization $P_s$ is given by

$$P_s = \sqrt{\frac{2}{3}} P_{so} \left[ 1 + \left\{ 1 + \frac{3}{4} \left( \frac{T_o - T}{T_c - T_o} \right) \right\}^{\frac{1}{2}} \right]^{\frac{1}{2}} \quad T < T_o \tag{76}$$

where $P_{so}$ is the spontaneous polarization at the phase transition $T_c$, and is given by

$$P_{so} = \sqrt{\frac{-3\xi}{4\zeta}}. \tag{77}$$

The difference between the Curie Weiss temperature $T_o$ and phase transition temperature $T_c$ in the Devonshire model is,

$$T_c - T_o = \frac{3}{16} \epsilon_o C \left( \frac{\xi^2}{\zeta} \right). \tag{78}$$

The low field dielectric permittivity $\epsilon$ above the phase transition obeys a Curie-Weiss law,

$$\epsilon = \frac{C\epsilon_o}{T - T_o} \tag{79}$$

$$\lim_{\delta \to 0} \epsilon = \frac{C\epsilon_o}{T_c - T_o} \tag{80}$$

$$\delta \geqq 0$$

$$T = T_c + \delta$$

$$\lim_{\delta \to 0} \epsilon = \frac{C\epsilon_o}{4(T_c - T_o)} \tag{81}$$

$$\delta = \leqq 0$$

$$T = T_c + \delta .$$

More generally below $T_o$, the permittivity can be written

$$\epsilon_{T<T_c} = \frac{\epsilon_o C}{(T_c - T_o)\left[ -\frac{4}{3} + \frac{16}{3}\gamma + \frac{20}{3}\gamma^2 \right] - (T_o - T)}, \tag{82}$$

where

$$\gamma \equiv \sqrt{1 + \frac{3}{4}\left(\frac{T_o - T}{T_c - T_o}\right)}. \tag{83}$$

The small signal permittivity at high fields is given by,

$$\epsilon_b \cong \frac{\epsilon}{1 + 3\epsilon\xi P_b^2 + 5\epsilon\zeta P_b^4}. \tag{84}$$

APPENDIX D

*Power Considerations*

Consider the circuit shown in Fig. 18. The current in the RF driver arm is given by

$$\begin{aligned} i_1 &= \frac{j\omega C_s V_{rf}}{\frac{C_s}{C_B} + j\omega C_s R_g + \left\{\frac{j\omega R_s C_s}{1 + j\omega C_s R_s + (R_s C_s / L_B)}\right\}} \\ &\cong \frac{\omega C_s V_{rf}(1 + jQ_s)}{Q_s} \end{aligned} \tag{85}$$

where $Q_s = \omega C_s R_s$ and the following inequalities are satisfied,

$$R_s, \omega L_B \gg \frac{1}{\omega C_s} \gg R_g, \frac{1}{\omega C_B}.$$

The instantaneous power delivered by the RF driver is,

$$\mathcal{P}_{\text{inst}}(\text{driver}) = \frac{\omega C_s V_{rf}^2}{2Q_s} + \frac{\omega C_s V_{rf}^2}{2Q_s}\cos 2\omega t - \frac{\omega C_s V_{rf}^2}{2}\sin 2\omega t. \tag{86}$$

In the same manner, we can obtain the instantaneous power delivered by the circuit to the modulator,

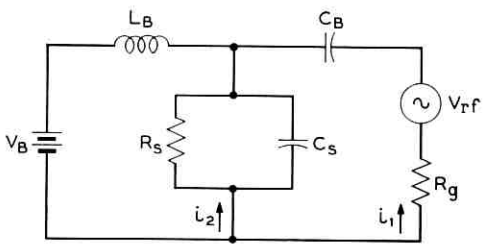



Fig. 18 — Equivalent circuit for biased operation.

$$\mathcal{P}_{\text{inst}}(\text{mod}) = \frac{V_B^2}{R_s} + \mathcal{P}_{\text{inst}}(\text{driver}) + \frac{2V_B V_{rf}}{R_s} \cdot \left(\cos \omega t - \frac{Q_s}{2} \sin \omega t\right). \tag{87}$$

The term $V_B^2/R_s$ is supplied by the bias supply. It can be readily shown that the last term is supplied by the dc blocking capacitor and RF blocking inductor. The dc blocking capacitor contributes a term $(V_B V_{rf}/R_s)$ $(\cos \omega t - Q \sin \omega t)$ and the RF blocking inductor a term $(V_B V_{rf}/R_s) \cos \omega t$. Since $Q_s$ is large the instantaneous power delivered by the driver is to a good approximation given by,

$$\mathcal{P}_{\text{inst}}(\text{driver}) \cong \frac{\omega C_s V_{rf}^2}{2} \sin 2\omega t \tag{88}$$

the dissipated power in the modulator $(\langle \mathcal{P}_{\text{inst}(\text{mod})} \rangle_{\text{time}})$ from (87) is

$$\mathcal{P}_d = \frac{\omega C_s V_{rf}^2}{2Q_s} + \frac{V_B^2}{R_s}. \tag{89}$$

Since the actual modulation waveform is nonsinusoidal, the analysis must be extended to include a pulse-type waveform. In addition, since we are driving a reactive load we need to characterize the driver power requirements. The time average of the reactive power is zero. The peak instantaneous power depends on the risetime of the voltage pulse. Neither of these factors is a satisfactory measure of the required driver power. A more realistic quantity from the point of view of the driver[22] is the energy delivered to the modulator per address pulse times the number of address pulses/second. From the foregoing analysis, (88), the power defined in this fashion is given by

$$\mathcal{P}_r \equiv (\tfrac{1}{2} C V_{rf}^2) \nu_r, \tag{90}$$

where $\nu_r = 2\nu$ (assuming that each half cycle represents an addressing pulse). The dissipated power for the case of pulse modulation can be obtained from a simple extension of the foregoing analysis. Expanding the pulse waveform in a Fourier series,

$$v = \frac{4}{\pi} V_{rf} \sum_{m=1,3,5,\cdots} \frac{1}{m} \sin m\omega t$$

it is easily shown[23] that the power dissipated in the sample is given by

$$\mathcal{P}_d = \frac{16}{\pi^2} \left(\frac{\omega C_s V_{rf}^2}{2Q_s(\omega)}\right) \sum_{m=1,3,\cdots} \left(\frac{Q_s(\omega)}{mQ_s(m\omega)}\right). \tag{91}$$

If the quality factor $Q_s$ is independent of frequency and the series is terminated at $m = 5$, (91) reduces to (29) of the text. The rise time of the pulse is approximately the rise time[24] associated with the highest frequency component $m_{\max}\omega$

$$\tau_{\text{rise}} \cong \frac{0.45}{m_{\max}\,\nu}. \tag{92}$$

It should be noted that the peak reactive power varies inversely with $\tau_{\text{rise}}$, and can be substantially larger than the reactive power defined in (84) if the rise time is much shorter than the period of an address pulse.

The pulse amplitude has been equated to the half-wave voltage in the linear electro-optic modulator and to the incremental half-wave voltage in the biased quadratic modulator. This implies that an addressing pulse is needed for only one of the two polarization states of the light beam (i.e., either the "0" or "1" state but not both).

An alternative scheme is to set the ambient state of the modulator to a point midway between the two desired states ($\frac{1}{4}$ wave bias) in which case *each* state requires an address pulse of one half the previous amplitude. The average power required is thus cut by a factor of two. We have not used this latter scheme since in the case of KTN it introduces a possible dc component into the voltage across the modulator. In the case of the linear electro-optic materials the presently conceived[20] driver circuitry does not permit any increase in the available power using this bias scheme.

REFERENCES

1. Nelson, T. J., Digital Light Deflection, B.S.T.J., *43*, 1964, p. 821.
2. Tabor, W. J., Use of Wollaston Prisms for a High-Capacity Digital Light Deflector, B.S.T.J., *43*, 1964, p. 1153.
3. Chen, F. S., Geusic, J. E., Kurtz, S. K., Skinner, J. G., and Wemple, S. H., J. Apply. Phys., *37*, 1966, p. 388.
4. Nye, J. F., *Physical Properties of Crystals*, Oxford Univ. Press, 1960.
5. *American Institute of Physics Handbook*, McGraw-Hill Book Co. Inc., New York, 1963, Sect. 6, p. 188.
6. Sliker, T. R. and Jost, J. M., Appl. Phys. Lett., J. Opt. Soc. Am. *56*, 130 (1966).
7. Peterson, G. D., Ballman, A. A., Lenzo, P. V., and Bridenbaugh, P. M., Appl. Phys. Lett., *5*, 1964, p. 234.
8. Devonshire, A. F., Phil. Mag., *40*, 1949, p. 1040; *42*, 1951, p. 1065.
9. Tabor, W. J., A High-Capacity Digital Light Deflector Using Wollaston Prisms, B.S.T.J., (to be published).
10. Bonner, W. A., Dearborn, E. F., and Van Uitert, L. G., Bull. Am. Ceramic Soc., January, 1965.
11. Huffstuttler, M. C., Jr., Curran, R., and Barns, R. L., private communication.
12. Warter, P. J., Jr., private communication.
13. von Hippel, A., Gross, E. P., Jelatis, J. G., and Geller, M., Phys. Rev., *91*, 1953, p. 568.; Kurtz, S. K. and Warter, P. J., Jr. Bull. Am. Phys. Soc, *11*, p. 34(A) (1966).
14. Macdonald, J. R., J. Chem. Phys., *29*, 1958, p. 1346.
15. Rose, A. *Concepts in Photoconductivity and Allied Problems*, Interscience Pub., New York, 1963.

16. Kaminow, I. P., Appl. Phys. Lett., *7*, 1965, p. 123.
17. Johnson, A. R., Appl. Phys. Lett., *7*, 1965, p. 195.
18. Chen, F. S., private communication.
19. Ashkin, A., Boyd, G. D., Dziedzic, J. M., Smith, R. G., Ballman, A. A., Levinstein, J. J., and Nassau, K., Appl. Phys. Letters *9*, 1966, p. 72.
20. Turner, E. H., Appl. Phys. Lett., *8*, June, 1966; Lenzo, P. V., Spencer, E. G., and Nassau, K., J. Opt. Soc. Amer., *56*, 1966, p. 633.
21. Mason, W. P., *Piezoelectric Crystals and Their Applications to Ultrasonics*, D. Van Nostrand Co., New York, 1950, Chap. III, p. 61 and Chap. XII, p. 303.
22. Petersen, R. C., private communication.
23. Lawrence, R. R., *Principles of Alternating Currents*, McGraw-Hill Book Co. Inc., New York, 1935, Chap. III, p. 94.
24. Lewis, I. A. D. and Well, F. H., *Millimicrosecond Pulse Techniques*, Pergamon Press, 2nd Ed., London, 1959, Chap. I, p. 7.

# On the Optimality of the Regular Simplex Code

By H. J. LANDAU and DAVID SLEPIAN

(Manuscript received May 25, 1966)

*We prove here the long conjectured fact that the regular simplex is the code of minimal error probability for transmission over the infinite-band Gaussian channel. The code is actually optimal for a rather wide class of assumed channel noises. We also establish the optimality of several other codes for the band-limited Gaussian channel.*

## I. INTRODUCTION

Since its introduction by Shannon[1] and Kotel'nikov[2] nearly 20 years ago, the geometric representation of signals has played an important role in communication theory.* By this scheme, a variety of physically different time-continuous communication systems can all be reduced to the same geometric model. The problem of finding optimal signals for transmission then becomes a geometric one. This paper solves one such problem.

In the model in question, signals to be transmitted are represented as points, or vectors from the origin, in a suitable finite dimensional Euclidean signal space $\mathcal{E}_n$ . The energy of any signal in $\mathcal{E}_n$ is proportional to the length of its representative vector; the bandwidth of the communication system is proportional to the dimension $n$ of the signal space. Received signals are also represented by vectors in $\mathcal{E}_n$ and the difference $\mathbf{Z} = \mathbf{Y} - \mathbf{X}$ between a transmitted signal $\mathbf{X}$ and the corresponding received signal $\mathbf{Y}$ is a vector random variable representative of the noise encountered during transmission. In a model commonly considered, the probability density of $\mathbf{Z}$ depends only on its magnitude, i.e.,

$$p(z_1, z_2, \cdots, z_n) = f(|\mathbf{Z}|), \tag{1}$$

* A detailed description of this viewpoint along with some references to the intervening literature can be found in Chapters 4 and 5 of the recent book[3] by Wozencraft and Jacobs.

and $f(\cdot)$ is an integrable nonnegative monotone decreasing function of its argument. We shall consider only this case in all that follows.

Suppose the transmitter has a list of $M$ signals, $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_M$ from which it selects the successive signals to be transmitted. We suppose these choices are made independently with equal probabilities and that the code, or list of possible sent signals, is known to the receiver. The receiver partitions $\mathcal{E}_n$ into $M$ disjoint regions $\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_M$ called decision regions. When the received signal lies in $\mathcal{R}_i$, the receiver asserts that $\mathbf{X}_i$ was transmitted. With this scheme, the probability of correct decoding is

$$Q = \frac{1}{M} \sum_{i=1}^{M} \int_{\mathcal{R}_i} f(|\mathbf{X} - \mathbf{X}_i|)\, dx_1\, dx_2 \cdots dx_n, \tag{2}$$

where $\mathbf{X} = (x_1, x_2, \cdots, x_n)$ is a generic point in $\mathcal{E}_n$.

With $M$ and $n$ given, how large can $Q$ be made by proper choice of the code and decision regions? For a given code it is well known (see Ref. 3, Section 4.2, for example) that $Q$ is maximized by choosing

$$\mathcal{R}_i = \{\mathbf{X} \,|\, |\mathbf{X} - \mathbf{X}_i| < |\mathbf{X} - \mathbf{X}_j|, \quad j \neq i\}, \tag{3}$$

$i = 1, 2, \cdots, M$. That is, the $i$th decision region consists of all points of $\mathcal{E}_n$ closer to $\mathbf{X}_i$ than to any other code word. Decision regions determined by (3) are known as maximum-likelihood regions.

The maximization of $Q$ over the code is more complicated. To obtain a meaningful problem it is necessary to put some restriction on the length of the code vectors, for without this, $Q$ can be made arbitrarily close to unity by choosing large enough vectors in distinct directions. Several different energy restrictions have been studied in the literature (see Ref. 4). Although optimal codes under these restrictions have not been found in general for fixed $M$ and $n$, much detail is known in the Gaussian case

$$f(x) = \frac{\exp(-x^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}}, \tag{4}$$

about the asymptotic form of $Q$ for such optimal codes, as $n \to \infty$ with $(1/n) \log M \to R$. These results are usually described in the channel capacity and reliability formulae terms of information theory.[3,4,5]

In this paper we restrict our attention to the case in which all code vectors are the same length. For convenience, we take

$$|\mathbf{X}_i| = 1, \qquad i = 1, 2, \cdots, M. \tag{5}$$

Such codes are called "equal energy codes".[6] The code optimization problem can then be stated as follows. Find $M$ points $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \cdots, \hat{\mathbf{X}}_M$

on the unit sphere* in $\mathcal{E}_n$ such that $Q$ as given by (2) and (3) attains its maximum value.

To our knowledge, the first investigation of particular codes from this geometric point of view was carried out in 1948 by L. A. MacColl[7] who investigated codes corresponding to the vertices of the three regular polytopes[8] in $n$-space. These are the regular simplex for which $M = n + 1$, the hypercube for which $M = 2^n$ and the cross-polytope or biorthogonal code for which $M = 2n$. MacColl wrote explicit expressions for $Q$ for these codes and evaluated them numerically for the Gaussian case (4) for a variety of values of $n$ and $\sigma$. Gilbert[9] continued this work and made comparisons with a variety of other point configurations. Balakrishnan[10] established a new expression for $Q$ in the Gaussian case, which permitted him to show that the regular simplex code is locally optimal (yields a larger $Q$ than nearby equal energy codes with $M = n + 1$). Later [11] he showed that as $\sigma \to \infty$ and as $\sigma \to 0$ the optimal code of $n + 1$ points approached the regular simplex. Weber[12] used Balakrishnan's form for $Q$ to show that for $n = 2$ the (globally) optimal code of $M$ points, $M = 3, 4, \cdots$, is the regular $M$-gon. For $n = 2, 3, \cdots$, he also showed the biorthogonal code to be a local optimum among equal energy codes with $M = 2n$, and described a family of locally optimal codes for $M = n + 1, n + 2, \cdots, 2n$.

In this paper, we at last lay to rest the longstanding conjecture that the regular simplex is optimal for $M = n + 1$ in the Gaussian case.† Specifically, we show that $Q$ as given by (2)–(3) is greater for the regular simplex than for any other equal energy code of $M = n + 1$ points in $\mathcal{E}_n$, $n = 3, 4, 5, \cdots$. This result is true for any monotone decreasing $f$. The method of proof is based on a generalization to higher dimensions of a theorem of Fejes-Tóth[13] concerning expressions related to the form (2) in 3 dimensions.‡ Our methods also establish that the optimal equal energy codes with parameters $M = 6$, $n = 3$, and $M = 12$, $n = 3$ are, respectively, the biorthogonal code and the code consisting of the midpoints of the faces of the regular dodecahedron. We conclude with some comments about the biorthogonal code and about the reliability of the infinite-band Gaussian channel.

## II. AN INEQUALITY FOR Q

For an equal energy code, the maximum-likelihood region $\mathcal{R}_i$ given by (3) can be determined as follows. Let $\mathcal{H}_{ij}$ denote the hyperplane that

* We shall hereafter use the caret ^ to denote unit vectors.

† It is incorrectly stated in Ref. 3, pp. 260, 364 that this result has been previously shown in the literature.

‡ We are indebted to E. N. Gilbert for calling Fejes-Tóth's work to our attention.

bisects perpendicularly the line segment joining $\hat{\mathbf{X}}_i$ to $\hat{\mathbf{X}}_j$. This plane passes through the origin and divides $\mathcal{E}_n$ into two half-spaces. We denote by $\mathcal{U}_{ij}$ the half-space containing $\hat{\mathbf{X}}_i$. It consists of all points of $\mathcal{E}_n$ closer to $\hat{\mathbf{X}}_i$ than to $\hat{\mathbf{X}}_j$. The region $\mathcal{R}_i$ is the intersection of $M$-1 such half-spaces,

$$\mathcal{R}_i = \bigcap_{\substack{j=1 \\ j \neq i}}^{M} \mathcal{U}_{ij}.$$

It is, therefore, a convex region bounded by a certain number of hyperplane faces that pass through the origin — a kind of flat-sided cone with vertex at the origin. We note that the various maximum-likelihood regions, $\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_M$, are disjoint and that together with their boundaries they exhaust $\mathcal{E}_n$.

Let us now call any convex region of $\mathcal{E}_n$ bounded by $k \geqq n$ hyperplanes through the origin a "flat-sided cone". We shall establish an upper bound for $Q$ as given by (2) when the $M$ decision regions $\mathcal{R}_i$ are any set of disjoint flat-sided cones (not necessarily maximum-likelihood regions of any code) that together with their boundaries exhaust $\mathcal{E}_n$. For our purposes it suffices to consider only the case in which $\hat{\mathbf{X}}_i$ lies in the interior of $\mathcal{R}_i$, $i = 1, 2, \cdots, M$.

We denote by $S$ the surface of the unit sphere in $\mathcal{E}_n$ with center at the origin. We denote by $R_i$ the intersection of $\mathcal{R}_i$ with $S$. The regions $R_i$ are "spherical polygons" that reticulate $S$ into a map or net. We shall evaluate $Q$ by first integrating over this net on $S$ and by then performing a radial integration.

Let $\mathbf{X}$ be a generic point in $\mathcal{E}_n$ distant $r$ from the origin (see Fig. 1)

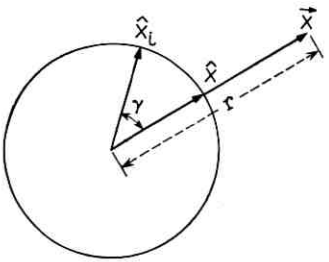



Fig. 1 — Reduction to unit sphere.

and let $\hat{\mathbf{X}}$ be a unit vector in the direction of $\mathbf{X}$, i.e., the terminus of $\hat{\mathbf{X}}$ is the radial projection of the generic point onto $S$. Then

$$|\hat{\mathbf{X}}_i - \mathbf{X}|^2 = 1 + r^2 - 2r \cos \gamma$$

and

$$|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|^2 = 2 - 2\cos\gamma$$

so that

$$|\hat{\mathbf{X}}_i - \mathbf{X}|^2 = (1 - r)^2 + r|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|^2.$$

We can thus write $f(|\hat{\mathbf{X}}_i - \mathbf{X}|) = g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|)$ where for each fixed $r$ the function $g_r(\cdot)$ is nonnegative and is monotone decreasing in its argument. The expression (2) in these terms becomes

$$Q = \int_0^\infty dr\, r^{n-1} U(r) \tag{6}$$

$$U(r) = \frac{1}{M}\sum_{i=1}^{M} \int_{R_i} g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|)ds \tag{7}$$

where $ds$ is the differential surface- or $(n-1)$-content of $S$ at the point $\hat{\mathbf{X}}$. Note that $U \geqq 0$. We proceed to find an upper bound for $U$. By (6) this will provide the desired bound for $Q$.

Let the terminus of the unit vector $\hat{\mathbf{Y}}$ determine a point $P$ on $S$ (see Fig. 2). The set of all points $\hat{\mathbf{X}}$ on $S$ such $\hat{\mathbf{X}}\cdot\hat{\mathbf{Y}} \geqq \cos\varphi \geqq 0$ will be called "the spherical cap of $S$ of angle $\varphi$ about $P$". Now let $\mathfrak{IC}$ be a hyperplane through the origin but not containing $P$ that intersects this spherical cap. That is $0 < \hat{\mathfrak{n}}\cdot\hat{\mathbf{Y}} < \sin\varphi$ where $\hat{\mathfrak{n}}$ is the unit normal to $\mathfrak{IC}$ directed positively toward the side on which $P$ lies. $\mathfrak{IC}$ divides the spherical cap into two parts. We denote by $W$ the part of the cap not containing $P$, and we denote by $w$ the content of $W$.

In what follows, the function

$$h_r(w) = \int_W g_r(|\hat{\mathbf{Y}} - \hat{\mathbf{X}}|)ds \tag{8}$$

will be of great importance to us. The notation suppresses the dependence of $h$ on $\varphi$, the angle of the spherical cap, and points out that with the

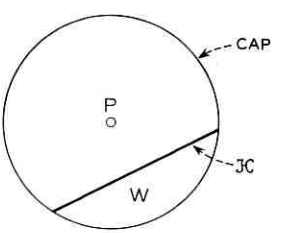



Fig. 2 — Cap cut by hyperplane.

geometry as described this integral is a function only of

$$w = \int_W ds \tag{9}$$

the content of $W$. We shall suppose $\varphi$ fixed in all that follows.

Two special properties of $h_r(w)$ are of particular concern. First, as shown in Appendix B, this function is increasing and convex. That is, if $w_2 > w_1$, then $h_r'(w_2) > h_r'(w_1)$ where $h_r'(w) = dh_r/dw$. This, of course, implies that

$$\sum p_i h_r(w_i) \geqq h_r(\sum p_i w_i), \tag{10}$$

where the $p_i$ are nonnegative weight factors summing to unity. Equality holds only when the $w_i$ are all equal.

The second property of $h_r(w)$ is somewhat more complicated to state, though in three dimensions it is intuitively obvious. Again let $\mathcal{H}$ be a hyperplane through the origin but not through $P$ that cuts off a piece $W$ of the spherical cap about $P$. Let $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_j$ be hyperplanes that each contain the origin and $P$. We denote by $V$ the portion of $W$ lying on the positive side of $\mathcal{H}_i$, $i = 1, 2, \cdots, j$, and we denote the content of $V$ by $v$. It is established in Appendix C that

$$\int_V g_r(|\hat{\mathbf{Y}} - \hat{\mathbf{X}}|)ds \geqq h_r(v), \tag{11}$$

where as before $\hat{\mathbf{Y}}$ is the vector from the origin to $P$ and $\hat{\mathbf{X}}$ is a generic point of $V$. Equality holds only if $\mathcal{H}$ is the sole hyperplane boundary of $V$ (i.e., if none of $\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_j$ form a part of the boundary of $V$).

With these two properties of $h_r(w)$ we can now establish the desired inequality for $U(r)$. We first "triangulate" each of the polygonal regions $R_i$ into "spherical pyramids" $R_{ij}$ having boundaries of $R_i$ as bases and $\hat{\mathbf{X}}_i$ as a vertex. More accurately described, the regions $R_{ij}$ are found as follows. The flat-sided cone $\mathcal{R}_i$ is bounded by pieces of $k_i$ (say) hyperplanes $\mathcal{H}_1^{(i)}, \cdots, \mathcal{H}_{k_i}^{(i)}$ through the origin. We denote by $\mathcal{B}_{ij}$ the portion of $\mathcal{H}_j^{(i)}$ that bounds $\mathcal{R}_i$. Now $\mathcal{B}_{ij}$ is itself bounded by a certain number $l_{ij}$ of $(n-2)$-flats through the origin. Through each of these $(n-2)$-flats we pass a hyperplane $\mathcal{H}_k^{(ij)}$, $k = 1, 2, \cdots, l_{ij}$, that contains $\hat{\mathbf{X}}_i$. These hyperplanes, along with $\mathcal{B}_{ij}$, determine a new flat-sided cone $\mathcal{R}_{ij}$ having $\mathcal{B}_{ij}$ as one face and the line containing $\hat{\mathbf{X}}_i$ as a one-dimensional boundary. The interiors of the $k_i$ flat-sided cones $\mathcal{R}_{i1}, \mathcal{R}_{i2}, \cdots, \mathcal{R}_{ik_i}$ are disjoint. Together with their boundaries they exhaust $\mathcal{R}_i$. The line through $\hat{\mathbf{X}}_i$ is common to the boundaries of all $k_i$ of these flat-sided cones. The spherical pyramid $R_{ij}$ is the intersection of $\mathcal{R}_{ij}$ with $S$.

We denote by $C_i$ the spherical cap of $S$ of angle $\varphi$ about $\hat{\mathbf{X}}_i$ (see Fig. 3).

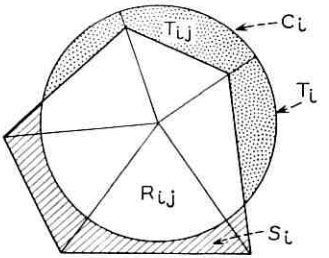



Fig. 3 — Cap and triangulated decision region.

Let $T_i$ be the portion of $C_i$ exterior to $R_i$ and let $S_i$ be the portion of $R_i$ exterior to $C_i$. A typical term of the sum (7) can then be written

$$\int_{R_i} g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|)ds = \left(\int_{C_i} + \int_{S_i} - \int_{T_i}\right) g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|)ds.$$

Now $T_i$ can be broken up into pieces $T_{ij}$ corresponding to the spherical pyramids $R_{ij}$. To accomplish this, we extend the sides of the pyramid beyond its base. Thus, if $\mathcal{R}_{ij}$ is on the positive side of $\mathcal{H}_j^{(i)}$ and $\mathcal{H}_k^{(ij)}$, $k = 1, 2, \cdots, l_{ij}$, then $T_{ij}$ is the part of the spherical cap on the negative side of $\mathcal{H}_j^{(i)}$ and the positive side of $\mathcal{H}_k^{(ij)}$, $k = 1, 2, \cdots, l_{ij}$. We now have

$$\int_{R_i} g_r ds = \int_{C_i} g_r ds + \int_{S_i} g_r ds - \sum_{j=1}^{k_i} \int_{T_{ij}} g_r ds. \tag{12}$$

Some of the regions $S_i$, $T_{ij}$ can, of course, be void.

We now sum (12) over the $M$ regions $R_i$. We write

$$k = \frac{1}{2}\sum_{i=1}^{M} k_i$$

for the total number of $(n-2)$-boundaries in the net on $S$. (Each boundary of $R_i$ is shared with one other spherical polygon.) There results

$$\begin{aligned} MU(r) &= \sum_{i=1}^{M} \int_{R_i} g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|)ds = M \int_{C_1} g_r(|\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}|)ds \\ &+ \sum_{i=1}^{M} \int_{S_i} g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|)ds - \sum_{i=1}^{M}\sum_{j=1}^{k_i} \int_{T_{ij}} g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}})ds. \end{aligned} \tag{13}$$

We next use (11) for the regions $T_{ij}$.

$$MU(r) \leqq M \int_{C_1} g_r(|\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}|)ds + \sum_{i=1}^{M} \int_{S_i} g_r ds - \sum_{i=1}^{M}\sum_{j=1}^{k_i} h_r(t_{ij})$$

where $t_{ij}$ is the content of $T_{ij}$. The convexity (10) of $h$ now gives

$$MU(r) \leqq M \int_{C_1} g_r(|\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}|)ds + \sum_{i=1}^{M} \int_{S_i} g_r ds - 2kh_r \left( \frac{1}{2k} \sum_{i=1}^{M} \sum_{j=1}^{k_i} t_{ij} \right). \tag{14}$$

Now denote by $s$, $c$, and $s_i$, respectively, the content of $S$, $C_i$, and $S_i$. A division of $S$ analogous to (13) (set $g = 1$ there) gives

$$\begin{aligned} s &= Mc + \sum_{1}^{M} s_i - \sum_{i=1}^{M} \sum_{j=1}^{k_i} t_{ij} \\ &= Mc + s' - \sum_{i=1}^{M} \sum_{j=1}^{k_i} t_{ij} \end{aligned} \tag{15}$$

where we write $s' = \sum s_i$ for the sum of the contents of all of the pieces of the polygons $R_i$ that fall outside their respective spherical caps. We then have from (14) and (15)

$$\begin{aligned} MU(r) &\leqq M \int_{C_1} g_r ds + \sum_{i=1}^{M} \int_{S_i} g_r ds - 2kh_r \left( \frac{Mc - s + s'}{2k} \right) \\ &= M \int_{C_1} g_r ds - 2kh_r \left( \frac{Mc - s}{2k} \right) \\ &\quad + \sum_{i=1}^{M} \int_{S_i} g_r ds - 2k \int_{K} g_r(|\hat{\mathbf{Y}} - \hat{\mathbf{X}}|)ds \end{aligned} \tag{16}$$

where $K$ is the region (see Fig. 4) of the spherical cap about $P$ that lies between hyperplanes through the origin that cut from the cap regions of content $(Mc - s + s')/2k$ and $(Mc - s)/2k$. This latter quantity will henceforth be assumed to be nonnegative. The normals to the two hyperplanes and the vector $\hat{\mathbf{Y}}$ from the origin to $P$ are chosen coplanar.

Note now that the sum of the last two terms in (16) cannot be positive, for we have

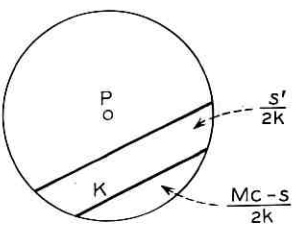



Fig. 4 — The region $K$.

$$\sum_{i=1}^{M} \int_{S_i} g_r(|\hat{\mathbf{X}}_i - \hat{\mathbf{X}}|)ds \leqq g_r(d) \sum_{i=1}^{M} \int_{S_i} dr = g_r(d)s' \tag{17}$$

and

$$2k \int_K g_r(|\hat{\mathbf{Y}} - \hat{\mathbf{X}}|)ds \geqq g_r(d)2k \int_K ds = g_r(d)s' \tag{18}$$

where $d = \sqrt{2 - 2\cos\varphi}$ is the distance from the center of the cap to the edge. Here we have used the fact that $g_r$ is monotone decreasing. Equality holds in (17) and (18) only when $s' = 0$.

In view of the above, the inequality in (16) can be continued by omitting the last two terms there and we then obtain our desired inequality

$$MU(r) \leqq M \int_C g_r(|\hat{\mathbf{Y}} - \hat{\mathbf{X}}|)ds - 2kh_r\left(\frac{Mc - s}{2k}\right) \tag{19}$$

where $C$ is the spherical cap of angle $\varphi$ about the terminus of $\hat{\mathbf{Y}}$ and we require $Mc - s \geqq 0$. Retracing all the inequalities used to derive (19), we see that the equality sign holds there if and only if $s' = 0$, all the regions $T_{ij}$ have equal content and each $T_{ij}$ is a region cut off from the spherical cap by a single hyperplane.

In closing this section we note one further fact. From the convexity of $h_r(x)$ it follows that $xh_r(\alpha/x)$ is monotone decreasing in $x$. For given $M$ and $c$, then, the right side of (19) is monotone increasing in $k$.

## III. OPTIMALITY OF THE REGULAR SIMPLEX AND CERTAIN OTHER CODES

Let a code of $M$ unit vectors in $\mathcal{E}_n$ have maximum-likelihood regions $\mathcal{R}_i$ that reticulate the surface of the unit sphere into a net having $k$ $(n-2)$-dimensional boundaries. We designate such a code by the symbol $\{n,M,k\}$. For certain values of the parameters $n$, $M$ and $k$, there may exist codes for which a spherical cap angle $\varphi$ can be found such that the conditions for equality hold in (19). We call such a code a *symmetric* $\{n,M,k\}$. By choosing $C$ so that (19) is an equality for such a symmetric code, we see that the probability $Q$ of no error for a symmetric $\{n,M,k\}$ is greater than the no-error probability of any nonsymmetric $\{n,M,k\}$. Indeed, the concluding remark of Section II shows that the no-error probability of a symmetric $\{n,M,k\}$ is greater than the no-error probability of every $\{n,M,k'\}$ if $k' < k$.

The regular simplex code consists of $M = n + 1$ unit vectors $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \cdots, \hat{\mathbf{X}}_{n+1}$ in $\mathcal{E}_n$ with $\hat{\mathbf{X}}_i \cdot \hat{\mathbf{X}}_j = -(1/n)$, $i \neq j$. The maximum-likelihood region $\mathcal{R}_i$ containing $\hat{\mathbf{X}}_i$ is bounded by $n$ hyperplanes. It is

readily verified that the regular simplex is a symmetric $\{n, n+1, n(n+1)/2\}$. Now no code of $n+1$ unit vectors in $\mathcal{E}_n$ can have more than $k = n(n+1)/2$ $(n\text{-}2)$-boundaries in its maximum-likelihood net, for, by the construction given in the first paragraph of Section II, each maximum-likelihood region can be bounded by at most $n$ hyperplanes. The regular simplex code then must have a $Q$ strictly greater than any other equal energy code of $n+1$ vectors in $\mathcal{E}_n$ except possibly another distinct symmetric $\{n, n+1, n(n+1)/2\}$, should such exist. But this latter eventuality cannot happen. That a symmetric $\{n, n+1, n(n+1)/2\}$ must be the regular simplex can be seen as follows. Since no $\mathcal{R}_i$ for a code of $M = n+1$ points can have more than $n$ hyperplane boundaries, then to have $k = n(n+1)/2$, every $\mathcal{R}_i$ must have exactly $n$ hyperplane boundaries. The $n$ hyperplanes $\mathcal{H}_1^{(1)}, \mathcal{H}_2^{(1)}, \cdots, \mathcal{H}_n^{(1)}$ that bound $\mathcal{R}_1$ bisect, respectively, the line segments from $\hat{\mathbf{X}}_1$ to $\hat{\mathbf{X}}_2$, from $\hat{\mathbf{X}}_1$ to $\hat{\mathbf{X}}_3$, $\cdots$, from $\hat{\mathbf{X}}_1$ to $\hat{\mathbf{X}}_{n+1}$. Since the code is assumed symmetric, these hyperplanes must be equidistant from $\hat{\mathbf{X}}_1$. Thus, all the other code points are equidistant from $\hat{\mathbf{X}}_1$. But a similar argument holds for each of the other regions $\mathcal{R}_2, \mathcal{R}_3, \cdots, \mathcal{R}_{n+1}$ and so all distances between pairs of code points are equal. But this property suffices to define the regular simplex.

The optimality of two other codes in $n = 3$ dimensions can readily be established by using (19). We note first that in 3 dimensions the conditions for equality to hold in (19) are such that the maximum-likelihood net on the sphere must be composed of congruent regular spherical polygons. A symmetric $\{3,M,k\}$ then must be the radial projection onto the unit sphere of a regular three-dimensional polyhedron. The code points are the centers of the faces of the polyhedron.

Consider now the code formed by the midpoints of the faces of a cube of edge length 2. This is the three-dimensional biorthogonal code. The maximum-likelihood net is given by the radial projection of the cube edges onto the inscribed unit sphere. The code is a symmetric $\{3,6,12\}$. There is no regular polyhedron with 6 faces other than the cube, so that we will have shown the three-dimensional biorthogonal code to be optimal if we establish that every $\{3,6,k\}$ must have $k \leqq 12$. To see this latter fact, note that for three-dimensional codes, at least three edges of the maximum-likelihood net must meet at each vertex of the net (since each $R_i$ is convex). Thus $3v \leqq 2k$ where $v$ and $k$ are, respectively, the total number of vertices and edges for the net. Euler's formula (Ref. 8, p. 9) $v - k + M = 2$ holds for the net, and so

$$k \leqq 3(M-2). \tag{20}$$

For the case at hand $M = 6$, and (20) gives $k \leqq 12$, so that the proof is completed.

Analogous reasoning shows that the centers of the faces of the dodecahedron give the best code with $M = 12$ points. The code is a symmetric $\{3,12,30\}$.

