THE BELL SYSTEM
TECHNICAL JOURNAL

VOLUME XLVI SEPTEMBER 1967 NUMBER 7

Statistical Analysis and Modeling of the High-Energy Proton Data From the *Telstar*® 1 Satellite

By J. D. GABBE, M. B. WILK and W. L. BROWN

(Manuscript received October 4, 1966)

This paper deals with the analysis of data from the omnidirectional high-energy proton detector on the Telstar® 1 satellite. The main accomplishment is the development of relatively simple (empirical) mathematical models which give a statistically accurate representation of the measured spatial distribution of intensity of protons with energies between 50 and 130 MeV.

These models depend upon the fitting of 8 (or 9 or 10) coefficients based on samples containing approximately 1000 of the nearly 80,000 experimental observations. The nature of the model for the average omnidirectional counting rate permits its closed form transformation to the equivalent equatorial pitch angle distribution.

Sufficiently accurate fits were achieved so that the residuals (equal to observed minus fitted) could be productively examined for possible dependence on variables other than the two magnetic coordinates used in the fitting. One consequence of this was the detection of instrumental susceptibility to temperature and bias voltage changes, which led to an objective partitioning of the data.

The present paper has several evolutionary aspects: In particular, a series of one-dimensional fits was employed as a base for developing a two-dimensional model; a preliminary analysis of all the data was used to guide the rejection of outliers; a first two-dimensional fit to all the data

led to a data-independent basis for partitioning the data; the mode of selection of a sample of data, to which the two-dimensional model was fitted, changed as deeper insight into the importance of this issue developed; and, after a very satisfactory fit to the data was attained, the model was improved by specialization and reparameterization so as to overcome some statistical defects and to achieve greater physical meaning.

The data cover the time period between July 1962 and February 1963, and the spatial region bounded by 1.09 $R_e \leqq R \leqq 1.95\ R_e$ and $0 \leqq \lambda < 58°$. Flux maps having a relative accuracy of about two percent are derived from the fit and presented. The temporal behavior of the intensity is examined and some changes are noted. The maximum value of the omnidirectional flux of protons with energies between 50 and 130 MeV is found to be $[5.7^{+1.4}_{-2.8}] \times 10^3$ protons/cm^2 sec at $L = 1.46$ on the magnetic equator, in good agreement with other experiments. Relative flux values and energy spectra are consistent with the generally accepted picture of the proton distribution.

TABLE OF CONTENTS

I. INTRODUCTION

This paper deals with the analysis of data from the omnidirectional high-energy proton detector on the *Telstar®* 1 satellite. The main accomplishment is the development of a relatively simple (empirical)

mathematical model which gives a statistically accurate representation of the measured spatial distribution of protons with energies between 50 and 130 MeV.

The *Telstar*® 1 satellite was launched into a 45°-inclined orbit with an apogee of 5600 km and a perigee of 950 km on day 191 (July 10), 1962. The period of precession of the apsis was 180 days. The satellite was instrumented to measure fluxes of energetic particles; in particular, counting rates of protons with energies above 50 MeV were recorded. Two thousand hours of telemetry was received during the active life of the satellite, which terminated on day 52 (February 21), 1963. The satellite and associated systems have been described in detail.[1] The particle-detection instruments have been documented[2] and some of the experimental results have been presented.[3, 4, 5]

The above-mentioned presentations of information concerning the earth's radiation belts have been principally graphical in format, owing to the complexity of the belts and the limited understanding of the details of the processes affecting them.

An accurate analytical representation of the data would enable convenient interpolation, extrapolation, and transformation. Thence it would be practical to make extensive comparisons with the results of other experiments and with various theoretical predictions and to summarize, analytically, such features as the equatorial omnidirectional counting rate and the approximate size of the equatorial loss cone. In addition, an empirical mathematical model would facilitate the study of temporal fluctuations in various regions of space. Of course, a good analytical representation, even though empirical, may also stimulate deeper physical insight and theories.

The present study was directed toward the development of a mathematical function which would, when fitted to the data, provide a convenient, concise and precise summary description. The mathematical model(s), which are herein presented, were empirically evolved, using the knowledge that the intensity distribution of these protons is, in the main, not rapidly variable in time. Even more specifically, the assumption has been that fluctuations in observed counting rates at a fixed point in space relative to the earth are independent random variables. Further, the main effort of the analysis has been to try to relate the observed counting rates to a two-dimensional magnetic coordinate system derived from three-dimensional spatial coordinates by mapping the known earth's magnetic field onto the field of a magnetic dipole.[6]

The mathematical models which are used depend upon fitting of between 8 and 10 coefficients based on samples containing approximately

1000 of the nearly 80,000 experimental observations. The nature of these models for the average omnidirectional counting rate permit their closed-form transformation to the equivalent equatorial pitch angle distribution.

The fitted models were sufficiently accurate so that the residuals (equal to observed minus fitted) of all the data could be productively examined for possible dependence on variables other than the two magnetic coordinates used in the fitting. One consequence of this was the detection of instrumental susceptibility to temperature and bias voltage change, which led to an objective partitioning of the data.

This article summarizes some of the productive aspects of the analysis of this body of data. A very large amount of "preliminary" work is not reviewed. Though not an historical description of the work, the present paper does have several evolutionary aspects. In particular, a series of one-dimensional fits were employed as a base for developing two-dimensional models; a preliminary analysis of all the data was used to guide the rejection of outliers; a first two-dimensional fit to all the data led to a data-independent basis for partitioning the data; the mode of selection of a sample of data, to which the two-dimensional model was fitted, changed as deeper insight into the importance of this issue developed; and, after a very satisfactory fit to the data was attained, the model was improved by specialization and reparameterization so as to overcome some statistical defects and to achieve greater physical meaning.

Readers with specific interests may wish to consult the Table of Contents, the summary (Section XIV) and the following overview for guidance.

Section II introduces the input data which have been analyzed. Coordinates and notation are tabulated, the distribution of the data is displayed, and the general quality and stability of the data are discussed. It is shown informally that the measurements may be usefully organized in the dipole magnetic coordinate system used.

In Section III, various alternative coordinate systems and scales are considered. The bases for choice of the x,L coordinate system for the independent variables and the square-root-of-counting-rate scale for the dependent variable are given.

Section IV brings together the ideas underlying the formulation and evolution of the models, and gives mathematical definitions and details. Some properties of the models which make them suitable smoothing functions for this body of data are indicated.

One-dimensional fits to the data in each of several L-slices (an

L-slice is a particular grouping of the data) are displayed on several scales and discussed in Section V. It is shown that L-slice fits suffer from fundamental deficiencies, in addition to being inconvenient to work with. The results of the L-slice fits are used to lead to a two-dimensional model.

Section VI contains the treatment of the preliminary fit of a two-dimensional model. This fit is of good quality and provides residuals which are used to help identify and eliminate extraneous sources of variability in the data and to serve as a basis for more refined sample selection.

The treatment of the two-dimensional fit to the data after it has been partitioned to reduce instrumental effects appears in Section VII. The method of sample selection is important, and some algorithms and their influence on the resultant fits are considered in Section 7.1. The advantages of selecting a sample on the basis of a preliminary fit are discussed. The fit itself is described and evaluated in the remainder of the section.

A more detailed statistical critique of the fit discussed in Section VII is contained in Section VIII; in particular, some remaining physical and statistical defects are pinpointed.

Section IX deals with a modified version of the model, which eliminates the remaining defects, and gives the results of fitting the most satisfactorily parameterized model of the proton distribution.

Residuals are used to study temporal effects in Section X. An increase in intensity near $L = 2$ is noted during October, 1962. An upper limit of 0.003 gauss is found for the diurnal variation of the earth's magnetic field near $L = 1.5$. A possible shift in the location of the atmospheric cutoff is examined.

The behavior of the radiation belt near the top of the atmosphere is the subject of Section XI. Although the data do not allow the position of the low-altitude cutoff to be accurately determined, the qualitative behavior precludes a simple atmospheric cutoff mechanism.

Section XII is devoted to a comparison of the *Telstar*® 1 results, presented as flux maps, with those obtained on Injuns 1 and 3, Explorers 4 and 15, and other satellites. Absolute flux values agree to within a factor of 2 in most cases, which is as well as can be expected. Very good agreement exists concerning the behavior of the intensity in the equatorial plane, on L-shells, and near the top of the atmosphere. Experimental results regarding the equatorial pitch angle (see Fig. 1) distribution are found to agree well with each other, but differ

appreciably from the published results of theoretical calculations.

Section XIII gives brief consideration to possible directions in which this work might be extended: improving the fit to the *Telstar*® 1 high-energy protons still further; approaching model development differently; employing the data more fully; and encompassing other more complex bodies of data.

Section XIV contains a brief summary of the results and Section XV contains acknowledgments.

Appendix A provides a detailed description of the particle detector and its calibration.

Appendix B gives some statistical background and details of the analysis, and Appendix C discusses statistical measures of the goodness of fit of the model over all the partitioned data.

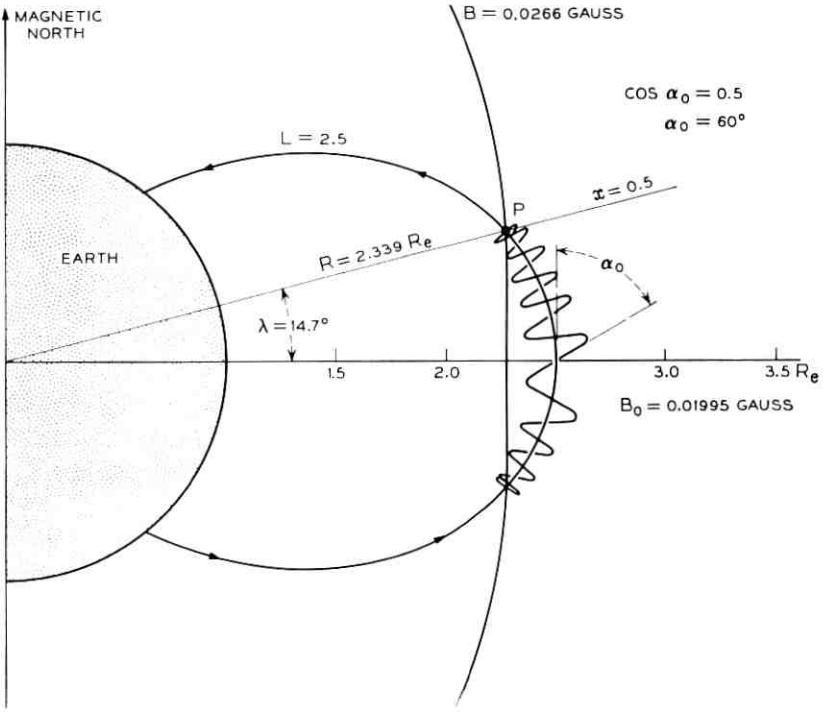


Fig. 1 — Magnetic coordinates of the point P. The spiral is the orbit of a particle trapped on the magnetic line of force $L = 2.5$ and mirroring at $B = 0.0266$ gauss. The equatorial pitch angle, α_0, is the angle between the velocity vector and the magnetic field vector at the equator.

II. THE DATA

The data which are studied in this paper were obtained with a detector on the *Telstar*® 1 satellite which measured protons with energies greater than 50 MeV. The sensitive detecting element is a semiconductor diode developed specifically for satellite experiments.[7] The effective geometric factor, g, of the detector depends upon proton energy, but over the region energy between 50 and 130 MeV the *average* geometric factor, $\bar{g}$, is relatively insensitive to the energy spectrum and an approximate value of 0.143 cm^2 steradian has been selected. These considerations are discussed in detail in Appendix A. The response of the detector is also dependent upon both temperature and electrical bias because of changes in the effective thickness of the active region of the detector. These effects are discussed in Section 6.8.

The primary input to our data reduction process consisted of: the telemetry record of the number of counts measured by the detector in an 11-second counting interval once every minute; the time at which the data were recorded (inserted by the recording station); and the ephemeris of the satellite position obtained from tracking data. These are supplemented by the satellite spin-axis orientation obtained from the mirror flash data[8] and by telemetered measurements of the satellite skin temperature near the detector and of the detector bias voltage.

During data reduction, the square root of the counting rate was computed for each recorded particle-counting interval and associated with the following information: date and time, geographic position, position in the earth's magnetic field, orientation of the detector relative to the magnetic field, bias voltage, and skin temperature.

The model developed in the present paper is based on the use of a two-dimensional magnetic coordinate system, in which the earth's magnetic field is mapped onto an axially symmetric dipole field using the adiabatic invariants of particle motion.[9] Any of a number of equivalent pairs of magnetic coordinates, including the B,L; R,λ and x,L sets[10] may be used to locate position in this dipole field. Briefly: The magnetic shell parameter, L, specifies a particular line of force (about which the trapped particle spirals) by the radial distance to the line in the equatorial plane of the dipole measured in units of one earth radius (see Fig. 1); position along the line of force is specified by either the magnetic induction (field strength), B, or by x, where $x=(1-B_0/B)^{\frac{1}{2}}$ is a convenient variable in the equations of the dynamics of charged particle motion. (B_0 is the magnetic induction at the equator on the line of force in question.) Magnetic dipole polar coordinates R and λ,

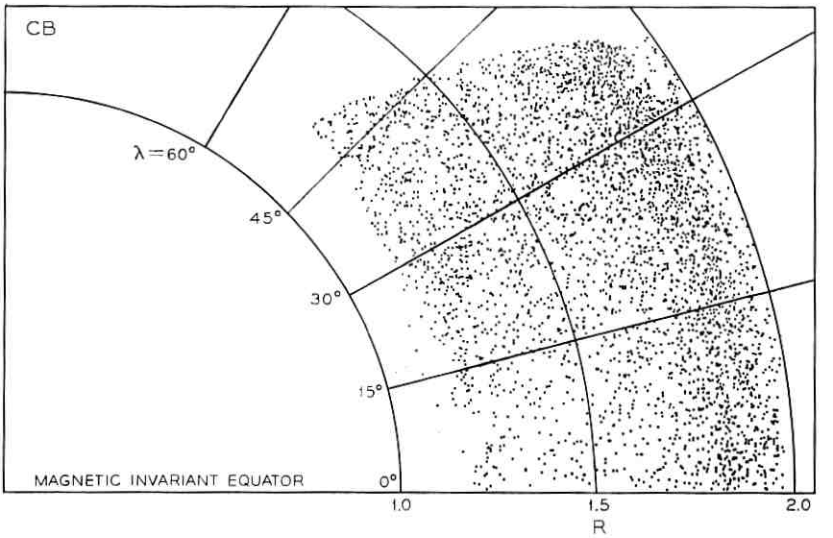


Fig. 2 — The spatial distribution of data for $L < 3$ in R, λ coordinates. Every twentieth point from the L-ordered data is plotted.

where R is the radial distance in earth radii and λ is the latitude angle, offer a sufficiently close analog to geographic coordinates to be convenient in many circumstances. The choice among these sets is discussed in Section III, as are the reasons for choosing the square root of the counting rate as the scale for the dependent variable.

The coordinates and variables, together with other symbols used in this analysis, are listed in Table I under the following headings: Radiation Intensity, Position and Orientation, Instrument and Energy Spectrum, Mathematical Model, Statistics, and Other. Summary information concerning units, constants, derivations, and sources is included.

The satellite was confined to the volume of space $\{1.09\ R_e \leqq R \leqq 1.95\ R_e, * 0 \leqq \lambda \leqq 58°\}$. For $\{L > 3, R < 1.95\ R_e\}$, the average counting rate is very nearly zero, and these data were not examined further. About 5 percent of the 50–130 MeV proton data for $L \leqq 3$ were associated with noise bursts which affected adjacent telemetry channels; these data were discarded. The study described below is based on the remaining 77,649 observations.

The spatial distribution of the data is indicated in Fig. 2 which is

* R_e = earth radius.

a plot in R,λ coordinates of the position of every twentieth point from the L-ordered data. Although data were not acquired continuously during the 226 days that the satellite was active, there are no time gaps in the data longer than two days in duration.

Fig. 3 is a plot of bands of constant counting rate made by plotting the R,λ coordinates at which certain specified numbers of counts were recorded during 11-second counting intervals. The data in Fig. 3 cover the entire seven-month life of the satellite. The narrowness of the contour bands demonstrates that the data are exceptionally well-behaved in both time and space, and that one may reasonably hope to describe radiation intensity in terms of R,λ coordinates or their equivalent.

Among the various sources of error in the data are: noise present in the received telemetry signal or introduced during the recording and processing of the telemetry; errors in the time as recorded by the ground station; errors in the satellite ephemeris; differences between the real magnetic field of the earth and the values of B and L calculated from the coefficients in the computer program INVAR (see Table I); and instrumental effects. In addition, one expects statistical fluctuations in the measured counting rate at a fixed position. The importance of these sources of error is discussed later.

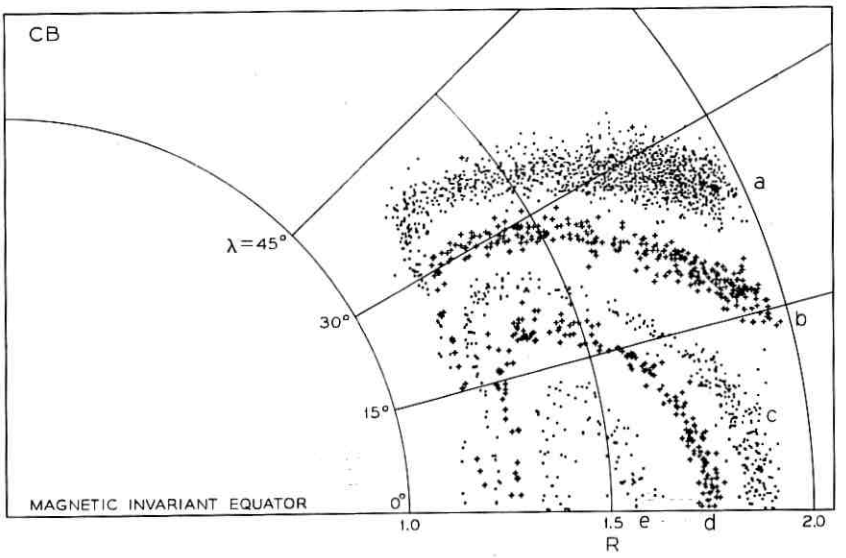


Fig. 3 — Bands of constant numbers of counts in 11 seconds in R,λ space: Band a, 4; Band b, 32; Band c, 127–129; Band d, 254–258; Band 3, 508–516 counts. All the data from the seven-month period are displayed.

III. CHOICE OF THE PRINCIPAL VARIABLES AND THEIR SCALES

The current state of knowledge of the earth's radiation belts suggests that the spatial distribution of high-energy protons may reasonably be organized on the basis of a two-dimensional magnetic coordinate system, except perhaps at very low altitudes near the South American magnetic anomaly, where longitude also becomes important. *Telstar*® 1 data plotted in Fig. 3 indicates that the observed counting-rate data does indeed depend principally on the magnetic coordinates, R and λ. The coordinates R,λ are defined in terms of the mathematically equivalent pair B,L.[9] A third equivalent set consists of L together with the coordinate x, suggested by Roberts,[10] defined in Table I.

We have primarily employed the x,L set in this study because of the following considerations: In the adiabatic theory, the mirror points of particles do not migrate between magnetic shells.[11] Within any shell, the coordinate x is approximately linear in λ for $\lambda < 30°$, and thus the near-equatorial data is not "crowded" into a small interval of the coordinate, as is the case for B. Moreover, we have been able to develop simple functional representations of the data in terms of x and L.

The flux of particles is the variable of greatest physical interest for comparing the results of different experiments, calculating physical effects of the radiation (such as radiation damage to devices in proposed orbits), deriving an energy spectrum from experimental measurements, examining the implications of various source and loss mechanisms, etc. However, the flux is not measured directly and requires for its calculation knowledge of the energy spectrum of the particles and of the energy dependence of the geometric factor of the detector. Even in the present circumstances where the conversion is (under the assumptions of Appendix A) quite insensitive to these, we prefer to carry out the bulk of the data analysis in terms mathematically equivalent to the directly observed counting rates.

From among the possible representations of the counting rate information (including counting rate, log counting rate, and square root of counting rate) the square root of the observed counting rate, Y, has been selected as the dependent variable. On the hypothesis that the number of counts in a given 11-second counting interval at any given position in space is a random variable with a Poisson distribution, it can be shown that the variance of Y is approximately constant, independent of its average value (see Appendix B.2). The least squares criterion has been used in all the estimating procedures; that is, coefficient estimates have been selected so that the sum of squares of dif-

TABLE I—COORDINATES, VARIABLES AND NOTATION

The redundent use of a few symbols is partly due to the decision to retain "standard" notation in both geophysics and statistics. The context should resolve any apparent ambiguities. Some symbols used locally in the text are not included in this table.

Symbol	Coordinate	Units	Source	Underlying variables	Remarks
Radiation Intensity					
J	Fitted average omnidirectional flux	protons/cm² sec	...	y, $\bar{g}$	Equation (21).
j	Predicted unidirectional flux	protons/cm² sec ster	...	y, $\bar{g}$	Equation (8).
Y, Y_j	Square root of observed counting rate	(counts/sec)$^{1/2}$	telemetry	Z	...
y	Fitted average value of Y	(counts/sec)$^{1/2}$	least squares fit	Y, x, L	Section IV. This symbol is used generically for all the models.
Z	Counts in an 11-second counting interval	counts	telemetry	...	Random variable.
Position and Orientation					
B	Magnetic induction	gauss	INVAR	r, θ, ϕ	Computer program INVAR by McIlwain[12] containing the Jensen and Cain[13] magnetic field coefficients for 1960. R_e = 6371.2 km.
B	Magnetic field strength	gauss	$\lvert\mathbf{B}\rvert$	r, θ, ϕ	...
B_0	Equatorial value of B	gauss	...	...	$B_e = 0.311653/L^3$ $= 0.311653/R^3$.
L	Magnetic shell parameter	ratio to earth radius (R_e)	INVAR	r, θ, ϕ	See **B**, above.

TABLE I — (Cont'd).

Symbol	Coordinate	Units	Source	Underlying variables	Remarks
Position and Orientation (Cont'd)					
L_m	Midpoint of an L-slice	same as L	...	...	Section V.
R	Magnetic dipole radial distance	R_e	B, L	r, θ, ϕ	$B = \frac{0.311653}{R^3}\left\{4 - \frac{3R}{L}\right\}^{1/2}$
R_e	Earth radii	km	Heiskanen[14]	...	6371.2 km.
r	Geocentric distance	earth radii (R_e)	ephemeris	tracking data	For geomagnetic calculations, r is corrected to altitude above the International Ellipsoid [Heiskanen[14]]. R_e = 6371.2 km.
T	Universal time	days	clock at telemetry receiving station	...	Measured in days from 0 hr 0 min. U. T., Jan. 0, 1962.
t	Local time	hours	...	T, ϕ	Apparent sun time (local mean time corrected for the equation of time taken from the American Ephemeris and Nautical Almanac[15]).
x	$\left[1 - \frac{0.311653}{BL^3}\right]^{1/2}$	dimensionless	B, L	r, θ, ϕ	See above for B and L.
α_o	Equatorial pitch angle	degrees	...	...	Fig. 1.
γ	Angle between satellite spin axis and local magnetic vector	degrees	**B**, ω	B, χ, δ	$\gamma = \mathbf{B}\cdot\omega/\lvert\mathbf{B}\rvert$, where ω is a unit vector parallel to the angular momemtum vector of the satellite.
δ	Declination of the satellite spin axis	degrees	mirror flash data	r, θ, ϕ, T, and astronomical data	Optical observations of the reflection of the sun from mirrors on the satellite, Courtney-Pratt, et al[8].

TABLE I—(Cont'd).

Symbol	Coordinate	Units	Source	Underlying variables	Remarks
Position and Orientation (Cont'd)					
θ	Colatitude	degrees	ephemeris	tracking data	geocentric angle.
λ	Magnetic dipole latitude	degrees	B, L	r, θ, ϕ	$\lambda = [\arccos(R/L)]^{1/2}$
μ_0	Cos α_0	dimensionless	...	...	Numerically equal to x.
ϕ	East longitude	degrees	ephemeris	tracking data	Geocentric angle.
χ	Right ascension of the satellite spin axis	degrees	mirror flash data	r, θ, ϕ, T, and astronomical data	See δ, above.
ω	Direction of the spin angular momentum vector of satellite	dimensionless	mirror flash data	χ, δ	See δ, above.
Instrument and Energy Spectrum					
E	Energy	MeV	...	...	...
E_0	e-folding energy	MeV	...	...	Used in energy spectrum, Appendix A.
g	Geometric factor of the detector	cm^2 ster	detector geometry	...	...
$\bar{g}$	Average geometric factor of detector	cm^2 ster	detector geometry	proton energy spectrum	Equation (20).
M	Exponent of integral power-law energy spectrum	dimensionless	...	...	...
N	Number of protons	dimensionless	...	...	...
n	Exponent of differential power-law energy spectrum	dimensionless	...	...	...
v_b	Bias voltage	bits	telemetry	resistor calibration	Each bit represents a step of -1.108 volts.
τ	Skin temperature	°C	telemetry	thermistor calibration	Measured near the detector.

TABLE I—(Cont'd).

Symbol	Coordinate	Units	Source	Underlying variables	Remarks
Mathematical Model					
A, A', A'', A''' A^{iv}	Equatorial value of y	(counts/sec)$^{1/2}$	fitting	L	The superscripts indicate various models, see Section IV. In particular A' indicates Model I and A'' indicates Model II. N.B. A is used generically for all the models, or when the distinction is unimportant or clear from the context.
A_p	Coefficient	(counts/sec)$^{1/2}$	fitting	...	Maximum value of A'' (and therefore y''), Model II, Equation (11).
a_1, a_2, a_3, a_4, a_5	Coefficients	...	fitting	...	Coefficients of A'' and A^{iv}, Equations (6) and (16).
b	...	dimensionless	...	x, (L)	Equation (18).
e_i	Coefficients	...	...	L	Equation (18).
F	...	dimensionless	...	L, (x)	Equation (19).
G, G', G'', G'''	...	dimensionless	...	x, (L)	Describes the x-dependence of y for the models indicated by the superscripts, see Section IV. N.B. G is used generically for all the models, or when the distinction is unimportant or clear from the context.

TABLE I—(Cont'd).

Symbol	Coordinate	Units	Source	Underlying variables	Remarks
Mathematical Model (Cont'd)					
L_e	Coefficient	same as L	fitting	...	Smallest value of L for which $y > 0$.
L_p	Coefficient	same as L	fitting	...	Position of A_p is at $(x, L) = (0, L_p)$.
L_1	Coefficient	same as L	fitting	...	Equation (16).
M	Coefficient	dimensionless	fitting	L	Model III, Equation (15).
P	Coefficient	dimensionless	fitting	L	Model III, Equation (15).
p_i	Coefficients	...	...	x	Equation (19).
Q	Coefficient	dimensionless	fitting	L	Model III, Equation (15).
R_c	R at cutoff	R_c	fitting	L	Equation (4).
$r_1, r_2, r_3, (r_4), (r_5)$	Coefficient	...	fitting	...	Coefficients of R_c, Equation (5).
S	Shape factor	dimensionless	fitting	L	Equation (3).
s_0, s_1	Coefficient	...	fitting	...	Coefficients of S, Equation (3).
x_c	Cutoff function	dimensionless	fitting	L	Smallest value of x for which $y = 0$, Equation (4).
y, y_L, y', y'', y'''	Fitted average value of y	$(\text{counts/sec})^{1/2}$	fitting	x, L	The subscript and superscripts indicate various models, see Section IV. In particular y' indicates Model I and y'' indicates Model II. N. B. y is used generically for all the models, or when the distinction is unimportant or clear from the context.
y_j	Fitted value	$(\text{counts/sec})^{1/2}$	fitting	x, L	Corresponds to the observation Y_j.
η	Coefficient	dimensionless	fitting	...	Shape factor, Equations (6) and (11).

TABLE I—(Cont'd).

Symbol	Coordinate	Units	Source	Underlying variables	Remarks
Other					
CB	Complete body (of data)	...	...	...	Designates all the data, see Section 4.5.
HTB	High temperature and high bias voltage (data)	... ...	... ...	... ...	Designates a subset of the data, see Sections 4.5 and 6.9.
Statistics					
Cov	Covariance	...	...	...	...
df	Degrees of freedom	dimensionless	...	...	See Wilks[16].
D_1^2, D_2^2	Squared distance	...	...	...	Appendix B.6.
g	Function	...	...	...	...
h	Function	...	...	...	...
n	Number of observations	...	...	...	...
R^2	Squared multiple correlation coefficient	...	...	...	...
Res, RES	Residual	...	...	...	Observed minus fitted.
SS	Sum of squares	...	...	...	...
u_j	Function of Y_j	...	...	...	...
$\bar{u}$	Mean of u_j	...	...	...	...
Var	Variance	...	...	...	Wilks[16].
w	Independent variable (vector)	...	...	...	...
w_i	Components of w	...	...	...	...
z	Values of the random variable Z	counts	...	...	...
α	Dependence coefficient	...	...	...	$\alpha = [1 - \sqrt{1 - \rho^2}]\,\mathrm{sign}\,(\rho)$, Wilk[17]. Equation (31).
β_i	Confidence coefficient	...	...	...	Wilks[16].
$\tilde{\delta}$	Correction to $\hat{\theta}$ (vector)	...	...	...	...

TABLE I—(Cont'd).

Symbol	Coordinate	Units	Source	Underlying Variables	Remarks
Statistics (Cont'd)					
$\tilde{\delta}_i$	Component of $\tilde{\delta}$	...	...	...	...
θ, θ'	Coefficient vector	...	...	...	...
θ_i	Component of θ'	...	...	...	...
$\hat{\theta}$, $\bar{\theta}$	Estimate of θ'	...	...	...	...
$\hat{\theta}_i$	Components of $\hat{\theta}$	...	...	...	Estimates of θ_i.
ν	Average value of a Poisson variable	counts	...	...	...
ρ	Correlation coefficient	dimensionless	...	...	...
σ	Standard deviation	...	...	...	...
MSD MSE MSR	The terms, mean square error (MSE), mean square residual (MSR) and mean square deviation (MSD) are used in this document to denote related but different entities, each measuring "goodness of fit" in relation to different situations. When a selected array of data is fitted by a model, the minimum sum of squares of residuals from the fit of those data divided by the degrees of freedom (number of selected observations minus number of coefficients fitted) is termed the MSE. When a fit based on a sample of data is used to generate residuals for all of the data, without refitting, the total sum of squares of these residuals divided by the number of residuals is termed the MSR. For defined "small" cells in x,L space, the sum of squares of deviations of observations from their average in the cell divided by the number of such deviations minus one is termed the MSD.				

ferences between observed and fitted values is minimized. The choice of the square root scale, Y, as the scale on which to represent the counting rate data makes troublesome differential weighting of the data in the least squares fitting unnecessary. Similarly, plots of Y versus various variables are convenient since the scatter in Y is approximately independent of the value of Y. In fact, the square root transformation will make the variance of the observation approximately independent of its average value whenever the variance is proportional to the mean. Thus, the procedure is more robust than the assumption of a Poisson distribution, for which the variance equals the mean. Further discussion and detail is given in Appendices B.2 and C.

The results were restored to counting rate and the flux was calculated using the best estimate of the average geometric factor, $\bar{g}$, (see Appendix A) to facilitate the discussion of the physical significance of the measurements.

IV. THE EVOLUTION OF THE MODELS

4.1 *General Approach*

This section provides a summary overview of the evolution of the models, the details and accomplishments of which are elaborated in the following sections and appendices.

The approach to model development in this study has been largely empirical. Theoretical physics considerations are currently too complex and speculative to do more than serve as a general guide and stimulus. We have proceeded on the presumption that an adequate model for the spatial distribution of the high-energy protons can be based on the mapping of the earth's magnetic field onto a two-dimensional axially symmetric dipole field, expressed, for example, in the coordinates x and L. This is supported by the plots of Fig. 3, the successful polynomial fits on L-lines of McIlwain,[18] Valerio,[19] and Fillius,[20] and by the results of the present study.

The ultimate justification of the mathematical models developed herein is that, when appropriate estimates of coefficients are inserted, good fits to the data are obtained. Various other mathematical, physical, and statistical considerations also provided guidance and evaluation.

The evolution involved successive interactions with the data and iteration on models. Roughly, the main stages included: grouping the

data into L-slices; inferring a mathematical function having adjustable coefficients which would fit a selected series of L-slices; developing a mathematical function to describe the dependence of the L-slice coefficients on L; thence fitting the two-dimensional model so-defined to a sample of the data; using this fit to screen outliers, to detect instrumental effects and, after partitioning the data, to select a representative sample of partitioned data for further fitting; after obtaining a very good fit to the partitioned data, some remaining physical and statistical defects of the model were overcome by a reparametrization and specialization. Further generalizations of the model were also tested.

4.2 *The L-slice Model*

As a developmental operational procedure (encouraged by the L-shell orientation of the adiabatic theory[11]) the data were grouped into a series of narrow bands according to L values (e.g., $1.849 \leqq L \leqq 1.851$) and plotted versus x. Retrospectively, there is every reason to believe that an initial approach based on grouping the data into x-slices would also have led to an effective analysis (see Section 13.2). Various functional forms, having adjustable coefficients dependent on L, were tested for adequacy of fit to the L-slices.

Initially, we employed the functional form

$$y_L(x) = \begin{cases} A \cdot G(x; x_c, S) & (x \leqq x_c), \\ 0 & (x > x_c), \end{cases} \tag{1}$$

where A, x_c and S are fitted coefficients for each L-slice, and

$$G(x; x_c, S) = \begin{cases} (1 - x^2)^{-\frac{1}{4}} \left[1 - \left(\frac{x}{x_c} \right)^2 \right]^{S+\frac{1}{4}} & (x \leqq x_c), \\ 0 & (x > x_c). \end{cases} \tag{2}$$

For this body of data from the region $\{R \leqq 1.95\, R_e\,,\ 1.15 \leqq L \leqq 3.0\}$, we have found this $y_L(x)$ function provides an adequately flexible model on L-slices, for appropriately fitted values of the coefficients A, x_c, and S. In this representation for given fixed L, the quantity A^2 may be interpreted as the average equatorial omnidirectional counting rate, since $x = 0$ on the equator, x_c represents a "cutoff" value for x, i.e., the cosine of the equatorial pitch angle corresponding to the "loss cone", and S has the effect of a shape factor in the y,x dependence.