The regular octahedron in $\mathcal{E}_3$ gives rise to a symmetric $\{3,8,12\}$ whose code points are the vertices of a cube. This is not the optimal configuration of 8 points in $\mathcal{E}_3$. By rotating one face of the cube 45 degrees about an axis perpendicular to the face and through its center, one obtains a $\{3,8,16\}$. By translating this face slightly toward the opposite face of the cube, and by slightly expanding both faces, one obtains a $\{3,8,16\}$ with minimum distance between code points strictly larger than the minimum for the cubic arrangement of points. There are then noise functions $f(|z|)$ of (1) for which this new code has a larger $Q$ than the cube-code.

It is not known whether the symmetric $\{3,20,30\}$ obtained from the regular icosahedron is an optimal code of 20 points.

## IV. THE BIORTHOGONAL CODE

The biorthogonal code is a symmetric $\{2n,n,n(2n-2)\}$. The $2n$ code points can be taken as the points on the coordinate axes unit distance from the origin. Alternatively, the code points can be described as the centers of the $(n-1)$-dimensional bounding cells of the unit $n$-cube. The radial projection of the cube onto the unit sphere with center at the center of the hypercube gives the maximum-likelihood net of the code.

We have seen that for $n = 3$ the biorthogonal code is optimal among codes of $M = 2n = 6$ points. It is natural to suspect that for all $n$ the biorthogonal code is optimal among codes of $2n$ points in $\mathcal{E}_n$. However, the methods used in this paper, based as they are on (19), will not suffice to settle this question, for, as will be shown below, when $n \geqq 4$, there exist $\{2n,n,n(2n-1)\}$ codes; i.e., codes with a larger $k$ value than the biorthogonal code.

It might be thought that this encumbering dependence of (19) on $k$ could be avoided — that an inequality for $Q$ independent of $k$ could be found which is attainable for optimal codes. The example already treated of the octahedron shows, however, that this dependence on $k$ is essential.

To construct a $\{2n,n,n(2n-1)\}$ for $n \geqq 5$, choose $2n$ distinct real numbers $\nu_1, \nu_2, \cdots, \nu_{2n}$. The vectors of the code are given by

$$\hat{\mathbf{X}}_i = (\alpha_i, \alpha_i\nu_i, \alpha_i\nu_i^2, \cdots, \alpha_i\nu_i^{n-1}), \tag{21}$$

where

$$\alpha_i = [1 + \nu_i^2 + \nu_i^4 + \cdots + \nu_i^{2n-2}]^{-\frac{1}{2}} \qquad i = 1, 2, \cdots, 2n$$

has been chosen so that $\hat{\mathbf{X}}_i$ is a unit vector. The code is closely related to the cyclic polytope described by Gale[14].

An important property of this code can be derived[14] by considering the polynomials

$$F_{ij}(\lambda) \equiv (\lambda - \nu_i)^2(\lambda - \nu_j)^2 = \sum_{p=0}^{4} A_{ij}^{(p)}\lambda^p \qquad i,j = 1, 2, \cdots, 2n \tag{22}$$

which are nonnegative. We define the $(2n)^2$ $n$-vectors

$$\mathbf{B}_{ij} \equiv (A_{ij}^{(0)}, A_{ij}^{(1)}, A_{ij}^{(2)}, A_{ij}^{(3)}, A_{ij}^{(4)}, 0, \cdots, 0).$$

We then have

$$\mathbf{B}_{ij}\cdot\hat{\mathbf{X}}_l = \alpha_l \sum_{p=0}^{4} A_{ij}^{(p)}\nu_l^p = F_{ij}(\nu_l) = \begin{cases} 0, & l = i \\ 0, & l = j \\ a_{ijl} > 0, & l \neq i, \quad l \neq j \end{cases} \tag{23}$$

where the positivity of the $a_{ijl}$ follows from the factored form (22) of $F_{ij}(\lambda)$.

To show that the points (21) determine a $\{2n,n,n(2n-1)\}$ we note first that they span $\mathcal{E}_n$. Indeed every choice of $n$ vectors $\hat{\mathbf{X}}_i$ from (21) yields an independent set, as can be seen by forming the determinant whose rows are the components of the vectors. These determinants are proportional to Vandermonde determinants and do not vanish. To show that $k = n(2n-1)$ for the code, consider the maximum likelihood region $\mathcal{R}_i$ containing $\hat{\mathbf{X}}_i$. By the construction described in the first paragraph of Section II, $\mathcal{R}_i$ is the intersection of the half-spaces

$$\mathcal{H}_j^{(i)}(\mathbf{X}) = (\hat{\mathbf{X}}_i - \hat{\mathbf{X}}_j)\cdot\hat{\mathbf{X}} \geqq 0 \qquad j = 1, 2, \cdots, 2n; \qquad j \neq i. \tag{24}$$

We assert that each of the $2n-1$ hyperplanes $\mathcal{H}_j^{(i)}$, $j = 1,2, \cdots, 2n$ with $j \neq i$, is indeed an $(n-1)$-dimensional boundary of $\mathcal{R}_i$. It will then follow that $k = \frac{1}{2}2n(2n-1)$ since there are $2n$ maximum likelihood regions. That $\mathcal{H}_j^{(i)}$ is an $(n-1)$-boundary of $\mathcal{R}_i$ results from the fact that there exists a point $\mathbf{X}_0$ contained in $\mathcal{R}_i$ that lies in $\mathcal{H}_j^{(i)}$ but not in $\mathcal{H}_k^{(i)}$, $k = 1, 2, \cdots, 2n$ with $k \neq i$ and $k \neq j$. From (23) we can choose $\mathbf{X}_0 = \mathbf{B}_{ij}$ since

$$\mathcal{H}_j^{(i)}(\mathbf{B}_{ij}) = 0$$

$$\mathcal{H}_k^{(i)}(\mathbf{B}_{ij}) = a_{ijk} > 0, \qquad k \neq i, \qquad k \neq j.$$

For $n = 4$, the configuration of eight points given by

$$\hat{\mathbf{X}}_k = \frac{1}{\sqrt{2}}\left(\cos k\frac{\pi}{4}, \sin k\frac{\pi}{4}, \cos k\frac{\pi}{2}, \sin k\frac{\pi}{2}\right) \qquad k = 1, 2, \cdots, 8$$

is a $\{4,8,28\}$. The proof is similar to that just given for the case $n \geqq 5$ with the role of the polynomial $F_{ij}(\lambda)$ being replaced here by the expression

$$F_{ij}^*(\lambda) = \left[1 - \cos\left(\lambda - i\frac{\pi}{4}\right)\right] \times \left[1 - \cos\left(\lambda - j\frac{\pi}{4}\right)\right].$$

We omit the details.

We close this section by noting that although we cannot show that the biorthogonal code has a largest $Q$ value for codes of $2n$ points, it does have largest nearest neighbor distance, 90° in angular terms. Indeed no collection of more than $n + 1$ vectors in $\mathcal{E}_n$ can have minimum angular distance between points greater than 90°. For consider* Fig. 5. Without

| | | | | | |
|---|---|---|---|---|---|
| + | − | − | . . . | − | − |
| o | + | − | | − | − |
| o | o | + | | − | − |
| . | . | . | | . | . |
| . | . | . | | . | . |
| . | . | . | | . | . |
| o | o | o | | + | − |

Fig. 5 — Table of component signs.

loss of generality the positive $x_1$-axis of a rectangular coordinate system can be chosen to lie along the first vector. The first column of the figure shows the sign of the components of this vector. The coordinate axes can be oriented so that $\hat{\mathbf{X}}_2$ lies in the $x_1 - x_2$ plane and the direction of the $x_2$-axis can be chosen so that the $x_2$-component of $\hat{\mathbf{X}}_2$ is positive. The second column of Fig. 5 shows the sign of the components of $\hat{\mathbf{X}}_2$. The first component must be negative since if the minimum distance is to be greater than 90° we must have $\hat{\mathbf{X}}_1 \cdot \hat{\mathbf{X}}_2 < 0$. Continuing in this manner we are forced to choose the components of the $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \cdots, \hat{\mathbf{X}}_{n+1}$ as shown. But now it is impossible to find an $(n + 2)$nd vector having a negative scalar product with these $n + 1$ vectors, for if the nonzero components of $\hat{\mathbf{X}}_{n+2}$ are all negative, it has a positive scalar product with $\hat{\mathbf{X}}_{n+1}$, whereas if the first positive component of $\hat{\mathbf{X}}_{n+2}$ is the $j$th, $\hat{\mathbf{X}}_j \cdot \hat{\mathbf{X}}_{n+2}$ is positive.

* This elegant proof was suggested by J. H. van Lint.

## V. THE INFINITE-BAND GAUSSIAN CHANNEL

When $M = n + 1$ and $f(x)$ is given by (4), the model discussed here describes the transmission of $M$ equally likely signals $s_i(t)$, $i = 1, 2, \cdots, n + 1$, of duration $T$ in white Gaussian noise of spectral power density $N/2$. Here the signals are constrained by

$$\int_0^T s_i^2(t)dt = PT.$$

When these signals are transmitted, the probability of no error using the best possible detection scheme is given by (2), where the $\hat{\mathbf{X}}_i$ must be chosen so that

$$\hat{\mathbf{X}}_i \cdot \hat{\mathbf{X}}_j = \frac{1}{PT} \int_0^T s_i(t)s_j(t)dt,$$

the $\mathcal{R}_i$ are the maximum-likelihood regions (3), and

$$\sigma^2 = \frac{N}{2PT}.$$

See Ref. 3, Sections 4.2 and 4.3 or Ref. 15 for a more detailed description of the correspondence between the geometric model and the physical one.

Our result that the simplex code is optimal means that *in communicating in infinite-band white Gaussian noise by means of* M *equally likely equal-energy signals of duration* T *(no bandwidth restrictions imposed) the error probability is minimized by choosing signals with normalized cross-correlation*

$$\frac{1}{PT} \int_0^T s_i(t)s_j(t)dt = -\frac{1}{M-1}, \qquad i \neq j \tag{25}$$

this being the value of $\hat{\mathbf{X}}_i \cdot \hat{\mathbf{X}}_j$ for the regular simplex.

The error probability with a best set of signals of form (25) is readily determined to be

$$P_e = 1 - \int_{-\infty}^{\infty} dx\, f(x)\Phi^{M-1}\left(x + \frac{1}{\sigma}\sqrt{\frac{M}{M-1}}\right), \tag{26}$$

where $f(x)$ is the Gaussian density (4) and $\Phi$ the cumulative

$$\Phi(y) = \int_{-\infty}^{y} f(x)dx.$$

When the transmission rate

$$R = \frac{\log M}{T}$$

is kept constant, along with $N$ and $P$, (26) becomes for large $T$ (and hence large $M$)

$$P_e = \exp [-E(R)T + o(T)], \tag{27}$$

where

$$E(R) = \begin{cases} \frac{C}{2} - R, & R \leqq C/4 \\ (\sqrt{C} - \sqrt{R})^2, & R \geqq C/4 \end{cases} \tag{28}$$

and $C = P/N$ is the capacity of the channel. That the minimal asymptotic error probability for this channel must have the form (27)–(28) was first proved by Wyner.[15]

## APPENDIX A

### *A Lemma*

The following lemma will be useful in establishing the main results of Appendices B and C.

*Lemma: Let $w_1(x)$ and $w_2(x)$ be integrable functions that satisfy*

$$\int_a^b w_1(x)dx = \int_a^b w_2(x)dx. \tag{29}$$

*Further, suppose there exists an $x'$, $a \leqq x' \leqq b$, such that*

$$\begin{aligned} w_2(x) &\geqq w_1(x), \qquad a \leqq x \leqq x' \\ w_2(x) &\leqq w_1(x), \qquad x' \leqq x \leqq b. \end{aligned} \tag{30}$$

*Then, if $m(x)$ is a nonnegative monotone increasing function,*

$$\int_a^b m(x)w_1(x)dx \geqq \int_a^b m(x)w_2(x)dx. \tag{31}$$

*If $m(x)$ is a nonnegative monotone decreasing function,*

$$\int_a^b m(x)w_1(x)dx \leqq \int_a^b m(x)w_2(x)dx. \tag{32}$$

*Equality holds in* (31) *and* (32) *only if $w_1(x) = w_2(x)$ for almost all $x$.*

*Proof*: If $m(x)$ is nonnegative and monotone increasing, then

$$\int_a^b m(x)[w_1(x) - w_2(x)]dx$$

$$= \int_a^{x'} m(x)[w_1(x) - w_2(x)]dx + \int_{x'}^b m(x)[w_1(x) - w_2(x)]dx$$

$$\geqq m(x') \int_a^{x'} [w_1(x) - w_2(x)]dx + m(x') \int_{x'}^b [w_1(x) - w_2(x)]dx$$

$$= m(x') \left[ \int_a^b w_1(x)dx - \int_a^b w_2(x)\,dx \right] = 0.$$

If $m(x)$ is nonnegative and monotone decreasing, the steps are the same with the inequalities reversed.

## APPENDIX B

*Convexity of $h_r(w)$*

Let $x_1, x_2, \cdots, x_n$ be the rectangular coordinates of a point in $\mathcal{E}_n$. The surface $S$ of the unit sphere centered at the origin can be given parametrically by

$$\begin{aligned}
x_1 &= \cos\theta_1 \\
x_2 &= \sin\theta_1 \cos\theta_2 \\
&\vdots \\
x_j &= \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{j-1} \cos\theta_j \\
&\vdots \\
x_{n-1} &= \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-2} \cos\theta_{n-1} \\
x_n &= \sin\theta_1 \sin\theta_2 \cdots \sin\theta_{n-2} \sin\theta_{n-1} \\
0 &\leqq \theta_i < \pi, \qquad i = 1, 2, \cdots, n-2 \\
0 &\leqq \theta_{n-1} < 2\pi
\end{aligned} \tag{33}$$

and the element of surface content is

$$ds = \sin^{n-2}\theta_1 \sin^{n-3}\theta_2 \cdots \sin\theta_{n-2}\, d\theta_1\, d\theta_2 \cdots d\theta_{n-1}. \tag{34}$$

We shall only be concerned with the case $n \geqq 3$.

The spherical cap of angle $\varphi$ about $P$, the end point of

$$\hat{\mathbf{Y}} = (1, 0, 0, \cdots, 0),$$

is given by $\theta_1 \leqq \varphi$. A hyperplane $\mathfrak{IC}$ that intersects this spherical cap is

$x_2 = x_1 \tan \alpha$ with $0 \leqq \alpha < \varphi$ and the intersection of $\mathfrak{IC}$ with the spherical cap is from (33)

$$\cos \theta_2 = \tan \alpha \cot \theta_1 .$$

We then have from the definition (8)

$$h_r(w) = \int_\alpha^\varphi d\theta_1 \int_\mu^\nu d\theta_2 \sin^{n-2} \theta_1 \sin^{n-3} \theta_2 g_r(\sqrt{2 - 2\cos\theta_1})$$
$$\int_0^\pi d\theta_3 \cdots \int_0^\pi d\theta_{n-2} \int_0^{2\pi} d\theta_{n-1} \sin^{n-4} \theta_3 \cdots \sin \theta_{n-2}$$

where $\nu = \arccos(\tan \alpha \cot \theta_1)$ and $\mu = -\nu$ if $n = 3$ but $\mu = 0$ if $n \geqq 4$. In either event, we can write

$$h_r(w) = k_n \int_\alpha^\varphi d\theta_1 \int_0^\nu d\theta_2 \sin^{n-2} \theta_1 \sin^{n-3} \theta_2 g_r(\sqrt{2 - 2\cos\theta_1}) \quad (35)$$

while for the content of the piece of the cap cut off by $\mathfrak{IC}$ we have

$$w = k_n \int_\alpha^\varphi d\theta_1 \int_0^\nu d\theta_2 \sin^{n-2} \theta_1 \sin^{n-3} \theta_2 \quad (36)$$

with $k_n > 0$ and independent of $\alpha$.

Straightforward differentiation of (35) yields

$$\frac{dh}{d\alpha} = -k_n \sec^2 \alpha \int_\alpha^\varphi d\theta_1 \sin\theta_1 \cos\theta_1$$
$$\cdot [1 - \sec^2 \alpha \cos^2 \theta_1]^{(n-4)/2} g_r(\sqrt{2 - 2\cos\theta_1}).$$

Now introduce

$$x = \cos^2 \theta_1 , \qquad a = \cos^2 \varphi, \qquad b = \cos^2 \alpha$$

and

$$\hat{g}(x) = g_r(\sqrt{2 - 2\sqrt{x}}).$$

We have $0 \leqq a < b \leqq 1$. Note that $\hat{g}(x) > 0$ is monotone increasing in $x$. In these terms

$$\frac{dh}{d\alpha} = -\frac{k_n}{2b} \int_a^b dx \left[1 - \frac{x}{b}\right]^{(n-4)/2} \hat{g}(x)$$

and

$$\frac{dw}{d\alpha} = -\frac{k_n}{2b} \int_a^b dx \left[1 - \frac{x}{b}\right]^{(n-4)/2} = -\frac{k_n}{n-2} \left(1 - \frac{a}{b}\right)^{(n-2)/2} .$$

Combining these results we find

$$\frac{dh_r(w)}{dw} = \frac{dh/d\alpha}{(dw/d\alpha)} = \int_a^b l_b(x)\hat{g}(x)dx > 0, \tag{37}$$

where

$$l_b(x) = \frac{\left(1 - \frac{x}{b}\right)^{(n-4)/2}}{\frac{2b}{n-2}\left(1 - \frac{a}{b}\right)^{(n-2)/2}}, \qquad a \leqq x \leqq b. \tag{38}$$

Note that

$$\int_a^b l_b(x)dx = 1. \tag{39}$$

When $n > 4$, the convexity of $h_r(w)$ can be established from (37)–(39) as follows. Consider two different $w$ values, say $w_2 > w_1$ with corresponding parameters $b_2 = \cos^2 \alpha_2$ and $b_1 = \cos^2 \alpha_1$. We have

$$1 \geqq b_2 > b_1 > a.$$

From (38) one readily finds that there is a unique real root $x'$ for which $l_{b_2}(x') = l_{b_1}(x')$, $a < x' < b_1$. For $a \leqq x \leqq x'$ we have $l_{b_1}(x) \geqq l_{b_2}(x)$. If we now define $l_b(x) = 0$ for $x > b$, we can also write $l_{b_2}(x) \geqq l_{b_1}(x)$ for $x \geqq x'$. From (39) we have

$$\int_a^{b_2} l_{b_1}(x)dx = \int_a^{b_2} l_{b_2}(x)dx.$$

The conditions of the lemma of Appendix A hold and we conclude from (37) that

$$w_2 > w_1 \Rightarrow \frac{dh_r(w_2)}{dw} > \frac{dh_r(w_1)}{dw}$$

which is the desired convexity.

When $n = 4$, (38) becomes $l_b(x) = (b - a)^{-1}$ for $a \leqq x \leqq b$. As before, we define $l_b(x) = 0$ for $x > b$. It is readily seen that the lemma again applies with $x'$ chosen as $b_1$. Convexity is then established in this case as well.

For $n = 3$, (37) and (38) give

$$\frac{dh_r(w)}{dw} = \int_a^b \frac{\hat{g}(x)dx}{2\sqrt{b-a}\sqrt{b-x}}$$
$$= -2\sqrt{b-x}\,\frac{\hat{g}(x)}{2\sqrt{b-a}}\Big|_a^b + \int_a^b \sqrt{\frac{b-x}{b-a}}\,d\hat{g}(x)$$
$$= \hat{g}(a) + \int_a^b \sqrt{\frac{b-x}{b-a}}\,d\hat{g}(x),$$

on integrating by parts. However, since $\hat{g}$ is increasing in $x$, it follows that the last integral is increasing in $b$ and hence also in $w$. The convexity proof is thus completed.

## APPENDIX C

*Proof of Equation (11)*

We shall be concerned here with two different regions, $V$ and $W$, cut off from the spherical cap of angle $\varphi$ about the point $P$ which we take as the terminus of the unit vector $\hat{\mathbf{Y}}$ in $\mathcal{E}_n$ (see Fig. 6). The region $V$ is the intersection of the spherical cap with a convex cone $\mathcal{U}$ having the origin as a vertex. It is assumed that $\mathcal{U}$ does not contain $P$. We denote by $Q$ a point of $V$ closest to $P$. The second region, $W$, is cut off from the cap by a single hyperplane $\mathcal{L}$ through the origin but not through $P$. $\mathcal{L}$ is chosen so that $W$ and $V$ have the same content, $w$ and $v$, respectively, and for purposes of our proof we restrict the normal to $\mathcal{L}$ to lie in the 2-plane through the origin, $P$ and $Q$. We wish to show that

$$I_V \equiv \int_V g_r(|\,\hat{\mathbf{Y}} - \hat{\mathbf{X}}\,|)ds \geqq I_W \equiv \int_W g_r(|\,\hat{\mathbf{Y}} - \hat{\mathbf{X}}\,|)ds \qquad (40)$$

with equality holding only if $V$ is cut off from the cap by a single hyperplane. Here, as in (11), $g_r$ is nonnegative and monotone decreasing and $\hat{\mathbf{X}}$ is a generic unit vector in $\mathcal{E}_n$. In the applications made of (40) in the main text, $\mathcal{U}$ is specialized to a type of flat-sided cone.

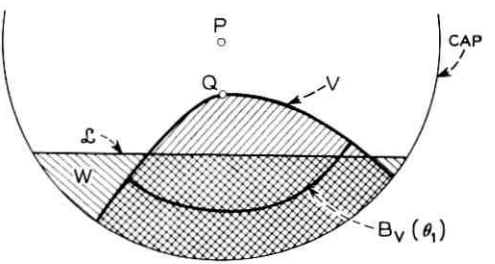



Fig. 6 — Regions involved in proof of (11).

Let us again adopt the spherical coordinates (33) with the pole $P$ of the cap located on the $x_1$-axis so that $\hat{\mathbf{Y}} = (1,0,0, \cdots, 0)$. We suppose the axes are oriented so that $Q$ lies in the $x_1$-$x_2$ plane. Then from (33) and (34)

$$I_V = \int_0^{\varphi} d\theta_1 g_r(\sqrt{2 - 2\cos\theta_1})\, v(\theta_1) \tag{41}$$

where

$$v(\theta_1) = \sin^{n-2}\theta_1 \int_{B_V(\theta_1)} d\theta_2 \cdots \int d\theta_{n-1} \sin^{n-3}\theta_2 \cdots \sin\theta_{n-1} \tag{42}$$

is the $(n-2)$-dimensional content of the intersection $B_V(\theta_1)$ of $V$ with the hyperplane $x_1 = \cos\theta_1$. Similarly,

$$I_W = \int_0^{\varphi} d\theta_1 g_r(\sqrt{2 - 2\cos\theta_1})\, w(\theta_1) \tag{43}$$

where $w(\theta_1)$ is the $(n-2)$-dimensional content of the intersection $B_W(\theta_1)$ of $W$ with the hyperplane $x_1 = \cos\theta_1$. By hypothesis we have

$$v = \int_0^{\varphi} d\theta_1 v(\theta_1) = w = \int_0^{\varphi} d\theta_1 w(\theta_1). \tag{44}$$

Since $g_r(\sqrt{2 - 2\cos\theta_1})$ is a nonnegative monotone decreasing function of $\theta_1$, all the hypotheses of the lemma of Appendix A will hold if we can show the existence of a $\varphi'$ such that

$$\begin{aligned} v(\theta_1) &\geqq w(\theta_1), \qquad 0 \leqq \theta_1 \leqq \varphi' \\ w(\theta_1) &\geqq v(\theta_1), \qquad \varphi' \leqq \theta_1 \leqq \varphi. \end{aligned} \tag{45}$$

The conclusion (32) of the lemma then is (40).

Our goal now, therefore, is to show that $v(\theta)$ and $w(\theta)$ cross only once as indicated in (45). Let $\alpha = \angle POQ$. If $Q^*$ is the nearest point in $W$ to $P$ and $\beta = \angle POQ^*$, then $\beta > \alpha$. For $0 \leqq \theta \leqq \alpha$, both $v(\theta)$ and $w(\theta)$ are zero. For $\alpha < \theta \leqq \beta$, $v(\theta) > w(\theta) = 0$. From (44) it then follows that there is a first point in $(0,\varphi)$ where $w(\theta)$ crosses up through $v(\theta)$, that is, where $v(\theta) = w(\theta)$ and $w'(\theta) > v'(\theta)$ where the prime denotes differentiation with respect to $\theta$. If there were a second crossing, at that point we would have $w' < v'$. We prove that there is only one crossing by demonstrating that

$$v(\theta_1) = w(\theta_1) \Rightarrow \frac{dw(\theta_1)}{d\theta_1} \geqq \frac{dv(\theta_1)}{d\theta_1} \tag{46}$$

for $0 \leqq \theta_1 \leqq \varphi$.

Let $\omega$ be such that $v(\omega) = w(\omega)$, $0 \leqq \omega \leqq \varphi$. Consider now the spherical pyramid $\Gamma_V$ having $Q$ as vertex and as base the set $B_V(\omega)$ defined below (42). $\Gamma_V$ is the set of all points $\hat{\mathbf{X}}$ of the form

$$\begin{gathered} \hat{\mathbf{X}} = \xi\hat{\mathbf{X}}_Q + \eta\hat{\mathbf{X}}_B, \\ \xi \geqq 0, \qquad \eta \geqq 0, \qquad \hat{\mathbf{X}}_Q = \overrightarrow{OQ}, \qquad \hat{\mathbf{X}}_B \in B_V(\omega) \end{gathered} \tag{47}$$

where, in order for (47) to be a unit vector, we have the additional restriction

$$|\hat{\mathbf{X}}|^2 = 1 = \xi^2 + \eta^2 + 2\xi\eta\hat{\mathbf{X}}_Q \cdot \hat{\mathbf{X}}_B . \tag{48}$$

Note that since $\mathcal{U}$ is convex and since $Q$ and $B_V(\omega)$ are contained in $\mathcal{U}$, it follows from (47) that $\Gamma_V$ is contained in $\mathcal{U}$ and $S$, hence $\Gamma_V$ is contained also in $V$.

Now let $\bar{v}(\theta_1)$ denote the $(n-2)$-dimensional content of the intersection of $\Gamma_V$ with the hyperplane $x_1 = \cos\theta_1$, where $\alpha \leqq \theta_1 \leqq \omega$. We have

$$\begin{aligned} \bar{v}(\omega) &= v(\omega) \\ \bar{v}(\omega - \delta) &\leqq v(\omega - \delta) \end{aligned} \tag{49}$$

where this last follows from the fact that $\Gamma$ is contained in $V$. One has then

$$\frac{v(\omega) - v(\omega - \delta)}{\delta} \leqq \frac{\bar{v}(\omega) - \bar{v}(\omega - \delta)}{\delta}$$

so that

$$\left.\frac{dv(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega} \leqq \left.\frac{d\bar{v}(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega} . \tag{50}$$

Consider next the spherical pyramid $\Gamma_W(Q)$ having $Q$ as vertex and as base the set $B_W(\omega)$ defined below (43). We denote by $\bar{w}(\theta_1)$ the $(n-2)$-dimensional content of the intersection of $\Gamma_W$ with the hyperplane $x_1 = \cos\theta_1$. As before, let $Q^*$ be the nearest point in $W$ to $P$. We denote by $\bar{w}^*(\theta_1)$ the $(n-2)$-dimensional content of the intersection of the spherical pyramid $\Gamma_W(Q^*)$ with the hyperplane $x_1 = \cos\theta_1$. Since $Q^*$ is contained in $\Gamma_W(Q)$, $\Gamma_W(Q^*)$ is also contained in $\Gamma_W(Q)$ and we have

$$\begin{aligned} \bar{w}(\omega) &= \bar{w}^*(\omega) \\ \bar{w}(\omega - \delta) &\geqq \bar{w}^*(\omega - \delta). \end{aligned} \tag{51}$$

From this it follows that

$$\left.\frac{d\bar{w}(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega} \leqq \left.\frac{d\bar{w}^*(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega}.$$

However, since $Q^*$ lies in the hyperplane,

$$\bar{w}^*(\omega) = w(\omega) \tag{52}$$

for all $\omega$, hence

$$\left.\frac{d\bar{w}(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega} \leqq \left.\frac{dw(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega}. \tag{53}$$

In the remaining paragraphs of this appendix we shall show that

$$\bar{v}(\omega) = \bar{w}(\omega) \Rightarrow \left.\frac{d\bar{v}(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega} \leqq \left.\frac{d\bar{w}(\theta_1)}{d\theta_1}\right|_{\theta_1=\omega} \tag{54}$$

which will establish (46) and complete our proof, for the hypothesis of (46) follows from that of (54) by (49), (51), and (52) and the conclusion of (46) follows from the conclusion of (54) by (50) and (53).

Let the spherical coordinates of a point $\hat{\mathbf{X}}$ in $\Gamma_V$ be denoted by the angles $(\varphi_1, \cdots, \varphi_{n-1})$ (see Fig. 7). We employ the angles $(\theta_1, \cdots, \theta_{n-1})$ to describe a point $\hat{\mathbf{X}}_B$ in $B_V(\theta_1)$. The content $\bar{v}(\mu)$ of the intersection $J(\mu)$ of $\Gamma_V$ with the hyperplane $x_1 = \cos\mu$, $\alpha \leqq \mu \leqq \omega$ is given by

$$\bar{v}(\mu) = \sin^{n-2}\mu \int d\varphi_2 \sin^{n-3}\varphi_2 \int d\varphi_3 \sin^{n-4}\varphi_3 \cdots \int_{J(\mu)} d\varphi_{n-1}. \tag{55}$$

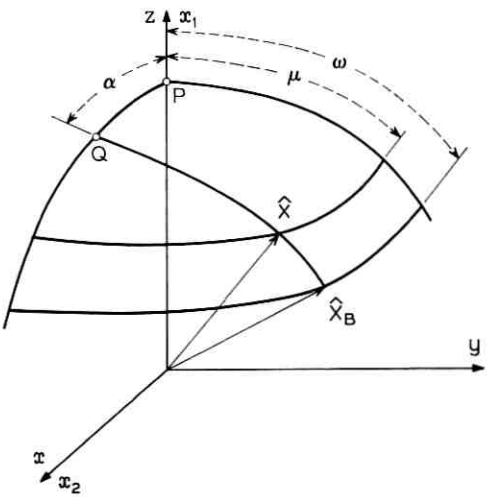



Fig. 7 — The mapping from $\hat{\mathbf{X}}$ to $\hat{\mathbf{X}}_B$.

The relationships (47)–(48), however, serve to define a one-one transformation between the coordinates $(\mu, \varphi_2, \cdots, \varphi_{n-1})$ of a point in $J(\mu)$ and the coordinates $(\omega, \theta_2, \cdots, \theta_{n-1})$ of a point in $B_V(\omega)$, so that $\bar{v}(\mu)$ can be expressed as an integration over $B_V(\omega)$ as well. Taking components of (47), we find successively

$$\begin{aligned}
\cos\mu &= \xi\cos\alpha + \eta\cos\omega \\
\sin\mu\cos\varphi_2 &= \xi\sin\alpha + \eta\sin\omega\cos\theta_2 \\
\sin\mu\sin\varphi_2\cos\varphi_3 &= \eta\sin\omega\sin\theta_2\cos\theta_3 \\
&\vdots \\
\sin\mu\sin\varphi_2\cdots\sin\varphi_{j-1}\cos\varphi_j &= \eta\sin\omega\sin\varphi_2\cdots\sin\varphi_{j-1}\cos\varphi_j \qquad (56) \\
&\vdots \\
\sin\mu\sin\varphi_2\cdots\sin\varphi_{n-2}\sin\varphi_{n-1} & \\
&= \eta\sin\omega\sin\varphi_2\cdots\sin\varphi_{n-2}\sin\varphi_{n-1} \\
j &= 3, 4, \cdots, n-1.
\end{aligned}$$

Dividing the $n$th equation by the $(n-1)$st yields $\varphi_{n-1} = \theta_{n-1}$. Dividing the $(n-1)$st equation by the $(n-2)$nd then yields $\varphi_{n-2} = \theta_{n-2}$. Proceeding in this manner, one finds $\varphi_j = \theta_j$, $j = 3, 4, \cdots, n-1$. The first two equations of (56) can be solved for $\xi$ and $\eta$. By substituting these expressions into (48) which now reads

$$\xi^2 + \eta^2 + 2\xi\eta\,(\cos\alpha\cos\omega + \sin\alpha\sin\omega\cos\theta_2) = 1, \qquad (57)$$

we obtain a single relationship connecting $\varphi_2$ and $\theta_2$ which we suppose solved in the form

$$\varphi_2 = \varphi_2(\theta_2, \mu). \qquad (58)$$

Equation (55) now becomes in the new variables

$$\begin{aligned}
\bar{v}(\mu) &= \sin^{n-2}\mu \int_{B_V(\omega)} d\theta_2 \frac{d\varphi_2}{d\theta_2}\sin^{n-3}\varphi_2 \int d\theta_3\sin^{n-2}\theta_3\cdots\int d\theta_{n-1} \qquad (59) \\
&= \sin^{n-2}\omega\int d\theta_2 G(\theta_2,\mu)h(\theta_2)
\end{aligned}$$

with

$$G(\theta_2,\mu) = \left[\frac{\sin\mu}{\sin\omega}\right]^{n-2}\frac{d\varphi_2}{d\theta_2}\left[\frac{\sin\varphi_2}{\sin\theta_2}\right]^{n-3} \qquad (60)$$

and

$$h(\theta_2) = \sin^{n-3}\theta_2 \int d\theta_3 \sin^{n-4}\theta_3 \cdots \int_{B_V(\omega)} d\theta_{n-1}, \tag{61}$$

where if $n = 3$ this latter expression is to be interpreted as unity. It is convenient now to define $h(\theta_2)$ to be zero if $\theta_2$ is not the second angle coordinate of a point in $B_V(\omega)$. In this notation, then, we have for $n \geqq 4$

$$\left.\frac{d\bar{v}(\mu)}{d\mu}\right|_{\mu=\omega} = \sin^{n-2}\omega \int_0^\pi d\theta_2 \left.\frac{\partial G(\theta_2,\mu)}{\partial\mu}\right|_{\mu=\omega} h(\theta_2) \tag{62}$$

$$\bar{v}(\omega) = \sin^{n-2}\omega \int_0^\pi d\theta_2 h(\theta_2). \tag{63}$$

If $n = 3$, the lower limits of integration here should be replaced by $-\pi$. It will be shown later that $\partial G/\partial\mu\,|_{\mu=\omega}$ is a nonnegative monotone decreasing function of $\theta_2$.

We next seek to determine the nature of the set $B_V(\omega)$ of given content $\bar{v}(\omega)$ that will maximize (62). We note first from (61) that for $n \geqq 4$

$$h(\theta_2) \leqq \sigma(\theta_2) \tag{64}$$

where $\sigma(\theta_2)$ is the surface content of a sphere of radius $\sin\theta_2$ in $\mathcal{E}_{n-2}$ since $\sigma$ is given by the integrals of (61) with the integration variables running through their maximum allowable range. Now let $B^*(\omega)$ be the set of points defined by $\theta_1 = \omega$, $0 \leqq \theta_2 \leqq \theta_2'$ where $\theta_2'$ is given by

$$\bar{v}(\omega) = \sin^{n-2}\omega \int_0^{\theta_2'} d\theta_2 \sigma(\theta_2).$$

For $B^*(\omega)$ we have

$$h^*(\theta_2) = \begin{cases} \sigma(\theta_2), & 0 \leqq \theta_2 \leqq \theta_2' \\ 0, & \theta_2' < \theta_2 \end{cases} \tag{65}$$

so that

$$\bar{v}(\omega) = \sin^{n-2}\omega \int_0^\pi d\theta_2 h^*(\theta_2). \tag{66}$$

We also have

$$\begin{aligned} h^*(\theta_2) &\geqq h(\theta_2), \qquad 0 \leqq \theta_2 \leqq \theta_2' \\ h^*(\theta_2) &\leqq h(\theta_2), \qquad \theta_2' \leqq \theta_2 \leqq \pi \end{aligned} \tag{67}$$

from (64) and (65).

The hypotheses of the lemma of Appendix A are thus met from (63), (66), (67), and the monotonicity of $\partial G/\partial\mu$. We conclude that among sets of equal content, $d\bar{v}/d\mu\,|_{\mu=\omega}$ is a maximum for the set $B^*(\omega)$. The set $B_W(\omega)$, however, coincides with $B^*(\omega)$. Equation (54) then follows for $n \geqq 4$. The modification necessary to treat the case $n = 3$ is trivial.

There remains only the demonstration that $\partial G/\partial\mu\,|_{\mu=\omega}$ is nonnegative monotone decreasing in $\theta_2$. Equation (57) and the first two equations of (56) are identical with the equations that would hold for the three 3-vectors $\overrightarrow{OQ}$, $\hat{\mathbf{X}}$ and $\hat{\mathbf{X}}_B$ of Fig. 7 constrained to satisfy (47). The relationship (58) between $\varphi_2$ and $\theta_2$ can most easily be written down by consulting this figure. The condition that the three points be coplanar with the origin is

$$\begin{vmatrix} x & y & z \\ x_Q & y_Q & z_Q \\ x_B & y_B & z_B \end{vmatrix} = 0 \tag{68}$$

where

$$\begin{aligned} x &= \sin\mu\cos\varphi_2 & y &= \sin\mu\sin\varphi_2 & z &= \cos\mu \\ x_Q &= \sin\alpha & y_Q &= 0 & z_Q &= \cos\alpha \\ x_B &= \sin\omega\cos\theta_2 & y_B &= \sin\omega\sin\theta_2 & z_B &= \cos\omega \end{aligned} \tag{69}$$

which serves to determine (58). Routine implicit differentiation of (68) and (69) and evaluation at $\mu = \omega$, $\varphi_2 = \theta_2$ yields

$$\left.\frac{d\varphi_2}{d\theta_2}\right|_{\mu=\omega} = 1 \tag{70}$$

$$\left.\frac{\partial\varphi_2}{\partial\mu}\right|_{\mu=\omega} = \frac{\sin\alpha\sin\theta_2}{\sin\omega(\cos\alpha\sin\omega - \sin\alpha\cos\omega\cos\theta_2)} \tag{71}$$

$$\left.\frac{\partial}{\partial\mu}\frac{d\varphi_2}{d\theta_2}\right|_{\mu=\omega} = \frac{\sin\alpha[\cos\alpha\sin\omega\cos\theta_2 - \sin\alpha\cos\omega]}{\sin\omega[\cos\alpha\sin\omega - \sin\alpha\cos\omega\cos\theta_2]^2}. \tag{72}$$

The denominators of (71) and (72) are positive since $\omega > \alpha$ implies $\tan\omega\cot\alpha > 1 \geqq \cos\theta_2$ which is the same as

$$\cos\alpha\sin\omega > \sin\alpha\cos\omega\cos\theta_2 .$$

The numerator of (72) is nonnegative for points $\hat{\mathbf{X}}_B$ of interest to us since we are concerned only with points in the portion of the cap cut off by the hyperplane that passes through $Q$ and through the origin $O$ and

has its normal lying in the plane $POQ$; i.e., points for which $x_2 \geqq x_1 \tan \alpha$. For points in this region on the sphere and in the hyperplane $x_1 = \cos \omega$ this inequality is

$$\sin \omega \cos \theta_2 \geqq \cos \omega \tan \alpha$$

or

$$\cos \alpha \sin \omega \cos \theta_2 - \sin \alpha \cos \omega \geqq 0.$$

Now from (60) and (70)

$$\left.\frac{\partial G}{\partial \mu}\right|_{\mu=\omega} = (n-2)\frac{\cos \omega}{\sin \omega} + \left.\frac{\partial}{\partial \mu}\frac{d\varphi_2}{d\theta_2}\right|_{\mu=\omega} + (n-3)\left.\frac{\cos \theta_2}{\sin \theta_2}\frac{\partial \varphi_2}{\partial \mu}\right|_{\mu=\omega}.$$

Using (71) and (72), it is readily seen that this expression is nonnegative and monotone decreasing in $\theta_2$.

REFERENCES


1. Shannon, C. E., Communication in the Presence of Noise, Proc. IRE, *37*, January, 1949, pp. 10–21.
2. Kotel'nikov, V. A., thesis, Molotov Energy Institute, Moscow, 1947, translated by R. A. Silverman as *The Theory of Optimum Noise Immunity*, McGraw-Hill Book Co., New York, 1959.
3. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering*, John Wiley & Sons, New York, 1965.
4. Shannon, C. E., Probability of Error for Optimal Codes in a Gaussian Channel. B.S.T.J., *38*, May, 1959, pp. 611–656.
5. Gallager, R. G., A Simple Derivation of the Coding Theorem and Some Applications, IEEE Trans., *IT-11*, January, 1965, pp. 3–18.
6. Slepian, D., Bounds on Communication, B.S.T.J., *42*, May, 1963, pp. 681–707.
7. MacColl, L. A., Signalling in the Presence of Thermal Noise, I, II, and III, Bell Laboratories internal memoranda issued May 27, June 30, and September 13, 1948.
8. Coxeter, H. S. M., *Regular Polytopes*, MacMillan Co., New York, 1963.
9. Gilbert, E. N., A Comparison of Signalling Alphabets, B.S.T.J., *31*, May, 1952, pp. 504–522.
10. Balakrishnan, A. V., A Contribution to the Sphere-Packing Problem of Communication Theory, J. Math. Anal. Appl., *3*, No. 3, December, 1961, pp. 485–506.
11. Balakrishnan, A. V., Signal Selection Theory for Space Communication Channels, Chapter 1 in *Advances in Communication Systems*, A. V. Balakrishnan, editor, Academic Press, New York, 1965.
12. Weber, C. L., On Optimum Signal Selection for M-ary Alphabets with Two Degrees of Freedom, IEEE Trans., *IT-11*, No. 2, April, 1965, pp. 299–300; New Solutions to the Signal Design Problem for Coherent Channels, IEEE Trans., *IT-12*, No. 2, April, 1966, pp. 161–167.
13. Fejes Tóth, L., *Lagerungen in der Ebene auf der Kugel und im Raum*, Springer-Verlag, Berlin, 1953, pp. 137–138.
14. Gale, D., Neighborly and Cyclic Polytopes, Proc. Symposia Pure Math., *VII*, pp. 225–232, Am. Math. Soc., Providence, Rhode Island, 1963.
15. Wyner, A., On the Probability of Error for Communication in White Gaussian Noise, to appear IEEE Trans., *IT-*.

# On the Use and Performance of Error-Voiding and Error-Marking Codes

By E. O. ELLIOTT

(Manuscript received May 25, 1966)

*In contrast to payroll or inventory data, which must reach the recipient in its entirety, there is another class of data that includes radar-tracking data, remote-sensory data or control data, etc., for which the requirement of completeness is not so stringent. Error control for this class of data may be accomplished by forward-acting error-correcting codes which void or mark any detected errors that they do not correct. In order to evaluate these error-voiding methods, the error rates for such codes are estimated in this paper using the error statistics of the Alexander-Gryb-Nast study.*

*A class of 18 (about 50 percent redundant) cyclic codes capable of correcting from one to five errors and having block lengths from 15 to 47 bits is examined. Only bounded-distance decoding is evaluated, but each code is assigned each possible decoding radius up to the maximum permissible radius determined by the capability of the code. Since interleaving generally reduces error rates, the error rates for this class of codes are estimated for interleaving constants from 50 to 300 in steps of 50.*

*It is concluded that:*

*(i) If voids are permissible (at a rate of about $10^{-4}$) then low undetected-error rates may be achieved by a code capable of correcting many errors but used to correct only two or three errors. Such a code might be about 50 percent redundant and have a block length between 25 and 50 bits.*

*(ii) It is impractical to obtain low void rates. If voids are not tolerable, then retransmission is required to obtain low error rates.*

*(iii) Interleaving is more effective with codes correcting three (or more) errors than with those correcting only single or double errors.*

## I. INTRODUCTION

In contrast to payroll or inventory data, which must reach the recipient in its entirety, there is another class of data that includes radar-tracking

data, remote-sensory data or control data, etc., for which the requirement of completeness is not so stringent. The distinction between these two classes of data is fundamental in the classification of customer requirements in data transmission and the selection of appropriate error-control methods.

If the complete data message is required at the receiving station then error control must either be carried out by error detection and retransmission or by forward-acting error correction. Of these two methods the former is the more economical to achieve low error rates. However, if completeness of the transmitted message is not essential and receipt of say 99.9 percent would be satisfactory, then very inexpensive error-voiding techniques may be employed to achieve the desired low error rates. With these techniques an error-detecting code is used to detect and then void (or mark) all detectable errors. If lower void rates are desired some error correction may be introduced and the remaining error-detection capability of the code used to void or mark errors. In order to evaluate these error-voiding methods, the error rates for such codes are estimated in this paper using the error statistics of the Alexander-Gryb-Nast study.[1]

Data for which completeness is an important requirement would include payrolls, inventories, orders, sales and banking records, etc. Since accuracy is a very important factor for this type of data, an automatic retransmission error-control system would probably be required. However, one can imagine cases in which manual retransmission would suffice. Errors could be marked or voided by the error-detecting code and when errors are so indicated in a message the recipient could initiate steps to obtain the missing data. This might be tolerable in some situations if only a small fraction of the messages required this special handling.

At the other extreme are messages which need not be received in their entirety to be effective. Among these we might list radar tracking data, remote sensory data, and remote control data. Again detected errors would be marked or voided. In some cases, the discarded data might be restored by some extrapolation or interpolation with neighboring blocks of the presumed error-free data. In other cases, it might suffice to operate with just the nonvoided blocks of data. For these procedures to work it is of course necessary that the void rate be low enough. The void rate itself however is not the sole factor determining the feasibility of the system. The time distribution of voids may also be very important. For example, with radar-tracking data a void rate of $10^{-4}$ (words/word) might be tolerable in itself but if it were realized on a channel on which

the voids tended to occur in runs or bunches it might not be tolerable since the tracking system cannot operate if too long a stretch of data is missing. This paper is concerned only with void rates and does not treat the time distribution of voids.

The splitting of a code's function to achieve both error correction and detection is accomplished by noting the distance of the received word from the nearest code word. If this distance is less than or equal to a given number which is called the employed correction radius then the received word is decoded as that nearest code word, otherwise it is not decoded and a detected error is announced. This is the method of bounded-distance decoding. Although there are other methods of combining error correction and detection, this one is considered here because several practical decoding algorithms conform to it.

The codes are also evaluated over a range of degrees of interleaving, because, if a given code is interleaved on a burst channel its performance improves. Interleaving may be thought of as though it were accomplished by reading the encoded data into a rectangular array row by row and then sending it on line column by column. The length of a row is the block length of the code. The number of rows is the interleaving constant $t$; two adjacent bits of an originally encoded block would be sent on line with $t - 1$ other bits between them.

This memorandum examines the effect that block length, redundancy, correction radius, and interleaving have on undetected error rates and void rates over the switched telephone network. Specifically, a class of 18 cyclic codes with block lengths ranging from 15 to 47 bits is examined. Among these are codes capable of correcting from one to five errors. Although a variety of redundancies is represented, codes with about 50 percent redundancy predominate. Using data from the Alexander-Gryb-Nast study, the error rates for this class of codes are estimated for each permissible correction radius with no interleaving and with interleaving with constants from 50 to 300 in steps of 50. A number of practical means for implementing many of these error-control systems are available. For this reason, the present investigation aims at giving a useful qualitative insight into the role of bounded-distance decoding for error control on the switched telephone network.

For a more complete understanding of error control, it would be necessary to consider alternative methods of decoding such as burst decoding, threshold decoding, etc. Also error statistics for other channels and modes of communication should be considered. This awaits further development of analytical techniques and the availability of additional error data.

## II. ESTIMATING ERROR RATES FOR BOUNDED-DISTANCE DECODING

A group code is commonly specified by the pair $(n,k)$ where $n$ is the block length of the code and $k$ is the number of information bits in a code word. When bounded-distance decoding is employed, the correction radius $a$ is added to this pair and the code is specified by the triplet $(n,k,a)$. Of course, $a$ is less than or equal to the maximum error-correcting capability $e$ of the code. Thus, if $x$ is the transmitted word and $y$ is the received word then if $y$ is at distance less than or equal to $a$ from some code word $z$, $y$ is decoded as $z$. If $z \neq x$ then an undetected error results, and if $y$ is not within distance $a$ from any code word a detected error results. To estimate the probabilities $P_E$ and $P_D$ of these two events (undetected error and detected error) the method of Ref. 2 is employed as it was in Ref. 3 to study permutation decoding which is a special case of bounded distance decoding.

Because of the perfect distance symmetries between the words of a group code, the probabilities $P_E$ and $P_D$ do not depend on which code word is transmitted. Therefore, to calculate $P_E$ and $P_D$ and simplify matters we assume the all zero word $\theta$ is transmitted. Let $C_a(m)$ be the total number of words of weight $m$ which are at a distance less than or equal to $a$ from some code word. As in Ref. 2 let $P(m,n)$ be the total probability that $m$ errors occur in a transmitted block of $n$ bits so that $P(m,n)/\binom{n}{m}$ is an approximation to the probability that a particular pattern of $m$ errors occur. Then, assuming $\theta$ is transmitted we see that, as an approximation,

$$P_E \cong \sum_{m=a+1}^{n} C_a(m) \frac{P(m,n)}{\binom{n}{m}}. \tag{1}$$

Similarly, if $D_a(m)$ is the total number of words of weight $m$ at a distance greater than $a$ from any code word then

$$P_D \cong \sum_{m=a+1}^{n} D_a(m) \frac{P(m,n)}{\binom{n}{m}}. \tag{2}$$

Clearly $D_a(m) = \binom{n}{m} - C_a(m)$ and now the problem is to obtain $C_a(m)$.

In Ref. 3 a formula is given for $C(m,j)$, the number of words of weight $m$ which are at a distance $j$ ($j \leqq e$) from some code word, and since

$$C_a(m) = \sum_{j=0}^{a} C(m,j)$$

the desired numbers are thus obtainable. Unfortunately, the formula in Ref. 3 involves a summation with terms of alternating sign which requires triple precision programming to obtain satisfactory accuracy on the computer. As a consequence of this requirement, the following alternate and more direct formula for $C(m,j)$ was derived. With it, double precision FORTRAN programming suffices.

$$C(m,j) = \sum_{i=0}^{j} w(m + 2i - j) \binom{m + 2i - j}{i} \binom{n - m - 2i + j}{j - i} \quad (3)$$

where $w(r)$ represents the number of code words of weight $r$. It is obtained from simple combinatorial considerations as follows.

Suppose $x$ is a code word of weight $r$ and $y$ is a word of weight $m$ and let $i$ be the number of bit positions in which $x$ is 1 and $y$ is 0, and $l$ be the number of bit positions in which $x$ is 0 and $y$ is 1. Let $j = i + l$ so $j$ is the distance between $x$ and $y$. Then $m = (r - i) + l = r + j - 2i$ and $i = (r + j - m)/2$. The total number of possible $y$'s of weight $m$ is then given by

$$\binom{r}{i} \binom{n - r}{j - i}.$$

Considering that each code word of weight $m + 2i - j$ $(= r)$ therefore has

$$\binom{m + 2i - j}{i} \binom{n - m - 2i + j}{j - i}$$

distinct words of weight $m$ at distance $j$ $(j \leqq e)$ from it and clearly $m - j \leqq r \leqq m + j$, i.e., $0 \leqq i \leqq j$, (3) then follows.

## III. ERROR RATES FOR A SAMPLE COLLECTION OF CODES

Cyclic codes or shortened cyclic codes have received a great deal of attention in the field of error control because of the ease in their implementation. For this reason a collection of 18 cyclic codes for which the spectral functions $w(r)$ were readily available was chosen for this study. Most of these codes and spectra are given in Ref. 3. Those which were not in Ref. 3 were included so that a wider range of redundancies would be represented. The codes are listed in Table I which gives their block length $n$, dimension $k$, minimum distance $d$, and maximum error-correcting capability $e$.

Using (1) and (2) and the $P(m,n)$ values from the Alexander-Gryb-Nast study, the probabilities $P_E$ and $P_D$ were estimated for each of these codes. Samples of the results are shown in Figs. 1, 2, and 3.

TABLE I—LIST OF CYCLIC CODES

| $n$ | $k$ | $d$ | $e$ |
|---|---|---|---|
| 15 | 11 | 3 | 1 |
| 15 | 10 | 4 | 1 |
| 15 | 7 | 5 | 2 |
| 15 | 6 | 6 | 2 |
| 15 | 5 | 7 | 3 |
| 15 | 4 | 8 | 3 |
| 15 | 2 | 10 | 4 |
| 17 | 9 | 5 | 2 |
| 17 | 8 | 6 | 2 |
| 21 | 12 | 5 | 2 |
| 21 | 11 | 6 | 2 |
| 23 | 12 | 7 | 3 |
| 23 | 11 | 8 | 3 |
| 31 | 21 | 5 | 2 |
| 31 | 20 | 6 | 2 |
| 31 | 16 | 7 | 3 |
| 31 | 15 | 8 | 3 |
| 47 | 24 | 11 | 5 |

Fig. 1 shows the effect which the correction radius has on error rates for two of the codes. The undetected error rate is noted to decrease about one order of magnitude for each unit decrease in correction radius. Also, the detected error rate, which is about $10^{-4}$, is rather insensitive to the correction radius.

In Fig. 2 the undetected error rate is plotted as a function of the efficiency of a 15-bit code. Each order of magnitude improvement in the error rate requires an increase of three redundant bits (which is 20 percent of the block length). An examination of the error rates for the codes with block length 31 (not shown here) reveals that the same change of three redundant bits is required to achieve an order of magnitude improvement with this longer block length code.

Error rates of double-error-correcting codes of about 50 percent redundancy are presented in Fig. 3 as a function of block length. Again the detected error rate is not a very sensitive parameter while the undetected error rate ranges over many orders of magnitude. Quite acceptable error rates are attainable with block lengths not much greater than 25 bits.

## IV. THE EFFECT OF INTERLEAVING ON ERROR RATES

Through interleaving (with constant $t$), the bits of a code word are separated when transmitted on line so each pair of adjacent bits have

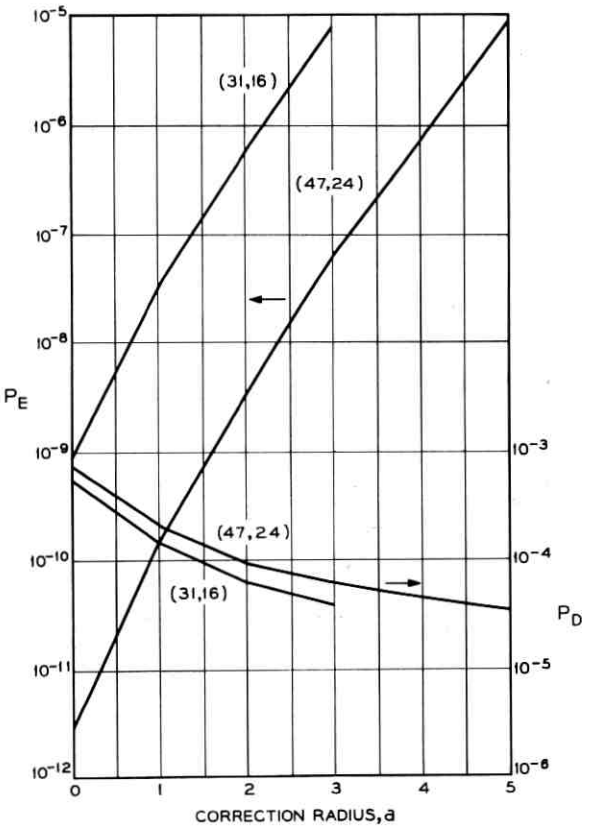



Fig. 1 — Error rates vs correction radius (no interleaving).

$t - 1$ bits from other code words between them. The effect of this is to decrease the error vulnerability dependence between the bits of a code word so that the interleaved channel is less of a burst channel and more like a memoryless channel. To examine the effect interleaving has on error control with random error-correcting codes, the $P_t(m,n)$ probabilities were approximated for interleaving constants $t = 50$, 100, 150, 200, 250, and 300, and the error rates for the 18 codes were estimated as in the previous section. To approximate the $P_t(m,n)$ values first the error autocorrelation function $a_t(k)$ of the interleaved channels is obtained from a smoothed version of the autocorrelation function $a(s)$ of the Alexander-Gryb-Nast data through the relation

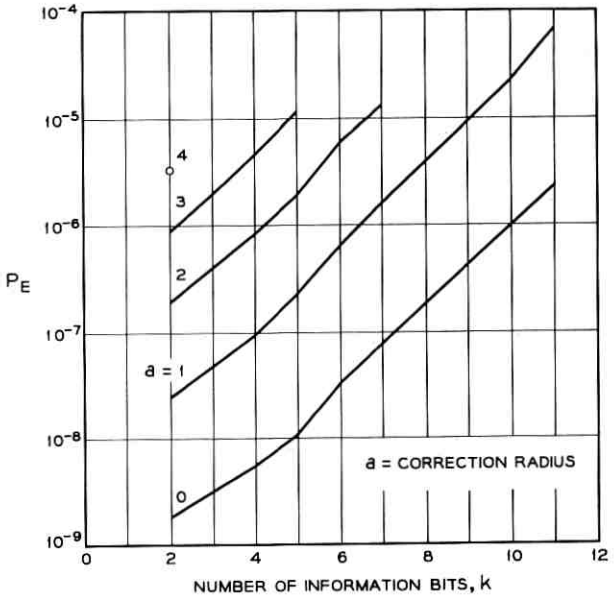



Fig. 2 — Error rates of $(15, k)$ codes.

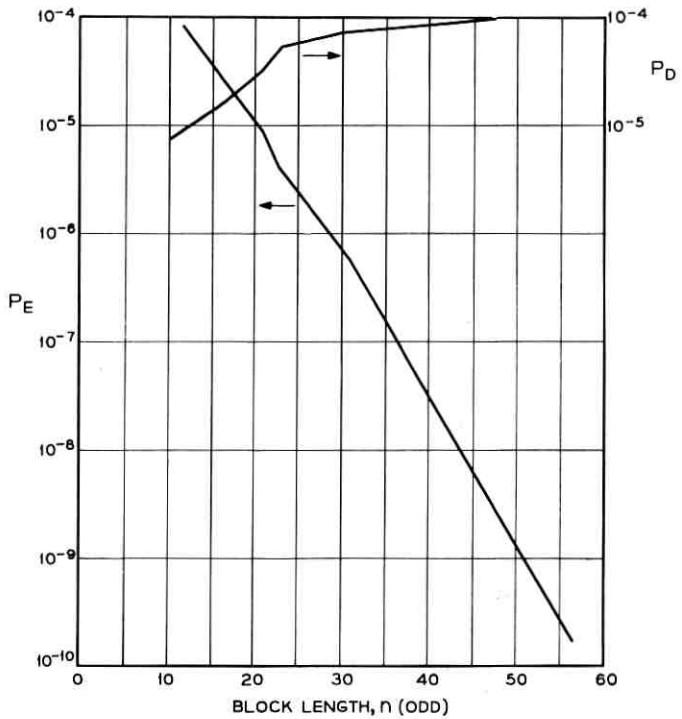



Fig. 3 — Error rates of double error-correcting $\{n, [(n + 1)/2]\}$ codes.

$$a_t(s) = a(t \cdot s).$$

Then the interleaved channels are assumed to have the property that the lengths of the error-free gaps before and after an error are independently distributed — i.e., the errors form a renewal process. (This seems to be a reasonable assumption since interleaving breaks down the memory in the error process. Its accuracy will be discussed below.) From the autocorrelation functions $a_t(s)$ the gap-length distributions $P_t(s)$ may then be calculated by the relations between them which are given in Ref. 4, and finally the $P_t(m,n)$ values are obtained from the $P_t(s)$ by the recursive methods of Ref. 4.

The undetected error rates of some codes used for forward acting error correction only are shown in Fig. 4 as a function of the interleaving constant $t$. There it is seen that interleaving is most effective on codes correcting four and five errors and is of only modest value on the codes correcting two or three errors. In fact, for the (31, 21) code it would

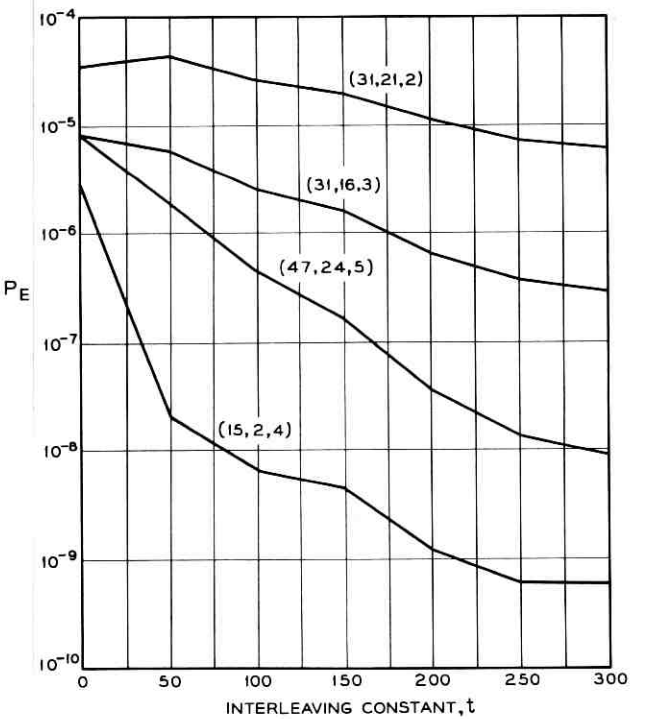



Fig. 4 — Error rate vs interleaving constant.

appear that the error rate increases in the range $t = 1 - 50$. This may result from the fact that the mathematical methods for obtaining $P_t(m,n)$ were different for $t = 1$ and $t = 50$ since the renewal assumption was not made in the case of no interleaving ($t = 1$). The renewal assumption would appear to be more appropriate the larger the interleaving constant $t$ becomes so our estimates of error rates would be more accurate for the larger values of $t$.

The conclusion to be drawn from Fig. 4 is that it takes a powerful code interleaved extensively to provide a low error rate. The price, in redundancy, interleaving or decoding complexity, paid to do this is high and it would take special circumstances to justify it.

In Fig. 5, block-error rates are plotted against block length to further show the effect of interleaving. The relationship between error rates and block length is linear and the slope is determined by the amount of interleaving. The equivalent memoryless channel is also shown for comparison. It appears that a very considerable amount of interleaving would be required to approach the memoryless channel.

Although it is not shown here, the detected error rates for codes

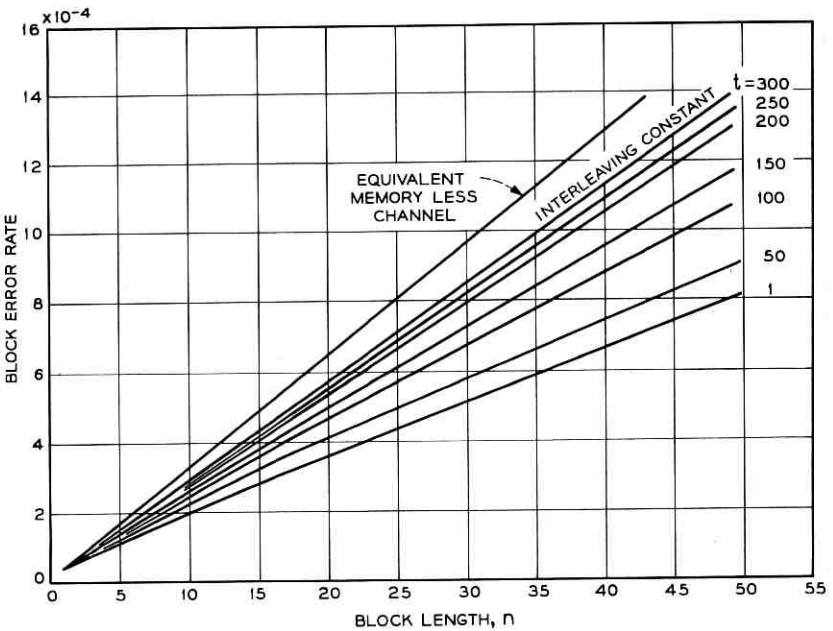



Fig. 5 — Block-error rate vs block length at various interleaving constant.

correcting several errors decrease a few orders of magnitude as the interleaving increases over the range considered here.

We have considered only random error-correcting codes. In Ref. 5, a burst-correcting code was studied at two levels of interleaving as part of an error-control experiment.

## V. CONCLUSIONS

Bounded-distance decoding has been examined as a means of utilizing codes to do both error-correction and detection on the switched telephone network with existing data sets. Data containing detected errors would be either voided or marked.

If voids in the received data are permissible (at a rate of about $10^{-4}$) then low undetected-error rates may be achieved by a code capable of correcting many errors but used to correct only two or three errors. Such a code might be about 50 percent redundant and have a block length between 25 and 50 bits.

The void rate is rather insensitive to correction radius, block length, and to a lesser extent, interleaving. It decreases with increasing correction radius, increases with increasing block length and decreases with increasing interleaving (for multi-error-correcting codes).

Interleaving is more effective with codes correcting three (or more) errors than those correcting only single or double errors.

If voids are not tolerable then retransmission is indicated as the means to obtain low error rates. A powerful interleaved and highly redundant error-correcting code is required to obtain low error rates. It would probably be called for in only very special cases.

Further work should be undertaken to investigate other methods of decoding codes, such as threshold or burst decoding, in order to gain a more complete insight in the realm of practical error control systems.

## REFERENCES

1. Alexander, A. A., Gryb, R. M., and Nast, D. W., Capabilities of the Telephone Network for Data Transmission, B.S.T.J., *39*, May, 1960.
2. Elliott, E. O., Estimates of Error Rates for Codes on Burst-Noise Channels,
3. MacWilliams, Jessie, Permutation Decoding of Systematic Codes, B.S.T.J., *43*, January, 1964, p. 485.
4. Elliott, E. O., A Model of the Switched Telephone Network for Data Communications, B.S.T.J., *44*, January, 1965.
5. Weldon, E. J., Performance of a Forward-Acting Error-Control System on the Switched Telephone Network, B.S.T.J., *45*, May, 1966.

# Duration of Fades Associated with Radar Clutter

By A. J. RAINAL

(Manuscript received June 2, 1966)

*The fluctuating envelope of the pulse-to-pulse radar echoes from a range cell consisting of a stationary target along with many independent, randomly moving scatterers is assumed to behave like a stationary Rayleigh process. In radar terminology this fluctuating or fading envelope of the pulse-to-pulse radar echoes is called signal plus clutter. The envelope of the pulse-to-pulse radar echoes may fade below some critical threshold level for a duration such that the performance of the radar becomes unsatisfactory. Theoretical approximations for the probability densities of both the duration of fades and the interval between fades of the underlying Rayleigh process are presented in graphs for various threshold levels and various signal-to-clutter power ratios. The corresponding exact results are at present unknown. The results of this paper apply to all other fields of science and technology for which a stationary Rayleigh process characterizes a fading phenomenon.*

## I. INTRODUCTION

Consider a pulsed radar system "viewing" a range cell consisting of a stationary target along with many independent, randomly moving scatterers as shown in Fig. 1. Each received echo consists of the vector sum of all the elementary echoes originating from within the range cell. The contributions from the randomly moving scatterers arrive at the radar receiver with random phases. As a result, each received echo will consist of a steady signal, from the stationary target, plus a Gaussian perturbation. Accordingly, samples of the envelope of the pulse-to-pulse radar echoes can be considered as samples of an underlying Rayleigh process. Thus, the effect of the randomly moving scatterers is to cause the envelope of the pulse-to-pulse radar echoes to fluctuate or fade in an irregular manner. The envelope of the pulse-to-pulse radar echoes may fade below some critical threshold level for a duration such that the performance of the radar becomes unsatisfactory.

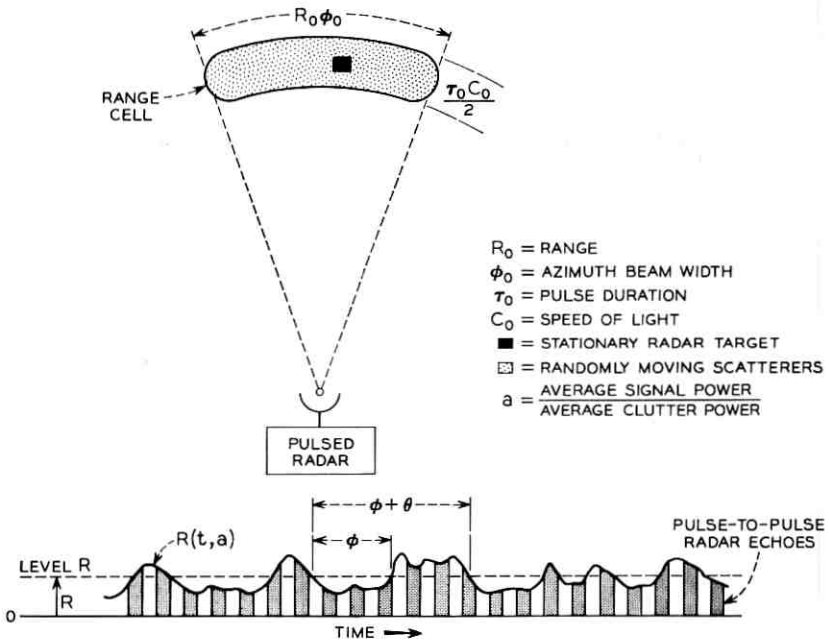



Fig. 1 — A model for studying the duration of fades associated with signal plus clutter. $R(t,a)$ represents the envelope of the pulse-to-pulse radar echoes from the range cell. At the level $R$, $\phi$ and $\phi + \theta$ represent the duration of a fade and the interval between fades, respectively.

In radar terminology the fluctuating or fading envelope of these pulse-to-pulse radar echoes is called signal plus clutter. Classical discussions of signal plus clutter were given by H. Goldstein and A. J. F. Siegert and can be found in Refs. 1 and 2. A well-known example of signal plus clutter is the envelope of the pulse-to-pulse radar echoes from a target surrounded by a great deal of "chaff". Some other examples may be the envelope of the pulse-to-pulse radar echoes from the aurora, the ionosphere, the sea, the ground, meterological precipitation, and a hypersonic object during reentry of the earths atmosphere.

A natural assumption for studying the duration of fades associated with signal plus clutter is that the random process underlying the fading is a Rayleigh process. However, only a few theoretical results are available concerning the duration of fades associated with Rayleigh processes. Thus, one is often unable to determine how well a Rayleigh process actually characterizes the duration of fades observed experimentally.

The purpose of this paper is to present some additional theoretical

results which characterize approximately the duration of fades one would expect when the random process underlying the fading is indeed a Rayleigh process. We shall assume that the envelope of the pulse-to-pulse radar echoes behaves like the Rayleigh process $R(t,a)$ sketched in Fig. 1. $R(t,a)$ represents the envelope of a stationary random process consisting of a sinusoidal signal of amplitude $\sqrt{2a}$ and frequency $f_o$ plus a Gaussian process of unit variance having a narrowband power spectral density $W_b(f - f_o)$ which is symmetrical about $f_o$. We assume that the radar pulse repetition frequency is several times greater than the bandwidth of the Rayleigh process $R(t,a)$ in order that an adequate number of radar echoes are used to form $R(t,a)$. Also, we shall assume that the variance of the receiver noise is negligible in comparison with the variance of the clutter.

Using notation consistent with Refs. 3, 4, and 5 we shall present theoretical approximations for the following probability functions for arbitrary signal-to-clutter power ratio "$a$":

(*i*) $P_o^-(\tau,R,a)$, the probability density of the duration of a fade of the Rayleigh process below the level $R$.
(*ii*) $P_1(\tau,R,a)$, the probability density of the interval between fades of the Rayleigh process below the level $R$.
(*iii*) $F_o^-(\tau,R,a)$, the probability that the duration of a fade of the Rayleigh process below the level $R$ lasts longer than $\tau$.
(*iv*) $F_1(\tau,R,a)$, the probability that the interval between fades of the Rayleigh process below the level $R$ lasts longer than $\tau$.

The model considered in this paper also has application in the study of the duration of fades in radio transmission. In fact Rice[6,7] led the way by analyzing the duration of fades in radio transmission assuming that the underlying random process was $R(t,0)$.

## II. INTEGRAL EQUATIONS AND EXPECTATIONS

Let us define the following auxiliary probability functions for arbitrary level $R$ and arbitrary signal-to-clutter power ratio "$a$":

(*i*) $Q^-(\tau,R,a)\,d\tau$, the conditional probability that an upward level-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given a downward level-crossing at $t$.
(*ii*) $[U(\tau,R,a) - Q(\tau,R,a)]\,d\tau$, the conditional probability that an upward level-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given an upward level-crossing at $t$.

Explicit expressions for these auxiliary probability functions were presented in Ref. 5.

Approximate theoretical results for $P_o^-(\tau,R,a)$ and $P_1(\tau,R,a)$ are given by the following integral equations:

$$P_o^-(\tau,R,a) = Q^-(\tau,R,a) - P_o^-(\tau,R,a) * [U(\tau,R,a) - Q(\tau,R,a)] \quad (1)$$

$$\begin{aligned} P_1(\tau,R,a) = {} & [U(\tau,R,a) - Q(\tau,R,a)] \\ & - P_1(\tau,R,a) * [U(\tau,R,a) - Q(\tau,R,a)] \end{aligned} \quad (2)$$

where $*$ denotes the convolution operator, that is,

$$f * g \equiv \int_{-\infty}^{\infty} f(t)g(\tau - t)dt.$$

Equations (1) and (2) were derived in Ref. 3 by applying McFadden's[8] "quasi-independence" idea to the Rayleigh process $R(t,a)$. Also, by definition we have that

$$F_o^-(\tau,R,a) = 1 - \int_0^{\tau} P_o^-(\tau,R,a)d\tau \quad (3)$$

and

$$F_1(\tau,R,a) = 1 - \int_0^{\tau} P_1(\tau,R,a)d\tau. \quad (4)$$

The exact expectations $E_o^-(\tau,R,a)$ and $E_1(\tau,R,a)$ associated with the respective densities $P_o^-(\tau,R,a)$ and $P_1(\tau,R,a)$ can be computed from the following equations:

$$\begin{aligned} E_o^-(\tau,R,a) &\equiv \frac{Pr\{R(t,a) < R\}}{N_R} \\ &= \frac{\sum_{n=1}^{\infty} (R^{2n}/2^n n!)_1F_1\ (n + \tfrac{1}{2};\ 2n + 1;\ -2R\sqrt{2a})}{\sqrt{(\beta/2\pi)}\ R_1F_1(\tfrac{1}{2};\ 1;\ -2R\sqrt{2a})} \end{aligned} \quad (5)$$

$$E_1(\tau,R,a) \equiv \frac{1}{N_R} = \frac{\exp\ [(R - \sqrt{2a})^2/2]}{\sqrt{(\beta/2\pi)}\ R_1F_1\ (\tfrac{1}{2};\ 1;\ -2R\sqrt{2a})} \quad (6)$$

where

$$\beta = (2\pi)^2 \int_0^{\infty} W_b(f - f_o)(f - f_o)^2\ df$$

$W_b(f - f_o)$ = narrowband power spectral density of the Gaussian process involved in the definition of $R(t,a)$

$$a = \frac{\text{average signal power}}{\text{average clutter power}}$$

$${}_1F_1(\alpha;\beta;x) = \text{the confluent hypergeometric function}$$

$$= 1 + \frac{\alpha}{\beta}x + \frac{\alpha(\alpha+1)}{\beta(\beta+1)}\frac{x^2}{2!} + \cdots .$$

$\Pr\{\ \}$ = probability of the event inside the brace

$N_R$ = average number of upward (or downward) crossings of the level $R$ per second.