The analysis using this $y_L(x)$ model is described in Section V.

4.3 *Dependence on L*

The $y_L(x)$ model was fitted to a series of L-slices, obtaining fitted values of A, x_c and S. These were each plotted against the nominal (mid-range) L value for the slice and a reasonably smooth variation with L obtained.

Thence we inferred the following functional dependence of the L-slice coefficient estimates on L:

$$S = S(L) = s_0 + s_1 L, \tag{3}$$

$$x_c = x_c(L) = \sqrt{1 - \left(\frac{R_c}{L}\right)^3 \left[4 - 3\frac{R_c}{L}\right]^{-1}}, \tag{4}$$

$$R_c = R_c(L) = L_0 + r_1(L - L_0) + r_2(L - L_0)^2 + r_3(L - L_0)^3, \tag{5}$$

$$A = A'(L) = \begin{cases} \dfrac{a_1(L - L_0)}{a_2 + (L - a_3)^\eta} & (L \geqq L_0), \\ 0 & (L < L_0), \end{cases} \tag{6}$$

where s_0, s_1, r_1, r_2, r_3, a_1, a_2, a_3, η and L_0 are fitted coefficients.

Equation (4) simply expresses the mathematical relationship between R (or R_c) and x (or x_c) in the magnetic dipole field (see Table I). The coefficient L_0, which occurs in $A'(L)$ and $x_c(L)$, may be interpreted as the lower bound of the L shells on which protons with energies above 50 MeV were measurable. The quantity $R_c(L)$ is such that $R_c(L) - 1$ is the equivalent dipole altitude at which the counting rate falls to zero.

4.4 *A Two-Dimensional Model—Model I*

The conjunction of (1) to (6) defines a two-dimensional model, referred to henceforth as Model I,

$$y'(x, L) = A'(L) \cdot G'(x, x_c(L), S(L)), \tag{7}$$

where G' is essentially the function G of (2), with x_c and S explicitly dependent on L.

Though empirical considerations mainly guided the choice of these functions, some physical and mathematical properties influenced the choice. In the present case, in which the geometric factor of the detector is considered to be independent of the energy spectrum (see Appendix A), $[y(x, L)]^2$ transforms in closed form to the equatorial pitch angle distribution, giving[10]

$$j(\mu_0, L) = \frac{4\pi}{\bar{g}} \frac{[A(L)]^2\left\{1 - \left[\frac{\mu_0}{x_c(L)}\right]^2\right\}^{2S(L)}}{2\pi x_c(L)\beta(\frac{1}{2}, 1 + 2S(L))}, \tag{8}$$

where $j(\mu_0, L)$ is the predicted equatorial unidirectional flux (protons/cm^2 sec ster) at equatorial pitch angle $\alpha_0 = \arccos \mu_0$, and β is the beta function,

$$\beta(p, q) = \int_0^1 u^{p-1}(1 - u)^{q-1}\, du. \tag{9}$$

In addition $y'(x, L)$ has good boundary behavior. The derivative at the magnetic equator, $\partial y'(0, L)/\partial x$, is 0, which provides continuity. When $\frac{1}{4} < S(L) < \frac{3}{4}$, then $\partial y'(x_c, L)/\partial x \to -\infty$ and $\partial[y'(x_c, L)]^2/\partial x = 0$. The estimated values of S do satisfy this constraint in the present case. The desirable consequences of this behavior of the derivatives will be discussed in Section V. The function $y'(x, L)$ gives smooth interpolation over regions sparse in data, and does not have any of the wild fluctuations often associated with polynomial fits.

The analysis of the data using Model I is described in Section VI.

4.5 *Summary Uses of Model I.*

The unspecified coefficients of Model I were estimated by nonlinear least squares fitting to a sample of about 1000 observations from the complete body of data. Thence this fit of Model I (the CB fit) was evaluated relative to all the data and to auxiliary variables, such as time, which were not included in the model. Outliers were thereby detected and screened. An instrumental effect was uncovered (see Section 6.8), and this led to an objective partitioning of the data, yielding a subset (HTB data) for further analysis. The CB fit of Model I was also used to specify a representative data sampling procedure for further fitting to the HTB data.

Though Model I produces a very good fit to the HTB data (see Section VII), it has certain physical and statistical defects. Specifically, though the quantities A and x_c in the L-slice model have a direct physical interpretation, most of the coefficients in $y'(x, L)$ do not. Additionally, the estimates of the coefficients in $A'(L)$ turn out to have exceedingly high statistical correlations and the model $y'(x, L)$, as a function of the coefficients, exhibits marked nonlinearities even in a close neighborhood of the least squares estimates (see Section 8.5).

Therefore, after clarifying the character of the data and obtaining a good fit, attention was given to additional improvements of the model.

4.6 *A Modified Model—Model II*

The statistical difficulties of Model I were entirely overcome by employing a specialized version of $A'(L)$, defined below. Furthermore, this specialized model, Model II, retains all the desirable properties of Model I while providing both aesthetic improvement and greater physical interpretability.

Model II is defined by

$$y''(x, L) = A''(L) \cdot G''(x, x_c(L), S(L)), \tag{10}$$

where G'' is as in (2), but with $S(L) = s_0$, and

$$A''(L) = \begin{cases} \dfrac{A_p(L - L_0)}{\dfrac{(\eta - 2)}{\eta}(L_p - L_0) + \dfrac{2}{\eta}\dfrac{[(L_p + L - 2L_0)/2]^\eta}{(L_p - L_0)^{\eta-1}}} & (L \geqq L_0), \\ 0 & (L < L_0), \end{cases} \tag{11}$$

where A_p, L_0, L_p and η are the coefficients to be estimated.

$A''(L)$ is a special case of $A'(L)$ and relates to it by the following transformations:

$$\begin{aligned} L_0 &= L_0 \\ \eta &= \eta \\ a_3 &= 2L_0 - L_p \\ a_2 &= 2^{\eta-1}(\eta - 2)(L_p - L_0)^\eta \\ a_1 &= 2^{\eta-1}A_p\eta(L_p - L_0)^{\eta-1}. \end{aligned} \tag{12}$$

Indeed, Model II is essentially defined by the following nonlinear constraint imposed on Model I:

$$a_2 = 2^{\eta-1}(\eta - 2)(L_0 - a_3)^\eta. \tag{13}$$

The coefficients of $A''(L)$ in Model II have the following physical interpretations:

- L_0 (as before) is the smallest value of L such that high-energy protons are measurable by the instrument;
- A_p is the square root of the maximum counting rate of high-energy protons in the radiation belt;
- L_p is the value of the magnetic shell parameter (on the equator, $x = 0$) at the highest radiation intensity;
- η may be interpreted as a shape factor for the equatorial (counting rate)$^{\frac{1}{2}}$ function, $A''(L)$.

The model $A''(L)$ has the form of a product, with the maximum value, A_p, being multiplied by a factor which decreases as L departs from L_p in either direction. Note that the factor multiplying A_p is dimensionless.

The other fitted coefficients of Model II are s_0, which is a shape factor for the dependence of (counting rate)$^{\frac{1}{2}}$ on x at constant L, and r_1, r_2 and r_3 which, with L_0, define the cutoff function $x_c(L)$.

The analysis of the HTB data using Model II and comparisons of Models II and I are considered in Section IX.

4.7 *Generalizations*

The previously defined models may be regarded as special cases of Model III defined by

$$y'''(x, L) = A'''(L) \cdot G'''(x, x_c(L), M(L), P(L), Q(L)), \tag{14}$$

where $A'''(L) = A'(L)$, defined in (6),

$$G''' = \begin{cases} \left[1 - \left(\frac{x}{x_c(L)}\right)^{M(L)}\right]^{P(L)} / (1 - x^2)^{Q(L)} & (x \leqq x_c), \\ 0 & (x > x_c), \end{cases} \tag{15}$$

$x_c(L)$ is as defined in (4), and $M(P)$, $P(L)$ and $Q(L)$ involve coefficients or functions to be fitted.

The function G' is a special case of G''', in which $M(L) = 2$ and $Q(L) = \frac{1}{4}$. This permits a closed form transformation to an equatorial pitch angle distribution. The function G'' additionally constrains $P(L) = s_0$, independent of L.

The more general G''' in Model III can be used on L slices to determine L-slice estimates of M, P, Q, as well as A and x_c, and these in turn inspected to infer functional dependence on L. Clearly, this more general form must lead to at least as good a fit as Models I or II. Work has been done with Model III[21] but no important improvement over Model II was obtained for this body of data.

Neither of the fitted models $y'(x, L)$ nor $y''(x, L)$ is applicable far outside the spatial and energy regions that include the data analyzed here. For example, Models I and II do not fit well to the 26–33 MeV protons measured by the *Telstar*® 1 satellite, nor are they suitable for fitting many of the electron distributions. Preliminary investigations indicate that these remarks may not apply to G''', whose additional coefficients allow more rapid changes in curvature as a function of x.

We have already shown for *Telstar*® 2 data[5] that $A(L)$ can be extended to include description of the plateau of high-energy protons reported by McIlwain[18, 22] near the equator at $R \approx 2.2\ R_e$, beyond the orbital extremes of the *Telstar*® 1 satellite. The extension was made by adding a term to $A'(L)$, (6), to give A^{iv} defined by

$$A^{iv} = A'(L) + a_4 \exp\left[-\frac{(L - L_1)^2}{a_5}\right], \tag{16}$$

where a_4, a_5, and L_1 are coefficients describing the equatorial distribution of the "excess" protons that give rise to the plateau. In the less stable parts of the radiation belts the early work on empirical time dependence presented by Gabbe and Brown[5] clearly requires extension.

V. FITS ON THE L-SLICES

The model of (1) and (2) was fitted to the data, on the scale of Y, in 92 individual L-slices, using a nonlinear, multidimensional, least squares, computer program (see Appendix B) to estimate the coefficients and produce various statistical measures. The procedure of fitting to L-slice data enabled one to test functional forms of $y_L(x)$ and then to evolve functional forms for the dependencies of the coefficients of the L-slice models on L.

Proceeding in this manner, however, has a number of possible pitfalls. In particular, the estimates of coefficients within an L-slice may be highly correlated, and the reliability of the actual values of the estimated coefficients also depends on the pattern of data points in the particular L-slice, e.g., whether or not there are points near x_c. Hence, the estimated values for any particular coefficient may not exhibit a smooth dependence on L.

The form of the L-slices whose middle values of L, called L_m, are 1.35, 1.801, 2.2015, and 1.79, respectively, are displayed in Figs. 4 to 7. The thin solid lines in the figures are the fits to the L-slice data (meaning of the dashed and thick solid lines will be taken up later). The numerical values of the coefficients of the fits, and the widths of the slices are given in Table II. Figs. 4 and 5 are examples of the high quality of fit which is typically obtained for L-slices having $L_m < 2$.

In Figs. 4(a) and 5(a), square root of counting rate is plotted against x. One sees that the fit to the data points (the thin solid line) is quite adequate. The cutoffs, x_c, are well-defined, the scatter in Y is approximately independent of y and the data are well-distributed in x.

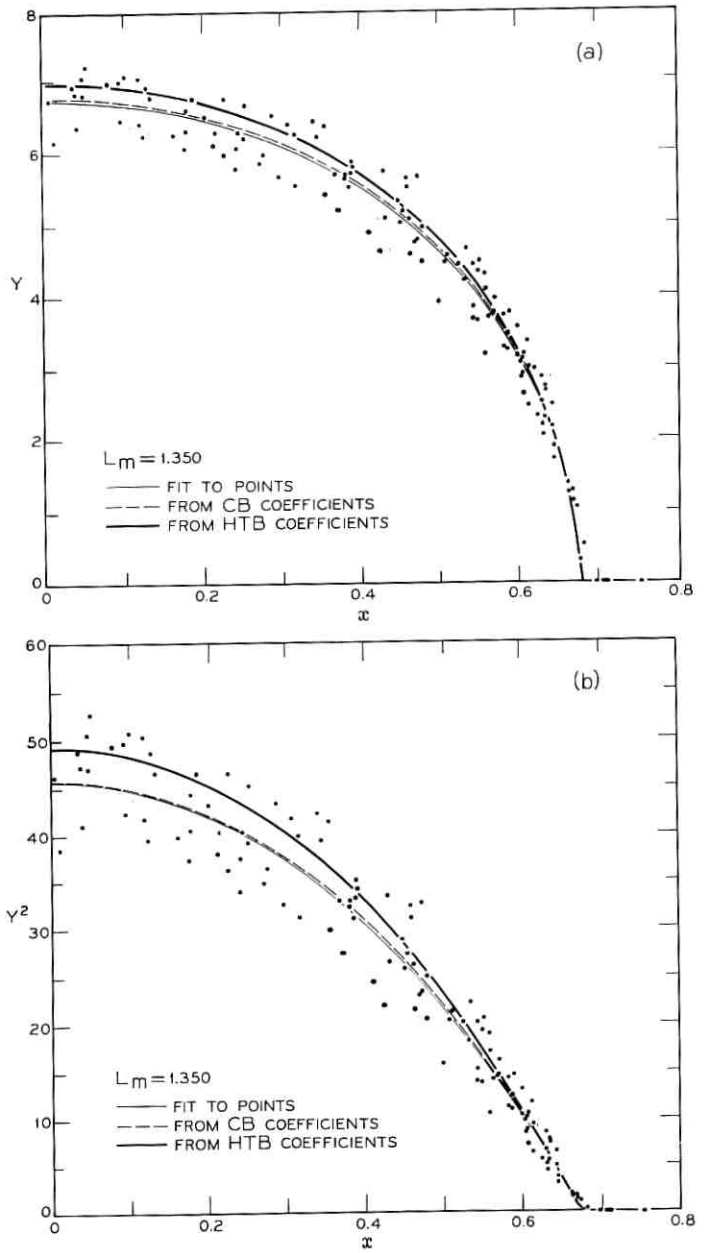


Fig. 4 — Data from the L-slice centered at $L_m = 1.35$ and the results of three fits shown on four scales.

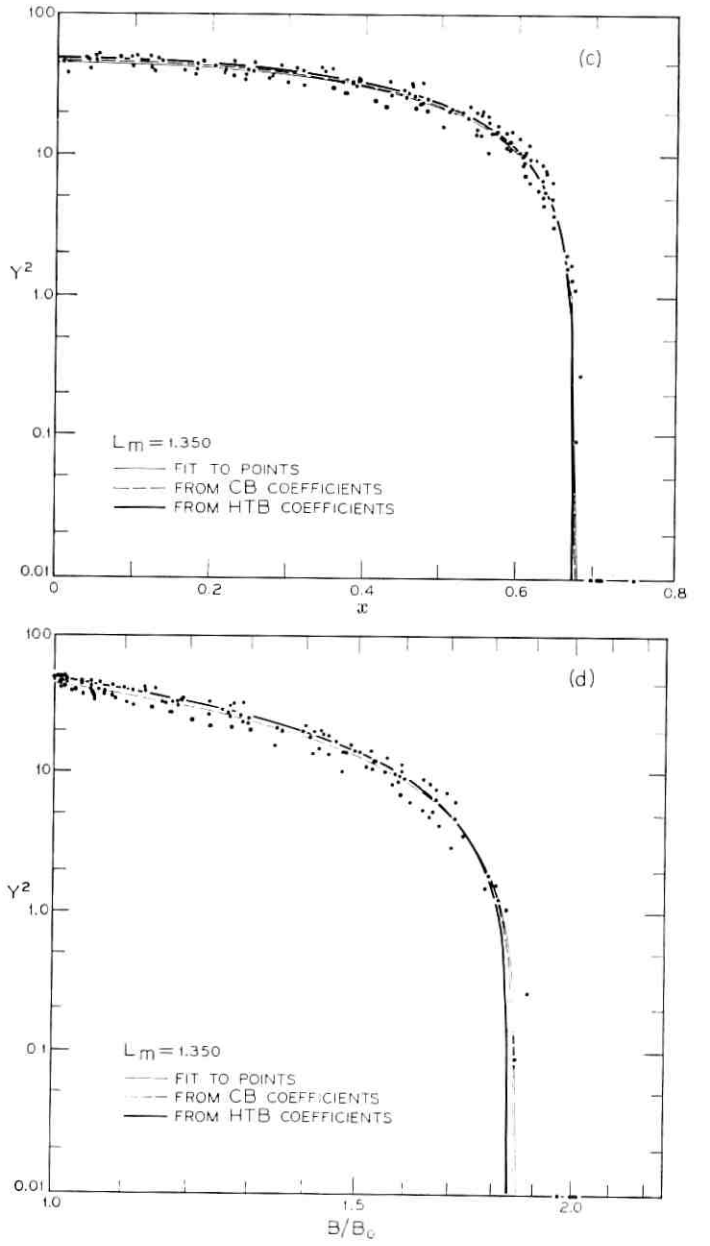


Fig. 4 — (continued)

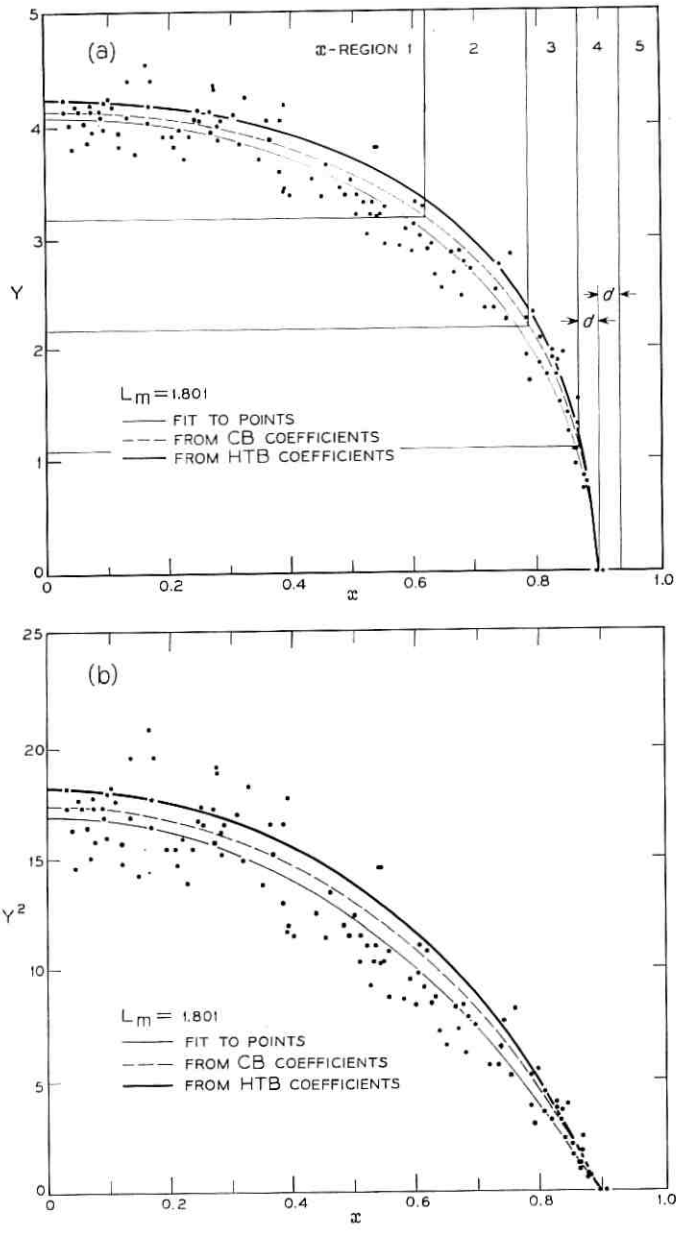


Fig. 5 — Data from the L-slice centered at $L_m = 1.801$ and the results of three fits shown on four scales. The partitioning in (a) is discussed in Section 7.1.

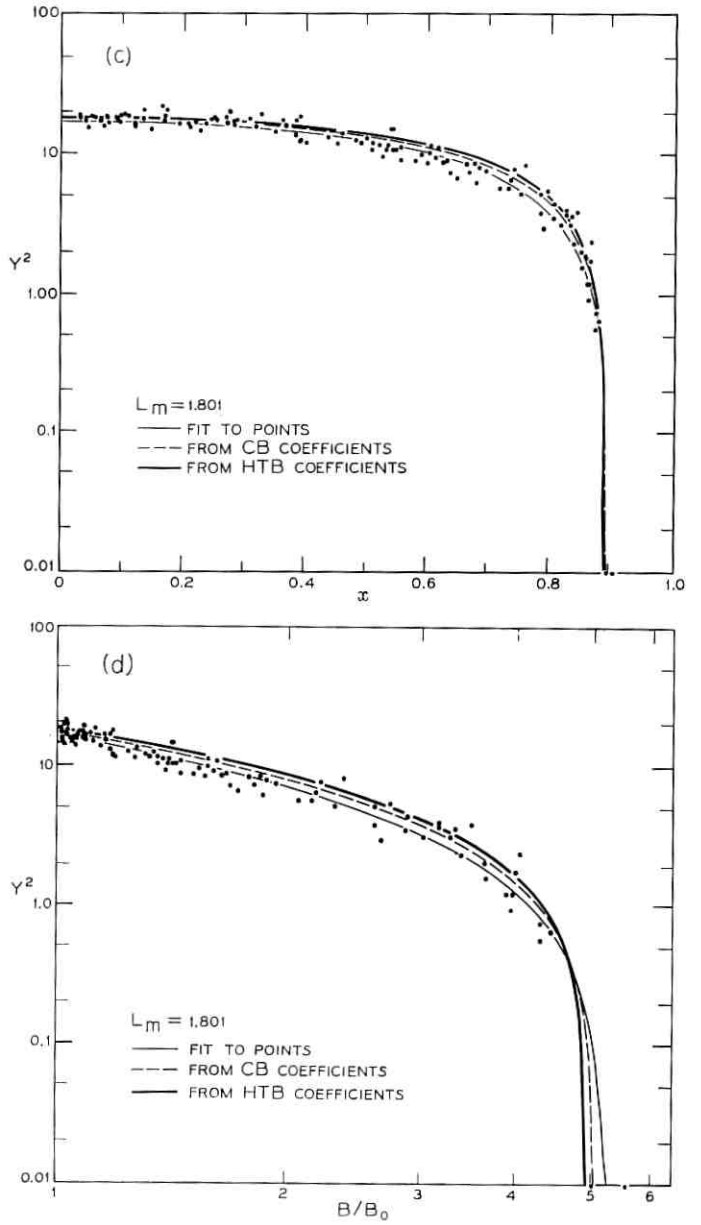


Fig. 5 — (continued)

TABLE II—COEFFICIENTS AND STATISTICS OF THE L-SLICE FITS.

L_m	1.35	1.801	2.2015	1.79
L_{MIN}	1.346	1.800	2.200	1.7895
L_{MAX}	1.354	1.802	2.203	1.7905
ΔL	0.008	0.002	0.003	0.001
A	6.757	4.109	1.70	4.324
$\sigma(A)$	0.053	0.031	0.12	0.043
x_c	0.6795	0.8998	0.954	0.923
$\sigma(x_c)$	0.0027	0.0044	0.011	0.015
S	0.324	0.390	0.58	0.478
$\sigma(S)$	0.018	0.024	0.10	0.060
Number of pts	140	129	144	65
MSE	0.1125	0.0497	0.0282	0.0478
Correlation coefficients				
A with x_c	0.281	0.309	0.724	0.408
A with S	0.605	0.561	0.940	0.548
x_c with S	0.774	0.820	0.890	0.944

As the cutoff is sharp on the scale of y, it is convenient to have a function which has an infinite derivative at x_c. Otherwise the exact x at which $y \to 0$ may have relatively little effect on the mean square error of the fit. This would lead to an ill-defined value for x_c, even though the data allows one to evaluate the position of the cutoff quite precisely for L values smaller than ≈ 1.9.

In Figs. 4(b) and 5(b), the counting rate, Y^2, is plotted against x. The thin solid lines represent the same fits as those in Figs. 4(a) and 5(a). One finds that the position of the cutoff is no longer well-defined on the plot. Instead the counting rate fades away as x increases. Having the derivative of y^2 equal zero at the cutoff (as noted in the previous section) is suitable in this situation. The scatter in Y^2 now changes with y^2, and is greater for large values of y^2 (small values of x). This nonuniform scatter makes it more difficult to judge the appropriateness of fit. If one wished to minimize the squared deviations between observed and fitted in terms of y^2 (or log y^2) the values of Y^2 (or log Y^2) would have to be weighted inversely as their estimated approximate variance, with a loss of intuitive appreciation of the quality of fit from a scatter plot and a substantial inconvenience in carrying out the fitting procedure.

In Figs. 4(c) and 5(c) the ordinate is log y^2. This choice of coordinate restores the ability to discriminate in the vicinity of the cutoff at the cost of a large loss of sensitivity in regions where the counting rate is higher.

Finally, Figs. 4(d) and 5(d) display the same data in the coordinate

system log y^2, log (B/B_0). This choice of abscissa expands the high-x region enormously, but contracts the low-x region to the point where it is impossible to see the details of the particle distribution in the vicinity of the equator ($x = 0$). This contraction would be even more severe if the abscissa were B or B/B_0.

In the region defined by $\lambda < 45°$, which covers the high energy proton data, the coordinate x provides adequate detail (see Ref. 10 for further discussion). If, however, the data had extended to $\lambda > 45°$ another choice of magnetic coordinate would have been desirable for $x > 0.95$, because all $\lambda > 45°$ are crowded into x values between 0.95 and 1.

The standard errors and correlations of the coefficients of the four L-slices under discussion, together with mean square error (MSE)* of fits, are listed in Table II. The standard error is in general a relatively small fraction of the estimate and the MSE is substantially greater at small values of L_m than at larger ones. This is further analyzed in Section VI.

At $L = 2.2$ the satellite gets no closer to the magnetic dipole equator than $\lambda = 20°$. This fact, which is associated with the problem of correlation of coefficient estimates within L-slices, is displayed more emphatically by choosing x as a coordinate, as in Figs. 6(a), (b), and (c), than by choosing log (B/B_0) as in Fig. 6(d). In addition, in Fig. 6(d) the expansion of the abscissa in the region of the cutoff makes it difficult to judge the physical appropriateness of the value of x_c which results from the least squares procedure. The same difficulty is encountered to a lesser degree with Fig. 6(b). However, in Figs. 6(a) and 6(c) one judges the x-intercept of the thin solid line to be too large, and Fig. 6(a) has the additional advantage of allowing one to make a better judgment of the quality of the fit at lower values of x. As might be surmised from the high values of the correlations for $L_m = 2.2$ in Table II, the value of x_c can be adjusted to a substantial extent without much change in the mean square error. These high correlations, which typically occur for $L_m > 2$, reduce confidence in the individual estimates of the coefficients for given L-slices. This difficulty also reduces the stability of the estimates of the coefficients as L_m is changed, and precludes basing the values of $x_c(L)$ and $S(L)$, for $L > 2$, on the fits to the L-slices.

A similar difficulty may be introduced when $L < 2$ by sampling fluctuations as illustrated in Fig. 7. In this case, there is a scarcity of

---

* Some statistical terms are defined in Table I.

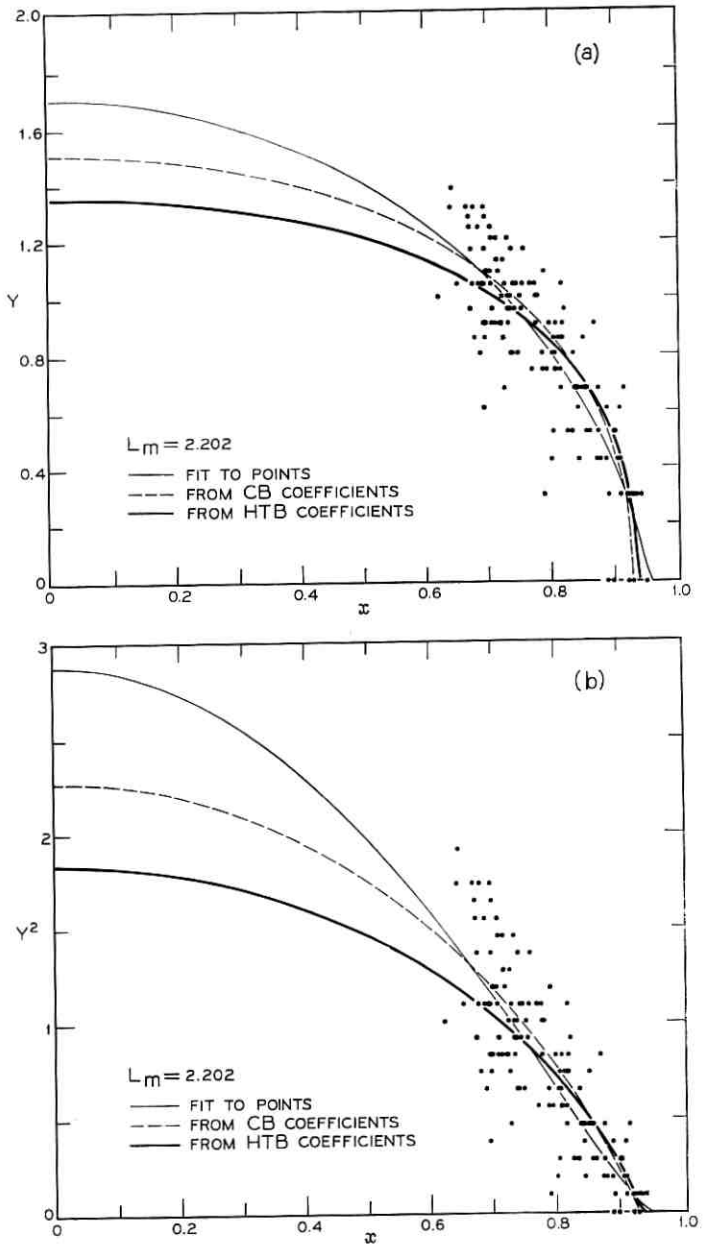


Fig. 6 — Data from the L-slice centered at $L_m = 2.202$ and the results of three fits shown on four scales.

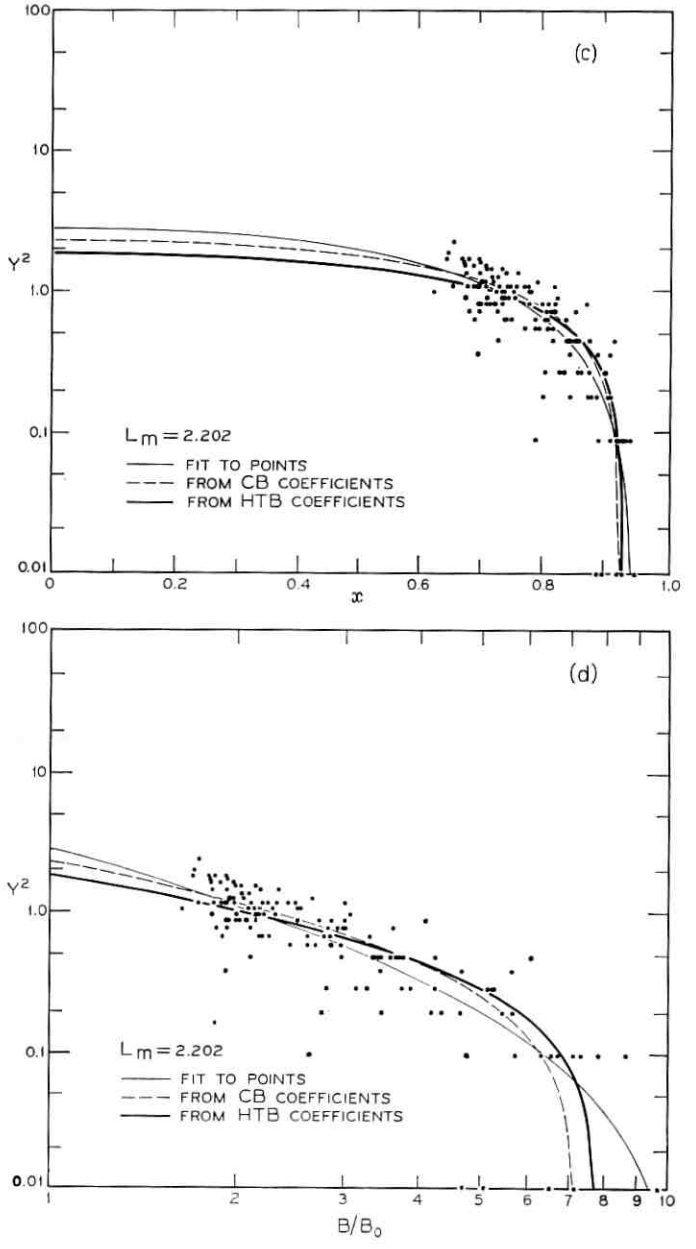


Fig. 6 — (continued)

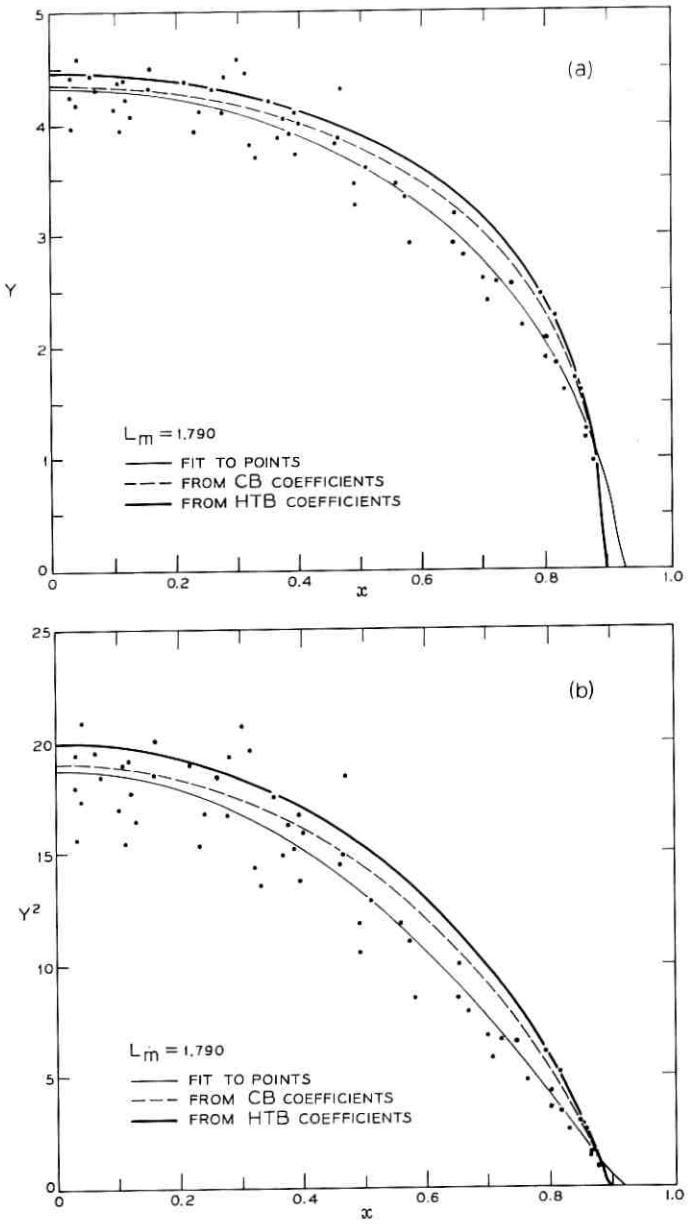


Fig. 7 — Data from the L-slice centered at $L_m = 1.790$ and the results of three fits shown on four scales.