Equations (5) and (6) were developed in Ref. 4, and they follow directly from some well-known results reported by Rice and Bennett. Each ${}_1F_1$ function appearing in (5) and (6) can be expressed in terms of a Bessel function of imaginary argument.

Thus, with the aid of a digital computer one can compute theoretical approximations for the probability functions of interest in this paper along with the exact theoretical expectations given by (5) and (6).

## III. RESULTS FOR A GAUSSIAN AUTOCORRELATION FUNCTION

In order to define the Rayleigh process $R(t,a)$ underlying the fading phenomenon we need to specify both $W_b(f - f_o)$ and the signal-to-clutter ratio "$a$". The normalized autocorrelation function $m(\tau)$ associated with $W_b(f - f_o)$ is given by

$$m(\tau) = \int_0^\infty W_b(f - f_o) \cos 2\pi(f - f_o)\tau df. \tag{7}$$

Thus, $m(\tau)$, rather than $W_b(f - f_o)$, can be used to define the Rayleigh process $R(t,a)$ underlying the fading phenomenon. Notice that $\beta$ appearing in (5) and (6) is merely $-m''(0)$. The primes denote differentiations with respect to $\tau$.

Ref. 1 points out that it is convenient to measure the normalized autocorrelation function of the fluctuating low frequency power $P(t) = R^2(t,a)$ and denotes this normalized autocorrelation function by $\rho(P,\tau)$. In explicit terms $\rho(P,\tau)$ is defined as

$$\rho(P,\tau) = \frac{E\{[P(t+\tau) - EP(t)][P(t) - EP(t)]\}}{Var\ P(t)} = \frac{EP(t+\tau)P(t) - E^2P(t)}{Var\ P(t)} \tag{8}$$

where

$$E = \text{Expectation}$$

$$Var = \text{Variance}.$$

Refs. 1 and 2 relate $m(\tau)$ and $\rho(P,\tau)$ as follows:

$$m(\tau) = \sqrt{a^2 + (1 + 2a)\rho(P,\tau)} - a. \tag{9}$$

Ref. 1 also points out that the appropriate value of "$a$" can be estimated by measuring the probability density of $P(t)$ and comparing the result with the theoretical probability density of $P(t)$. Thus, (9) indicates that the Rayleigh process $R(t,a)$ underlying the fading phenomenon can also be defined from measurements of the normalized autocorrelation function $\rho(P,\tau)$ and the value of "$a$."

For purposes of computation we shall take $W_b(f - f_o)$ and $m(\tau)$ as

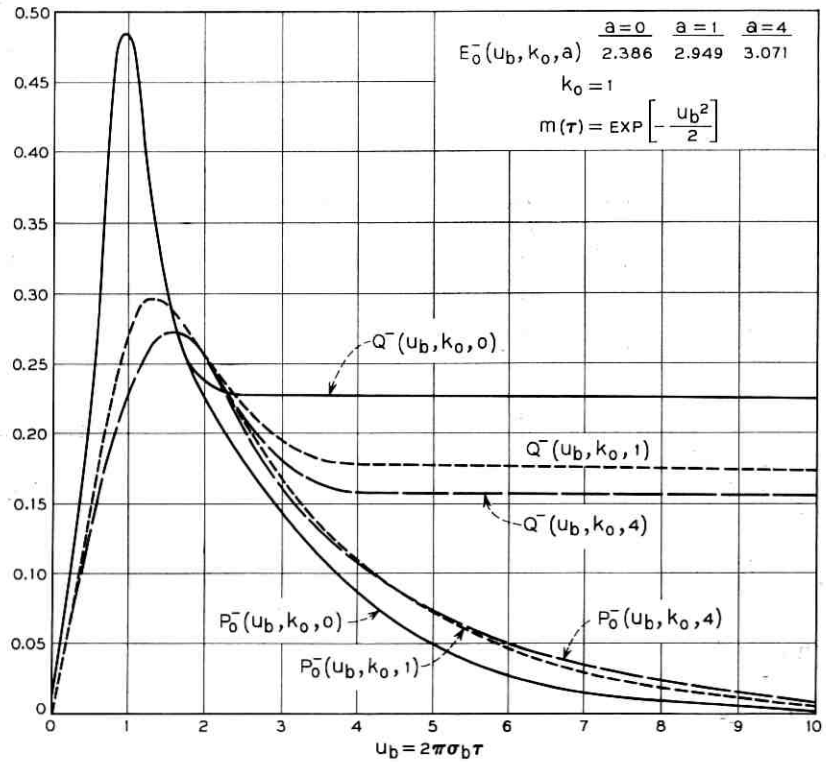



Fig. 2 — $P_0^-(u_b, k_0, a)$ is the probability density of the duration of a fade of the Rayleigh process below the normalized level $k_0$. The autocorrelation function of the Gaussian process involved in the definition of the Rayleigh process is $m(\tau)$ and the signal-to-clutter power ratio equals "$a$."

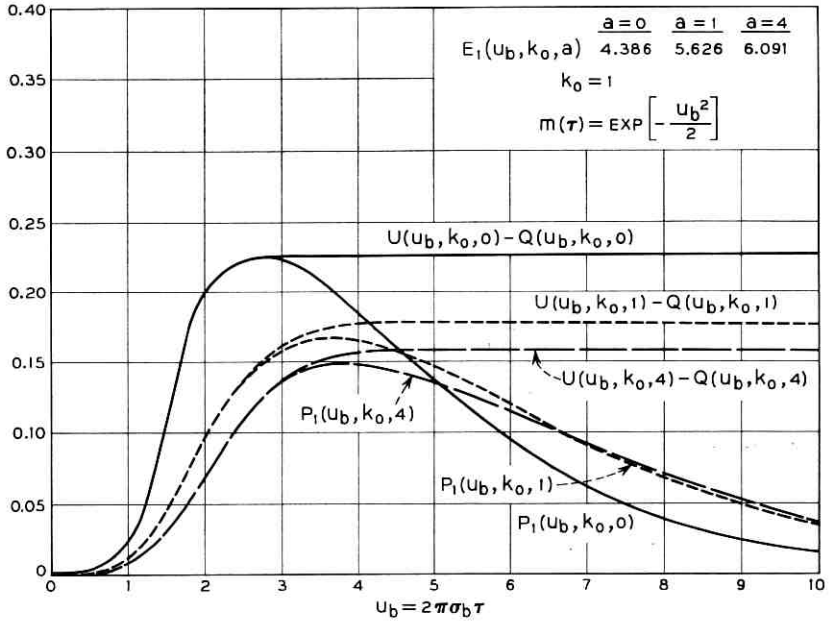



Fig. 3 — $P_1(u_b , k_0 , a)$ is the probability density of the interval between fades of the Rayleigh process below the normalized level $k_0$. The autocorrelation function of the Gaussian process involved in the definition of the Rayleigh process is $m(\tau)$ and the signal-to-clutter power ratio equals "$a$."

follows:

$$W_b(f - f_o) = \frac{1}{\sigma_b\sqrt{2\pi}} \exp\left[\frac{-(f - f_o)^2}{2\sigma_b^2}\right] \tag{10}$$

and

$$m(\tau) = \exp\left[\frac{-(2\pi\sigma_b\tau)^2}{2}\right]. \tag{11}$$

This particular choice tends to characterize the radar clutter fluctuations observed experimentally.[1,2] From (11) we see that it is convenient to define normalized time as $u_b = 2\pi\sigma_b\tau$.

For the experimenter it is convenient to normalize the threshold level with respect to the average value, $ER(t,a)$, of the Rayleigh process. We shall consider three such normalized levels $k_o$

$$\frac{R}{ER(t,a)} \equiv k_o = 1, \sqrt{\frac{2}{\pi}}, \frac{1}{\sqrt{2\pi}}. \tag{12}$$

The expectation $ER(t,a)$ was derived by Rice[9] and is given by

$$ER(t,a) = \sqrt{\frac{\pi}{2}}\,{}_1F_1\left(-\frac{1}{2}; 1; -a\right). \tag{13}$$

When $a = 0$ we have that $R = \sqrt{(\pi/2)}$, 1, $\frac{1}{2}$. These latter two values of $R$ were also considered by Rice[7] for the case $a = 0$.

Figs. 2 through 10 present the computed results for $a = 0$, 1, 4, and $k_o = 1, \sqrt{2/\pi}, 1/\sqrt{2\pi}$. The numerical evaluation of $Q^-(\tau,R,a)$ and $U(\tau,R,a) - Q(\tau,R,a)$ was carried out by using Simpson's rule. Integral equations (1) and (2) were solved numerically by using the trapezoidal rule. All results are plotted with respect to normalized time $u_b$. The corresponding experimental results for $k_o = 1$ and $a = 0, 1, 4$ were presented in Ref. 4, and they agree well with the approximate theoretical results presented in Figs. 2, 3, and 4.

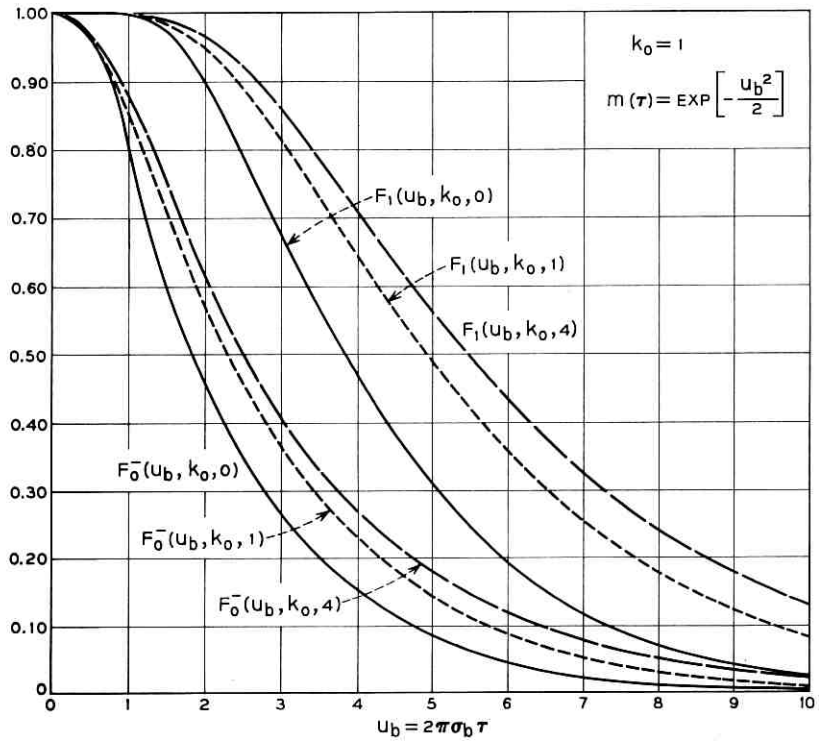



Fig. 4 — $F_0^-(u_b, k_0, a)$ is the probability that the duration of a fade of the Rayleigh process below the normalized level $k_0$ lasts longer than $u_b$. $F_1(u_b, k_0, a)$ is the probability that the interval between fades of the Rayleigh process below the normalized level $k_0$ lasts longer than $u_b$.

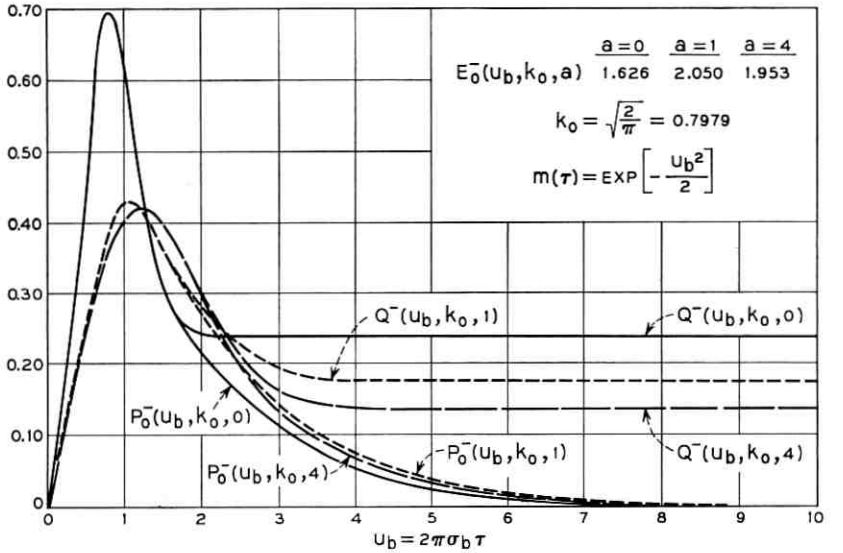



Fig. 5 — $P_0^-(u_b, k_0, a)$ is the probability density of the duration of a fade of the Rayleigh process below the normalized level $k_0$. The autocorrelation function of the Gaussian process involved in the definition of the Rayleigh process is $m(\tau)$ and the signal-to-clutter power ratio equals "$a$".

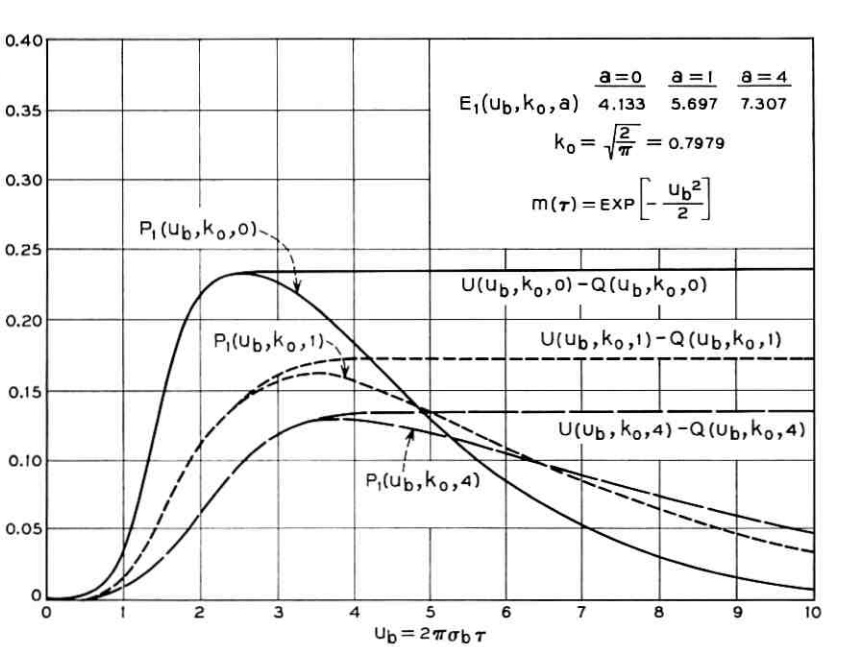



Fig. 6 — $P_1(u_b, k_0, a)$ is the probability density of the interval between fades of the Rayleigh process below the normalized level $k_0$. The autocorrelation function of the Gaussian process involved in the definition of the Rayleigh process is $m(\tau)$ and the signal-to-clutter power ratio equals "$a$".

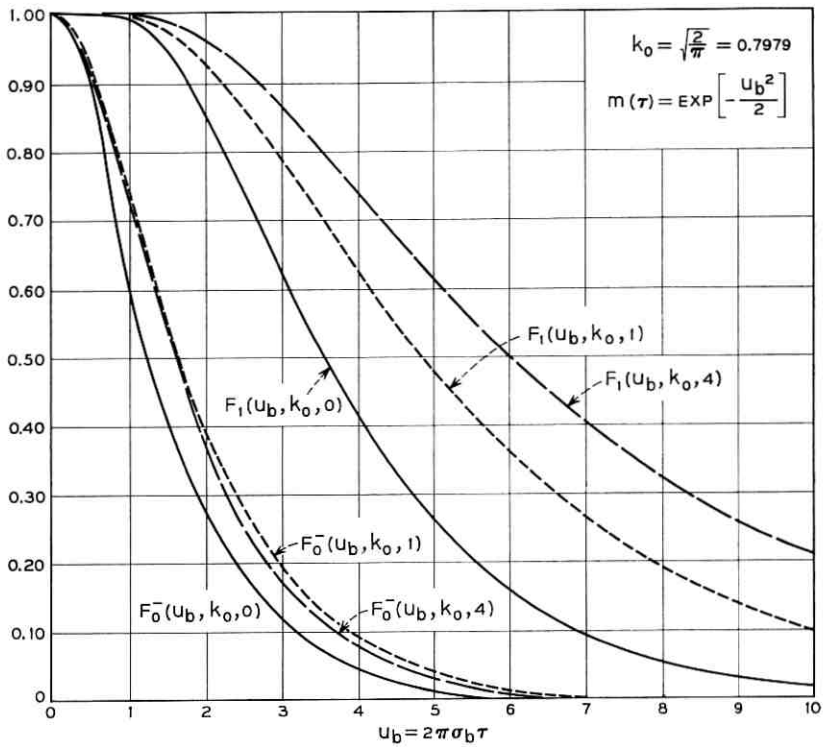



Fig. 7 — $F_0^-(u_b, k_0, a)$ is the probability that the duration of a fade of the Rayleigh process below the normalized level $k_0$ lasts longer than $u_b$. $F_1(u_b, k_0, a)$ is the probability that the interval between fades of the Rayleigh process below the normalized level $k_0$ lasts longer than $u_b$.

For deep fades and large signal-to-clutter power ratio "$a$", one would expect $P_o^-(\tau,R,a)$ to approach a Rayleigh probability density. For as "$a$" gets large the Rayleigh process $R(t,a)$ tends to behave much like a Gaussian process, see (3.6) of Rice,[9] and the durations of deep fades of Gaussian processes are known to be characterized by a Rayleigh probability density.[7,10] Figs. 8 and 10 show that this is, approximately, the case when $k_o = 1/\sqrt{2\pi}$ and $a = 4$. Thus, for $k_o \leqq 1/\sqrt{2\pi}$ and $a \geqq 4$ we have the following approximate results:

$$P_o^-(\tau,R,a) = \frac{\pi}{2E_o^-}\left(\frac{\tau}{E_o^-}\right)\exp\left[-\frac{\pi}{4}\left(\frac{\tau}{E_o^-}\right)^2\right] \tag{14}$$

and

$$F_o^-(\tau,R,a) = \exp\left[-\frac{\pi}{4}\left(\frac{\tau}{E_o^-}\right)^2\right]. \tag{15}$$

The value of $E_o^-$ appearing in (14) and (15) is given by (5) with $R = k_oER(t,a)$.

Equations (14) and (15) are useful approximations when $k_o$ is small and "$a$" is large for an arbitrary normalized autocorrelation function $m(\tau)$ such that $m'''(0^+) = 0$, although we have been treating the restrictive Gaussian autocorrelation function defined by (11). The condition $m'''(0^+) = 0$ leads to $Q^-(0^+,R,a) = 0$, and thus the approximation given by (14) is exact at $\tau = 0^+$. As a partial check on this generalization we also verified that (14) and (15) begin to be useful approximations when $k_o = 1/\sqrt{2\pi}$, $a = 4$ for the normalized autocorrelation functions

$$m(\tau) = \frac{\sin 2\pi f_c\tau}{2\pi f_c\tau} \tag{16}$$

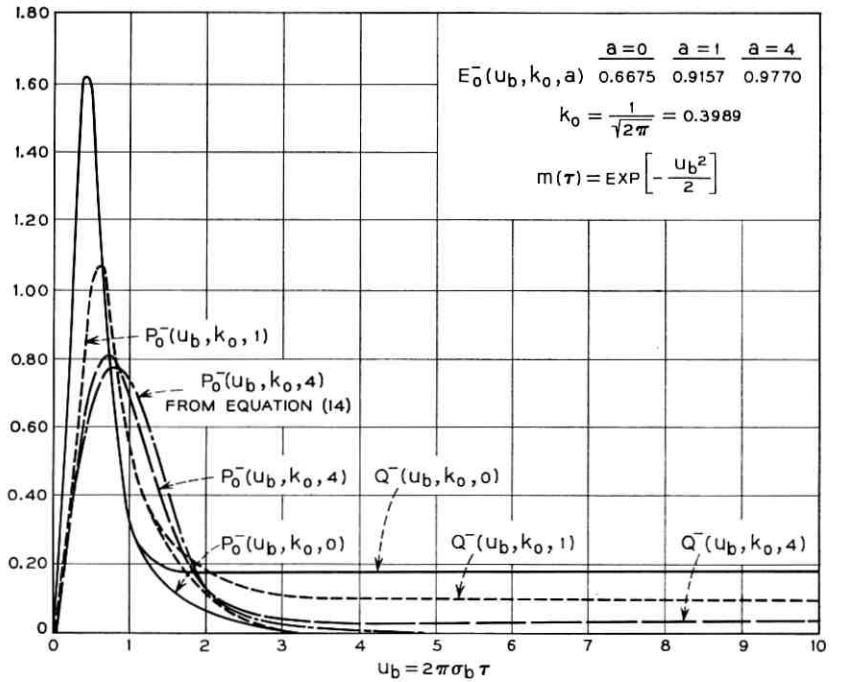



Fig. 8 — $P_0^-(u_b, k_0, a)$ is the probability density of the duration of a fade of the Rayleigh process below the normalized level $k_0$. The autocorrelation function of the Gaussian process involved in the definition of the Rayleigh process is $m(\tau)$ and the signal-to-clutter power ratio equals "$a$."

and

$$m(\tau) = \left[1 + \omega_2 \mid \tau \mid + \frac{(\omega_2 \tau)^2}{3}\right] \exp\ (-\omega_2 \mid \tau \mid). \tag{17}$$

Equation (16) corresponds to an ideal bandpass power spectral density $W_b(f - f_o)$ given by

$$W_b(f - f_o) = \begin{cases} (2f_c)^{-1} & \text{for } f_o - f_c \leqq f \leqq f_o + f_c \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Equation (17) corresponds to a power spectral density $W_b(f - f_o)$ given by

$$W_b(f - f_o) = \frac{8/(3\pi f_2)}{\left[1 + \left(\frac{f - f_o}{f_2}\right)^2\right]^3}, \tag{19}$$

where

$$\omega_2 = 2\pi f_2 .$$

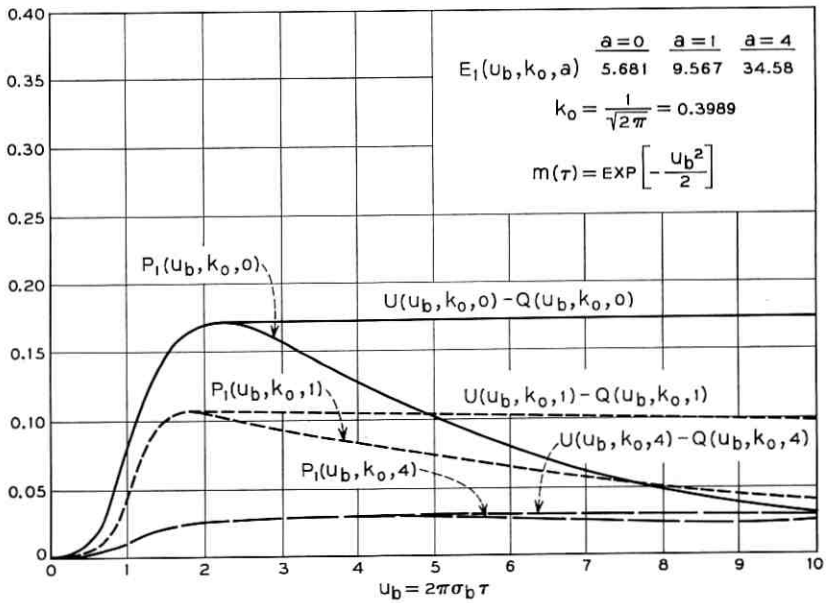



Fig. 9 — $P_1(u_b, k_0, a)$ is the probability density of the interval between fades of the Rayleigh process below the normalized level $k_0$. The autocorrelation function of the Gaussian process involved in the definition of the Rayleigh process is $m(\tau)$ and the signal-to-clutter power ratio equals "$a$".

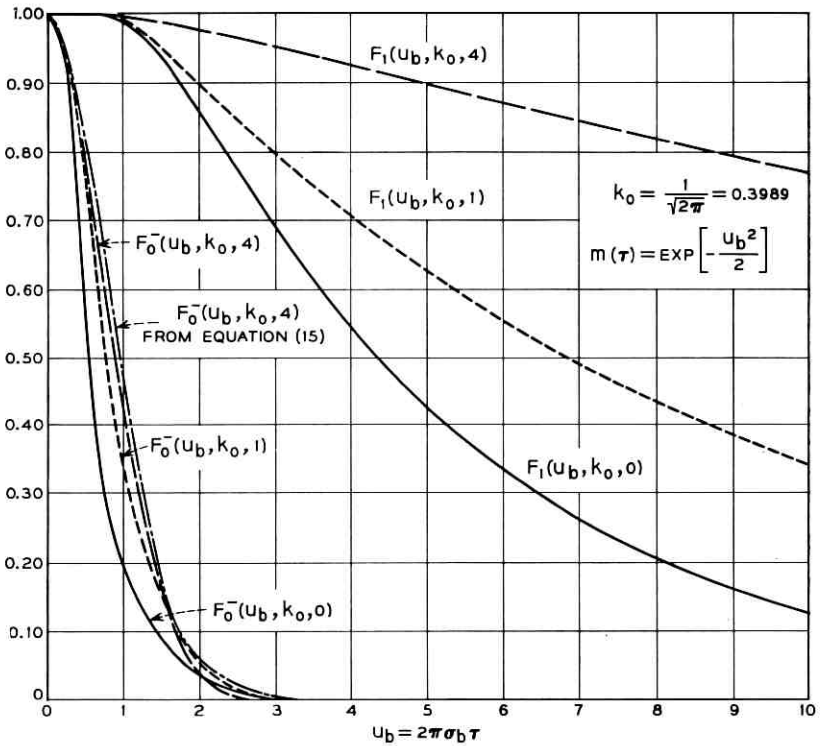



Fig. 10 — $F_0^-(u_b, k_0, a)$ is the probability that the duration of a fade of the Rayleigh process below the normalized level $k_0$ lasts longer than $u_b$. $F_1(u_b, k_0, a)$ is the probability that the interval between fades of the Rayleigh process below the normalized level $k_0$ lasts longer than $u_b$.

For a given $m(\tau)$ with $m'''(0^+) = 0$ along with $a \geqq 4$, $k_o = 1$, the duration of fades and the interval between fades of the Rayleigh process $R(t,a)$ behave as if they were generated at the mean value level of a Gaussian process having a normalized autocorrelation function of $m(\tau)$. For example, compare $P_o^-(u_b, k_o, 4)$ of Fig. 2 with the experimental points plotted in Fig. 2 of Ref. 3. Also compare $P_1(u_b, k_o, 4)$ of Fig. 3 with the experimental points plotted in Fig. 3 of Ref. 3.

## IV. CONCLUSIONS

Assuming that the random process underlying a fading phenomenon is a stationary Rayleigh process, one can compute useful theoretical approximations for the probability functions which characterize the duration of fades and the interval between fades. The corresponding exact results are at present unknown.

For deep fades and large signal-to-clutter power ratio the duration of fades is characterized, approximately, by a Rayleigh probability density.

For large signal-to-clutter power ratio the duration of fades and the interval between fades of the Rayleigh process below the mean value level behave as if they were generated at the mean value level of a certain Gaussian process.

The results of this paper apply to all fields of science and technology for which a stationary Rayleigh process characterizes a fading phenomenon.

## V. ACKNOWLEDGMENT

It gives me great pleasure to acknowledge stimulating discussions with S. O. Rice. I am also indebted to R. T. Piotrowski for programming the digital computer.

## REFERENCES

1. Kerr, D. E., Propagation of Short Radio Waves, M.I.T. Radiation Laboratory Series, *13*.
2. Lawson, J. L., Uhlenbeck, G. E., Threshold Signals, M.I.T. Radiation Laboratory Series, *24*.
3. Rainal, A. J., Zero-Crossing Intervals of Envelopes of Gaussian Processes, Technical Report No. AF-110, DDC No. AD-601-231, The Johns Hopkins University, Carlyle Barton Laboratory, Baltimore, Maryland, June 1964. Abstracted in IEEE Trans. Inform. Theor., *IT-11*, No. 1, January, 1965, p. 159.
4. Rainal, A. J., Zero-Crossing Intervals of Rayleigh Processes, Technical Report No. AF-108, DDC No. AD-600-393, The Johns Hopkins University, Carlyle Barton Laboratory, Baltimore, Maryland, May 1964. Abstracted in IEEE Trans. Inform. Theor., *IT-11*, No. 1, January, 1965, p. 159.
5. Rainal, A. J., Axis-Crossing Intervals of Rayleigh Processes, B.S.T.J., *44*, July–August, 1965, pp. 1219–1224.
6. Rice, S. O., Radio Field Strength Statistical Fluctuations Beyond the Horizon, Proc. IRE, *41*, February, 1953, pp. 274–281.
7. Rice, S. O., Distribution of the Duration of Fades in Radio Transmission, B.S.T.J., *37*, May, 1958, pp. 581–635.
8. McFadden, J. A., The Axis-Crossing Intervals of Random Functions — II, IRE Trans. Inform. Theor., *IT-4*, March, 1958, pp. 14–23.
9. Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, B.S.T.J. *27*, January, 1948, p. 123, equation 3.13.
10. Kac, M. and Slepian, D., Large Excursions of Gaussian Processes, Annals Math. Statistics, *30*, December, 1959, pp. 1215–1228.

# A Geometric Interpretation of Diagnostic Data from a Digital Machine: Based on a Study of the Morris, Illinois Electronic Central Office

By J. B. KRUSKAL and R. E. HART

(Manuscript received June 7, 1966)

*Using the diagnostic data collected for the Morris Central Control malfunction dictionary, we devise a natural concept of "distances" between malfunctions. Ten thousand malfunctions were placed as points in six-dimensional space in such a way that the Euclidean interpoint distances approximately equaled the diagnostic "distances". The remarkable fact that this is possible has many implications.*

*By finding circuit characteristics common to a cluster of neighboring malfunctions, we are able to associate these characteristics with the region of the six-dimensional space which holds these malfunctions. By this means, we characterize various regions of space according to functional troubles. This suggests a technique for locating malfunctions and also suggests some longer-range possibilities.*

TABLE OF CONTENTS



## I. INTRODUCTION

To identify a malfunction is always difficult. The great size, complexity, and speed of modern digital machines render this difficulty severe. An age-old method is to observe the symptoms and deduce their cause. For larger machines, this is exceedingly impractical.

A familiar aid is the use of tests. The record of tests passed and tests failed provides many more clues to the trouble. However, even the use

of tests does not avoid time consuming analysis by a highly trained expert, which is slow and expensive.

One very successful and ingenious approach to alleviating this difficulty (proposed by Werner Ulrich) is to make a large dictionary listing many malfunctions with corresponding test results. For each of many known malfunctions, we obtain a pattern of 0's and 1's which indicate the test results. For example:

| Test Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | ··· |
|---|---|---|---|---|---|---|---|---|---|
| Result | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | ···. |

A 0 indicates a correct result and a 1 an incorrect result. These patterns are then arranged in some systematic order. Together with each pattern we include identification of the malfunction. When we wish to find a malfunction, we simply locate the pattern of test results in the dictionary. If a sufficiently comprehensive set of tests is used and a sufficiently comprehensive set of known malfunctions is included, such a dictionary can achieve a high degree of success. We note that to collect the test patterns for the dictionary, the only practical procedure may be actually to insert the malfunctions in a real model of the machine.

Not all malfunctions can be found by using such a dictionary. Some conceivable malfunctions will not be listed in the dictionary, and other malfunctions produce different test results on different occasions (inconsistent results). However, it is not necessary to use a dictionary only for exact pattern matching. If a malfunction produces different patterns on different occasions, we may expect these different patterns to be "similar" to each other and, indeed, this has been found to be true of the data from the Morris, Illinois electronic central office. Broadly speaking, we feel that patterns are similar if they differ in only a few places. We call the number of places in which two patterns differ the "Hamming distance" between the patterns. This is a rough measure of dissimilarity between patterns.

However, some tests are more important than others. Therefore, we have refined the idea of Hamming distance by weighting the various tests and using weighted Hamming distance (WHD). We have found that the WHD between two patterns is a good and meaningful measure of dissimilarity between the malfunctions which yield those patterns.

"Distances" suggest a geometric model. Is it possible, for example, to represent each malfunction by a point in a plane, in such a way that the (ordinary Euclidean) distance between two malfunctions is approximately equal to the WHD between the corresponding patterns? If true, it would be tremendously significant. For it would mean that the patterns and hence the malfunctions somehow form a two-dimensional

set, that each malfunction can be represented by two coordinates in a way that contains the information in the patterns.

It would be equally significant if we could represent the malfunctions by points, not in the two-dimensional plane, but rather in three-dimensional space, or even by points in $n$-dimensional space, as long as $n$ is reasonably small.

In fact, the Morris data can be represented by such a geometric model. For these data, six dimensions were found to be appropriate. In six dimensions the typical deviation for our data between WHD and Euclidean distance (ED) is reasonably small (about 7 percent).

We emphasize the fact that the small number of dimensions is not something that would happen with just any data, nor could it happen by chance. Random data might have fitted into 100 dimensions by chance. But the smaller the number of dimensions needed, the more significant the result. Six dimensions are remarkably few to represent 10,000 patterns of 657 bits each.

It should also be understood that the number 6 is approximate, and that 5 or 7 are also reasonable values. Fewer dimensions can be used at the cost of larger deviations between WHD and ED, while more dimensions can be used to further reduce these deviations. However, the value 6 was chosen by following the principle of parsimony, which recommends that data be represented by as few numbers as are needed to fit the data satisfactorily.

Not only the malfunctions have a geometrical interpretation (as points in six-dimensional space): it appears that a test can be represented as a "hyperplane" (that is, a flat cut of all space) that separates the malfunctions that fail from the malfunctions that pass the test.

A convincing demonstration of the meaningfulness of the geometric interpretation of malfunctions as points in space lies in the way malfunctions with some common characteristic cluster together in space. For example, malfunctions internal to a single register may cluster together in a small region; malfunctions which often affect the logical state of a single lead may cluster together; and malfunctions which affect the common function of a group of related operations may cluster together. In this paper, we discuss several such clusters of malfunctions in the Morris data. It is not easy to predict in advance how the malfunctions will cluster together, but by examining the geometric model we may observe which characteristics describe clusters of malfunctions. It should be obvious that since closeness in the geometric model is based on WHD, malfunctions which cluster together are those which affect the functioning of the machine in similar ways.

It should be noted that to predict how a particular malfunction affects

the functioning of the machine is exceedingly difficult. Thus, while two malfunctions in the same circuit might be thought to have similar effects, and in many cases do, it also happens that two malfunctions which might be expected to react similarly turn out to be quite different. Detailed analysis of such cases reveals unexpected facts about how the machine reacts to malfunctions. Such analysis has given us new insights into the nature of malfunctions.

The WHD has a definite utility in locating a malfunction if we are not able to locate it by exact pattern match in the dictionary. For given the pattern of the unknown malfunction, and some pattern in the dictionary, we may judge the likelihood of the unknown malfunction being the dictionary malfunction by the WHD between the two patterns. The smaller the WHD, the greater the likelihood of the two malfunctions being the same.

The utility of the geometric model lies partly in the fact that it gives a more compact (or parsimonious) way of representing much of the information which is contained in the patterns and WHD's. For example, to provide a way of locating those patterns which lie within some small WHD of the pattern of the unknown fault is very difficult without the geometric model. But with the geometric model we can simply cut (six-dimensional) space into cells, and list the faults within each cell; this serves the same purpose.

Another potential utility of the geometric model is the possibility that it may reveal some underlying truths about malfunctions. Since each malfunction can be represented by six coordinates, it is natural to ask whether each malfunction is characterized by the degree to which it possesses each of six hypothetical underlying characteristics. If we could find such underlying characteristics, there would surely be many benefits. However, we have not isolated such characteristics.

## II. BACKGROUND

### 2.1 *The Electronic Central Office at Morris, Illinois*

All the data and illustrations in this paper are associated with the electronic central office which was in commercial use for over a year at Morris, Illinois, between 1960 and 1962. A duplicate system was built and tested during the same period at Whippany, New Jersey, and the dictionary data we shall describe were actually collected on the Whippany laboratory model. We shall give an extremely brief description of these systems. We hope that readers already familiar with the system will excuse the omissions and extreme simplification necessary in such a

brief account. For a good general description of the system, see Ref. 1. For more detail, see Refs. 2 and 3. The final report (Ref. 4) gives a good account of the results of the whole experiment.

The electronic central office (ECO) contains the central control (CC), the flying spot store (FSS), the barrier grid store (BGS), as well as the subscriber lines, trunk circuits, the switching network itself, and other important units. Our attention is focused on the CC, which controls all the other units (see Fig. 1). The CC is a stored program machine. One memory device for it is the FSS, which provides semipermanent memory for the stored program and for large tables of "translation" information. The other memory device is the BGS, which provides changeable memory in which the CC records calls in progress, numbers being dialed, etc. The CC communicates directly with the switching network.

To assure continuous operation of the central office, certain subsystems are provided in duplicate. At any given moment, one unit of each duplicate pair of units has "active" (controlling) status, while the other unit has either "standby" (ready to take control) or "out-of-service" (malfunctioning) status. The system is so organized that either unit of a duplicate pair can be made active, independently of the status of the other pairs. To insure that the standby units will be ready to assume active status when needed, they are continuously exercised. Even in the absence of any malfunction, the roles of active and standby are exchanged periodically.

To prevent machine malfunctions from propagating large amounts of wrong information in the changeable memory, malfunctions must be detected quickly and the processing of telephone traffic interrupted until the faulty unit is taken out of service and replaced by its standby. To achieve very rapid detection of machine malfunctions, the outputs

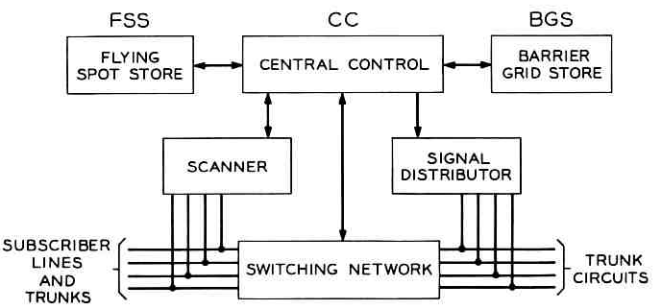



Fig. 1 — Simplified block diagram of the electronic switching system at Morris, Illinois.

of the active and standby units of each duplicate pair are continually compared. In the case of units other than the CC, it is the CC which does the comparing and which takes the proper action. This includes checking to determine whether the trouble symptoms will recur, and if so which unit is responsible (for example, one particular BGS or one particular FSS). The CC then performs diagnostic tests on the unit and reports the results to the maintenance craftsman for corrective action.

The treatment of malfunctions occurring in a CC is necessarily different. Since the CC is the "doctor" who decides which unit has the malfunction, trouble in the CC leads to a situation in which the "doctor" must diagnose himself. This is a complex situation and one that requires special precautions and handling. Briefly, the procedure is this. Under normal circumstances, the two CC's are performing identical operations at the same time. Through a limited number of interconnections between them, each CC receives the results of certain operations in the other CC. These are compared, by means of matching circuits in each CC, with its own results. In the event of a disagreement, the active CC proceeds to check itself by programmed tests. These same tests are run in the standby CC at the same time. If the active CC decides that it is functioning correctly, then the match circuits are interrogated to determine whether the standby CC had the same test results. Ordinarily, when the active CC decides it is sick, it will turn control over to the other CC. To guard against a CC which is so sick that it cannot do this, a timer is provided in each CC which must be reset periodically. If it is not reset, it will turn control of the system over to the other CC regardless of any CC operations that may be going on at the time.

Once the malfunctioning CC has "out of service" status, the active CC requests corrective action by the maintenance craftsman on the system teletypewriter. The active CC then proceeds (in the free intervals between taking care of normal telephone traffic) to perform a diagnostic program which provides information to help the maintenance craftsman locate the trouble.

### 2.2 *The Diagnostic Program for the Central Control*

Each CC consists of several thousand small pluggable circuit cards. Each card contains an assembly of diodes, transistors, resistors, etc. to form a digital "building block". One card may contain up to 5 AND gates, or 5 OR gates, or one flip-flop, or up to two amplifiers, etc. Practically every malfunction which occurs during normal use is due either to failure of a component on a card, or to a bad connection between a

card and its connector. Thus, the primary problem of maintenance is to isolate a malfunction to the card involved.

To aid the maintenance craftsman in this very difficult task, the CC is provided with a large diagnostic program containing nearly a thousand tests. A typical test consists of a short sequence of operations (performed in parallel by both CC's). The matching circuits are turned on only at certain critical points during the sequence. Since the active CC has previously been found to be good, it is the standard by which we test the standby CC. Therefore, any (unintended) mismatch of information between CC's is assumed to be due to a malfunction in the standby CC. Any test whose result in the standby CC matches that of the active CC is recorded as passed, otherwise it is recorded as failed. A 0 denotes a test passed, and a 1 denotes a test failed.

The diagnostic tests are divided into eight logical phases (*A* through *H*). Within each phase the tests are numbered (in octal) from 0 to a maximum (in some phases) of 177. Phases *D* and *H* each require more than 177 tests and therefore, each is divided into two physical phases. The additional phases in *D* and *H* are denoted *DP* and *HP* (for *D* Partial and *H* Partial).

The tests of the diagnosis were organized as much as possible so that phase *A* tests the most basic equipment and phase *B* tests the next most basic equipment. The remaining phases test the remainder of the CC. In normal operation, test failures during phases *A* or *B* cause all the remaining phases to be omitted. However, it should be realized that many malfunctions in the basic equipment exercised by phases *A* and *B* do not cause test failures during phases *A* and *B*. Also many malfunctions outside the basic equipment do cause test failures during phases *A* and *B*.

Although the test results were actually presented on the system teletypewriter in a quite different manner, it is best for theoretical purposes to visualize the test results from a single pass of the diagnostic program as a long row of 0's and 1's. The first digit represents the result of the first test in phase *A*; the remaining digits represent the rest of the phase *A* tests, followed by the phase *B* tests, etc. In this paper, we shall sometimes refer to the results of a single run-through of the diagnostic program as "the test pattern" or "the pattern of test results".

### 2.3 *The Central Control Maintenance Dictionary*

The dictionary described here was constructed under the supervision of S. H. Tsiang, and was completed prior to our use of the data in it.

The whole dictionary project is very well described by Tsiang and Ulrich (Ref. 5).

The large size and complexity of the CC made it necessary to find a better technique than direct reasoning for using the diagnostic test results to locate the malfunctioning circuit card. It was decided to make a dictionary which shows the actual test results for a large number of possible malfunctions that may occur in the system during normal use. The only practical method of compiling such a dictionary was actually to insert the malfunctions in an operating machine and perform the diagnostic tests.

The dictionary was prepared on the Whippany Laboratories model, which was a duplicate of the system at Morris, Illinois. Approximately 50,000 malfunctions were inserted in the CC including such troubles as a shorted diode, an open diode, an open resistor, a flip-flop permanently set or reset, etc. These malfunctions were introduced into every card in the CC.

For each malfunction introduced, the results of the diagnostic tests together with the identification of the malfunction were punched on paper tape by the system. This information, which represents the raw data from which the dictionary was constructed, was transferred to a magnetic tape and sorted on an IBM 704 computer.

The data that we have just described can be visualized as a large matrix with about 50,000 rows and nearly 1,000 columns. Each row corresponds to a malfunction from which the test pattern was obtained. Each column corresponds to a particular diagnostic test. An entry of 0, for example, indicates that the malfunction on that row passed the diagnostic test for that column, while an entry of 1 indicates the malfunction failed for that test.

To facilitate looking up a particular pattern of test results, it is necessary that the rows of the matrix be sorted into some systematic order. Basically, the dictionary consists of the sorted matrix (each row represented in a condensed form) with the malfunction identification pertaining to each row. In many cases the same test pattern appeared in more than one row, that is, different malfunctions produced the same test pattern. In such cases the test pattern was listed only once for all the rows, but with all the corresponding malfunctions, so that each test pattern appeared only once in the dictionary. The dictionary, listing about 30,000 malfunctions, required 1300 pages.

The dictionary just described is a good measure of the effectiveness of the diagnostic tests. The diagnostic program was designed with two basic objectives: (*i*) to contain tests sensitive to virtually every mal-

function that might occur spontaneously in the CC and (*ii*) to produce distinct test failure patterns that identify each malfunction and distinguish it from all others. The data collected yielded the startling fact that the diagnostic programs fell quite short of the first objective. Of the 50,000 malfunctions inserted, approximately 20,000 of them resulted in the all 0's test pattern (all tests passing).

There are a variety of causes for these undetected malfunctions. Some are due to components which serve only a protective purpose, so that their malfunction is observable only in the presence of the trouble being protected against. Others were due to auxiliary equipment which was installed but was neither used, tested, nor covered by the diagnostic programs. Other undetected malfunctions in the CC could have been detected with more diagnostic program or more hardware. No doubt other causes operated as well.

However, we view the problem of undetected malfunctions as a solvable problem which is outside our scope of interest. We presume that in other digital machines to which our ideas might apply, this problem will have been solved. We restrict our attention to the 30,000 detected malfunctions.

The second objective was met fairly satisfactorily (for the detected malfunctions). The extent to which it was met is measured by how many circuit cards are listed in the dictionary for each pattern. The average number was less than 3, which is quite satisfactory but a few patterns had hundreds of associated circuit cards, which is unfortunate.

### 2.4 *An Evaluation of the CC Dictionary*

The CC dictionary was intended to be used by finding in the dictionary the exact test failure pattern obtained in the field. The dictionary entry indicates the list of circuit packages to replace. When this works, it is an easy method of locating malfunctions. Unfortunately, two facts complicate this technique. Many malfunctions yield different test patterns on different occasions. Other malfunctions, though always yielding the same test pattern in the field, yield a pattern that does not appear in the dictionary.

The major reason that test patterns differ from one test run to the next is that the test runs start with the machine in different configurations. Most notably, various flip-flops may have different states. Although the test program attempts to place the machine in a uniform initial state before each test sequence, the malfunction may prevent this being done. One reason that field test patterns may consistently

differ from the dictionary pattern are the intermachine differences, both in electrical parameter values due to manufacturing variability and in logic due to the inevitable program and hardware modification required for dictionary construction.

After the dictionary was prepared, many informal experiments were performed to test its efficiency. Much practical knowledge was gained as to the detailed manner in which test results might differ from the dictionary test results for the same malfunction. This information was partly formalized by S. H. Tsiang (in an unpublished memorandum) who developed "empirical rules" for use with the dictionary. When the test pattern was found in the dictionary but replacement of the listed circuit cards failed to correct the malfunction, or when the pattern was not found in the dictionary at all, these rules were used to alter the pattern to a likely candidate. Use of these rules significantly improved the use of the dictionary.

One formal experiment to evaluate the CC dictionary was performed on the Morris, Illinois model by R. N. Breed and described in an unpublished memorandum. Approximately 600 faults were selected for this evaluation. Malfunctions of various types were chosen in proportion to their frequency of occurrence in certain failure records, and so as to represent all parts of the CC. However, malfunctions known to produce no test failures were avoided.

Of the 600 malfunctions, 30 were eliminated (for unstated reasons) at the time the data was collected, 47 more were eliminated from the summary figures in the memorandum because they "probably could have been found by diagnosis of some unit other than the central control". These 47 malfunctions would have belonged to the last two categories below. The remaining 523 malfunctions were divided into categories as shown:

47 % findable with perfect match to dictionary results,
13 % findable using the empirical rules,
21 % not findable because all the tests were passed,
19 % not findable, even with the aid of the empirical rules.

The third category is of interest to logic circuit designers and diagnostic programmers. Our interest is primarily with the fourth and second categories. Our methods offer real possibilities for identifying the malfunctions responsible for otherwise mysterious test patterns, and for more easily identifying malfunctions which would otherwise require the use of complex empirical rules.

## III. DATA

### 3.1 *The Data Used in our Study*

Of the 30,000 malfunctions with test patterns not all 0, about 10,000 malfunctions failed one or more tests in phase $A$ and/or phase $B$. When the dictionary was originally prepared, it was observed that malfunctions of this sort, besides consistently failing tests in phase $A$ and/or $B$, generally failed a great many tests in the other phases, and did so in an inconsistent way. For this reason the CC dictionary suppresses the test results of the other phases for this group of test patterns.

As we wished to reduce the scope of our study (to cut down computation time and cut down the bulk of printed results), we removed these 10,000 malfunctions (failing phase $A$ and/or $B$) from the data.

However, we were eager to reduce the bulk of the data still further. We realized that for test patterns with few test failures, our methods offer less potential advantage than for patterns with many failures. This is true for several reasons. The direct deductive method tends to work well for test patterns with few test failures. Furthermore, so very many other malfunctions yield test patterns within a very small WHD of the observed pattern that it may not be practical to use the dictionary for nonexact matching in the way we shall discuss.

We found that the patterns with 3 or less test failures constituted about half of the remaining 20,000 malfunctions. Certain of our calculations would have been distorted by 2800 malfunctions which all gave exactly the same pattern of 3 test failures (namely, phase $H$ tests 100, 101, and 102). (These malfunctions caused the standby CC to "lock up", that is, stopped its master-clock; the tests are part of a special set of tests comprising the "CC lockup diagnosis.") To avoid this distortion and to reduce the data to a reasonable quantity, it seemed natural to restrict ourselves to patterns with 4 or more test failures.

Thus, we finally used a matrix with 10,937 rows (malfunctions). As we have eliminated test failures in phases $A$ and $B$, the only columns with 1's in them can be those for phases $C$ through $H$. Of these, just 657 columns actually contained 1's.

A few facts about this matrix may be of interest. Fig. 2 shows graphically the number of rows with exactly $k$ 1's in them, as a function of $k$. The few rows with the greatest number of 1's in them have, respectively, more than 511, exactly 466, 441, 325, 310, 252, and 247 1's in them. Fig. 3 shows the number of columns with $k$ 1's in them as a function of $k$. The few columns with the most 1's in them have, respectively, 4335,

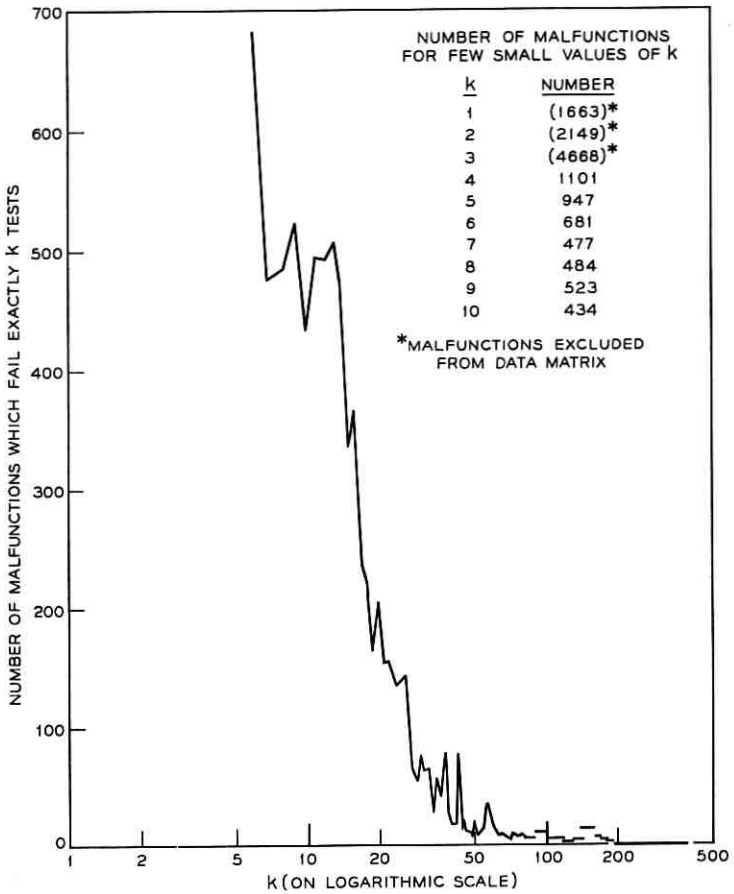



Fig. 2 — Number of malfunctions which fail exactly $k$ tests, that is, the number of rows of the matrix with exactly $k$ 1's. For larger $k$, the data have been grouped.

2978, 2711, 2653, 2383, 1933, and 1744 1's in them. The column with 4335 1's in it corresponds to test HP 61; thus, this test was failed by about 40 percent of the malfunctions in the matrix.

### 3.2 *The Test Weights*

We now restrict attention to the data we used in our study. It consists of a matrix with 10,937 rows and 657 columns, whose entries are 0 or 1. In the introduction we referred to weights $w_i$ which we associated with the diagnostic tests, or in other words, with the columns of the matrix. These weights are all positive, and the largest possible value is 1.

Before describing the meaning of these weights and the formulas used to obtain them, we shall mention a few facts about the weights actually obtained from the data. Just one weight is greater than 0.999. There are 35 weights greater than 0.95. The weights in the interval from 0.25 up to 0.95 are very sparse, while below 0.25 the weights are densely but erratically distributed. A graph of the density of weights versus $w$ is shown in Fig. 4. The brief table below summarizes the same information.

| Dividing points | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.95 |
|---|---|---|---|---|---|---|
| Number of weights | 272 | 73 | 119 | 72 | 34 | 52 | 35 |

The smallest weight is 0.0097.

The weights are intended to reflect the extent to which the information given by the entries in one column of the matrix is independent of the information given by other columns. Thus, a column which does not at all resemble any of the other columns would have a full weight of 1, while a column which is almost the same as a great many other columns would have a very small weight.

For example, suppose 10 columns are identical. Then they all contain the same information, so it is natural to give each one a weight of 1/10

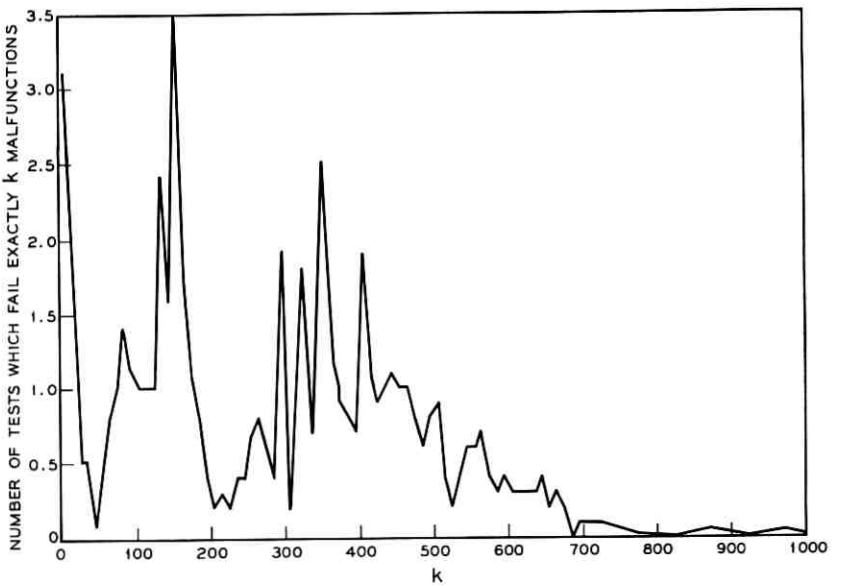



Fig. 3 — The number of tests which fail exactly $k$ malfunctions, that is, the number of columns of the matrix with exactly $k$ 1's in them. Curve has been smoothed by grouping 10 or 50 values of $k$.

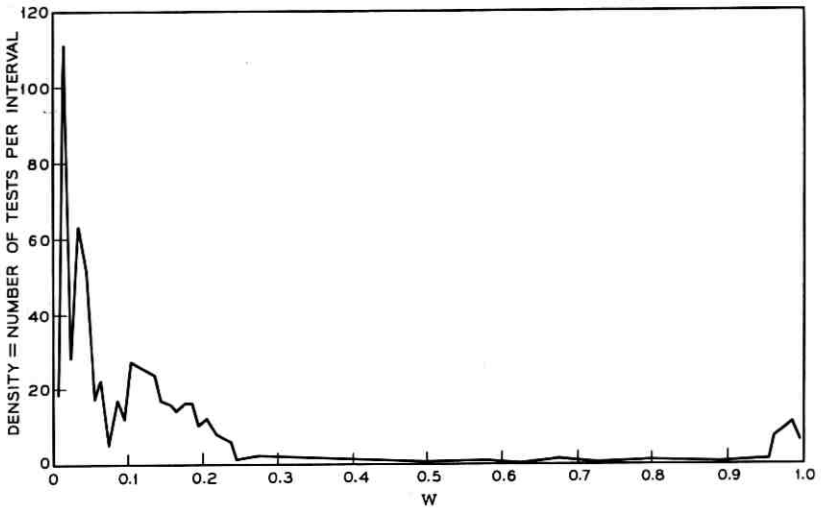



Fig. 4 — The density of test weights as a function of weight $w$. The vertical scale is the number of tests in an interval of 0.01 on the $w$ axis In some cases, the curve has been smoothed over longer intervals.

to reduce the importance of that information to its proper value. However, it is not enough to consider columns which are exactly identical; in our data there are a great many columns which are almost the same and which must be taken into account.

Suppose we have a way to measure how much alike two columns are. In particular, suppose $L_{ij}$ is the amount of likeness or similarity between columns $i$ and $j$. We suppose that $L_{ij}$ lies between 0 and 1, with $L_{ij} = 1$ if the two columns are identical, and $L_{ij} = 0$ if the two columns are not at all alike. Of course $L_{ij} = L_{ji}$ . Then a natural formula for the weights is

$$w_i = \frac{1}{L_{i1} + L_{i2} + \cdots + L_{i,\,657}} = \frac{1}{\sum_j L_{ij}}.$$

For example, if columns 1 through 10 are identical to each other but do not resemble the other columns at all, then the weight of each of these columns is 1/10. (In more detail, consider say, column 7. Then $L_{7j} = 1$ for 10 values of $j$, and $L_{7j} = 0$ for all other values of $j$, so the denominator is 10.)

As another example, suppose there are just three columns (instead of 657) and that the values of $L_{ij}$ are those given below:

| | | $j$ = 1 | 2 | 3 |
|---|---|---|---|---|
| $i$ | 1 | 1.0 | 0.5 | 0.1 |
| | 2 | 0.5 | 1.0 | 0.7 |
| | 3 | 0.1 | 0.7 | 1.0 |

Then

$$w_1 = \frac{1}{1.0 + 0.5 + 0.1} = \frac{1}{1.6} = 0.625,$$

$$w_2 = \frac{1}{0.5 + 1.0 + 0.7} = \frac{1}{2.2} = 0.455,$$

$$w_3 = \frac{1}{0.1 + 0.7 + 1.0} = \frac{1}{1.8} = 0.555.$$

How shall we measure the likeness of columns? One practical answer, which we used, is the square of the correlation coefficient between the columns. For readers who are not familiar with this widely used statistical quantity, we tell a little about it. The correlation coefficient itself lies between $-1$ and $+1$. It is $+1$ for two identical columns, $-1$ for two columns which are exact opposites, and 0 if the 1's are arranged as if they had been sprinkled randomly and independently in the two columns. The correlation coefficient has intermediate values in intermediate situations. To be very concrete, suppose the matrix has 12 rows (so each column has 12 entries). Suppose one column has three 1's and the other column has four 1's. Then the correlation coefficient depends on the number of rows in which both columns have 1's at the same time. The following table gives the actual values.

| Number of 1's in common | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Correlation coefficient | $-0.408$ | 0 | $+0.408$ | $+0.816$ |
| Likeness value | $+0.167$ | 0 | $+0.167$ | $+0.667$ |

The formula for the correlation coefficient which applies to the present circumstances involves

$N$ = the number of rows in the matrix,
$n_i$ = the number of 1's in column $i$,
$n_j$ = the number of 1's in column $j$,
$n_{ij}$ = the number of 1's in common to columns $i$ and $j$.

Then we have

$$\text{correlation coefficient between columns } i \text{ and } j = \frac{Nn_{ij} - n_i n_j}{\sqrt{n_i(N - n_i)} \cdot \sqrt{n_j(N - n_j)}}.$$

The suitability of this way of measuring likeness is established by its successful application to this data. However, we may justify it in part by appealing to intuition and numerical experimentation.

The correlation coefficient is $+1$ if and only if the two columns are identical. It is clearly appropriate that the likeness should be 1 in this case.

The correlation coefficient is $-1$ if and only if two columns are complementary, that is, if one column has 1's precisely where the other column has 0's. In this case, we may say that the two columns carry the same information even though their entries are opposite. For this situation means that one test fails just when the other passes. Clearly if we know this, it is enough to perform just one of the tests; we can predict the result of the other. Thus, it is reasonable to assign a likeness of $+1$ to this situation, as our likeness measure does.

If the 1's in the two columns appear as if they are independently located (in the statistical sense), so that knowledge of the entry in one column has no predictive power at all for the entry in the other column, then it is reasonable to assign likeness 0. This situation occurs if and only if

$$\frac{n_{ij}}{N} = \frac{n_i}{N}\frac{n_j}{N},$$

also if and only if the correlation coefficient is 0. Thus, we assign likeness 0 in this case.

This pins down the value of our likeness value for two extreme situations and one intermediate situation. If we are considering the correlation coefficient as the basis for measuring likeness, this still leaves many possibilities. For example, the likeness can be the absolute value of the correlation coefficient, the square of the correlation coefficient, any positive power of the absolute value of the correlation coefficient, to mention only a few possibilities. We experimented with various possibilities including the absolute value, the square and three functions of the correlation coefficient whose graphs are made up of straight-line segments.

Some tests are so individualistic that they surely deserve weights of

almost 1. On the other hand, there is a large cluster of perhaps 100 tests (that is, columns) which are so nearly alike that at least the most typical of them deserve weights as small as 0.01. As the absolute value formula yielded weights ranging from about 0.1 to 0.01, it was clearly inappropriate. The reason was equally clear; the effect of statistical fluctuations on the many small likenesses is cumulative and noncancelling and leads to unduly large denominators. Since there appears to be no way to arrange for cancellation, it is desirable to reduce the effect of small likenesses.

Two of the segmented-straight-line functions did this very successfully, and yielded weights ranging from virtually 1 to slightly under 0.01. Then Colin Mallows pointed out that the squared correlation coefficient lies very neatly between these two functions. When tried, the square yielded very similar weights, with the smallest one even a trifle smaller.

## IV. GEOMETRY

### 4.1 *Weighted Hamming Distance*

Ordinary Hamming distance between two test patterns (two rows in the matrix) is just the number of places in which they differ, in other words, the number of tests for which they have different results. To calculate this, we accumulate 1 for each position in which the test patterns differ.

Weighted Hamming distance is similar except that instead of accumulating 1's we accumulate the weight associated with that position. For example, if two test patterns differ only in the results of tests 3, 5, and 17, then the WHD (weighted Hamming distance) between them is given by

$$\text{WHD} = w_3 + w_5 + w_{17} .$$

We note that this a true distance in the mathematical sense of the word. In particular, WHD satisfies the triangle inequality: the WHD between patterns 1 and 2, plus the WHD between patterns 2 and 3, is always greater than or equal to the WHD between patterns 1 and 3.

We measure the dissimilarity between malfunctions by the WHD between their test patterns. If two malfunctions yield test patterns between which the WHD is small, we consider the malfunctions similar, but if the WHD is large we consider them dissimilar. With this in mind, let us consider the intuitive meaning of the test weights. Suppose that

tests 1 through 100 are all very much like each other; that is, these tests generally fail together or pass together. This means that a test pattern will generally fail almost all or pass almost all these tests: it is unlikely that a test pattern will fail approximately half these tests. Thus, in comparing the dissimilarity of two test patterns with regard to this group of tests, the main information we get is whether they are the same or opposite. If we used ordinary Hamming distance, then test patterns which are opposite would have a distance of at least 100 just from these tests alone. Yet "same" or "opposite" on this group of tests may be no more significant than same or opposite on a single test which is an "individualist". By down-weighting like tests and using WHD, we prevent large groups of like tests from swamping the information contained in tests which are very "individualistic".

Now it is true that for test patterns which are the same for most of the tests in this large group, the few tests in the group which yield different results may be very significant. We view this as fine-grain information, however, in contrast to the broadbrush information contained in the group as a whole. We do not know of any practical way to have the WHD based on a single set of weights reflect both kinds of information.

Nevertheless, there is a way within our general scheme to make use of this fine-grain information, though it is not an idea which we have actually attempted. The technique is this. We would collect together some group of test patterns in our matrix which are fairly similar to each other; for example, we might arbitrarily pick some test pattern as the "center", then form the group of all test patterns in the matrix which are within some fixed WHD of the center. Presumably we would arrange things so as to get a group of several hundred test patterns. Using this group of test patterns we would have a new data matrix (actually a submatrix of the original, with all the columns but only a selected set of rows). Using this submatrix we would calculate new weights, using the same formulas but applying them to this new smaller matrix. We could call these "local" weights as they apply only to this one local group of test patterns *when compared with each other*.

These local weights could differ very greatly from the original "global" test weights. The global weight of the test could be high or low, independently of the local weight. Furthermore, the local weights for this local group of test patterns might be entirely different from the local weights we would derive from some other local group of test patterns.

Using the local WHD (based on local weights) to measure dissimilarity between test patterns in some group is probably a good way to make use of the fine-grain information.

4.2 *The Geometric Model*

Once a meaningful concept of distance between test patterns exists (such as WHD), it is natural to ask whether these distances can be realized in a geometric model. For example, can we represent each pattern by a single point in the plane, in such a way that the ordinary Euclidean distance (ED) between the points is equal to the WHD between the corresponding patterns?

First, we remark that there is nothing inherent in the concept of distance which will force this to happen. Thus, if this happens it tells us something about the data. It tells us that in some sense or other the test patterns form a two-dimensional set. What this means is not clear. But, that it means something important is indicated by the tremendous information compression which is achieved.

To understand this, let us suppose that we have 10,000 test patterns. Between these test patterns there are

$$(10{,}000)(9{,}999)/2 \doteq 50{,}000{,}000$$

WHD's. If we can represent each test pattern in the plane, that requires two coordinates per pattern, so that we require 20,000 numbers to represent the patterns. Since the ED's are of course computable from these 20,000 coordinates (by the usual formula learned in high school), and since the ED's equal the WHD's, we have compressed the information from 50,000,000 numbers into 20,000 numbers. In other words, from the 20,000 numbers required to represent the patterns, we can recover by simple arithmetic all the 50,000,000 WHD's.

Any model which achieves such compression is bound to be useful, for it permits us to handle information in a much more concentrated manner. Beyond its direct utility, however, any model which achieves such compression is trying to tell us something about the data.

(The classic example of this are the 20 years worth of extremely accurate astronomical observations made by Tycho Brahe in the sixteenth century. Kepler found a model consisting of his three famous laws from which it was possible to explain these observations. Basically, his model represented each planet's motion by an ellipse. Thus, using his model it was possible to explain all of Tycho's observations of one planet from 12 numbers — 6 for the planet's motion and 6 for the earth's motion. It is clear that the enormous compression of information in itself was useful in this situation. It is also clear that the model was trying to say something, however, even if it took Newton to hear it.)

Without trying to compare ourselves to Kepler, we feel that the in-

formation compression of our model is a striking phenomenon which demands investigation, and must produce something of value.

We do not get a representation by points in the plane, nor by points in three-dimensional space, but only by points in six-dimensional space. We can represent each pattern by six coordinates in such a way that the ED's are approximately equal to the WHD's. This applies to not quite all the 10,937 patterns in our matrix — there were three exceptions that did not fit. (These three exceptions probably result from malfunctions which in fact cause $A$ or $B$ phase test failures, but were not excluded from our data due to some recording failure which dropped the $A$ and $B$ phase results.)

We notice first that the compression of information goes down as the number of dimensions goes up. For 10,000 patterns represented in 6 dimensions, the same 50,000,000 WHD's are recoverable not from 20,000 coordinates but from 60,000 coordinates. The compression is slightly less.

We notice second that the value of the compression depends on how good the approximation is. The more accurately the ED's represent the WHD's, the more valuable the compression is. In our case the typical difference between matching ED and WHD is about 7 percent. More exactly,

$$\sqrt{\sum (\text{ED} - \text{WHD})^2 / \sum \text{WHD}^2}$$

is in the neighborhood of 7 percent. Though not striking, it seems entirely adequate when matched with the compression we have.

A scatter diagram of WHD's against ED's is shown in Fig. 5. Each point displays the WHD and ED between one pair of malfunctions. The figure contains almost 5000 points, corresponding to all possible pairs from among the list of 100 malfunctions referred to in the next section, and is impressive testimony to how well the ED's match the WHD's.

### 4.3 *How to Compute the Geometric Model*

In this section, we describe the necessary computation very briefly, just enough to take the mystery out of it. Suppose then that we wish to place 10,000 points in some space — we will use the plane to make it easier to visualize, though exactly the same procedure works in three-dimensional space or six-dimensional space. The information we have consists of the approximately 50,000,000 WHD's between these points.

We start by placing the 10,000 points in the plane in any arbitrary

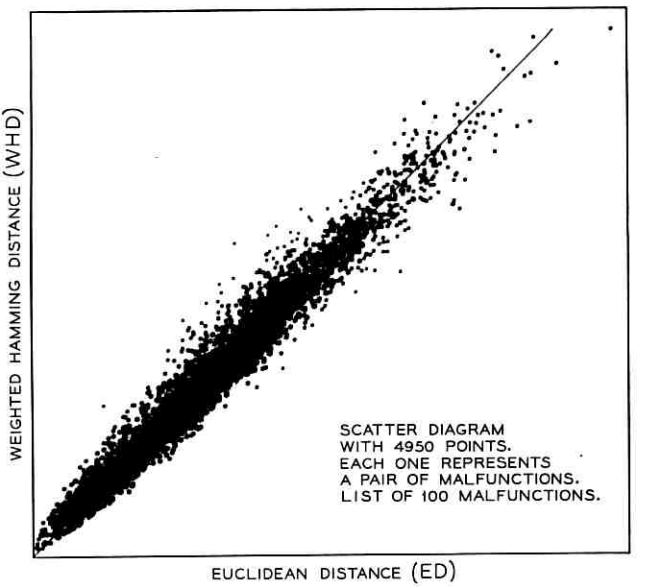



Fig. 5 — Scatter diagram of weighted Hamming distance (WHD) against Euclidean distance (ED). Each point represents one pair of malfunctions. All pairs from among 100 malfunctions are displayed.

configuration, and pinning them down so that they cannot slide about. Next, we buy 50,000,000 ideal springs. These springs are massless and all have the same restoring force ratio (Young's modulus) — let us suppose the common value is 1. However, the springs are all of different lengths. In fact, each spring has a length equal to one of the WHD's. Now we fasten each spring between the two appropriate points. Thus, between points $i$ and $j$ we attach the spring whose unstretched length is the WHD between patterns $i$ and $j$. Of course it is necessary to stretch or compress the springs, and we do so as required. Naturally these ideal springs do not buckle when compressed, and furthermore several of these springs can occupy the same space at the same time, so that we do not need to worry about how they cross each other.

After all the springs are attached, we suddenly pull out all 10,000 pins, permitting the points to slide about (but only on the plane — we do not permit them to fly up in the third dimension). If there is some dissipative force, such as air resistance or friction, the springs and points will eventually come to rest. By the laws of physics, they will come to rest at a minimum energy configuration, that is, a configuration at which the potential energy stored in the springs is a minimum.

What can we say about this configuration? The potential energy in each spring, according to our assumptions, is $(\mathrm{ED} - \mathrm{WHD})^2$. The total energy is the sum of all these. Thus, the final resting configuration is one which minimizes this sum, or almost the same which minimizes

$$\sqrt{\sum (\mathrm{ED} - \mathrm{WHD})^2 / \sum \mathrm{WHD}^2}.$$

Our computation is basically an imitation of the spring motion. We also start with an arbitrary configuration. Then we figure the net force on each point exerted by the springs, and move all the points where they would be a short time later. We again figure the net forces, and again move the points. After enough repetitions, the net forces reach 0, and we know that we have reached the minimum energy configuration.

This is a good intuitive description of what we do, but we would not like to leave the impression that our computation is in any way unrigorous. In the language of numerical computation, we are seeking to minimize the expression given above. To do so, we perform an iterative process known as the method of gradients (or the method of steepest descent). Thus, we start with an arbitrary configuration, and compute the (negative) gradient, which is just the same thing in this case as the net forces on all the points. We move a little in the direction of the (negative) gradient — that is, just the motion along the force vectors. Then we again calculate the gradient and again move. When the gradient is zero, we have reached a minimum.

Of course, we cannot really perform this computation as described on all 10,000 points at once. To perform one single movement would require nearly 30 hours (on the IBM 7090), even if we could manage to keep all the numbers required in the internal memory.

To get around this difficulty we hoisted ourselves by our own boot straps. We started with 33 points, and performed the computation exactly as described. Then we "pinned down" these 33 points, and introduced 67 more points, and $33 \times 67$ "springs." During this computation only the 67 new points were allowed to move. Thus, we had 100 points located, though not quite perfectly. We then performed the original computation on these 100 points, starting with the configuration we had already achieved. This just moved the 100 points a little — it was a "polishing" operation. We then picked a set of 20 from these 100 points, in such a way that these 20 are well spaced over the region of space covered by the 100 points, with no pair of the 20 points too close together. We "pinned down" these 20 points very firmly, and introduced $20 \times 10{,}000$ "springs". During this computation only the 10,000

new points were allowed to move. This located all 10,000 points, though not quite perfectly. We simply tolerate this imperfection (though there are practical ways to reduce it if it should seem intolerable).

The reason this computational scheme is practical is that when we pin down the 20 points and introduce 10,000 new points, we can handle them one by one. Thus, at any time we only need to deal with 20 fixed points, one movable point, and 20 springs from the movable point to the fixed points. After the one movable point comes to rest, we remove it before introducing the next.

### 4.4 *Why Six Dimensions?*

As we have noted, we do not achieve perfect equality between the ED's and WHD's. The typical difference is about 7 percent. Obviously in seven dimensions we can reduce this figure while in five dimensions it would be larger. The more dimensions, the better we can make the ED's match the WHD's.

It would be possible to draw a curve of the typical error versus the dimension. (We would put dimension on the horizontal scale and error on the vertical scale.) We would then get a descending curve. On this basis, the more dimensions the better. On the other hand, from the point of view of information compression, the more dimensions the worse. Thus, we wish to strike a balance.

The principle of parsimony advocates obtaining the highest compression possible while retaining "satisfactory fit". In other words, use as few dimensions as possible with the typical error satisfactorily small.