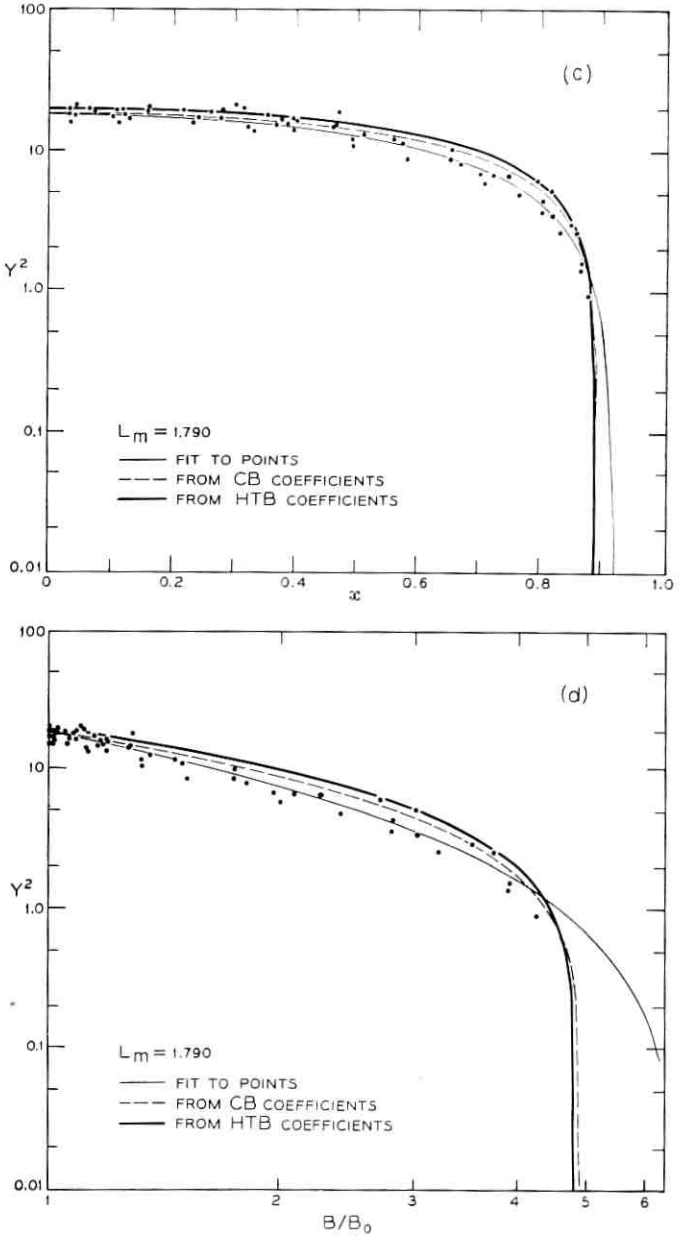


Fig. 7 — (continued)

data near and beyond the cutoff, unlike the slice with $L_m = 1.801$ illustrated in Fig. 5. The paucity of data near the cutoff in the L-slice centered on $L_m = 1.79$ both correlates and distorts the values of x_c and S. In this particular case, the width of the L-slice can be increased to avoid this difficulty, but, in general, increasing the width of the slice to include enough data may introduce a serious L-dependence within the slice. As a result, x_c may be determined by points near one extreme of L within the slice, A by points at the other extreme and S by some combination. This problem is especially severe below $L = 1.3$ where data begin to become sparse.

The plotted points in Figs. 8 to 10 summarize the dependencies of the estimates of the L-slice coefficients A, x_c, and S, respectively, on L_m, for all 92 slices. More than one value of the coefficients is plotted for some values of L_m because on occasion the width of the L-slice was

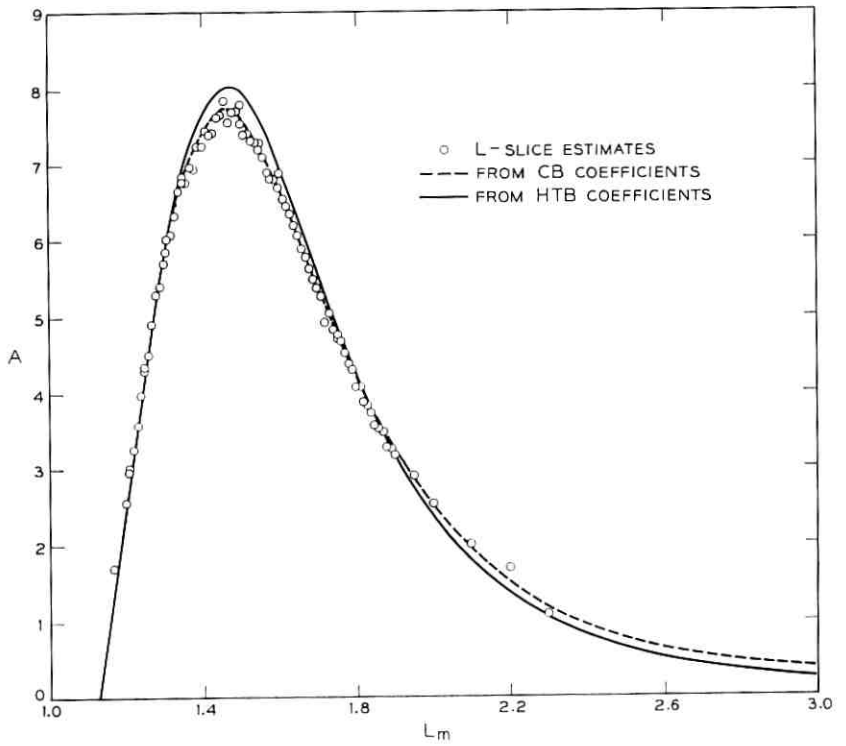


Fig. 8 — Three estimates of A as a function of L.

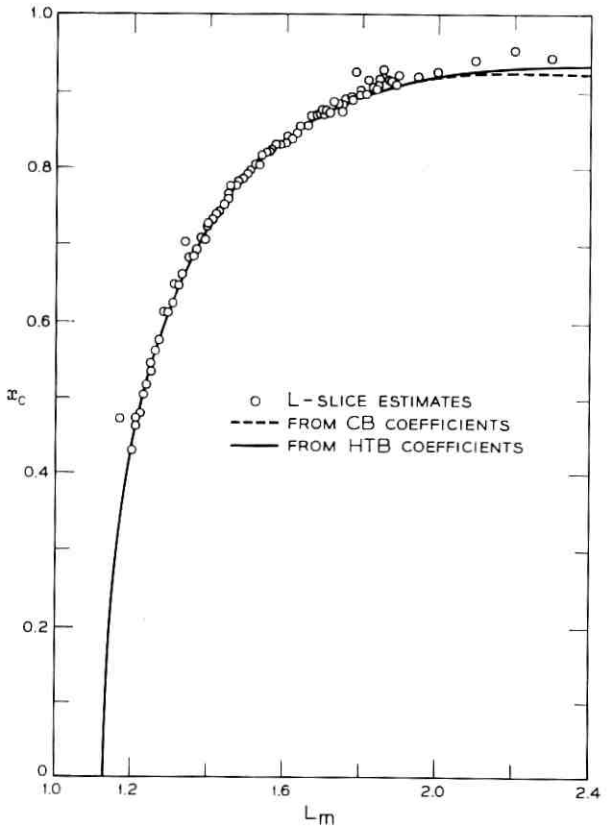


Fig. 9 — Three estimates of x_c as a function of L.

varied without changing L_m. Although there are local fluctuations in the estimates that arise from the way a narrow L-slice samples the data, the estimates exhibit a smooth dependence on L. The fluctuations are particularly pronounced near $L_m = 1.8$ in Figs. 9 and 10, and $L_m = 1.3$ in Fig. 10.

The standard errors of the L-slice estimates of A are typically 1 percent for $L < 1.95$, but become as large as 6 percent where there are no equatorial data, as is the case for $L > 1.95$. Fox x_c estimates, the standard errors are typically 0.5 percent. The estimates of S have a standard error of about 5 percent (± 0.015) near $L = 1.5$ and about 15 percent (± 0.05) near $L = 1.2$ and $L = 2$. The meanings of the curves in Figs. 8 to 10 will be discussed in the following sections.

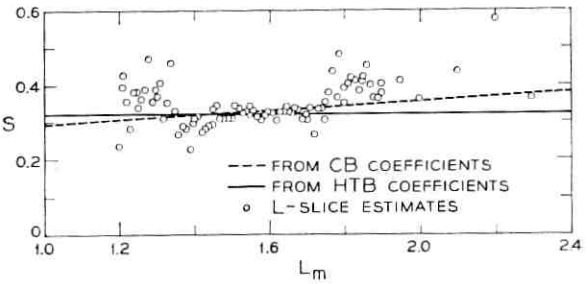


Fig. 10 — Three estimates of S as a function of L.

In summary, the L-slice approach enables one to infer a functional dependence of L-slice coefficients on L and to obtain an intuitive appreciation of the quality and nature of fit. The fitting procedure requires refinement by being carried out as a simultaneous two-dimensional process in x and L jointly. This overcomes the "grouping" inaccuracy in the L-slice approach and in addition makes good use of the data in those regions where data are scarce. The resultant function also provides convenient and excellent interpolation of data over the entire x,L region while employing a relatively small number (8, 9, or 10) of fitted coefficients.

VI. THE TWO-DIMENSIONAL FIT FOR THE COMPLETE BODY OF DATA

The analysis of this section is a precursor to the more refined parallel analysis of Section VII. This preliminary analysis produces the following results of consequence: Model I (see Section 4.4) is shown to be satisfactory; instrumental effects are identified and an objective algorithm for partitioning the data to reduce these effects is formulated; outliers are screened; and a more adequate basis for sample selection is provided. Many statistical details are omitted from this section, and statistical matters are dealt with more fully in Sections VII, VIII, and IX and in Appendices B and C.

6.1 *Sample Selection and Fit*

It was necessary, for practical computing reasons, to make a selection of approximately 1000 observations on which to carry out the simultaneous two-dimensional (in x and L) nonlinear (in the coefficients) least squares fit. In this preliminary phase, the nearly 80,000 data points were sampled by dividing the L-range from 1.15 to 3.00

into 925 contiguous intervals, each 0.002 wide. One data point was selected from each interval. As the data are approximately uniformly distributed in x (in the x-range covered by the satellite) in each L-slice (see Figs. 4 to 7), no effort was made at this point to influence the x distribution of the observations in this subset. The question of the "design" of the sample to be used as a basis for fitting the model is rather important, however, since the fit obtained with the empirical model is responsive to the distribution of data in x,L space. Other bases of sampling were employed later (see Section 7.1 and Appendix B.3).

Model I, described in Section 4.4, was fitted to the 925-point sample from the complete body (CB) of data. As this serves only as a preliminary fit, the values of the CB coefficients and other statistics are not presented here.

The quality of this fit was examined from various viewpoints: (*i*) by its behavior along the boundaries of the belt; (*ii*) by comparison with the L-slice fits; (*iii*) by plotting the residuals (observed value minus fitted value) versus the x and L coordinates; and (*iv*) by examining the mean square residuals (MSR) in various regions of magnetic coordinate space. Though the coefficients of the model were estimated from 925 sampled data points, the evaluation of quality of fit was based on all the nearly 80,000 observations.

6.2 *Evaluation of Fit at Equator*

The points in Fig. 11 are the values of Y (square root of observed counting rate) plotted against L for all data points for which x is near 0, specifically $x < 0.037$ (i.e., $\lambda < 1°$). For a given L, $y'(x, L)$ changes very little between $x = 0$ and $x = 0.037$ (see Figs. 4 and 5) and the points in Fig. 11 may be regarded as approximate equatorial points. The curve in Fig. 11 gives the fitted values of $A'(L) = y'\ (0, L)$ using the CB coefficients, and appears to represent the data very well. Note that $A'(L)$ has not come from a fit to the equatorial data as such, but rather is the equatorial value of y' as predicted by the two-dimensional fit. That is, the fitted $A'(L)$ does not minimize the sums of squares of deviations for just the equatorial points, but is, rather, the optimum fit in the least squares sense to the 925-observation sample, and these observations are distributed through x,L space. The excessive scatter in the equatorial value of Y between $L = 1.35$ and $L = 1.55$ which shows in Fig. 11 will be taken up in the next section.

The values of $A'(L)$ are also plotted for reference as the dashed

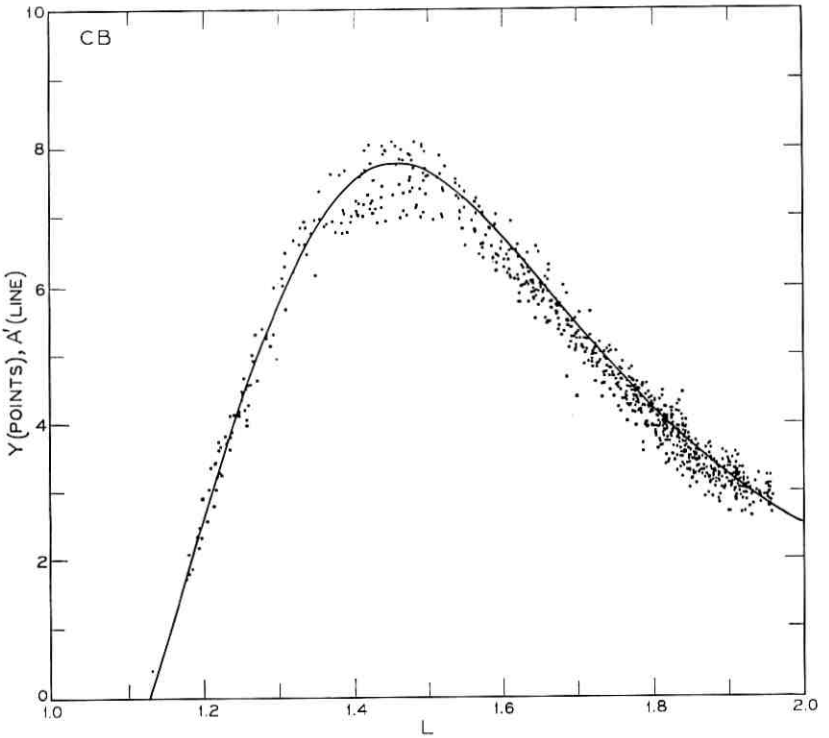


Fig. 11 — All data for $x < 0.037$ (i.e., within 1° of the magnetic invariant equator) and the equatorial value estimated from the CB coefficients plotted against L. A' and Y are in units of (counts/sec)$^{1/2}$.

line in Fig. 8. One sees that the L-slices give quite good estimates for A, although these estimates tend to be a little erratic and to favor the lower values rather too much in the neighborhood of $L = 1.4$.

6.3 *Evaluation of Fit at Cutoff*

The cutoff may be thought of as the position of the outer envelope of the nonzero counting rate, or the inner envelope of the zero counting rate. Thus, in practice the location of the cutoff is associated with the sensitivity of the detector, rather than with the absence of particles. For $L \geqq 2$, there is a wide range of x over which there are many instances of either zero or one count occurring during the 11-second counting interval, and as a result the cutoff is not well-defined. This is exemplified in Fig. 6. The overlapping of the region in which no count is

observed with that in which one count is observed shows clearly in Fig. 12. The locations of occurrences of zero counts are plotted in R,λ coordinates in Fig. 12(b) and in x,L coordinates in Fig. 12(d). Figs. 12(a) and (c) show the locations at which one count (one, two, and three counts for $L < 1.5$) was recorded. (The density of points has been reduced at high L to improve the clarity of the display.)