Actually, we did not draw such a curve with the complete data. We did draw such a curve, however, with a small sample of the data (using a similar but more complicated model than the one we have described). For this sample, we computed the typical error for 2, 4, 6, and 8 dimensions. The typical error in 4 dimensions seemed too large, while in 6 dimensions it was tolerable. Going to 8 dimensions produced little reduction. Thus, we decided to use 6 dimensions.

Another reason for using 6 dimensions was the fact that when W. Thomis used a different scheme for coordinatizing faults, he needed 6 coordinates, which seemed to point to 6 dimensionality also. It is clear from this discussion, however, that 5 or 7 dimensions would also be satisfactory, but 4 or 8 would probably not be. Though it is difficult to say which is best, we see that 6 dimensions represents a reasonable compromise value for these data.

### 4.5 *Tests and Hyperplanes*

The geometric model may have considerably more meaning than we have indicated so far. It may be possible to represent tests as well as malfunctions. We do not represent tests by points, however, but by "hyperplanes".

In general, a hyperplane is an infinite flat cut which divides space into two parts. In three-dimensional space a hyperplane is an ordinary flat plane.

In two-dimensional space, that is, in the plane, a hyperplane is a straight line. In one-dimensional space, that is, in the line, a hyperplane consists of a single point. In $n$-dimensional space, a hyperplane is an $(n - 1)$-dimensional flat subspace. (Hyperspace is an old-fashioned name for a higher dimensional space, and the hyperplane is the analogue in these spaces of the plane in three dimensions.)

In the following geometric discussion it would be well to have a mental picture of either two or three-dimensional space. Each hyperplane is then visualized as a line or a plane.

Thus, suppose we have space (actually six dimensional but visualized as two or three-dimensional). In it are 10,000 points representing malfunctions. Now pick some particular test. Every malfunction which fails this test we color red; there are relatively few of them. Every malfunction which passes this test we color black; these are the majority of malfunctions. How are the red points situated? Are they scattered among the black ones?

There is reason to believe that in most cases the red points and the black points may be separated by a hyperplane. That is, the red points and the black points are not all mixed up. If space is two-dimensional, this means that a straight line can be drawn with the red points on one side and the black points on the other. If space is three dimensional, then a plane exists with the red points all on one side and the black all on the other.

In a sample from the Morris data involving 27 malfunctions and about 200 tests placed in 6 dimensions, we found this to be true. In some cases, the hyperplane did not quite perfectly separate the two kinds of points; a few points would be slightly on the wrong side, but the amount by which the points were on the wrong side was extremely small.

We believe that most of the tests would be representable as hyperplanes in the main body of data analyzed. If a few tests are not representable, that would probably say something interesting about these tests. Among other things, it might suggest reducing their weight.

If we suppose that the tests can, in fact, be represented by hyperplanes, then we can calculate the information compression of the model in a different way. The original data consists of about

$$10{,}000 \times 650 \doteq 6{,}500{,}000$$

bits. To represent both the malfunctions and the tests in 6 dimensions requires about

$$(10{,}000 + 650) \times 6 \doteq 64{,}000$$

numbers. From these numbers we can reconstruct the original data (though not perfectly), for to find whether a particular bit is 0 or 1 we merely need to check which side of some hyperplane some point lies on. The imperfections result from the fact that the hyperplanes from some tests do not perfectly separate the malfunctions which pass from those which fail.

From this viewpoint, the information compression consists of representing 6,500,000 bits by 64,000 numbers. This viewpoint probably provides a more meaningful measure than the simpler one presented before.

### 4.6 *Utility of the Geometric Model*

There are several kinds of utility for the geometric model. One kind is theoretical and long range. By examining the data in the model, we hope to learn something about the structure of the data. It is basic procedure in data analysis to look at the data with one's common sense on the alert. Where the data can be represented in compact form, this is much more useful.

Another utility of the geometric model is very immediate. To have the malfunctions represented by coordinates simplifies the process of finding nearby malfunctions. To illustrate this most clearly, let us suppose for the moment that the malfunctions could be represented in two dimensions instead of six. Imagine the malfunctions placed on a "map". This would resemble a photograph of the starry night-sky. Now suppose a new malfunction is to be identified. We would calculate its coordinates, plot its position on the map, and look for the nearest few points. Suppose on the other hand that we wished to find the nearest few points without the aid of the representation in two dimensions. In principle this is easy enough. It is only necessary to run through every malfunction in the dictionary one by one, and compute its WHD from the unknown malfunction, and finally pick out the smallest few WHD's. Computationally,

however, this is very much more difficult than use of the map. We see from this that the map very much simplifies the computation necessary to pick out the nearest few malfunctions.

Unfortunately we cannot use the map in six dimensions. Other techniques relying on the coordinates, however, are available. For example, we can cut space up into small cells, and list the malfunctions which occur in each cell. Then to find the nearest malfunctions to an unknown one, we look in the same cell and the neighboring cells.

We may summarize this value of the geometric model as reducing the computation required to select nearby malfunctions.

## V. CLUSTERS

### 5.1 *The Region Containing all Malfunctions*

It will be helpful to know something about the "galaxy" of malfunctions, that is, the region of 6-dimensional space in which the 10,937 malfunctions lie. The "healthy machine" (no malfunction, or a malfunction which yields the test pattern of no failures) corresponds to a point with coordinates approximately

$$1.0, \quad 0.0, \quad 0.7, \quad 0.1, \quad -0.8, \quad 1.3.$$

Rounded off to the nearest integer, these are

$$1 \quad 0 \quad 1 \quad 0 \quad -1 \quad 1.$$

This appears to be fairly near the edge of the "galaxy". The center of the galaxy is approximately at

$$2 \quad 0 \quad 2 \quad 0 \quad -1 \quad 1.$$

(By center, we mean the center of gravity, or average position.) If we exclude seven outlying malfunctions from consideration, then the extreme values of each coordinate are

| | | | | | | |
|---|---|---|---|---|---|---|
| minima | −3 | −9 | −4 | −8 | −8 | −4; |
| maxima | 11 | 5 | 10 | 9 | 5 | 8. |

Thus, the range of each coordinate is roughly 14. Of course, the malfunctions are not at all evenly scattered. A tremendously heavy concentration exists near the "healthy machine". Other dense spots also exist.

To further study the distribution of the malfunctions in six-dimensional space, we split space up into cells of uniform size and shape. (We avoided cubic cells because in six dimensions the corners of the cube

"stick out" rather far. Instead, we used the so-called "Voronoi regions" associated with the "body-centered cubic lattice". Each cell can be thought of as a six-dimensional cube with the corners chopped off.) There are 417 cells which contain malfunctions. Thus, the average "populated" cell contains about 24 malfunctions. The cell containing the healthy machine, however, contains 1883 malfunctions. Altogether three cells contain more than a 1000 malfunctions each, while 162 cells contain only a single malfunction each. The median number of malfunctions per populated cell is 2.

The Euclidean distance of a malfunction from the healthy machine is a measure of how severe the malfunction is. It is interesting to compare this measure with the more primitive measure consisting simply of the number of tests failed. In Fig. 6 there is a little circle for each cell. The horizontal coordinate is the Euclidean distance of the cell center from the healthy machine. The vertical coordinate is the average number of tests failed for the malfunctions in the cell. The dense region displays a definite relationship, even though the great scatter shows that it is a loose one.

### 5.2 *Clusters of Malfunctions*

Suppose you had a map of a city, and a list of the people who live there, together with various census information — address, income, national origin, age, education, etc. Suppose you wished to understand the social structure of the city. One obvious approach would be to identify different neighborhoods. You might find that one neighborhood had mostly high income residents, another might contain people mostly of one national origin, near a university you might find many people with higher education, and so forth. You would be seeking to identify clusters of people who live near each other and who share some common characteristic.

We face exactly this situation. We have a six-dimensional "map", and we have a list of malfunctions. About each malfunction we know whether it is a diode, amplifier, or flip-flop trouble, etc. Also, we know the specific nature of the trouble: what circuit it is in, how it operates there, and so forth. It is natural to look for clusters of malfunctions which are near each other in space (that is, which have similar test patterns) and which share some circuit characteristic.

One reason for seeking clusters of malfunctions is to learn how malfunctions affect the operation of the system. We may find that apparently similar malfunctions (for example, flip-flops in a single register which are stuck in the 1 position) do not form a cluster, that they produce

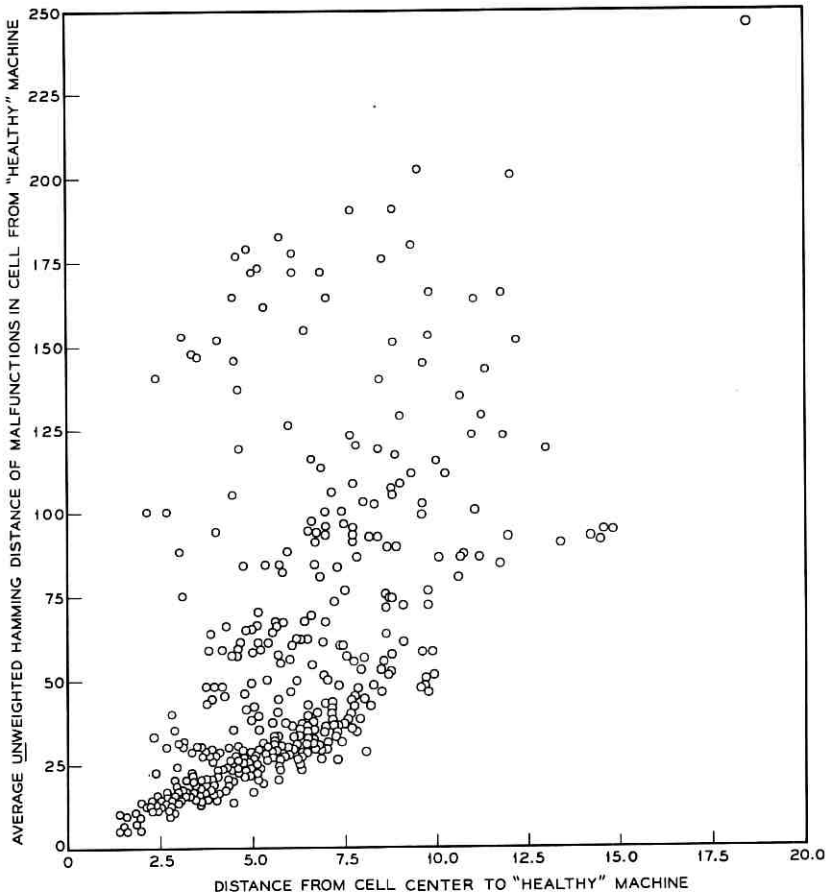



Fig. 6 — Each circle represents one of the 417 populated cells in 6-dimensional space.

very different test patterns. When we inquire into why this happens, we learn something about the nature of the system (and of the diagnostic test program). On the other hand, we may find that apparently rather different malfunctions produce very similar results. For example, a flip-flop stuck in the 1 position may yield very much the same test pattern that the same flip-flop stuck in the 0 position does (or it may not — we have observed both situations frequently). If these two malfunctions are near each other, then we learn that the significant aspect of these malfunctions is merely that this particular flip-flop is out of order, and that the precise nature of its failure is not important.

Of course, whether a malfunction is in a diode or in a resistor (say) has no direct bearing on its position in space, because its position results solely from the diagnostic test data, which, in turn, reflect its effect on the overall operation of the machine.

A cluster generally occupies a region of space without sharp boundaries. The malfunctions of the cluster are heavily concentrated at the center of the region, more lightly spread further out, and may sprinkle themselves out to a considerable distance. Thus, the boundaries of the region cannot be precisely located. It often makes sense to talk about the center of the cluster, however, which means the point of high concentration.

In a city, different clusters of people may overlap. For this reason, one neighborhood may contain, say, two nationality groups together with a sprinkling of artists. In the same way, different clusters of malfunctions often overlap, so the one small region may contain a mixture of malfunctions from several clusters. Occasionally, a cluster may totally dominate the region it occupies, so that practically every malfunction in its region belongs to it.

In the most common case, when clusters overlap, there should be some explainable reason. We discuss some cases of this sort.

Sometimes a very narrowly defined cluster may be a subcluster to a more broadly defined one. For example, we might have a small cluster of malfunctions whose common circuit characteristic is that they hold a particular wire down to a low voltage (prevent the lead from carrying the digit 1) under certain logical conditions (not necessarily identical among the malfunctions of the cluster). If this wire is one of a related group of wires, we may be dealing with a subcluster of a larger cluster of similar malfunctions involving any wire in the group.

### 5.3 *How Clusters are Found*

We wish to emphasize that clusters are not to be found by following preconceived notions as to which malfunctions resemble which other malfunctions. Instead one must try to look at the data with an open mind, or listen to what the data is trying to say.

Concretely, the procedure used was to examine a small region of space which contains not too many malfunctions. (When the data was still unfamiliar, we chose to examine regions with only 5 or 10 malfunctions, though later when our procedures improved we could handle many more.) We analyzed in detail the effects on the circuitry of every malfunction in this region. We then asked ourselves what common element there was to all or most of the malfunctions involved. If we found what

appeared to be a common element, we then traced out by means of the circuit diagrams all of the malfunctions which shared this common element, and noted their locations. If we found these to lie in a single compact region, we considered that we had indeed identified a cluster of malfunctions. Of course the region involved would include the smaller region from which we started. On the other hand, if we found the malfunctions with these characteristics to lie in several distinct regions, only one of which contained the original region, we knew that the malfunctions formed not one but several clusters. In this case, it was necessary to ask what characteristics differentiated the malfunctions in different regions.

Having identified a cluster as above, we did not always rest content with its description. We examined other malfunctions which lie in its region and asked whether a broader definition of the common characteristics would include some of these (without including malfunctions in other regions). Thus, by a process of referring back and forth between spatial locations and circuit effects, we arrived at brief meaningful descriptions for clusters.

### 5.4 *The Clusters We Found*

We have identified and described 23 clusters. It would require too much space to discuss them all, so we just illustrate our results briefly by a few examples. (Although specific circuits are named to avoid vagueness, readers unacquainted with the circuits of the CC should have no difficulty following the discussion.) A few more clusters are described in the appendix to illustrate some other aspects.

The 23 clusters each contain from 6 up to about 350 malfunctions. The median number of malfunctions is 65, and the quartiles are 27 and 220.

One cluster of 58 malfunctions is associated with two intertwined circuits which are called "Add 1 C" and "Add 1 D". (These circuits are used to add 1 to the C and D addresses in an instruction.) In Fig. 7 we show this cluster geometrically. As coordinates 3 and 6 vary most within this particular cluster, we use them to display the 57 points.

In six dimensions, the center of this cluster is approximately at $(1\frac{1}{2}, -1, 2, \frac{1}{2}, -1, 2)$. Extracting coordinates 3 and 6, we see that the center of the displayed set of points should be at (2, 2), which indeed it is. In the figure, we see that the points lie along a straight line of slope about $\frac{1}{3}$. (Experienced statisticians will cover up the 3 or 4 most deviant points to strengthen the visual impression, knowing that this aids the eye in forming a more valid impression of the goodness of fit. This is because the eye weights isolated points more heavily than points in a dense re-

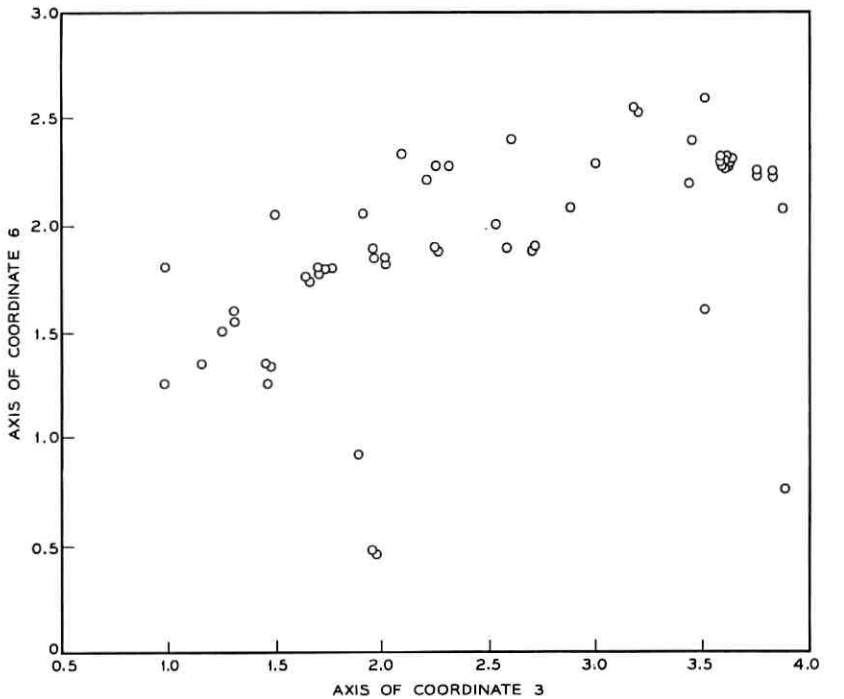



Fig. 7 — The 57 malfunctions of the "Add 1C Add 1D" cluster, displayed in the plane of coordinates 3 and 6.

gion, whereas equal weight should be attached to all points.) In fact, in six dimensions, the points lie more or less along a straight line. While we do not know the significance of this, it tends to indicate that the cluster has some internal structure.

The clusters are generally associated with "actions" rather than with circuits. By this we mean that the malfunctions in a cluster are often spread over many circuits which are considered as quite separate functional components by the circuit designers. The malfunctions in the cluster, however, always have their major disruptive effect on what circuit designers would consider as a specific action. The cluster above is unusual in this respect because it can be interpreted either way. The following examples are more typical.

One action consists of reading one or two bits from the BGS (outside the CC) and placing them in one or two flip-flops of the access register (in the CC). A cluster of about 230 malfunctions is associated with this action. The most typical malfunctions in this cluster cause the bits to be

placed in extra flip-flops of the register, or to be placed in the wrong flip-flops, or not to be placed in any flip-flops at all, under various conditions. The malfunctions in this cluster are often in one circuit (the CD memory), but occur in several other circuits as well.

Another action consists of reading one or two bits from the BGS and making a decision which depends on their values. A different cluster of about 220 malfunctions is associated with this action. The malfunctions in this cluster are scattered over many circuits and are of diverse types. Within this cluster we could pick out three subclusters, associated with much more specific actions. One such action consists of reading a 0 from the physical BGS tube numbered 0; the associated cluster has about 33 malfunctions. Another action consists of placing a bit from the BGS into a flip-flop (in the CC) called BG0, which holds it temporarily; the associated cluster has about 27 malfunctions. Another action consists of pulsing a lead called D06 from the D translator; the associated cluster has about 23 malfunctions. (The reason that this subcluster belongs in this cluster is too complex to explain here.)

## VI. BY-PRODUCTS

### 6.1 *The Main Reason for Inconsistent Diagnostic Patterns*

During the cluster analysis we discovered several interesting by-products. The most significant one is the main reason for inconsistent diagnostic results.

It was discovered very early by those making the dictionary that the same malfunction could produce quite different diagnostic results on different occasions, that is, the diagnostic results are inconsistent from one occasion to another. There is great variability among malfunctions in this respect. Some are very badly inconsistent, and were never observed to produce precisely the same diagnostic pattern twice. Other malfunctions are very stable, and are never known to produce any inconsistencies at all. Also, there is great variability among the diagnostic tests. Some participate in many inconsistencies, others in none.

There has been much speculation as to the cause of these inconsistencies, and many possible explanations have been offered. However, it has been difficult to decide which explanations actually are correct.

We analyzed large numbers of recorded inconsistencies in detail. We believe that we have established the dominant reason for the actual observed inconsistencies. It is the differences in the state of the CC at the time the diagnostic test is performed. One source of such differences is externally controlled flip-flops which signal the state of external

circuits (for example, whether the ringing signal is on or off). Still another source consists of flip-flops which the CC attempts to initialize to a certain state before diagnosis but which are not actually initialized due to the malfunction. Since diagnosis is interspersed with the normal processing of telephone traffic, failure to initialize a flip-flop means that its value at the start of the diagnostic sequence will vary in an unpredictable manner from one diagnostic run-through to another.

During construction of the dictionary, the CC had no telephone traffic to process. Moreover, the dictionary making program was present. These two factors, operating through the sources of inconsistency just mentioned, caused some fairly regular differences between the dictionary patterns and the field test patterns. For example, during dictionary construction many externally controlled flip-flops did not ever change state because the corresponding circuits were not used.

These observations do not solve the program of how to handle inconsistent diagnostic patterns, but they do perhaps provide a framework within which it is easier to attack the problem.

### 6.2 *Three Incidental Discoveries*

We have suggested that browsing through the dictionary data can reveal unexpected conclusions, *if* the browsing is facilitated by methods which permit this enormous body of data to be examined incisively. Our geometric model is one such method. We briefly mention three easily describable discoveries by way of example.

One incidental discovery was that the relay point which simulated a particular shorted diode in a particular AND gate (malfunction 24 in package F52618) developed a high resistance during much of the time that the dictionary data was being taken. Thus, the malfunction closely resembled an open diode, rather than a closed diode.

A second discovery was an error in the test program by which the data was developed. In particular, the three instructions which constitute test HP11 were misarranged.

A third discovery was that one of the (hardware) malfunction simulators stepped through the malfunctions in the wrong order. This particular simulator could be substituted for any one of the current-supplying OR gate cards. It had the capability of acting as a properly functioning card, an unplugged card, or a card with either a shorted diode, an open diode, or an open resistor on any one input lead.

For example, for card type F52626, which consists of 3 two-input gates and has 6 input leads, Table I shows both the intended order and the actual order in which the 19 malfunctions were stepped through.

TABLE I

| Input lead | Nature of malfunction | Should have been malfunction number | But was actually malfunction number |
|---|---|---|---|
| | unplugged card | 1 | 1 |
| 1 | shorted diode | 2 | 4 |
| | open diode | 3 | 2 or 3 |
| | open resistor | 4 | 2 or 3 |
| 2 | shorted diode | 5 | 6 |
| | open diode | 6 | 5 or 7 |
| | open resistor | 7 | 5 or 7 |
| 3 | shorted diode | 8 | 10 |
| | open diode | 9 | 8 or 9 |
| | open resistor | 10 | 8 or 9 |
| 4 | shorted diode | 11 | 14 |
| | open diode | 12 | 11 or 12 |
| | open resistor | 13 | 11 or 12 |
| 5 | shorted diode | 14 | 18 |
| | open diode | 15 | 13 or 15 |
| | open resistor | 16 | 13 or 15 |
| 6 | shorted diode | 17 | 19 |
| | open diode | 18 | 16 or 17 |
| | open resistor | 19 | 16 or 17 |

### 6.3 *"Forward" and "Backward" Acting Malfunctions*

It seems worthwhile here to emphasize the important difference between "forward" and "backward" acting malfunctions. While this distinction is not original with us, its importance was made abundantly clear by the great complexity of test results for backward acting malfunctions versus the relative simplicity for forward acting malfunctions.

Suppose information flows between circuits $A$, $B$, $C$, and $X$ as shown:

$$A \rightarrow \begin{matrix} \nearrow X \rightarrow B. \\ \searrow C \end{matrix}$$

Suppose circuit $X$ has a malfunction. If this malfunction causes $B$ to misoperate, we say that the malfunction acts "forward"; if it causes $A$ or $C$ to misoperate, we say that it acts "backward". For example, in the Morris CC circuitry, an open diode in an AND gate was forward-acting, as it only altered the output of the AND gate. However, a shorted diode in an AND gate was often backward-acting (as well as forward-acting) depending on the circuit configuration, as it could prevent the

input lead voltage from rising, thereby preventing other branches of the input lead from performing their intended functions.

It hardly seems possible to design a circuit which avoids forward-acting malfunctions. For if the information which flows along normal paths is wrong, the recipient circuit cannot be expected to act as intended.

On the other hand, one might hope to build a circuit in which backward-acting malfunctions are kept to a minimum. (Though, of course, this feature might have to be balanced against other desirable features.) Avoidance of backward-acting malfunctions would surely simplify the diagnostic problem greatly, not only during normal maintenance, but also during the process of debugging the first model of the machine.

## VII. CONCLUSIONS

We list several conclusions which are surely true for the data described in this paper, and which might well hold for similar diagnostic data from other digital machines.

(*i*) If different diagnostic tests are weighted suitably, then the weighted Hamming distance between test patterns is a meaningful measure of dissimilarity between malfunctions.

(*ii*) It is possible to represent the malfunctions geometrically as points in a space of low dimensions in such a way that the Euclidean distances between the points approximate the weighted Hamming distances between the corresponding patterns.

(*iii*) There may also be a geometric representation of the diagnostic tests as hyperplanes (flat cuts) in the same low dimensional space, such that each hyperplane separates most of the malfunctions which fail from the malfunctions which pass the corresponding test.

(*iv*) Representation of diagnostic patterns as points in low dimensional space offers immediate possibilities as a tool for locating malfunctions.

(*v*) This representation and the concurrent representation of tests as hyperplanes offer longer range possibilities for selecting good diagnostic tests, for eliminating redundant or useless tests, for improving diagnostic procedures, and for generally studying the relationship of malfunctions to diagnostic tests. The possible value of these representations results both from the data compression they yield and from their possible validity as models of nature.

(*vi*) Studying the diagnostic results in detail, which is made much easier by the techniques discussed in this paper, can reveal weaknesses

in the diagnostic programs and in the malfunction-simulation hardware. Such study also leads to insight and understanding which is not easily acquired by other means.

## APPENDIX

### A.1 *"Repeat Order" Cluster*

There is a special circuit for repeating certain orders up to a maximum of 32 times, with the address in the order being incremented by 1 each time. In some cases, a second part of the instruction which specifies a flip-flop in one of the access registers is also incremented by 1 each time. The major use of this repeat facility is in writing or reading a whole word between the BGS (which has single bit readout) and an access register.

One cluster consists of about 200 malfunctions in the repeat counter, which counts the repetitions. Stuck flip-flops in the counter, bad carries, and bad input are typical.

The center of gravity of this cluster is about at

$$2, \quad -1, \quad 2\tfrac{1}{2}, \quad 1, \quad -1, \quad 1.$$

As this center is quite close to the center of the first cluster in Section 5.4, a great deal of overlapping might be expected. On examination this turns out to be correct.

The intimate connection between these two clusters of malfunctions is natural because both circuits involved are used only during repeat orders. In fact, it might be more natural to treat the two clusters as a single cluster of malfunctions whose common element is that they disturb the functioning of repeat orders.

### A.2 *"Zero Flip-Flop Reading" Cluster*

There are about 80 "miscellaneous flip-flops" which can be explicitly read by the "read flip-flop" order. This cluster consists of 311 malfunctions whose common characteristic is that when any flip-flop whose value is 0 is read in this way, the answer is frequently 1.

The "read flip-flop" order operates through a large "flip-flop reading" circuit which is shown in Fig. 8. The action of this circuit is to transfer the value of the selected miscellaneous flip-flop into a control flip-flop known as FF, and to use the value from there. There are about 230 so-called "isolation" diodes through which the various miscellaneous flip-flop values are funnelled into FF. The circuit is so arranged that

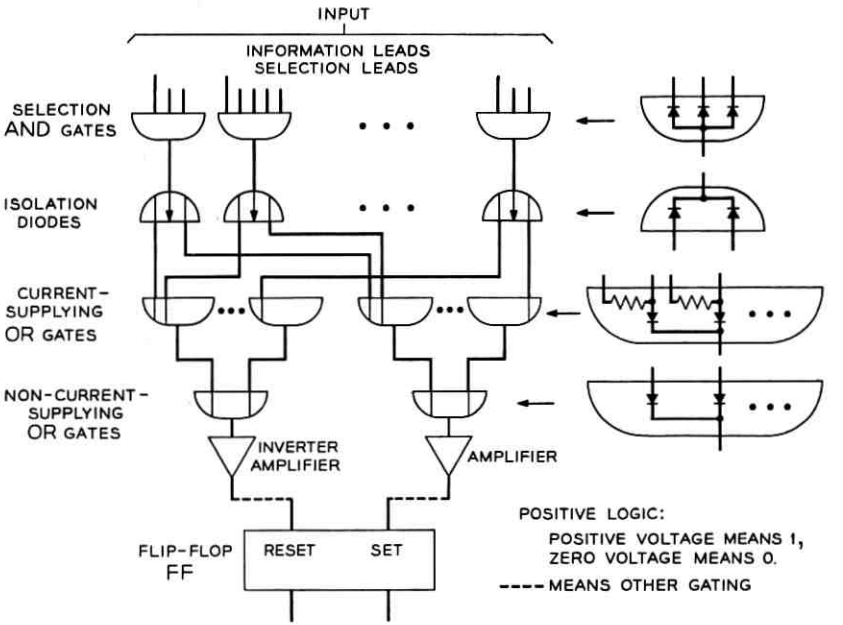



Fig. 8 — Simplified block diagram for flip-flop reading circuit.

open-circuiting *any* one of these diodes causes trouble when reading a 0 value from *any* miscellaneous flip-flop (though a 1 value is read correctly). For half these diodes, an open circuit causes the value read to be the value left in FF from before. For the other half of these diodes, an open circuit causes both the "set" and "reset" leads of FF to be pulsed simultaneously; we do not know what effect this has, but we believe that it usually leaves FF unchanged.

Besides open-circuited isolation diodes, removal of a package which contains several of these diodes, or removal of a gate package which feeds into these diodes, has the effect of causing a 0 value *always* to be read as a 1. Malfunctions of this sort also belong to this cluster.

There are a variety of other malfunctions which belong to this cluster. For example, there is an amplifier which feeds the "set" input to FF during the "read flip-flop" operation. If its output is stuck in the high voltage state, or if it is removed (which has almost identical circuit results), or if the diode through which it feeds is open-circuited, we should and do obtain malfunctions in the cluster. If the amplifier which feeds the "reset" input to FF during "read flip-flop" operation is stuck at low voltage output, we obtain a malfunction in the cluster. If the

gating lead (called RFFT), which gates the "set" and "reset" impulses into FF during "read flip-flop" operation, is prevented from operating by a short-circuited diode just on those occasions when a 0 value is being read, we again obtain a malfunction in the cluster.

A few of the malfunctions which belong to the cluster require deeper explanations. For example, if the RFFT lead (referred to above) *never* operates, because the amplifier feeding it is stuck at low voltage output, we again obtain a malfunction in the cluster. The effect of this malfunction is that FF *always* retains its previous value during a "read flip-flop" operation, regardless of whether the value being read is 0 or 1. It is not immediately clear why this malfunction should give test patterns closely resembling others in the cluster. However, by analyzing the diagnostic test program, we find that FF (which is also used by other operations) is most often left with a 1 in it upon entering the critical diagnostic test operations. Thus, this malfunction most often causes errors in reading the value 0.

This cluster has relatively sharp boundaries. Also, it is a "pure" cluster, that is, all the malfunctions in the region of space it occupies belong to it; other clusters do not overlap. The center of gravity of the cluster is approximately at

$$2\tfrac{1}{2},\quad 1,\quad 1,\quad 0,\quad 2,\quad 0.$$

The extreme values of the various coordinates are as shown:

| | | | | | | |
|---|---|---|---|---|---|---|
| minima | 1, | 0, | 0, | −2, | 0, | −1, |
| restricted minima | 1, | 0, | 0, | −1, | 1, | 0, |
| restricted maxima | 4, | 2, | 2, | 1, | 2, | 1, |
| maxima | 6, | 4, | 2, | 2, | 3, | 4. |

By "restricted maxima" we mean the maximum values for the 291 most centrally located of the 311 malfunctions; the other 20 are rather thinly sprinkled.

No two malfunctions in this cluster have identical test patterns. This may seem strange, for those package removals which cause a 0 value always to be read as a 1 have identical circuit effects. Also, those isolation-diode open-circuits that cause a 0 value to be read as whatever was left in FF from before have identical circuit effects. Why should the malfunctions within one of these groups produce diverse test patterns?

In the case of the package removals, almost all the variations in the test patterns result from reading flip-flops whose value is controlled from outside the CC and varies from time to time. Some examples are flip-flops RTA ("ringing tone active"), BSYT ("busy tone"), RNGS

("ringing scan") and 10MSCK ("10 millisecond clock"). When the value happens to be 1 the value is read correctly and the corresponding diagnostic tests are passed; when the value is 0, it is read incorrectly, and the contrary happens. Several of these flip-flops are read (in effect) four times during the diagnostic program, and we are able to follow any changes which take place. The slower changing ones like RTA are indeed observed to change either not at all or only once during the course of a single diagnostic run-through, while a faster changing one like RNGS is sometimes observed to change more often.

In the case of the isolation-diode open-circuits whose effect is to leave in FF its previous value, the test results are subject to the same source of variability. However, they are also subject to the additional variability of depending on the previous contents of FF. While this is more often 1 than 0, it is 0 significantly often. Thus, a test which the previous group fails may be passed by this group, and vice versa. All the tests ever failed by this group, however, include all the tests ever failed by the previous group, and more as well.

An interesting sidelight concerns the difference between even-numbered and odd-numbered tests. The diagnostic tests are conducted in pairs, with nondiagnostic work intervening between pairs. For this reason, each even-numbered test is entered from within the diagnostic program; analysis reveals that in this case FF contains a 1 when the individual test sequence is entered. As the relevant program was not available, we were unable to determine the situation for odd-numbered tests, but we infer indirectly that FF could have either value depending on unknown circumstances. As a consequence, certain tests on different flip-flops which are exactly similar to each other (including the fact that the flip-flop normally has value 0 at the time) fail or pass depending on whether the test is even or odd-numbered. There are several examples of this sort.

The reader may suspect that our circuit analysis is incomplete, and may suspect that the pattern differences for different malfunctions actually reflect different circuit effects. Fortunately we have at least three cases in which the same identical malfunction in this cluster was diagnosed twice. The differences between the test patterns for the self-same malfunction were quite as great as between test patterns for different malfunctions of one type in the cluster.

### A.3 *Two Clusters Affecting the BGS Address Register*

These two clusters give a useful insight into the process of cluster analysis. It is probably true that they should be merged into one larger cluster which includes them both.

One consists of about 60 malfunctions affecting the EPO (execute program order) gating pulse which comes through a transformer into the BGS address circuit. This transformer serves only the BGS address circuit and an associated circuit. Almost all the malfunctions in this cluster are shorted diodes which short circuit this EPO lead. One or two are malfunctions in the transformer which prevent the pulse from appearing.

The second cluster consists of about 65 malfunctions which cause one or both of the gating leads BSBGX or BSBGY to operate or fail to operate. These leads are the leads which enable input to the X and Y halves of the BGS address register.

The centers of these two clusters are approximately at

$$3\tfrac{1}{2}, \quad -1, \quad 2\tfrac{1}{2}, \quad -\tfrac{1}{2}, \quad -\tfrac{1}{2}, \quad 1\tfrac{1}{2}$$

and

$$4, \quad -1, \quad 2, \quad -\tfrac{1}{2}, \quad 0, \quad 2.$$

Thus, they are fairly close together by comparison with the sizes of the clusters, which suggests that the clusters may overlap a good deal. Closer examination reveals that they do overlap a great deal.

This suggests that we do not have two distinct clusters but two types of malfunctions in one cluster. The next natural step would be to speculate that the common element to both clusters is a serious input difficulty to the BGS address register. To carry the analysis further we would then examine the other malfunctions in the region of space which these two clusters occupy, and see if many of them fit this new description. If so, we would systematically trace out (from the circuit diagrams) all like malfunctions. If essentially all of them should lie in this same region, then we would consider that we had arrived at a satisfactory cluster.

REFERENCES

1. Joel, A. E., Quirk, W. B., et al, An Experimental Electronic Telephone Switching System: A conference at Morris, Illinois, October 12, 1960. (Part of the 1960 General Fall Meeting of the AIEE.) (Unpublished document.)
2. Joel, A. E., Jr., An Experimental Switching System Using New Electronic Techniques, B.S.T.J., *37*, 1958, pp. 1091–1124.
3. Seckler, H. N. and Yostpille, J. J., Functional Design of a Stored-Program Electronic Switching System. B.S.T.J., *37*, 1958, pp. 1327–1382.
4. Haugk, G., Greenwood, T. S., and Yostpille, J. J., Morris Electronic Switching System — Final Report. (Unpublished document, March, 1963.)
5. Tsiang, S. H. and Ulrich, Werner, Automatic Trouble Diagnosis of Complex Logic Circuits, B.S.T.J., *41*, 1962, pp. 1177–1200.

# Comparison Between a Gas Lens and Its Equivalent Thin Lens

By D. MARCUSE

(Manuscript received June 23, 1966)

*Gas lenses can be replaced by equivalent thin lenses. This paper shows a comparison between ray traiectories through 100 gas lenses and 100 equivalent thin lenses. The agreement is good enough to warrant the use of equivalent thin lenses for the study of the transmission properties of beam waveguides made of gas lenses.*

## I. INTRODUCTION

Gas lenses have been studied for their potential use as focusing elements in beam waveguides.[1,2,3,4] Two earlier papers[2,3] were concerned with the study of the optical properties of a particular gas lens (see Fig. 1) and came to the conclusion that certain types of gas lenses behave as optically thin lenses. The equivalent thin lens approximating the optical properties of the gas lens is not flat but deformed to fit the shape of the principal surface of the gas lens.

The definition of the equivalent thin lens is based on the optical properties of the gas lens for input rays parallel to the optical axis. For those rays the two lenses are optically equivalent by definition. This equivalence need not necessarily hold true for arbitrary input rays. To show that the equivalent lens can replace the gas lens for arbitrary input rays is the purpose of this paper. For the purpose of optical waveguides a gas lens can be replaced by an equivalent thin lens if the ray trajectories through many gas lenses coincide reasonably closely with the ray trajectories through the equivalent thin lenses. A computer simulated experiment was conducted to determine the ray trajectories through 100 gas lenses and through 100 equivalent lenses and to compare their results. It will be shown in this paper that the two ray trajectories are very nearly the same. This result allows us to use the equivalent thin lenses to study the light guidance properties of gas lenses. This replacement is particularly desirable to examine the wave optics properties

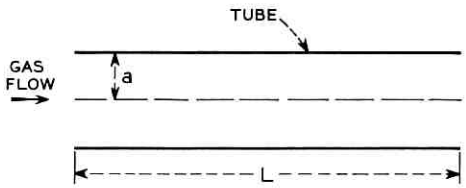



Fig. 1 — Tubular gas lens. A cool gas is blown into a warm tube.

of gas lenses since it would be prohibitively complicated to solve the problem of wave propagation through the actual gas lens.

## II. RAY TRACING THROUGH GAS LENSES AND THIN LENSES

The details of determining the principal surface and focal length of a gas lens are discussed in Ref. 3. Typical results of the principal surface and the dependence of the focal length on ray position are shown in Figs. 2 and 3. Strictly speaking there are two principal surfaces. Since they coincide rather closely, however, only one will be considered.

The equivalent thin lens is assumed to have the shape of the principal surface of the gas lens, as shown in Fig. 2, and is assigned the focal length $f$ of the gas lens with its dependence on radius as shown in Fig. 3.

Ray tracing through the gas lens is accomplished by numerical integration of the ray equation. Since rays are being traced through 100 gas lenses in succession, high accuracy is required. For that reason I used the exact ray equation instead of the approximation which was sufficient for the purpose of Ref. 3. The ray trajectory in the gas lens is obtained by numerical integration of the ray equation. This trajectory, however,

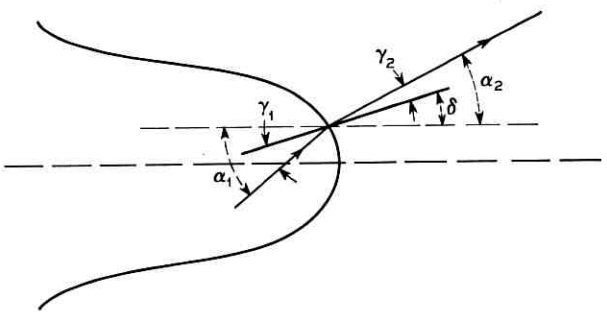



Fig. 2 — The principal surface of the tubular gas lens. The angles used for ray tracing are indicated.

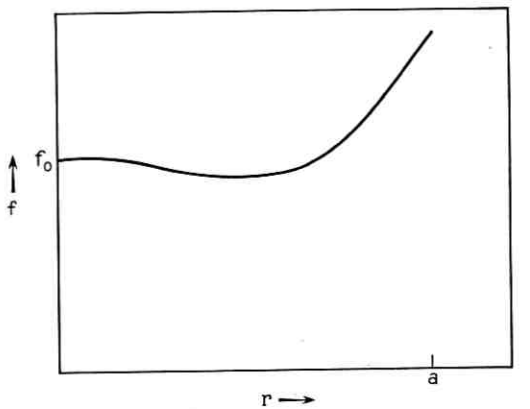



Fig. 3 — Focal length dependence on radius for the tubular gas lens under typical operating conditions.

cannot be used for comparison with the ray trajectory through the equivalent lenses. To compare the two trajectories, the ray entering each gas lens was extended in a straight line into the lens to find the point at which it intercepted the principal surface. This point was used for comparison with the ray trajectory through the thin lenses.

Ray tracing through warped thin lenses has to be done with care since it is easy to violate laws of nature. One might be tempted to use the usual procedure for straight thin lenses and simply break each ray entering the lens at a distance $r$ from the optical axis by an angle $\beta$, which is independent of the input angle, according to

$$\tan \beta = -\frac{r}{f}. \tag{1}$$

It was pointed out in Ref. 5 that (1) violates Liouville's theorem of statistical mechanics and that one has to use the equation

$$\sin \gamma_1 = \sin \gamma_2 + F(r). \tag{2}$$

The angle $\gamma_1$ is formed between the input ray and the direction normal to the lens surface and $\gamma_2$ is the angle between the normal direction and the output ray, Fig. 2. To compute the ray trajectory through the thin lens we have to determine the angle $\gamma_1$ from the input angle $\alpha_1$ of the ray with respect to the optical axis and the angle $\delta$ of the lens normal with respect to the optical axis,

$$\gamma_1 = \alpha_1 - \delta. \tag{3}$$

Then we determine $\gamma_2$ from (2) and obtain $\alpha_2$ from the equation

$$\alpha_2 = \gamma_2 + \delta.$$

The function $F(r)$ in (2) is determined from the known focal length of the lens. If $\alpha_1 = 0$, we obtain from (3) $\gamma_1' = -\delta$. The angle $\alpha_2$ for an input ray parallel to the optical axis is known from the focal length of the lens

$$\tan \alpha_2' = -\frac{r}{f},$$

so that

$$\gamma_2' = \alpha_2' - \delta.$$

The function $F(r)$ is therefore, determined from

$$F(r) = \sin \gamma_1' - \sin \gamma_2'. \tag{5}$$

This complicated procedure does not lend itself easily to the formulation of a difference equation to determine the ray trajectories. An analytical solution for the ray trajectories through warped thin lenses cannot be obtained as easily as for thin straight lenses.[6] However, numerical ray tracing with the help of an electronic computer is only slightly more involved and time consuming as for thin straight lenses.

The results of ray tracings through gas lenses and equivalent thin lenses are shown in Figs. 4 and 5. The solid curve is the gas lens ray trajectory, the broken curve is the corresponding ray trajectory through the equivalent thin lenses. The points entered in these curves are the points of intersection of the (extended) rays with the principal surface of the gas lens or with the equivalent thin lens. These points were connected by straight lines. This procedure represents the ray trajectory through the thin lenses exactly. For the gas lenses it gives the exact ray trajectory only outside of the lenses. The two figures show the ray trajectories only from lens 62 to 100, the agreement is better at the beginning of the trajectory.

The two trajectories agree very well in Fig. 4. If the radius of the gas lens tubes is assumed as $a = 3$ mm the ratio of lens spacing $D$ to lens radius $a$ ($D/a = 1200$ for Fig. 4) corresponds to lenses spaced 3.6 $m$ apart. Fig. 5 was computed with a ration $D/a = 330$ so that with $a = 3$ mm the lens spacing would be $D = 0.99$ m. Even for lenses spaced that close the concept of equivalent thin lenses works quite well.

These results show that the gas lenses can be replaced by equivalent

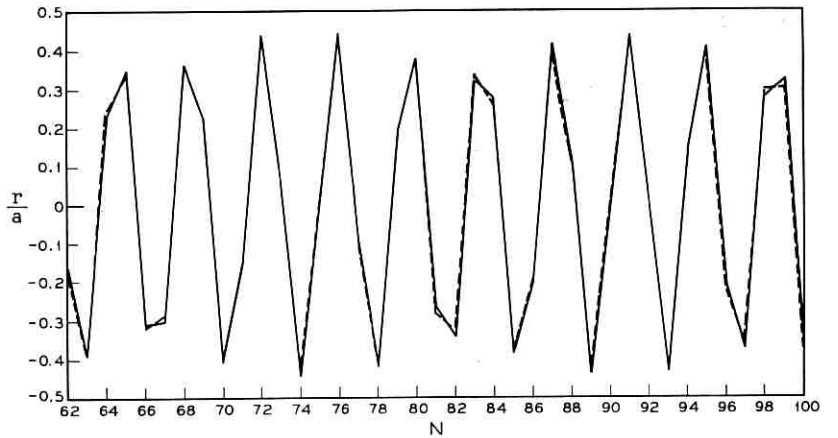



Fig. 4 — Comparison of ray trajectories through gas lenses and equivalent thin lenses. $n$ = lens number, $a$ = radius of gas lens, $D$ = lens spacing, $f_o$ = focal length of rays close to the optical axis, $L$ = length of gas lens. $D/a = 1200$, $D/f_o = 2.16$, $L/a = 50$.

thin lenses. This replacement does not simplify the problem of ray tracing or of tracing wave field through the gas lenses sufficiently to make it accessible to an analytic treatment but it simplifies the numerical treatment greatly and reduces the time of numerical calculations to an economically acceptable level.

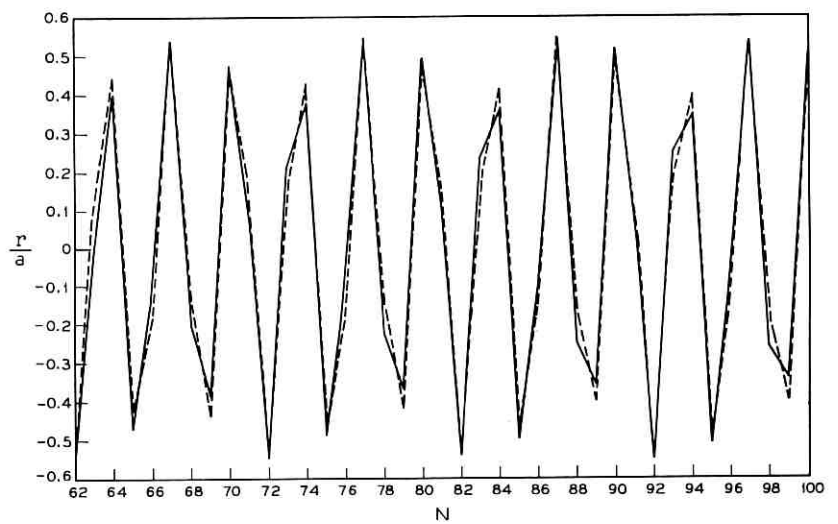



Fig. 5 — Same as Fig. 4 with $D/a = 330$, $D/f_o = 2.74$, $L/a = 50$.

REFERENCES

1. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., *43*, July, 1964, pp. 1469–1475.
2. Marcuse, D. and Miller, S. E., Analysis of a Tubular Gas Lens, B.S.T.J., *43*, July, 1964, pp. 1759–1782.
3. Marcuse, D., Theory of a Thermal Gradient Gas Lens, IEEE, Trans. MTT, *13*, November, 1965, pp. 734–739.
4. Marcuse, D., Properties of Periodic Gas Lenses, B.S.T.J., *44*, November, 1965, pp. 2083–2116.
5. Marcuse, D., Physical Limitations on Ray Oscillation Suppressors, B.S.T.J., *45*, May–June, 1966, pp. 743–751.
6. Hirano, J. and Fukatsu, Y., Stability of a Light Beam in a Beam Waveguide, Proc. IEEE, *52*, November, 1964, pp. 1284–1292.

# Deformation of Fields Propagating Through Gas Lenses

By D. MARCUSE

(Manuscript received June 23, 1966)

*The concept of a thin lens equivalent to a gas lens is used to calculate distortions of off-axis Gaussian fields in beam waveguides composed of gas lenses. A computational method for the numerical solution of this problem based on the Kirchoff-Huygens diffraction integral is developed. It is shown that off-axis Gaussian fields deform considerably as they travel through a sequence of gas lenses. These deformations are substantial even though the lens distortions may be small. If the light beam deforms it is hard, if not impossible, to steer it back on-axis. This problem can be avoided if some means of beam redirection are used to keep the field on-axis, thus preventing the occurrence of significant beam deformation.*

## I. INTRODUCTION

Interest in optical communications has stimulated research to find a suitable optical transmission medium. The beam waveguide first suggested by Goubau[1] appears to be an efficient optical waveguide. It is composed of lenses which periodically refocus the light beam, counteracting its tendency to spread apart by diffraction.

Gas lenses have been suggested as focusing elements of beam waveguides.[2,3,4] Of the various types of gas lenses, the tubular gas lens, Fig. 1(a), has been studied in some detail.[3,4] This gas lens can be represented by an equivalent thin lens which is warped to fit the shape of the principal surface of the gas lens and which is given its focal length with the proper dependence on its radius. It was shown in Ref. 5 that ray trajectories through 100 gas lenses coincide closely with ray trajectories through the corresponding equivalent lenses. Replacing the complicated gas lens with the equivalent thin lens simplifies considerably the study of beam waveguides composed of gas lenses.

In this paper, we will make use of the equivalent thin lens concept to investigate the propagation of wave fields through a beam waveguide of

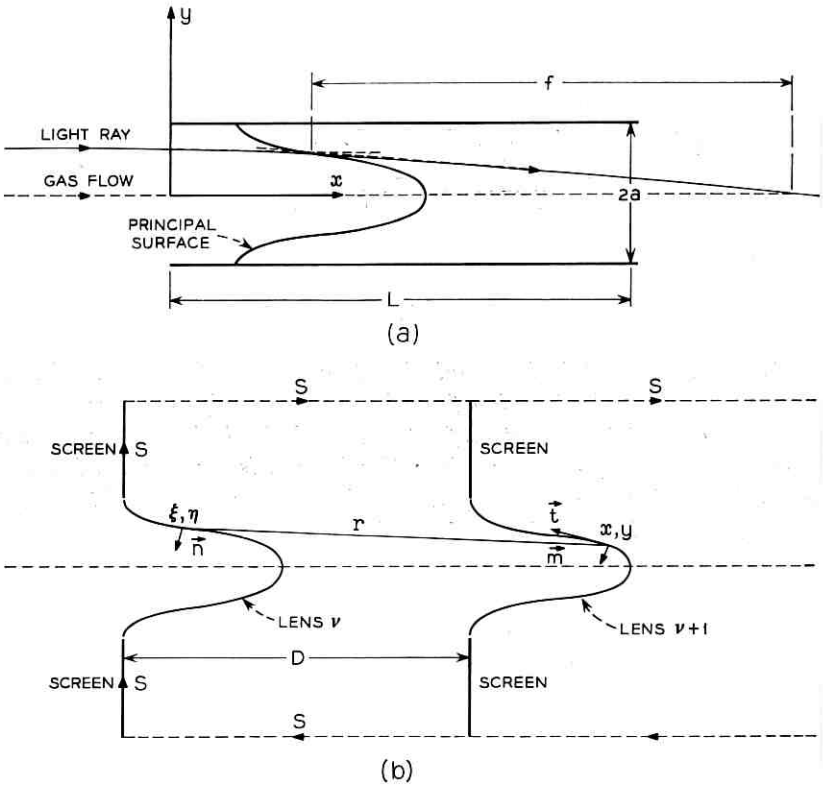



Fig. 1— (a) Schematic of the gas lens indicating the definition of principal surface and focal length. (b) The equivalent warped, thin lenses representing the gas lens beam waveguide.

gas lenses. The justification for replacing the gas lenses with equivalent lenses comes from geometric optics.[5] One might wonder if the argument based on geometric optics can be carried over into wave optics. The geometric optics description neglects diffraction effects. Inasmuch as diffraction effects can be neglected as the field passes through the lens, the geometric optics description should give the correct answer. Based on this line of reasoning, one may expect the equivalent thin lens to be a good approximation, as long as the gas lenses are short compared to their spacing.

The wave optics properties of the beam waveguide composed of gas lenses are obtained using a two-dimensional version of the scalar Kirchoff-Huygens diffraction integral. The problem had to be limited to two

dimensions to make it tractable for computer calculations. This simplification can be visualized as replacing the actual lenses by cylindrical lenses.

We study how off-axis field distributions with a Gaussian intensity profile propagate through the beam waveguide. Unfortunately there are further limitations on the physical problem we can compute, imposed by the limited size of the available computer memory. The calculations are accelerated if as much of the integral kernel as possible can be stored in the machine without having to recalculate it each time it is needed. The IBM 7094 used for these calculations has 24,000 storage locations available in its memory. Since we are dealing with a complex kernel, 100 integration points across the (linear) lens require 20,000 storage locations. This means that we can use no more than 100 integration points to compute our problem. This limits the ratio of lens aperture to field extension across the lens which we can use. Either we use the full lens aperture and launch a field which fills an appreciable part of it or we use a very narrow field distribution and limit the aperture to a size which allows us to approximate the narrow field reasonably well with the 100 integration points at our disposal. This limitation forced me to calculate the field distribution in the gas lens either at a much lower frequency than that of the visible 6328Å line of a He-Ne laser or to take the actual laser frequency but use only a small fraction of the actual lens aperture.

In spite of all these limitations imposed by computer economics, some interesting results can still be obtained.

In a beam waveguide composed of ideal lenses no field distortion results as an off-axis Gaussian beam travels through the waveguide. In a beam waveguide composed of gas lenses, off-axis Gaussian beams break up into double humped shapes and deform so much that it is hard to locate the initially well defined field distribution. This result is important for beam waveguides using electronic control mechanisms to reposition a beam when it has wandered away from the waveguide axis.[6] If the beam breaks up into several beams, repositioning becomes impossible. This problem can be minimized by using two gas lenses back-to-back close together. The resulting combined lens has far less principal plane distortion as the individual lenses and leads to far less field distortion.

The field distortion observed in these simulated gas lenses can be attributed in part to the distortion of the principal plane. A fictitious lens with the same focal length aberration as the gas lens but an undistorted principal plane shows less field distortion. However, the focal length aberration also contributes its share of field distortions.

A large part of this paper is taken up with the description of the calculation procedure. This is justified since the development of a workable and logical procedure is perhaps the main contribution of this work. The reader who is interested only in the numerical results may skip over the following two sections to the section entitled "Discussion of Numerical Results."

## II. THE TWO-DIMENSIONAL DIFFRACTION INTEGRAL

The Kirchoff-Huygens diffraction integral is a solution of the scalar wave equation.

$$\Delta\Psi + \beta^2\Psi = 0. \tag{1}$$

As explained in the introduction, we are not interested here in the three-dimensional case usually treated but in its two-dimensional counterpart. The two-dimensional Kirchoff-Huygens integral is

$$\Psi(x,y) = \frac{i}{4}\int_s \left\{\frac{\partial\Psi}{\partial n} H_o^{(1)}(\beta r) - \Psi \frac{\partial}{\partial n} H_o^{(1)}(\beta r)\right\} dS. \tag{2}$$

The integral is to be extended over a closed curve $S$, $n$ indicates the direction of the normal to the curve $S$ which counts positive if it points outward of the area enclosed by $S$. $H_o^{(1)}$ is the Hankel function of zero order and first kind. The variable $r$ is the distance between the observation point $x,y$ inside of $S$ and the integration point $\xi$, $\eta$ on $S$,

$$r = \sqrt{(x-\xi)^2 + (y-\eta)^2}. \tag{3}$$

$dS$ is the line element along the curve $S$. The constant $\beta$ is related to the wavelength $\lambda$ of the radiation field by

$$\beta = \frac{2\pi}{\lambda}. \tag{4}$$

We are dealing with an optical radiation field. The observation point $x,y$ will always be far enough from the line $S$ so that

$$\beta r \gg 1.$$

It is, therefore, possible to replace the Hankel function by its approximation for large argument and write (2)

$$\Psi(x,y) = \frac{\exp\left(i\frac{\pi}{4}\right)}{\sqrt{8\pi\beta}} \int_s \left\{\frac{\partial\Psi}{\partial n}\frac{\exp(i\beta r)}{\sqrt{r}} - \Psi\frac{\partial}{\partial n}\left(\frac{\exp(i\beta r)}{\sqrt{r}}\right)\right\} dS. \tag{5}$$

Equation (5) relates the values of the field $\Psi(\xi,\eta)$ on $S$ to its values inside $S$. We want to use this expression to calculate the field at lens $n+1$ if the field at lens $n$ is known. Our lenses are the equivalent thin lenses of Fig. 1 (b) which represent the gas lens of Fig. 1 (a). The fields have to be known over the surface of the lens which is not plane. We assume that the lens is apertured by an opaque screen and follow the usual practice of setting

$$\Psi(\xi,\eta) = 0 \quad \text{and} \quad \frac{\partial\Psi}{\partial n} = 0 \tag{6}$$

on the screen. We use as the curve $S$ the line formed by the lens surface. the opaque screen which extends from $-\infty < \eta < \infty$, and close it by a suitable curve at infinity. The following lens of the beam waveguide lies thus inside $S$, Fig. 1 (b).

The Kirchoff-Huygens integral presents a problem. It requires us to know not only $\Psi$ on $S$ but also $\partial\Psi/\partial n$. It is not sufficient, therefore, to simply evaluate the integral (5) but also the integral which follows from it by differentiation with respect to the normal $\mathbf{m}$ to the surface of the next lens in the beam waveguide.

A substantial simplification results if instead of $\Psi$ we use a function $\Phi$ defined by the equation

$$\Psi = \Phi e^{i\beta x}. \tag{7}$$

This transformation serves the following purpose. The field propagating in the beam waveguide can be expected to have phase fronts which are not too different from that of plane waves. Since we collect the field over the curved surface of the lenses we have a substantial phase variation simply because the curved surface crosses many phase fronts of the almost plane wave. The transformation (7) displays explicitly the plane wave part of the phase variation. The remaining phase variation left in $\Phi$ is much less rapid and therefore much easier to calculate. Substituting (7) into (5) leads to an equation for $\Phi$. We also replace the phase constant $\beta$ by

$$\beta = 2\pi N \frac{D}{a^2} \tag{8}$$

with

$$N = \frac{a^2}{D\lambda}. \tag{9}$$

$N$ is the Fresnel number which is often used to characterize optical

resonators and beam waveguides. $D$ is the distance between lenses and "$a$" the half-width of their apertures.

Replacing $\Psi$ by $\Phi$ introduces the term $\exp[i\beta(r + \xi - x)]$ under the integral sign. We make use of the fact that $x - \xi$ is almost as large as $r$ and write approximately

$$r + \xi - x = \frac{1}{2}\frac{(y-\eta)^2}{x-\xi}\left\{1 - \frac{1}{4}\left(\frac{y-\eta}{x-\xi}\right)^2\right\}. \tag{10}$$

Using (7), (8), and (10) we can rewrite (5)

$$\begin{aligned}\Phi_{\nu+1}(y) = \frac{\sqrt{ND}}{2a}\exp\left(-i\frac{\pi}{4}\right)\int_{-a}^{a}\left\{\left(\frac{\partial r}{\partial n} - \frac{\partial \xi}{\partial n}\right)\Phi_\nu(\eta)\right.\\ \left. + \varphi_\nu(\eta)\right\}\frac{\sqrt{1+\left(\frac{d\xi}{d\eta}\right)^2}}{\sqrt{r}}\\ \cdot\exp\left[i\pi N\frac{D}{a^2}\frac{(y-\eta)^2}{x-\xi}\left\{1 - \frac{1}{4}\left(\frac{y-\eta}{x-\xi}\right)^2\right\}\right]d\eta.\end{aligned} \tag{11}$$

The line element $dS$ was expressed by

$$dS = \sqrt{1+\left(\frac{d\xi}{d\eta}\right)^2}\,d\eta \tag{12}$$

where $\eta = \eta(\xi)$ or $\xi = \xi(\eta)$ is the function describing the curved lens. The function $\varphi_\nu(\eta)$ is defined by

$$\varphi_\nu(\eta) = \frac{i}{\beta}\frac{\partial \Phi_\nu(\eta)}{\partial n}. \tag{13}$$

The subscripts $\nu$ and $\nu + 1$ have been added to underscore the iterative nature of the process.

The iterative equation for the calculation of $\varphi_{\nu+1}$ follows from $\Phi_{\nu+1}$ by differentiation. Neglecting certain small terms under the integration sign results in

$$\begin{aligned}\varphi_{\nu+1}(y) = \frac{\sqrt{ND}}{2a}\exp\left(-i\frac{\pi}{4}\right)\int_{-a}^{a}\left\{\frac{1}{2}\left(\frac{y-\eta}{x-\xi}\right)^2\frac{\partial x}{\partial m} - \frac{y-\eta}{x-\xi}\frac{\partial y}{\partial m}\right\}\\ \cdot\left\{\left(\frac{\partial r}{\partial n} - \frac{\partial \xi}{\partial n}\right)\Phi_\nu(\eta) + \varphi_\nu(\eta)\right\}\frac{\sqrt{1+\left(\frac{d\xi}{d\eta}\right)^2}}{\sqrt{r}}\\ \cdot\exp\left[i\pi N\frac{D}{a^2}\frac{(y-\eta)^2}{x-\xi}\left\{1 - \frac{1}{4}\left(\frac{y-\eta}{x-\xi}\right)^2\right\}\right]d\eta.\end{aligned} \tag{14}$$

The symbol $m$ was used to designate the normal of the $(\nu + 1)$th surface $y = y(x)$.

For reasons explained later, we also need the derivation of $\Phi$ in tangential direction $t$. Defining

$$\chi = \frac{i}{\beta} \frac{\partial \Phi}{\partial t} \tag{15}$$

we get the integral expression for $\chi_{\nu+1}$ by replacing $\partial/\partial m$ by $\partial/\partial t$ in (14), it is unnecessary to write this expression down since it is exactly the same as that for $\varphi_{\nu+1}$ except for the change just mentioned.

The three integrals for $\Phi$, $\varphi$, and $\chi$ have a substantial part of their integrands in common. This similarity facilitates the machine calculations of these integrals greatly.

The power flow through the lenses can be computed from the expression[7]

$$P_\nu = \frac{\omega}{2} \int_{S_\nu} Im(\Psi \nabla \Psi^*) dS \tag{16}$$

with $\omega$ being the angular frequency of the radiation field and $Im$ denoting the imaginary part of the expression in parenthesis. Or replacing $\Psi$ by $\Phi$ and the line element by (12) we get with the help of (13)

$$P_\nu = \frac{\omega\beta}{2} \int_a^a \left\{ \mathrm{Re}\,(\Phi_\nu \varphi_\nu^*) - \frac{\partial \xi}{\partial n} |\, \Phi_\nu \,|^2 \right\} \sqrt{1 + \left(\frac{d\xi}{d\eta}\right)^2}\, d\eta. \tag{17}$$

Equation (17) can be used to compute the power flow through the lenses and observe power loss due to diffraction caused by the finite lens apertures.

The reader who is familiar with the work of Fox and Li[8] might wonder why the present case is so much harder to compute than the resonators studied by these authors. Fox and Li used only one integral to describe the field distribution over one mirror in terms of the field distribution over the other, they did not calculate the integral for $\partial\Psi/\partial n$ simultaneously with that for $\Psi$. The reason for the success of the much simpler theory in their case was the fact that the surfaces over which they had to integrate were either perfectly flat or very nearly plane. The normal derivatives occurring in (2) involve the cosines of angles $\varepsilon$ between the normal to the surface of integration and the normal to the phase fronts of the wave. As long as this angle is small

$$\cos \varepsilon \approx 1$$

and the angle can be ignored. For the purpose of the normal derivative

the wave can be treated as perfectly plane and the derivative can be written as

$$\frac{\partial \Psi}{\partial n} = i\beta\Psi. \tag{18}$$

However, the angle $\alpha$ between the direction normal to the surface of our lenses and the optical axis is not small. If $\varepsilon$ is again the angle between the normal to the phase front of the wave and the optical axis then $\alpha + \varepsilon$ is the angle entering the cosine. But even if $\varepsilon$ is small

$$\cos(\alpha + \varepsilon) \approx \cos\alpha - \varepsilon \sin\alpha.$$

The departure of the phase front from a plane wave can no longer be neglected but enters in first order. The expression (18) is no longer a valid approximation and the whole calculation becomes much more difficult.

### III. FIELD TRANSMISSION THROUGH THE LENSES

So far we have considered the transmission of the field from the surface of one lens to that of the next. However, the lenses have so far not even entered the picture other than to force us to calculate the field over the surface of the lens. The process of calculating the effect of the lens on the field is also rather complicated. In the case of plane, thin lenses it is sufficient to regard the lens simply as a phase transformer which retards the phase of the field differently in different parts of the lens. This simple picture is inapplicable in our case of curved lenses.

Liouville's theorem of statistical mechanics is the guide to the proper description of a thin lens. I have shown in two earlier papers[5,9] how rays pass through thin lenses. The ray gets broken by the lens by an angle which depends not only on the part of the lens which the ray intersects, but also by the angle between the ray and the normal to the lens surface. If $\gamma_1$ is this angle for the entering ray and $\gamma_2$ that for the ray leaving the lens the dependence between these two angles is given by[9]

$$\sin\gamma_2 = \sin\gamma_1 + F(y). \tag{19}$$

The function $F(y)$ is determined by the lens. The focusing property of the lens determines the angle $\gamma_2'$ if $\gamma_1'$ corresponds to a ray incident parallel to the optical axis. $\gamma_1'$ and $\gamma_2'$ are known from the desired focal length of the lens and its shape. $F(y)$ is determined by substituting $\gamma_2 = \gamma_2'$ and $\gamma_1 = \gamma_1'$ into (19).

These ray optics properties of the lens have to be used to determine its influence on the field. The normal directions to the phase fronts coin-

cide with the rays associated with the field, they have to be determined from the derivatives of the field function. Let us assume that we split the field function $\Psi$ into its magnitude $G$ and phase angle $\beta\vartheta$

$$\Psi = G \exp (i\beta\vartheta)$$

or using $\Phi$ rather than $\Psi$

$$\Phi = G \exp [i\beta(\vartheta - x)]. \tag{20}$$

The function $\vartheta(x,y)$ is the eikonal of geometric optics and satisfies the eikonal equation of free space[10]

$$| \nabla\vartheta | = 1. \tag{21}$$

We take the tangential derivative of $\Phi$

$$\frac{\partial\Phi}{\partial t} = \left[ i\beta \left( \frac{\partial\vartheta}{\partial t} - \frac{\partial x}{\partial t} \right) + \frac{\partial G}{\partial t} \frac{1}{G} \right] \Phi. \tag{22}$$

The term $\partial\vartheta/\partial t$ can be expressed in the following way

$$\frac{\partial\vartheta}{\partial t} = \nabla\vartheta \cdot \frac{\partial \mathbf{s}}{\partial t} = | \nabla\vartheta | \left| \frac{\partial \mathbf{s}}{\partial t} \right| \cos (kt);$$

$\partial\mathbf{s}/\partial t$ is a unit vector in the tangential direction $t$, $(kt)$ is the angle between the direction normal to the phase front of the wave and the tangential direction. Using (21) and the property of the unit vector we obtain

$$\frac{\partial\vartheta}{\partial t} = \cos (kt). \tag{23}$$

With the help of (15) and (23) we get from (22)

$$\cos (kt) = \frac{\partial x}{\partial t} - \frac{\chi}{\Phi} + \frac{i}{\beta} \frac{1}{G} \frac{\partial G}{\partial t}.$$

The left-hand side of this equation is real by definition and so is $G$ and its derivatives. This means that the imaginary parts of the right-hand side have to cancel each other and we obtain

$$\cos (kt) = \frac{\partial x}{\partial t} - \mathrm{Re} \left( \frac{\chi}{\Phi} \right). \tag{24}$$

Re designates the real part of the expression in parentheses. The derivative $\partial x/\partial t$ is known from the geometry of the lens, and $\chi$ as well as $\Phi$ have been computed from their integral expressions. The angle between the rays associated with the field and the tangential direction $t$ of the

curve describing the lens shape is thus determined. The angle $(kt)$ is related to $\gamma_1$, the angle between the ray and the normal to the lens surface, by

$$\gamma_1 = \frac{\pi}{2} - (kt) \tag{25}$$

so that

$$\cos (kt) = \sin \gamma_1 .$$

The angle $\gamma_2$ between output ray and lens normal is obtained from (19). Indicating by a prime the angles and field quantities of the field after leaving the lens we have

$$\cos (kt)' = \sin \gamma_2$$

and from (23)

$$\vartheta' = \int_{t_1}^{t} \cos (kt)' \, dt \tag{26}$$

or

$$\Delta\vartheta = \vartheta' - \vartheta = \int_{-a}^{y} [\cos (k't) - \cos (kt)] \sqrt{1 + \left(\frac{dx}{dy}\right)^2} \, dy. \tag{27}$$

The transformed field after it has passed the lens can now be calculated

$$\Phi_{\nu+1}' = \Phi_{\nu+1} \exp (i\beta\Delta\vartheta). \tag{28}$$

Finally, we need to know the normal derivative $\varphi_{\nu+1}'$ of $\Phi_{\nu+1}'$ before we are ready for the next iteration step. Replacing derivatives with respect to $t$ by the normal derivatives with respect to $m$ in (22) and multiplying by $i/\beta$ we obtain

$$\varphi_{\nu+1} = \left[\frac{\partial x}{\partial m} - \frac{\partial \vartheta}{\partial m} + \frac{i}{\beta}\frac{1}{G}\frac{\partial G}{\partial m}\right]\Phi_{\nu+1} . \tag{29}$$

The derivative $\partial\vartheta/\partial t$ was equal to $\cos (kt)$, similarly we can write

$$\frac{\partial \vartheta}{\partial m} = \cos (km). \tag{30}$$

The angle $(km)$ is related to $\gamma_1$ by

$$(km) = \pi - \gamma_1 .$$

The reader might wonder why I bothered introducing the angle $(kt)$ and the derivative $\chi$ since $\partial\vartheta/\partial m$ which is determined by $\varphi$ gives the

angle $\gamma_1$ directly. However, $\partial\vartheta/\partial m$ only determines $\cos\gamma_1$. The conversion of $\cos\gamma_1$ to $\sin\gamma_1$ leaves the sign of $\gamma_1$ ambiguous. No such ambiguity arises if $(kt)$ is computed.

The angle $(km)'$ belonging to the output field can now easily be obtained with the help of (19) and (24)

$$\frac{\partial\vartheta'}{\partial m} = \cos(km)' = -\sqrt{1-\sin^2\gamma_2}. \tag{31}$$

Substituting (31) into (29) written for the primed quantities we get

$$\varphi_{\nu+1}' = \left[\frac{\partial x}{\partial m} - \cos(km)' + \frac{i}{\beta}\frac{1}{G}\frac{\partial G}{\partial m}\right]\Phi_{\nu+1}'$$

or using (29) once more and keeping in mind (28)

$$\varphi_{\nu+1}' = \varphi_{\nu+1}\exp(i\beta\Delta\vartheta) + [\cos(km) - \cos(km)']\varphi_{\nu+1}'. \tag{32}$$

The transformed field quantities of (28) and (32) are certain to conform with the requirements of ray optics. However, this is not quite sufficient to satisfy all the wave optics requirements. Numerical results have shown that the fields $\Phi'$ and $\varphi'$ substituted into the power formula (17) yield a different number for the power flow than the one obtained from using (17) with $\Phi$ and $\varphi$. The fields $\Phi'$ and $\varphi'$ after having passed the lens should carry the same amount of power as the input fields. The transformation procedure, outlined so far, takes into account the phase of the field and the change in slope of the phase fronts in accordance with physical principles but it does not account for any change in field amplitude which the physics of the (lossless) lens might also require. In fact, the failure of this transformation to obey conservation of energy points to a need to readjust the field amplitudes. To correct the amplitudes of the field quantities $\Phi'$ and $\varphi'$ locally, I computed the ratio of the integrals of (17) taken with the two fields. Letting $I$ be the integrand of (17) calculated with the use of $\Phi$ and $\varphi$, and $I'$ the corresponding value obtained using $\Phi'$ and $\varphi'$, I calculated

$$R = \sqrt{\frac{I}{I'}} \tag{33}$$

and introduced

$$\Phi'' = R\Phi'$$

and (34)

$$\varphi'' = R\varphi'.$$

This last transformation does not affect the phase of the field or its slope but adjusts the field amplitude so that using $\Phi''$ and $\varphi''$ the power is conserved in the process of transmitting the field through the infinitely thin and lossless lens. This last transformation does not transform away diffraction losses, however, since those occur in passing the field from one lens to the next.

This completes the description of the iteration procedure. It is surprising how much the calculation is complicated by the simple fact that the lenses are not plane but curved. One might regard the simplicity of the plane lenses as a lucky break. The present procedure naturally is more time consuming. To pass the field through 100 lenses of the lens waveguide with plane lenses using the simple procedure of Fox and Li takes 0.023 hours of 7094 computer time. The procedure described above takes 0.13 hours for the same number of lenses or 5.65 times as long. The present procedure is that much more involved.

## IV. DISCUSSION OF NUMERICAL RESULTS

The calculation procedure described on the previous pages was used to study the fate of an off-axis field distribution as it propagates through the beam waveguide. In a beam waveguide composed of ideal, thin lenses the field would suffer no distortions as it travels through the lenses provided that its shape corresponds to a mode of this structure. A mode, even if displaced from the axis, keeps its shape in a perfect beam waveguide. The center of gravity of such an off-axis mode follows the ray trajectory of geometric optics. The field may look somewhat different as it passes different lenses. But whenever its path brings it back to its original position on the lens it assumes the original shape.

This property of ideal lens guides is no longer true for beam waveguides composed of distorting lenses. Now the original field distribution is changed even if the field returns to its original position. These field distortions are best displayed in a motion picture. However, in a paper one has to limit oneself to the display of a few representative frames of such a motion picture.

To launch the field into the waveguide I started with an ideal lens whose focal length corresponded to twice the on-axis focal length of the simulated gas lenses. This procedure was chosen since the modes of the ideal beam waveguide have plane phase fronts right on the lens or in other words after the field has traversed one-half of the lens. A plane phase front and the flat starting lens allow us to take

$$\varphi_o = \frac{i}{\beta}\frac{\partial \Phi_0}{\partial n} = 0$$

so that $\varphi_\nu$ is known initially and the field can get started. On all the following lenses $\Phi_\nu$ as well as its derivations are calculated.