Because the cutoff is increasingly difficult to define from the data as L increases beyond ≈ 2, the position of the cutoff predicted by the fitted model is not a good boundary condition to use in judging the quality of the two-dimensional fit. Instead the locus of positions for which exactly one count per counting interval is predicted is superimposed as the solid lines in Figs. 12(a) and (c) upon the array of points giving the band of positions at which one count per counting interval was observed. The data are represented quite satisfactorily by the solid lines particularly in the region ($L \leqq 1.90$) where the belt ends abruptly. The fit is least satisfactory near $L = 2$ ($\lambda = 40°$). Adding the terms $r_4(L - L_0)^4$ and $r_5(L - L_0)^5$ to the expansion for $R_c(L)$ in (5) does not appreciably improve the fit near $\lambda = 40°$.

The line $x_c(L)$, representing the cutoff itself, is plotted as the dashed line in Fig. 12 and is seen to be a reasonable outer envelope for the nonzero counts.

The present estimate of $x_c(L)$ is also shown as the dashed line in Fig. 9. Below $L \approx 1.8$, the estimates of x_c from the individual L-slices are in good agreement with estimates from the two-dimensional fit. However, above $L \approx 1.8$ the L-slices give erratic values for x_c. As demonstrated in Fig. 7, the L-slice estimates may be biased toward high values, a circumstance which makes it difficult to extract a satisfactory fit for $x_c(L)$ from the estimates of x_c produced by fitting the L-slices.

6.4 *Behavior of $S(L)$*

The values of the function $S(L)$ generated by the two-dimensional fit cannot be subjected to a simple boundary comparison with the data. The function $S(L)$ is plotted as the dashed line in Fig. 10 along with the L-slice estimates. It will be seen that the L-slice estimates tend to be somewhat higher than the values given by $S(L)$ in the neighborhoods of $L = 1.3$ and $L = 1.9$. However, if the form of $S(L)$ is taken to provide a better fit to the points in Fig. 10, then the resulting two-dimensional fit yields a physically less satisfactory fit of the cutoff function $x_c(L)$ to the boundary data without substantial improve-

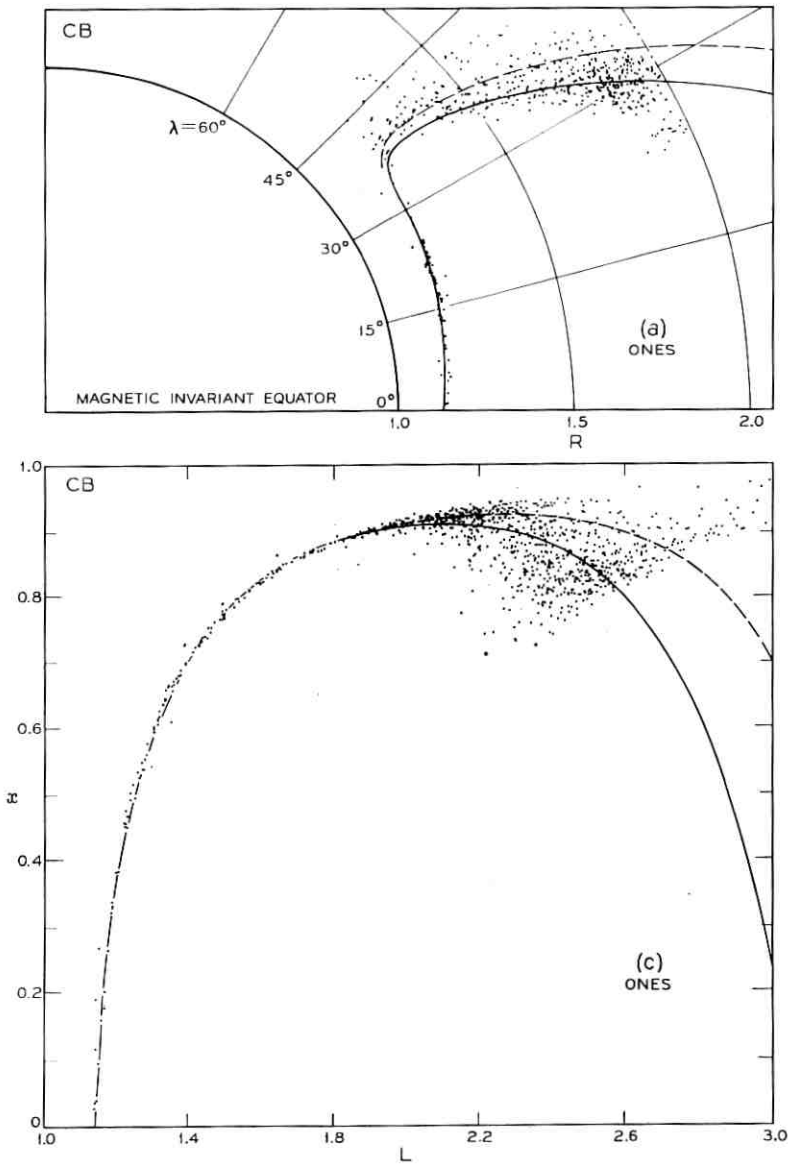


Fig. 12 — All positions in R, λ space (a) and x, L space (c) at which one count (one, two, and three counts for $L < 1.5$) was observed in an 11-second counting interval, and all positions in R, λ (b) and x, L space (d) at which zero counts were observed in an 11-second counting interval. The solid lines are the loci of positions at which the CB coefficients estimate one count in 11 seconds. The dashed lines are the loci of the cutoff function $x_c(L)$ or $R_c(L)$ calculated from the CB coefficients. The trace $R = 2.0\ R_e$, which explains the absence of data

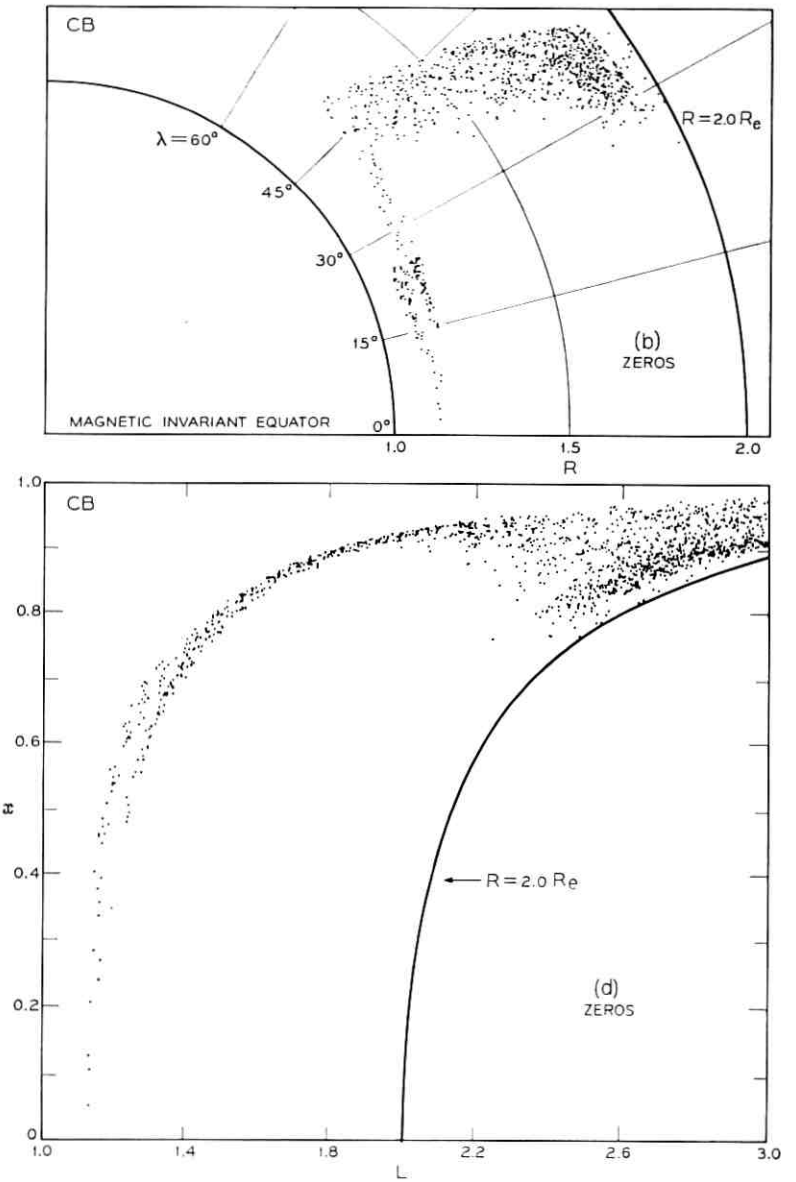


Fig. 12 — (continued)

in the lower right-hand corner of the x,L plots, appears in part (d). The cluster of points near $R = 1.1$ and $\lambda = 20°$ in part (b) of the figure is data acquired by the telemetry station at Woomera, Australia. It represents observations made near perigee when the satellite was below the bottom edge of the proton belt, which is high over the western Pacific Ocean.

ment in the overall fit (see also Section 4.7). Admittedly, this judgment is subjective because it is made in regard to regions where the cutoff is poorly defined by the data because of the insufficent sensitivity of the detector. The high values of S near $L = 1.9$ appear to arise from the correlation problem discussed in Section V in connection with Fig. 6 and Table II.

6.5 *Behavior of the Fit on Several L Slices*

The dashed lines in Figs. 4 to 7 are the values predicted by the CB coefficients superimposed on the L-slice data along with the previously derived L-slice fit. In Figs. 4 and 5, the difference between the thin solid and the dashed lines is insignificant, and this is generally the case for $L < 1.95$. At $L_m = 1.79$, the predictions from the CB coefficients differ importantly from the fit to the L-slice only for x values at which there are no data.

For $L_m = 2.2$, however, the two predictions are noticeably different as may be seen in Fig. 6. The fit to the L-slice gives the estimate $x_c = 0.954$ (see Table II); the two-dimensional fit yields $x_c = 0.928$; and the difference exceeds two standard deviations. The question as to which of the two lines is a better representation of the data in this L-slice in the physical sense, rather than in the least squares sense applied to these points by themselves, is connected with criteria which will be discussed in the following sections. The basic fact is that the two-dimensional fit provides a mechanism by which the data on every L-slice can influence the fit on every other L-slice and thereby provides a fit that is more satisfactory overall than the collection of individual L-slice fits.

6.6 *Residuals in x,L Space*

The data were also examined for dependencies on x and L over and above those provided for by the fitted mathematical model. This is accomplished by studying the residuals, i.e., $(Y - y)$, for all the nearly 80,000 observations. The residuals provide a very sensitive basis for judging the quality of the fit. The removal of the principal dependence on x and L by subtracting the fitted function from the observations has the effect of allowing small systematic differences to be prominently displayed.

Fig. 13 shows a 3100-point sample of the residuals, $Y - y$, plotted against L, where, to keep the density of the points reasonable, only one point has been plotted from each of the nearly 3100 contiguous

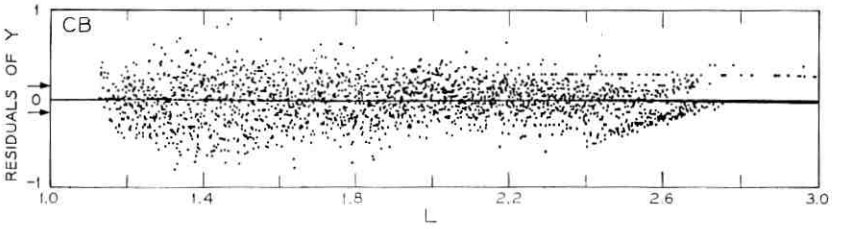


Fig. 13—CB residuals of Y (i.e., $Y - y$ calculated from the CB coefficients) plotted against L. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed. No more than one point is plotted for an L increment of 0.0006.

L-intervals, of width $\Delta L = 0.0006$, between $L = 1.15$ and $L = 3$. Ideally, the residuals should scatter randomly about 0, without any perceivable pattern. For $L < 2.4$ there is only a little indication of a nonrandom trend. However, for $L > 2.4$ there is a distinct pattern. This pattern is associated with the quantization error, which becomes important where the number of counts per counting interval is very small. When $0 < y < \sqrt{1 \text{ count}/11 \text{ sec}}$ and $Y = 0$ or $\sqrt{1 \text{ count}/11 \text{ sec}}$, the result is the tailing upward toward the residual $= 0$ axis that starts at $L \approx 2.4$. When $y = 0$ and $Y = 0$ or $\sqrt{1 \text{ count}/11 \text{ sec}}$, one gets the two-line pattern (0 and $0.0310 = \sqrt{1/11}$) seen clearly in Fig. 13 for $L \gtrsim 2.7$. (The thickening of the zero axis indicates the presence of data points.)

Fig. 14 is a plot of the residuals against x for all points for which $1.4 < L < 1.6$. The residuals in Fig. 14 show no structure; however, their average value is a little less than zero. This dip is confirmed by the points in the range $1.4 < L < 1.6$ in Fig. 13, and means that the value of y is slightly high relative to the data in this region. However, the lack of structure in Fig. 14 indicates that the bias is independent of x in this region.

Fig. 15, the plot of the residuals vs x for $1.85 < L < 1.90$, shows the region in which the fit is poorest. The residual points are not symmetrically distributed about zero and the asymmetry seems to depend on x. Notice that the value of y is slightly too large near $x \approx 0.05$ and $x \approx 0.65$. The discussion of these trends is continued below, after some further analysis has been described.

6.7 *Mean Square Residuals in x,L Space*

Another way of gauging the quality of fit is to compute the mean square of the residuals (MSR) separately for various regions of x,L space. Trends in these quantities may indicate regional varia-

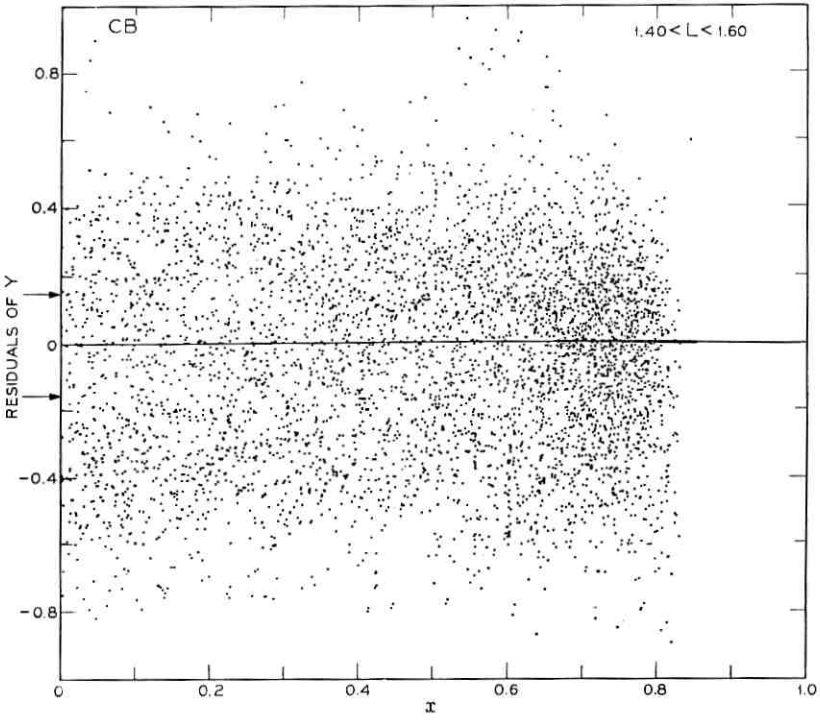


Fig. 14 — CB residuals of Y (i.e., $Y - y$ calculated from the CB coefficients) plotted against x for $1.40 < L < 1.60$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

tions in the adequacy of fit. The data and residuals were divided into three groups. Group I contains all the "good" data points "within" the boundaries of the > 50 MeV proton belt. These points are defined as those not included in Groups II and III. Group II consists of the "good" data points "outside" the boundaries of the belt. These are points which meet two criteria: they have values of (x, L) for which x is greater than $x_c(L) + 0.001$, and they are not in Group III. Group III comprises the outliers or "bad" data points, defined as those points whose residuals are greater than three times the overall root mean square residual of the points in all *three* groups together.* The most probable origin of a point in Group III is a telemetry error.

If the number of counts in a counting interval behaves like a

* Note that only 0.5 percent of the data fall in Group III.

Poisson random variable, then the variance of Y^2 would be equal to the average value of Y^2. As noted in Appendix B, when Y is not near zero, the variance of Y would then approximately equal 0.023, independent of the average value of Y. This value then might approximately represent the average value of the mean square residual, MSR, on the scale of Y. Thus, the number 0.023 provides a baseline for the comparisons discussed below.

Table III lists the mean square residuals (MSR) by L range and by Group. For Group II, Y is frequently zero and, as $x > x_c$ implies $y \equiv 0$, one finds that the residual is zero very often. Of course, under the Poisson assumption the variance of Y when its average value is 0 or very close to 0 will be less than 0.023 (see Appendix B.2) and the appearance of MSR values smaller than 0.023 in Group II is thus not surprising. A similar circumstance exists in Group I for $L > 2.6$.

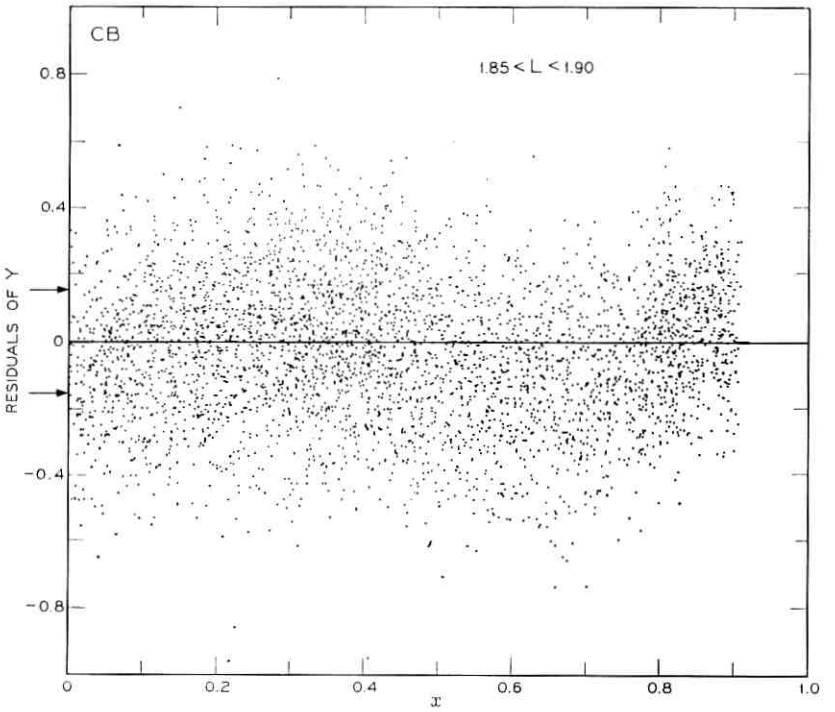


Fig. 15 — CB residuals of Y (i.e., $Y - y$ calculated from the CB coefficients) plotted against x for $1.85 < L < 1.90$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

TABLE III—MEAN SQUARE RESIDUALS AND MEAN SQUARE ERRORS.

L-Range		All data. CB coefficients					
		Group I		Group II		Group III	
L_{MIN}	L_{MAX}	No. of points	MSR	No. of points	MSR	No. of points	MSR
1.1	1.2	148	0.039	31	0.009	0	0.
1.2	1.3	1147	0.053	68	0.019	9	4.171
1.3	1.4	1608	0.106	99	0.027	20	2.265
1.4	1.5	1939	0.106	120	0.010	19	6.743
1.5	1.6	2974	0.083	101	0.004	22	9.079
1.6	1.7	3835	0.056	104	0.004	26	4.617
1.7	1.8	5233	0.055	87	0.001	29	4.356
1.8	1.9	8487	0.054	92	0.001	54	4.110
1.9	2.0	8880	0.041	98	0.011	55	4.280
2.0	2.1	6261	0.043	106	0.033	24	8.031
2.1	2.2	5354	0.032	183	0.049	18	6.509
2.2	2.3	4717	0.030	313	0.047	16	6.982
2.3	2.4	4040	0.034	477	0.030	21	9.478
2.4	2.5	3769	0.044	716	0.021	22	8.296
2.5	2.6	2987	0.038	1000	0 014	15	6.462
2.6	2.7	2066	0.023	1696	0.010	15	17.098
2.7	2.8	225	0.011	3104	0.007	24	13.343
2.8	2.9	0	0.0	2784	0.006	11	14.545
2.9	3.0	0	0.0	2394	0.005	6	15.908
1.1	3.0	63670	0.048	13573	0.011	406	7.011
1.1	3.0	925	0.045	MSE of CB Sample (Group I + Group II)			
1.1	2.0	975	0.065	MSR of equatorial points, $\lambda < 1°$ (Group I)			

L-Range		HTB data. Model I coefficients (see Table IV)					
		Group I		Group II		Group III	
L_{MIN}	L_{MAX}	No. of points	MSR	No. points of	MSR	No. of points	MSR
1.1	1.2	111	0.037	28	0.010	0	0.
1.2	1.3	650	0.045	49	0.028	8	4.435
1.3	1.4	633	0.059	78	0.043	6	1.253
1.4	1.5	693	0.050	56	0.0	7	5.892
1.5	1.6	926	0.039	43	0.019	1	0.816
1.6	1.7	1342	0.036	38	0.002	6	5.472
1.7	1.8	2161	0.037	39	0.005	8	5.474
1.8	1.9	4708	0.037	30	0.003	38	4.981
1.9	2.0	5585	0.046	28	0.013	40	4.184
2.0	2.1	3728	0.049	38	0.021	16	9.037
2.1	2.2	3258	0.033	38	0.036	10	9.716
2.2	2.3	2857	0.030	80	0.034	10	6.581
2.3	2.4	2335	0.032	135	0.027	14	9.094
2.4	2.5	2193	0.043	212	0.020	11	8.922
2.5	2.6	1831	0.041	278	0.011	11	6.638
2.6	2.7	1520	0.027	464	0.007	9	17.120
2.7	2.8	1083	0.014	765	0.007	18	15.845
2.8	2.9	146	0.009	1433	0.007	9	16.145
2.9	3.0	0	0.0	1317	0.005	4	20.590
1.1	3.0	35760	0.038	5149	0.009	226	7.926
1.1	3.0	960	0.036	MSE of 960-point HTB Sample (Group I + Group II)			
1.1	2.0	429	0.035	MSR of equatorial points, $\lambda < 1°$ (Group I)			

Poisson approximation: variance $\approx$ 0.023. Note conditions in text and Appendix B.2.

For the overall fit, the MSR of Group I (L range from 1.1 to 3.0) is only twice 0.023. However, for $1.3 < L < 1.6$ the Group-I MSR is four times 0.023. This L range is associated with the large scatter in the equatorial data plotted in Fig. 11, and Fig. 14 shows that this scatter is independent of x, rather than just an equatorial phenomenon. This issue is pursued further below.

6.8 *Dependence of Residuals on Other Variables*

Studies were made of the possible dependence of the residuals on observed variables other than x and L. Indeed, it will appear that some of the excess scatter exhibited in Table III and in Figs. 11 and 14 is associated with instrumental effects.

The regularities inherent in the orbit and orientation of a satellite, the motion of the earth, and the location and operation of the telemetry receiving stations lead to systematic interrelations among the various coordinates listed in Table I. A simple example concerns temperature. The satellite cools when its enters the earth's shadow. This eclipse occurs only on the night side of the earth. Thus, if the detector is temperature sensitive, one would see a false day-night effect in the counting rate. If, because of additional dependencies, data are available during eclipse for only a limited span of days, a false secular effect might also be observed. Because of the implications of the preceding discussion, a careful study was made of the behavior of the residuals with respect to a large number of coordinates, and attention was given to the details of the relationships among the coordinates during the search for contributors to the inflation of the MSR.

We present below the evidence that has led us to the conclusion that two instrumental effects, variations in bias voltage and changes in temperature of the detector, are principal causes of inflation of the MSR.

There was no temperature sensor on the particle detector. The instrument is not exposed to sunlight and is relatively well-insulated thermally from the skin and frame of the satellite. Consequently, temperature measurements of the skin are not closely related to the temperature of the detector. However, a good indicator of detector temperature is elapsed time since entering or since leaving eclipse. Fig. 16 gives plots of the residuals, $Y - y$, against time in minutes measured from the more recent of the two events, entered shadow or entered sunlight. Residuals associated with periods during which the

satellite did not enter eclipse once per orbit are segregated at the far right-hand side of the plots, labeled A on the abscissa.

Figs. 16(a) and (b) are for $1.4 < L < 1.6$. The points in Fig. 16(a) are those for which the bias voltage was between 95.3 and 97.5 volts, while Fig. 16(b) contains those for bias voltages between 92.0 and 95.3 volts. The decrease in the residuals (and also in the observed counting rate) after the satellite enters eclipse (and the temperature falls) and the increase after the satellite leaves eclipse (and the temperature rises) may be seen distinctly in both figures. In addition the residuals are noticeably more negative for the low (92.0 to 95.3 V) bias range. Both low bias voltage and low temperature are known to decrease the efficiency of the detector and one expects an appreciable effect to be introduced into the counting-rate data. In the present case the scatter is about ± 15 percent of the counting rate. A consequence of this is the excess scatter that has been noted particularly with reference to Fig. 11 and Table III.

Figs. 16(c) and (d) are analogous to Figs. 16(a) and (b), but the residuals are for the L range 1.85 to 1.90. Again, the systematic influence of low temperatures and low bias voltages is unmistakable.

6.9 *Partitioning the Data*

Two ways of responding to these instrumental effects might be: (*i*) to try to correct the data, or (*ii*) to disregard the affected data. It is not possible to make a correction to the counting rate that is properly independent of the experimental results because; (*i*) the bias voltage was measured in steps of 1.11 V, which is not sufficiently fine-grained; (*ii*) it would be necessary to estimate the temperature of the instrument using a complicated hypothetical relationship between the instrumental temperature, skin temperature, and time after entering eclipse (or sunlight); and (*iii*) we have an insufficient knowledge of the temperature and bias-voltage sensitivity of the detector.

Though an ad hoc correction based on the observed counting rates could have been attempted, it was decided for practical reasons to eliminate both the low-temperature and low-bias points and use only that data which was gathered under the following conditions:

(*i*) The satellite had been in sunlight for the previous 50 minutes, and thus had attained temperature equilibrium reasonably well (see Fig. 16).
(*ii*) The bias voltage was between 95.3 and 97.5 volts.

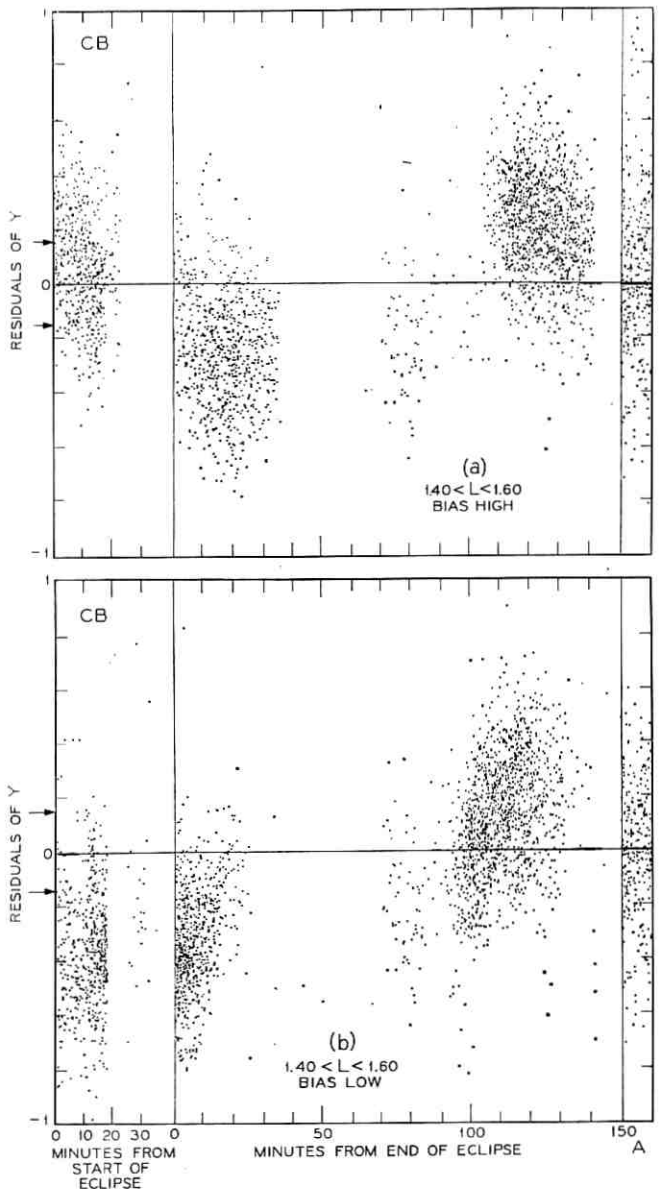


Fig. 16 — CB residuals of Y (i.e., $Y - y$ calculated from the CB coefficients) plotted against time in minutes from the most recent of the two events, entered eclipse or entered sunlight. Data taken on days during which no eclipse occurred are plotted

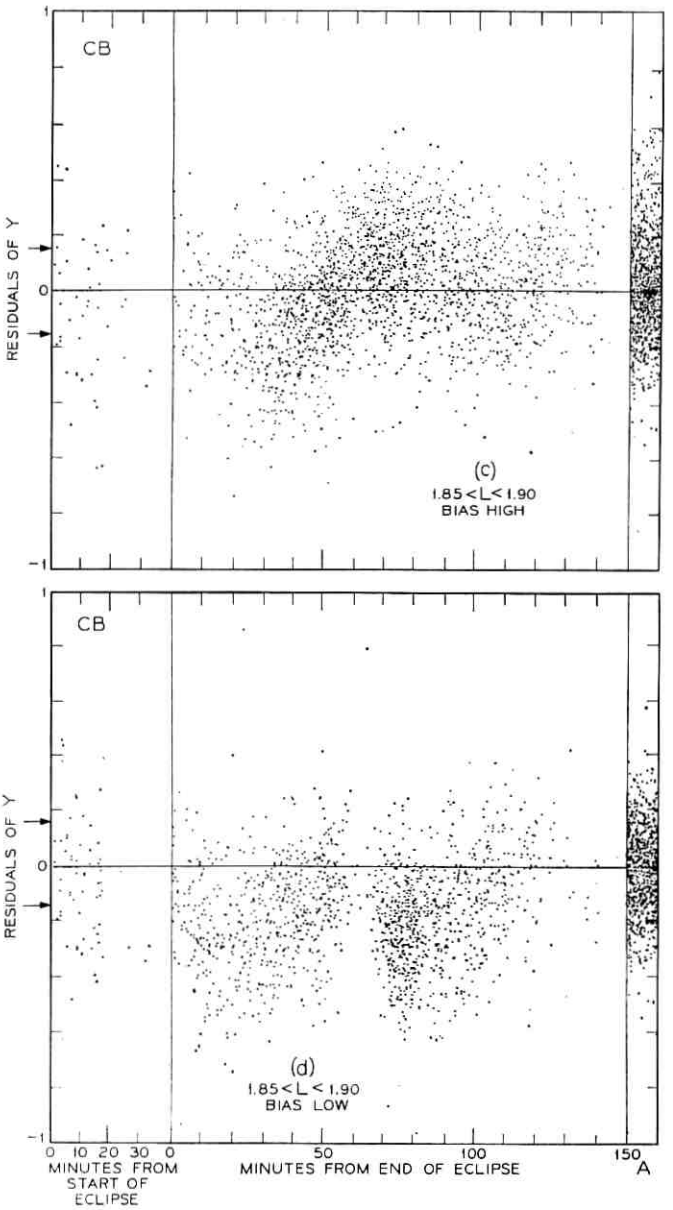


Fig. 16 — (continued)
within the region marked "*A*" at arbitrary values of the abscissa. The arrows indicate ± the approximate standard deviation if Y^2 were Poisson distributed.

This selection yields a homogeneous body of 41,135 points, henceforth referred to as high temperature-high bias (HTB) observations. The remaining 36,500 points, which represent a mixture of temperature and bias conditions, were used only occasionally in further analyses. This selection process coincidentally produces one unfortunate associated circumstance, namely, the exclusion, as low-bias data, of all measurements made between days 325 and 373.

Further analysis and model fitting and development based on, and directed towards, this HTB data is detailed in the following sections and Appendix C.

VII. THE TWO-DIMENSIONAL FIT FOR THE SELECTED (HTB) DATA

7.1 *Sample Selection*

The distribution of the HTB data in magnetic space is indicated in Fig. 17, which gives the R,λ coordinates of every tenth point from the 41,135 L-ordered HTB observations. The data provide reasonably adequate, though uneven, coverage. As a practical requirement for the fitting procedure, a "representative" sample of about 1000 observations must be selected.