Figs. 2(a) and 2(b) show the shape of the principal surface $p$ and the focal length $f$ of the lens as functions of position $y/a$. The function $p$ as well as the focal length $f$ are displayed normalized with respect to the length $L$ of the gas tube. The coordinate $y$ is plotted normalized with respect to the radius $a$ of the tube. These curves correspond to a gas lens operated with a gas velocity which minimizes the focal length at an input gas temperature $T_o = 300°K$, wall temperature of gas tube 355°K, an index of refraction of $n = 1 + 4.210^{-4}$ and a ratio* of $L/a = 50$.

We consider a beam waveguide composed of gas lenses of this type spaced so that $D/f_o = 2$, where $D$ is the distance between adjacent lenses and $f_o$ is the value of the focal length at $y = 0$. Into this beam waveguide we launch a field with a Gaussian intensity profile whose center of gravity is shifted off the optical axis as shown in Fig. 3(a). This field distribution corresponds to a mode of the ideal confocal beam waveguide which is shifted off-axis. The position and shape of this field on the next two lenses is given in Figs. 3(b) and 3(c). Since the beam waveguide is nearly confocal, the center of gravity of the field moves like a ray in a confocal waveguide. No field distortion is yet discernible. Jumping 100 lenses ahead in the beam waveguide we see in Figs. 4(a), 4(b), and 4(c) that the field begins to distort from its original shape. After having traversed 150 lenses the field shows a distinct break-up into two peaks, Fig. 5(a). The appearance of the field on two adjacent lenses can be quite different, Fig. 5(b). Finally, we see the wave field on the lenses 209 and 210 in Figs. 6(a) and 6(b). The distortion has changed somewhat but is not basically different.

The field of Fig. 3(a) fills one-third of the gas lens between the points where it carries more than exp (−2) of its peak power. If we assume a tube with $a = 0.317$ cm (0.125 inch) a waveguide mode of that width corresponds to a light wavelength of $\lambda = 4.60 \times 10^{-4}$ cm which is 7.26 times as long as the wavelength of the 6328Å line of the HeNe laser.

I mentioned in the introduction that the width of the field distribution with respect to the tube radius cannot be made arbitrarily narrow. To consider fields which are similar to modes of the beam waveguide at $\lambda = 6.328 \times 10^{-5}$ cm forces us to reduce the lens aperture. The ratio of field extension and waveguide aperture is maintained if we reduce the wavelength from $\lambda = 4.60 \times 10^{-4}$ cm to $\lambda = 6.328 \times 10^{-5}$ cm and aperture the lens at a value of $y/a = 0.371$ of Figs. 3 through 6. Using only that part of the waveguide between $-0.371 \leqq y/a \leqq 0.371$ and

* These values correspond to $v_0/V = 6.45$ and $C(L/a) = 0.192$ with $v_0/V$ and $C(L/a)$ defined in Ref. 4.

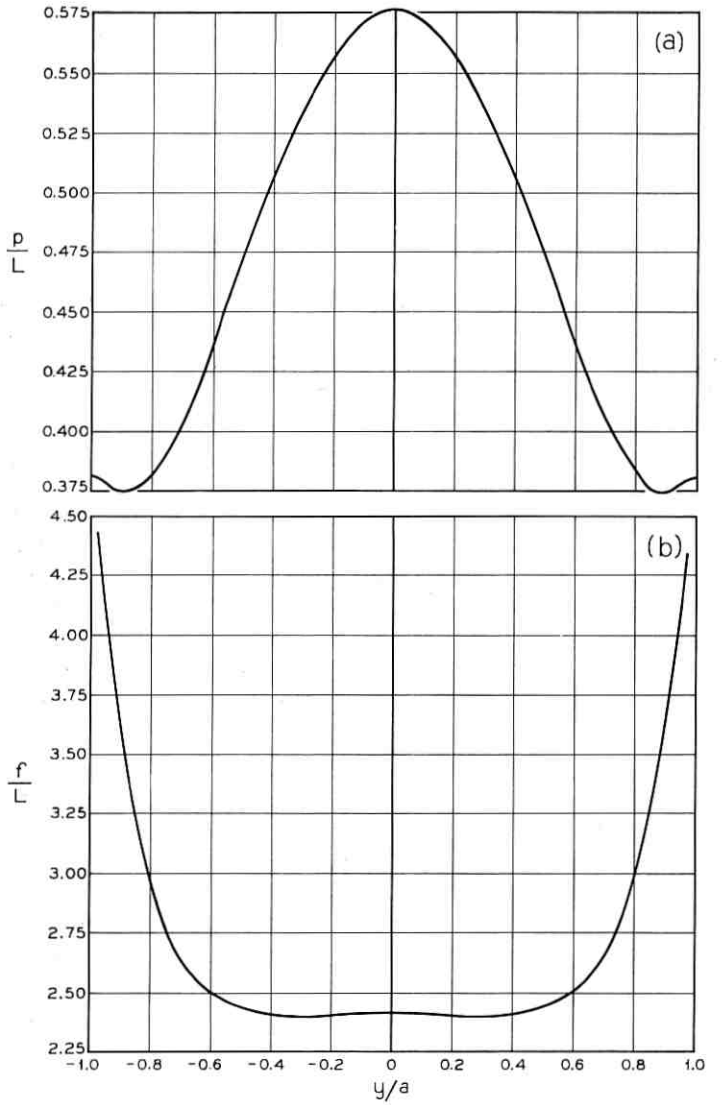



Fig. 2—(a) The principal surface $p$ of the gas lens normalized with respect to the tube length $L$ as a function of $y/a$. (b) The focal length $f$ of the gas lens normalized with respect to the tube length $L$ as a function of $y/a$.

renormalizing the $y$-coordinate so that these boundaries again correspond to $-1 \leqq y/a \leqq 1$ leads to the shape of principal surface and focal length as shown in Figs. 7(a) and 7(b). This is still the same lens, with the only difference that we expanded its center portion. The center portion of the lens has far less distortion as the whole lens of Fig. 2. Figs. 3(a),

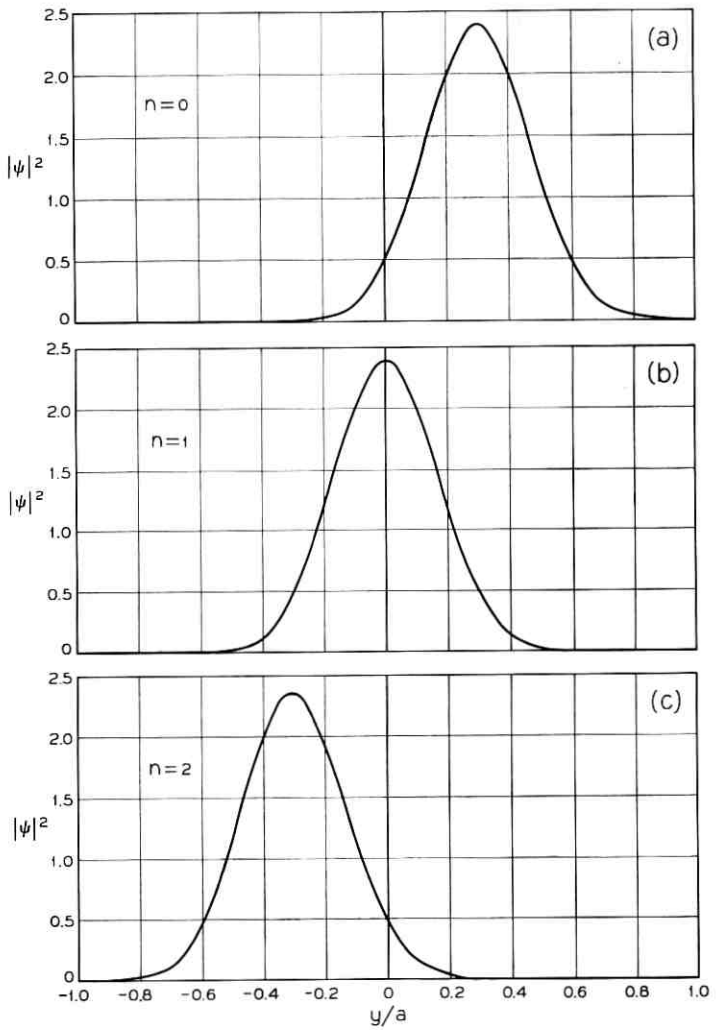



Fig. 3 — The Gaussian field distribution on the first three lenses represented by Fig. 2. The power $P$ carried by the field is $P = 1$.

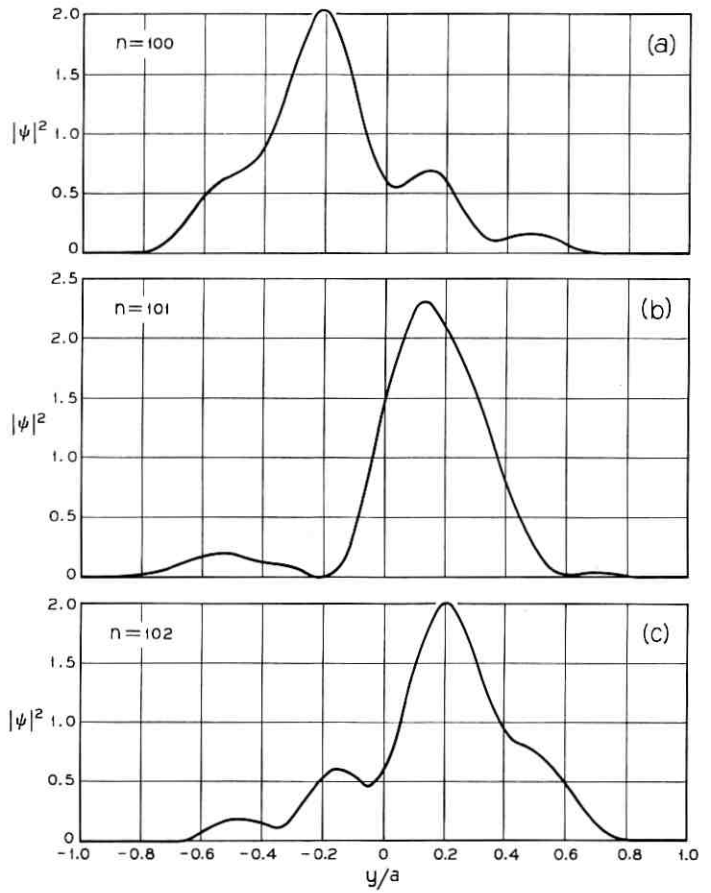



Fig. 4 — The distorted field after passing through 100 lenses $P = 0.969$.

3(b), and 3(c) show again the field distribution on the first three lenses at the wavelength of $\lambda = 6.328 \times 10^{-5}$ cm and the apertured lens. After traversing 120 lenses this field suffered noticeable distortions shown in Figs. 8(a) and 8(b), even though it "sees" now only the center portion of the lens where the focal length depends only very little on $y$ and where the principal surface is much closer to a plane. The dotted curves also shown in these and all remaining figures of field configurations were obtained by maintaining the focal length of the equivalent gas lens, but using a lens with a perfectly flat principal plane. The comparison between the solid and dotted curve shows that the field distortion

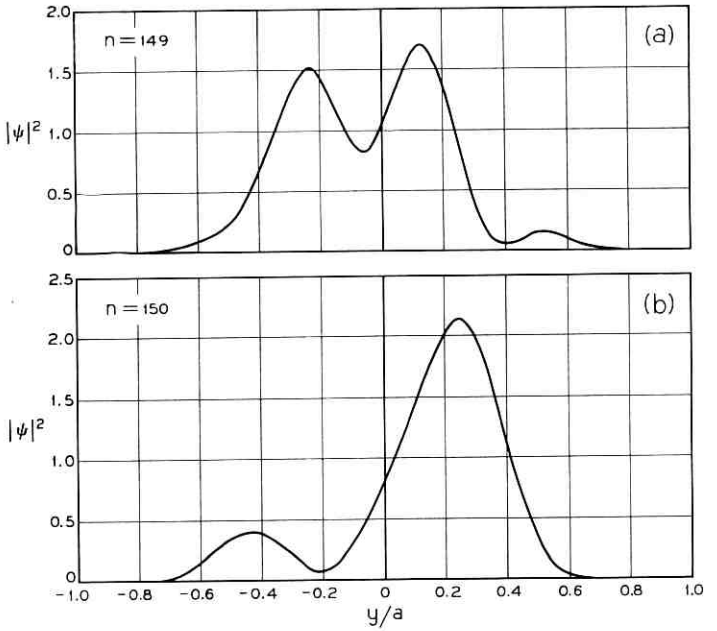



Fig. — 5 Field distortion after 150 lenses. $P = 0.956$.

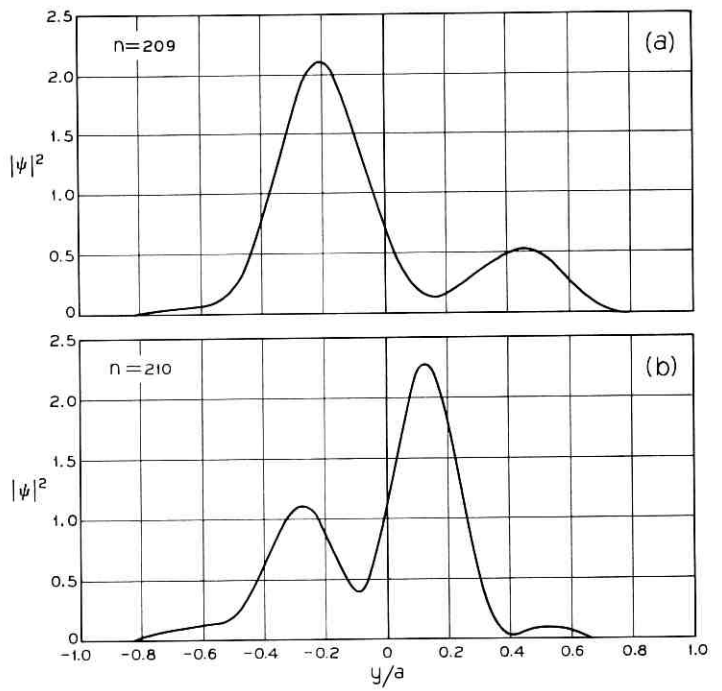



Fig. 6 — Field distortion after 210 lenses. $P = 0.941$.

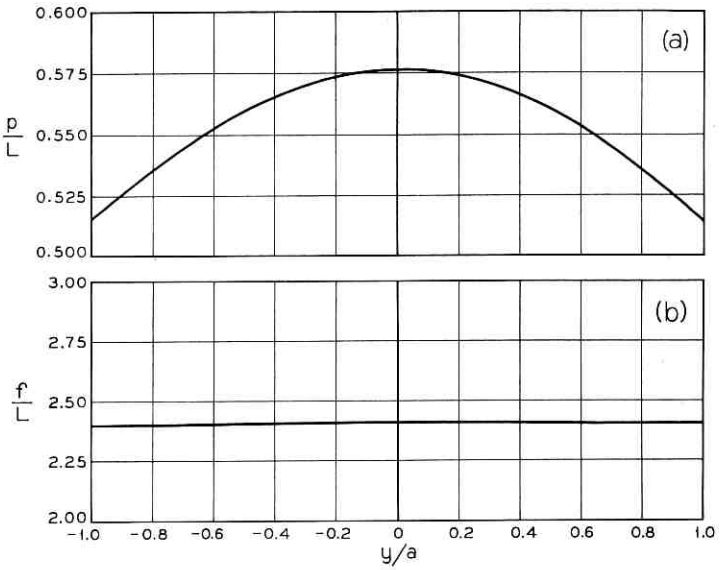



Fig. 7 — (a) Principal surface of the apertured lens. (b) Focal length of the apertured lens.

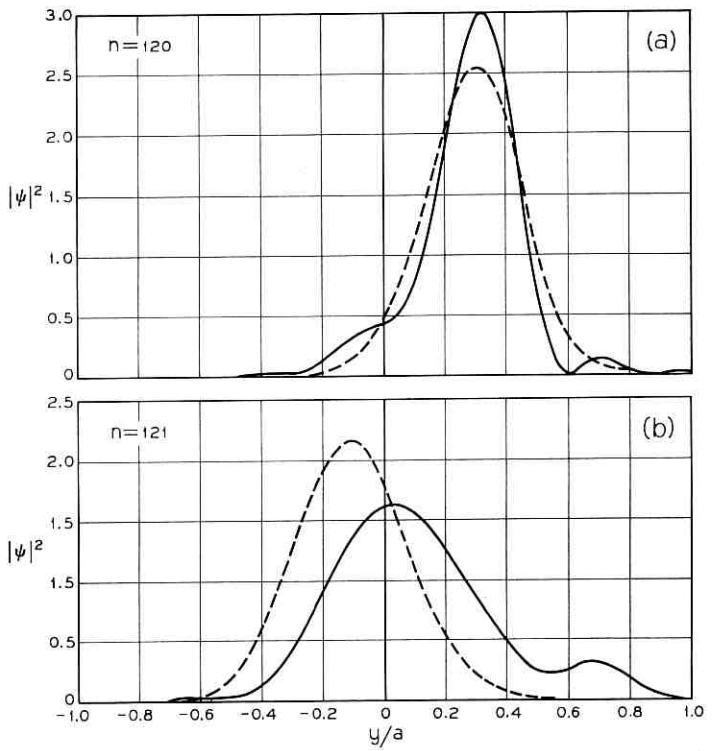



Fig. 8 — Field distortion of the field in the apertured lens after traversing 120 lenses. Solid curves represent the warped thin lens, $P = 0.976$. Dotted curves represent a fictitious plane lens with the same focal length aberration, $P = 1.000$.

can be attributed mainly to the distorted principal plane of this gas lens. The change in width of the field distribution on adjacent lenses as seen in Figs. 8(a) and 8(b) is caused by the departure of the beam waveguide from exact confocality. Figs. 9 and 10 show how bad the field distortions get after 250 and about 400 lenses. Most surprising is the fact that the field distortions of Figs. 8 through 10 are only slightly less severe than those of Figs. 4 through 6, in spite of the substantial improvement of lens aberrations.

To study this point further I constructed a gas lens with even less principal plane distortion by using two gas lenses back-to-back as shown in Fig. 11. The center portion of the principal surface and focal length curve is shown in Figs. 12(a) and 12(b). The expansion and renormalization of these curves is the same as that of Figs. 7(a) and 7(b). The principal surface of this lens, Fig. 11, approximates a plane even better than Fig. 7(a) however, there is more focal length distortion apparent in Fig. 12(b) than in Fig. 7(b). This lens distorts substantially less than

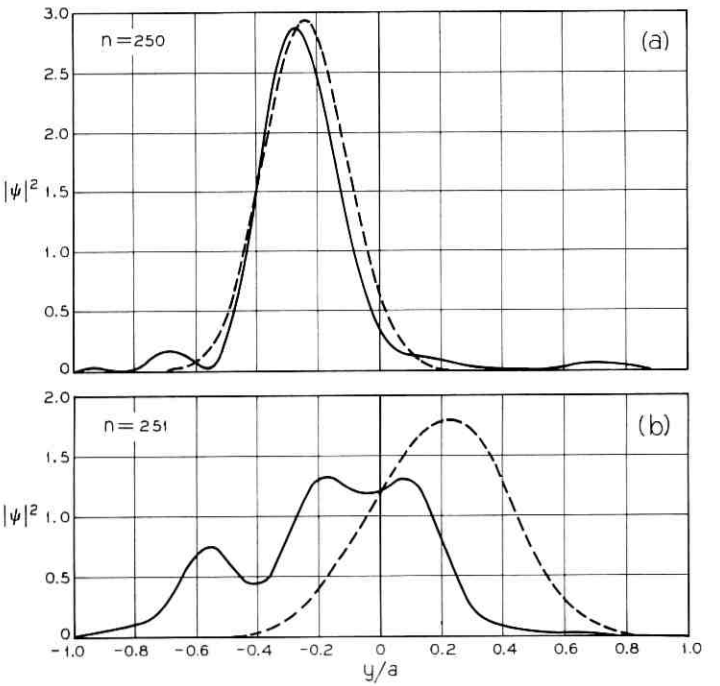



Fig. 9 — Field distortion of the field in the apertured lens after traversing 250 lenses. Solid curve, $P = 0.929$. Dotted curve, $P = 1.000$.

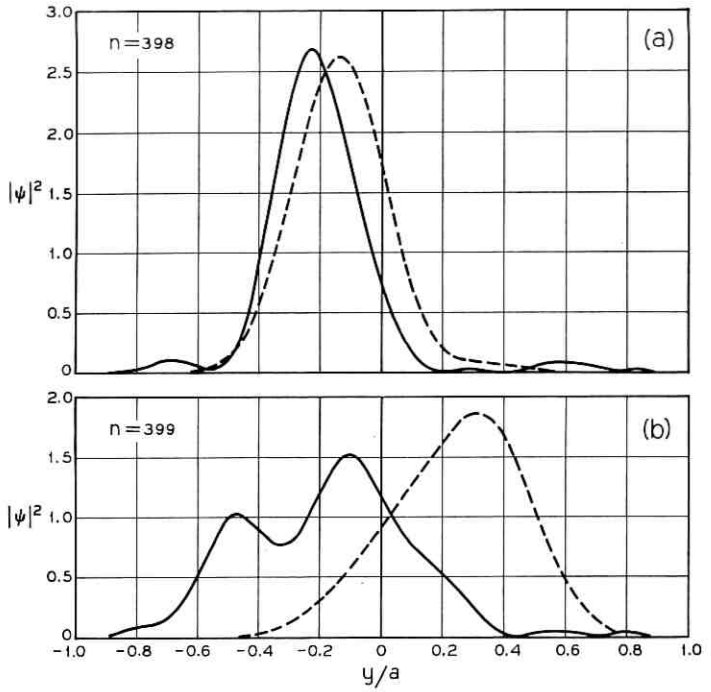



Fig. 10 — Field distortion after 400 lenses. Solid curve, $P = 0.894$. Dotted curve, $P = 1.000$.

the simple lens of Fig. 1, as a comparison of Figs. 8 through 10 with Figs. 13 through 15 indicates. However, even a lens with the characteristics of those shown in Figs. 12(a) and 12(b) causes the field to break up into the double-humped shape of Fig. 16 after traversing 295 lenses.

It is interesting to note the difference between the solid curve and the

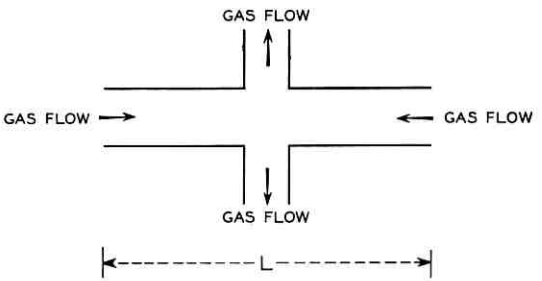



Fig. 11 — Two gas lenses operated back-to-back minimize principal plane distortion.

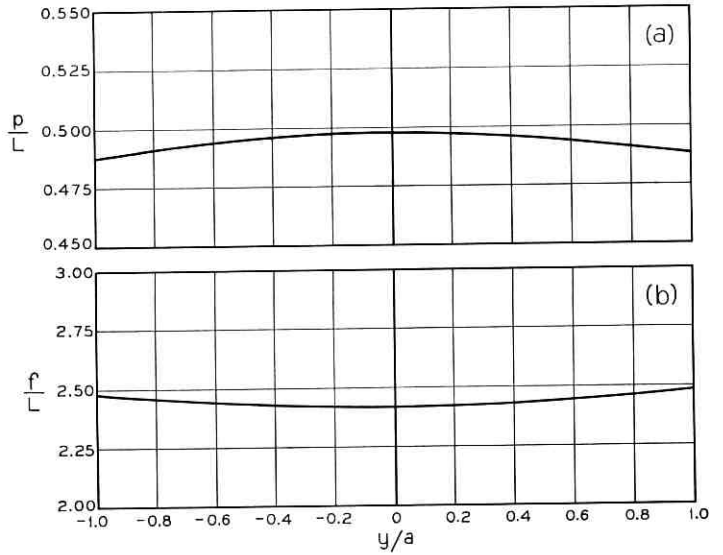



Fig. 12 — (a) Principal surface of gas lens of Fig. 11. (b) Focal length of gas lens of Fig. 11.

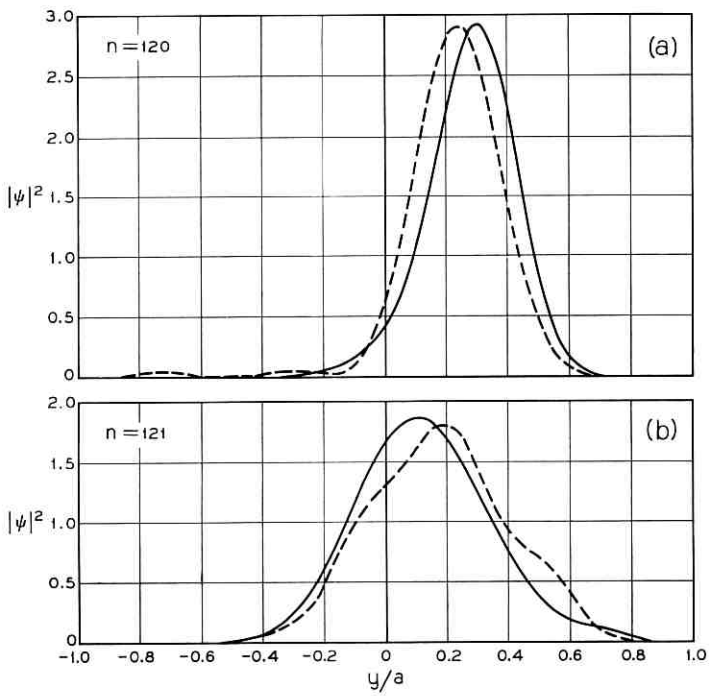



Fig. 13 — Field distortion after 120 lenses. Solid curve represents the lens of Fig. 11, $P = 0.9996$. Dotted curve represents a fictitious plane lens with focal length of Fig. 12(b), $P = 1.0000$.

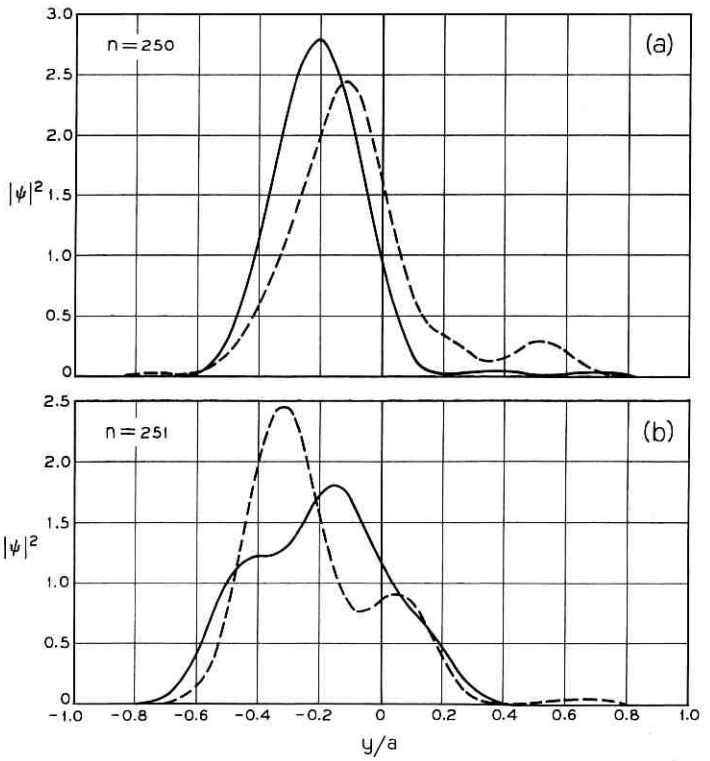



Fig. 14 — Field distortion after 250 lenses. Solid curve, $P = 0.9992$. Dotted curve, $P = 1.0000$.

dotted curve of Figs. 13 through 15. Both curves show field distortions. Those of the solid curves are caused by the combined action of principal plane and focal length distortions, while those of the dotted curves are due to focal length aberration only. It appears that the two distorting influences cancel out to some extent since the solid curves of Figs. 13(b), 14(a), and 15(a) show less distortion than the corresponding dotted curves.

Figs. 14(b) and 15(a) show that even the plane lens with only focal length aberration (dotted curve) has the tendency to distort the field into a multiply-humped shape. Theoretical work by E. A. J. Marcatili and further computer simulations have established a periodicity in this behavior. Plane lenses with focal length distortion cause an off-axis field to break up into a double-humped shape which becomes perfectly symmetrical after some distance. After twice this distance the field re-

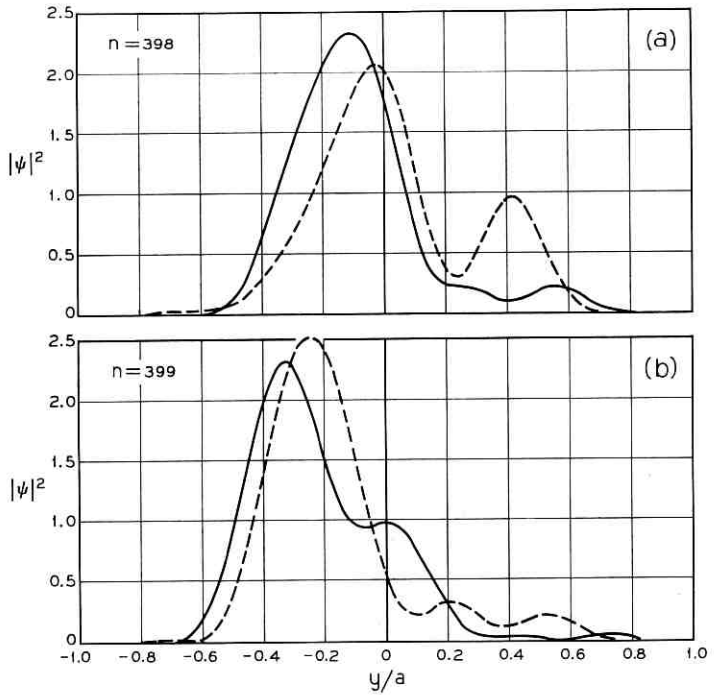



Fig. 15 — Field distortion after 400 lenses. Solid curve, $P = 0.9988$. Dotted curve, $P = 1.0000$.

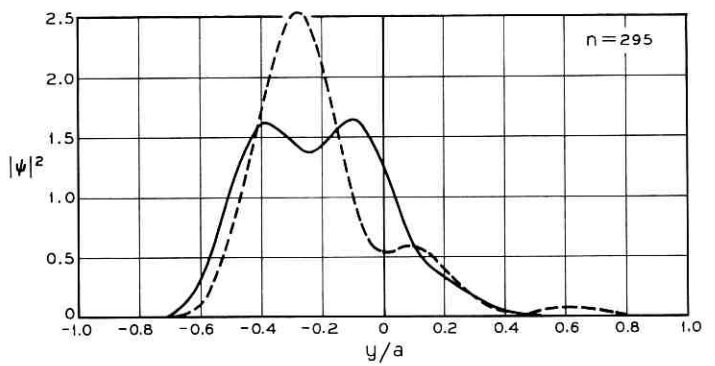



Fig. 16 — The field assumes a double-humped appearance after 295 lenses.

turns to its original shape, etc. No such periodicity seems to exist for distortions caused by a warped principal plane. The periodicity of field distortions caused by focal length aberrations gives a clue to the problem of why so little lens distortion can lead to such serious field distortions. In principle, the field always breaks up into a perfectly symmetric double-humped shape if it is allowed to travel far enough in the beam waveguide. The required distance depends on the amount of focal length aberration but the final field distortion does not. Similarly, it is possible that arbitrarily small distortions of the principal plane may always lead to serious field distortions if given enough length of waveguide. It is still surprising, however, that the slight aberration shown in Fig. 12(a) and 12(b) causes the field to become double humped after only 295 lenses.

REFERENCES

1. Goubau, G. and Schwering, F., On the Guided Propagation of Electromagnetic Wave Beams, IRE Trans., *AP-9*, May, 1961, pp. 248–256.
2. Berreman, D. W., A Lens or Light Guide Using Convectively Distorted Thermal Gradients in Gases, B.S.T.J., *43*, July, 1964, pp. 1469–1475.
3. Marcuse, D. and Miller, S. E., Analysis of a Tuublar Gas Lens, B.S.T.J., *43*, July, 1964, pp. 1741–1758.
4. Marcuse, D., Theory of a Thermal Gradient Gas Lens, IEEE Trans. MTT, *13*, November, 1965, pp. 734–739.
5. Marcuse, D., Comparison Between a Gas Lens and its Equivalent Thin Lens. This issue, pp. 1339–1344.
6. Marcatili, E. A. J., Ray Propagation in Beam Waveguides with Redirectors, B.S.T.J., *45*, January, 1966, pp. 105–115.
7. Courant, R. and Hilbert, D., *Methods of Mathmatical Physics, II*, Interscience Publishers, 1962, p. 317.
8. Fox, A. G. and Li, T., Resonant Modes in a Maser Interferometer, B.S.T.J., *40*, March, 1961, pp. 453–488.
9. Marcuse, D., Physical Limitations on Ray Oscillation Suppressors, B.S.T.J., *45*, May–June, 1966, pp. 743–751.
10. Born, M. and Wolff, E., *Principles of Optics*, Pergamon Press, New York, 1959, p. 112, Equation (15).

# Contributors to This Issue

EDWIN O. ELLIOTT, A.B., 1949, M.A., 1951, Ph.D., 1959, University of California, Berkeley; Operations Evaluation Group of M.I.T., 1954–1958; Stanford Research Institute, 1958–1959; Assistant Professor of Mathematics, University of Nevada, Reno, 1959–1960; Bell Telephone Laboratories, 1960—. At Bell Laboratories Mr. Elliott has been engaged in mathematical analysis of error-control methods for digital data communication systems and in the application of measure-theoretic techniques in the study of stochastic processes. He has also worked on problems in the congestion theory of traffic. Member, American Mathematical Society, Operations Research Society of America, Pi Mu Epsilon, Sigma Xi, Phi Beta Kappa.

RICHARD E. HART, 1948–1950, Fairleigh Dickinson College; 1954–1956, Michigan State University; Bell Telephone Laboratories, 1956—. Mr. Hart has worked in the System planning area on the Morris experimental electronic telephone central office. He is currently concerned with the system planning and maintenance requirements for No. 1 ESS.

JOSEPH B. KRUSKAL, Ph.B., 1948, B.S., 1948, M.S., 1949, University of Chicago; Ph.D., 1954, Princeton University; Logistics Research Project, George Washington University, 1950–1953; Analytical Research Group, Princeton University, 1954–1956; Mathematics Department, University of Wisconsin, 1956–1958; Mathematics Department, University of Michigan, 1958–1959; Bell Telephone Laboratories, 1959—. Mr. Kruskal has done research in several areas of mathematics, including combinatorics, statistics, and computer applications. Currently he is working primarily in statistics, both theoretical and applied. Member, American Mathematical Society, Mathematical Association of America, Society for Industrial and Applied Mathematics, Psychometric Society, Sigma Xi, Pi Mu Epsilon.

STEWART K. KURTZ, B.S., 1956, M.S., 1957, Ph.D., 1960, Ohio State University; Bell Telephone Laboratories, 1960—. During the first two years at the Laboratories, Mr. Kurtz worked in the microwave maser

group studying paramagnetic resonance phenomena. He is presently concerned with light modulation techniques and electro-optic materials research. Member, American Physical Society, Sigma Xi, Phi Beta Kappa.

HENRY J. LANDAU, A.B., 1953, A.M., 1955, Ph.D., 1957, Harvard University; Bell Telephone Laboratories, 1957–1959; Institute for Advanced Study 1959–1960; Bell Telephone Laboratories, 1960—. Mr. Landau has been engaged in mathematical research in function theory and harmonic analysis. Member, American Mathematical Society, Phi Beta Kappa, Sigma Xi.

DIETRICH MARCUSE, Diplom Vorpruefung, 1952, and 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 is presently working on the transmission aspect of a light communications system. Member, IEEE.

J. R. PIERCE, B.S., 1933, M.S., 1934, Ph.D., 1936, California Institute of Technology; Bell Telephone Laboratories 1936—. His earlier work included microwave tubes and communication, including satellite communication. Dr. Pierce is now Executive Director, Research—Communications Sciences Division, with responsibility in such fields of research as radio, electronics, acoustics and vision, mathematics and psychology. He has published books on electron beams, traveling-wave tubes, speech and hearing, information theory and quantum electronics. Member, National Academy of Sciences, National Academy of Engineering, Air Force Association; Fellow, American Academy of Arts and Sciences, Institute of Electrical and Electronics Engineers, American Physical Society, Acoustical Society of America.

ATTILIO J. RAINAL, University of Alaska, University of Dayton, 1950–52; B.S.E.Sc., 1956, Pennsylvania State University; M.S.E.E., 1959, Drexel Institute of Technology; Dr. ENG., 1963, Johns Hopkins University; Bell Telephone Laboratories, 1964—. Mr. Rainal is engaged in research on noise theory with application to radar theory. Member, Tau Beta Pi, Eta Kappa Nu, Sigma Tau, Pi Mu Epsilon, Sigma Xi, IEEE.

DAVID SLEPIAN, 1941–43, University of Michigan; M. A., 1947, Ph.D., 1949, Harvard University; Bell Telephone Laboratories, 1950—. He has been engaged in mathematical research in communication theory, switching theory, and theory of noise, as well as various aspects of applied mathematics. Mr. Slepian has been mathematical consultant on a number of Bell Laboratories projects. During the academic year 1958–59, he was Visiting Mackay Professor of Electrical Engineering at the University of California at Berkeley. Member, AAAS, American Math. Society, Institute of Math. Statistics, IEEE, SIAM, URSI Commission 6.