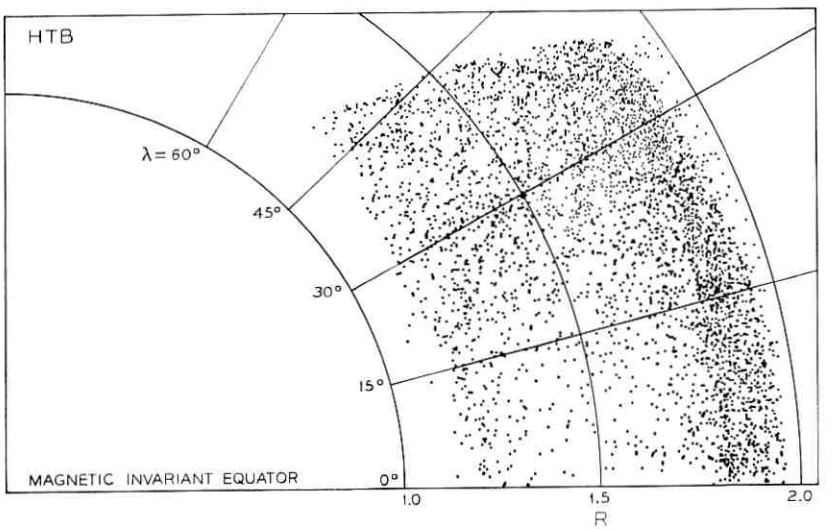


Fig. 17 — The spatial distribution of the HTB data for $L < 3$ in R,λ coordinates. Every tenth point from the L-ordered data is plotted.

It is intuitively clear from preliminary knowledge of the radiation distribution that some sample configurations will be far more effective than others in defining the functional form of the proton flux.

The sample selection is important because: (*i*) nothing more than a sophisticated smoothing function is being fitted and we want this function to be broadly applicable over the entire space; (*ii*) an optimum fit in one region of space does not necessarily imply a good fit elsewhere; (*iii*) the spatial distribution of data points depends on the satellite orbit and the position of the telemetry stations; (*iv*) even with the square root transformation, there remains some differential variance among the data.

These considerations argue against using a simple random sample or even a random sample in x with a systematic sample in L such as in the CB fit. Indeed, they also argue against fitting *all* (unweighted) HTB data, even if this were practical. Alternatively, points might be chosen on the basis of a simple geometric grid in magnetic space. Such a procedure would be easy to use, but it is arbitrary with respect to the radiation belts.

Sampling procedures might be based on particular physical features of the radiation belts to emphasize the goodness of fit, for example, where the flux is high or where diffusion across L lines might be important. However, such fits would be too biased for our present general objective.

One is thus led to a sampling process based on properties of the radiation belt itself, as described for example by the preliminary CB fit. In particular, a high density of data points is desirable in regions where the value of y is changing rapidly, while a low density will suffice where the function is changing slowly. A realization of this criterion would be to define about 1000 x,L cells, within each of which the range of y from the preliminary fit would be the same. However, there are appreciable practical difficulties in defining the boundaries of such cells.

Thus, the following hybrid procedure was used to define the 960-point HTB sample on which the subsequent fitting was done: The L-range from 1 to 3 was divided into about 120 L-slices of equal ($\approx$ 0.017) width in L. Each L-slice was then divided into eleven x,L cells using a scheme that depends on the preliminary fit. The first ten cells were chosen so that within each cell the range of y predicted by the CB model is closely 1/10 of the equatorial value of y at the center of the L-slice. The eleventh cell lies beyond x_c. The

method of partitioning in the x direction is illustrated by the partition of the L-slice in Fig. 5(a) into five x-regions by the horizontal lines. (The distance d is added to x_c to define the lower-x boundary of the last cell.)

To take some account of differential variances remaining after the square root transformation, the following procedure was employed: The mean square deviation from the mean (MSD*) was calculated for all the HTB data in each x,L cell defined above; thence, after visual inspection of the results (see Appendix C), three groupings of contiguous x,L cells were made according to whether the MSD's were generally below 0.013, between 0.013 and 0.020, or above 0.020; the corresponding regions were then given relative weights of 2, $1\frac{1}{2}$, and 1, respectively. The weight 1 implies that one point was sampled from the cell.

These weights were assigned on the basis of a judgment which considered: (*i*) the desire to increase the weight of low variance (i.e., near-zero counting rate) observations and thus to aid the definition of the cutoff; and (*ii*) the desire to keep from "wasting" sample points in the region $x \gg x_c$ since such data will add little to the specification of $x_c(L)$ and virtually nothing to the estimation of $A(L)$ and S.

Fig. 18 shows the distribution in x,L space of the 960-point sample which was used. The number 960 came about because a number of the defined cells had no data in them. Our experience with several other samples of the HTB data gives us confidence in both the rationale behind, and the results obtained with, this 960-point set, henceforth referred to as the HTB sample. However, sampling procedures tailored to the requirements of special purpose fits will give better results in some regions of x,L space.

Some additional discussions relevant to sample selection and data usage are given in Section 13.3 and Appendices B.3 and C.2.

7.2 *The HTB Fit*

A slightly constrained version of Model I of Section 4.4 was fitted to the 960-point HTB sample. The results are referred to as the HTB fit. The constraint is $s_1 = 0$, in (3). Most of the values of s_1 obtained in preliminary fits to various samples of the HTB data differed from zero by less than two standard deviations. Also, the points in Fig.

---

* See Table I for definition of MSD, MSR and MSE.

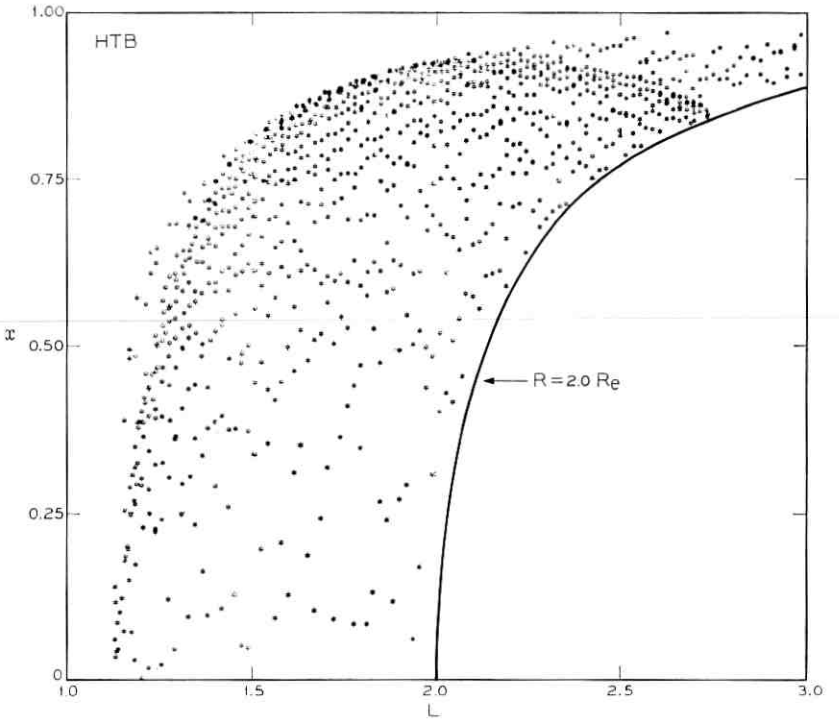


Fig. 18 — The distribution of the 960-point HTB sample in x, L space. The trace $R = 2.0\ R_e$ explains the absence of data in the lower right-hand corner of the figure.

10 do not suggest a linear dependence of S on L.* The effect of this constraint on the value of the fitted cutoff function was examined and found to be unimportant.

The estimated HTB coefficients (obtained by fitting the constrained model to the HTB sample) appear in Table IV. The physical interpretation of L_0 as the lowest L on which > 50 MeV protons were measurable was noted in Section 4.3. The standard error of 0.001 ($\approx$6 km in altitude) is no larger than the uncertainties inherent in the calculation of L itself.

The interpretation of S as a shape factor (see Section 4.2) is straightforward in the present case, i.e., where $s_1 = 0$. The standard error of 0.005 is much smaller than the standard errors of the estimates of S generated from the fits to L-slices (Table II) and is

* Some higher-order models for $S(L)$ were tried but proved unsatisfactory (see also Sections 6.4 and 9.2).

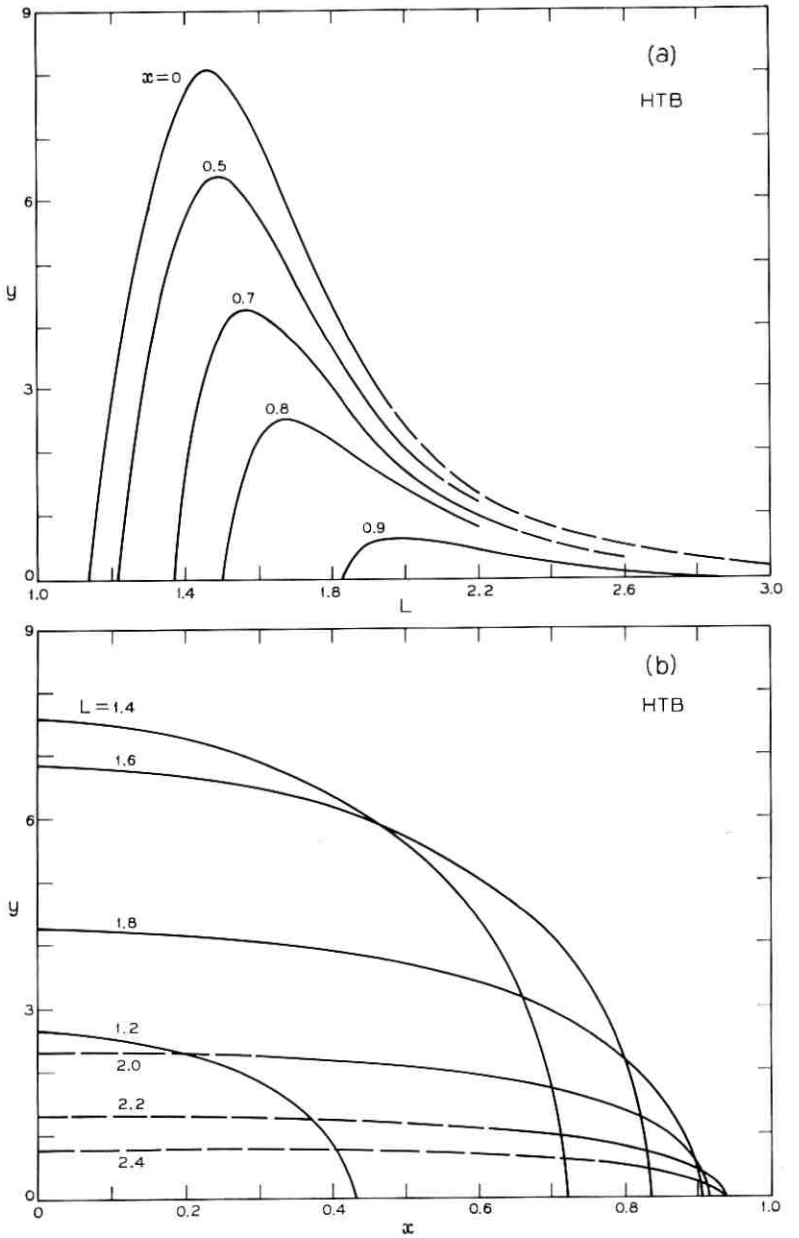


Fig. 19 — Graphical summary of the HTB fit, (a) curves of y' vs L for constant x, (b) curves of y' vs x for constant L, (c) contours of constant y' in x,L space.

also small compared to the scatter in Fig. 10. This implies that a substantial fraction of the scatter may be associated with the high correlation between S and x_c on the L-slice fits. Further consideration of standard errors and correlations of the fitted coefficients and detailed statistical evaluation of the fit is deferred to Section VIII.

Fig. 19 presents a graphical summary of the function $y'(x, L)$. Part (a) of the figure shows y' vs L for (several) constant x. Physically, these curves correspond to values of the intensity of radiation vs L for constant magnetic dipole latitude, because x = constant implies λ = constant. The nesting of the curves in Fig. 19(a) is a consequence of the fact that $G'(x; x_c, S)$ decreases monotonically with x [see (2) and Fig. 19(b)]. The shape of the curves changes smoothly with L, and the position of the maximum shifts smoothly toward higher L as the value of x (and therefore λ) increases.

The nesting property does not hold for plots of y' vs x at constant L. This general consequence of the existence of a maximum in $A'(L)$ is displayed in Fig. 19(b). All the curves in Fig. 19(b) have similar dependences on x.

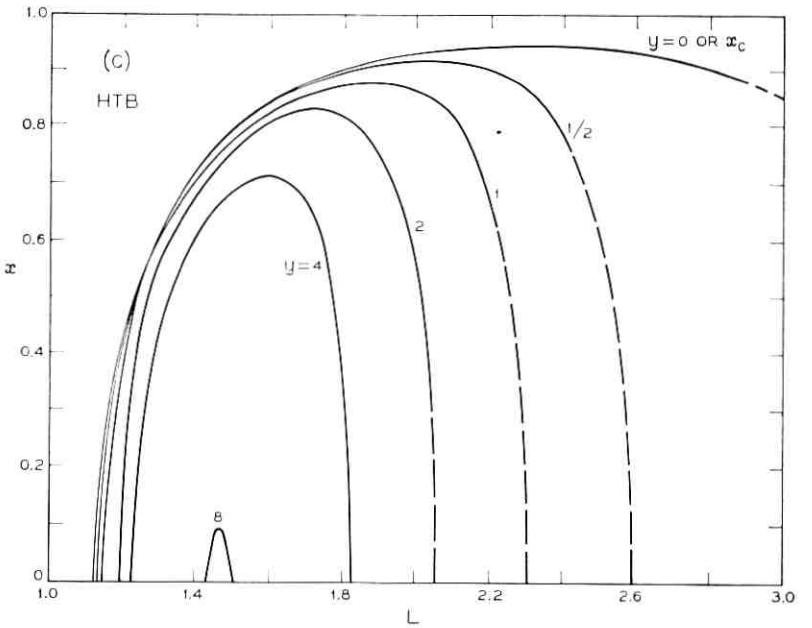


Fig. 19 — (continued)

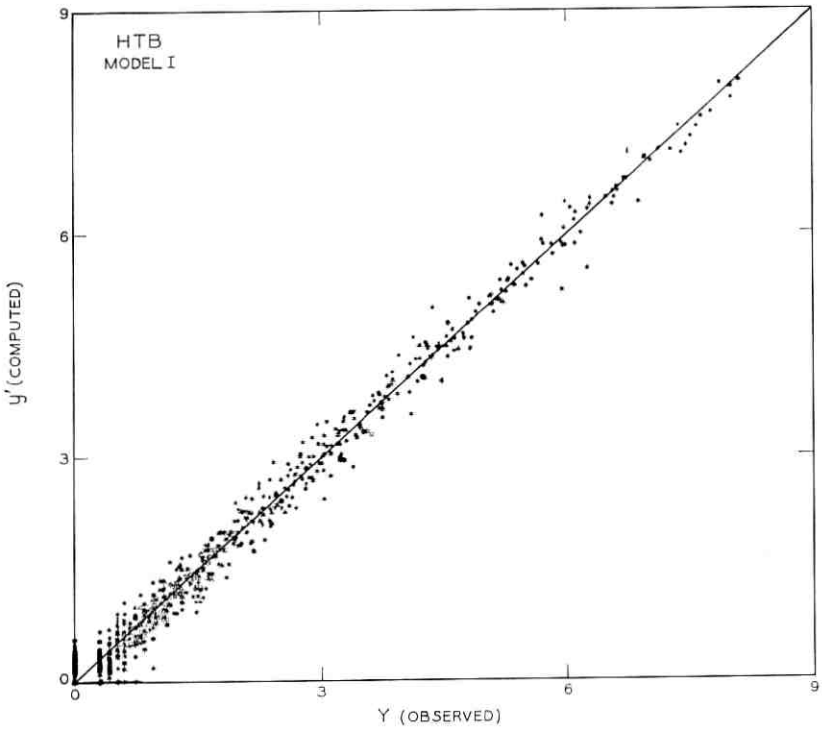


Fig. 20 — The value of y' computed from the HTB coefficients of Model I vs the observed value, Y, for the 960-point HTB sample.

Fig. 19(c) contains contours of constant y' plotted in x,L space and completes the graphical summary. The contours surround the point $x = 0$, $L = 1.46$ at which the peak intensity occurs.

7.3 *Evaluation of Fit to the HTB Sample*

A summary indication of the quality of the fit of the 9-coefficient Model I to the HTB sample is given in Fig. 20, in which the fitted (computed) value, y', is plotted against the corresponding observed value, Y. The solid straight line would represent the case of a perfect fit. This is impossible on the basis of a model using only x,L coordinates since different Y values were observed for the same x,L pairs. It is seen, however, that the scatter of the plotted points about the line of perfect fit is reasonably uniform and that the horizontal width of the "scatterband" is roughly constant over the entire range of y'.

In the following subsections, the quality of fit to the entire body of HTB data is scrutinized, using many of the procedures used in the previous section to evaluate the CB fit.

7.4 *Evaluation of Fit on Equator*

The HTB fit along the equatorial boundary is displayed in Fig. 21. The points are the values of observed Y plotted against L for all HTB data for which $0 \leqq x < 0.037$ (i.e., $\lambda < 1°$), and the plotted curve is $A'(L)$, defined in (6), using the HTB coefficients of Table IV. Comparing Fig. 21 with Fig. 11, it is seen that most of the excess scatter has been eliminated. The curve in Fig. 21 does not deviate noticeably from the center line of the points (except for $1.5 < L < 1.6$, where the curve is a trifle high and for $L \approx 1.95$, where the curve is a trifle low).

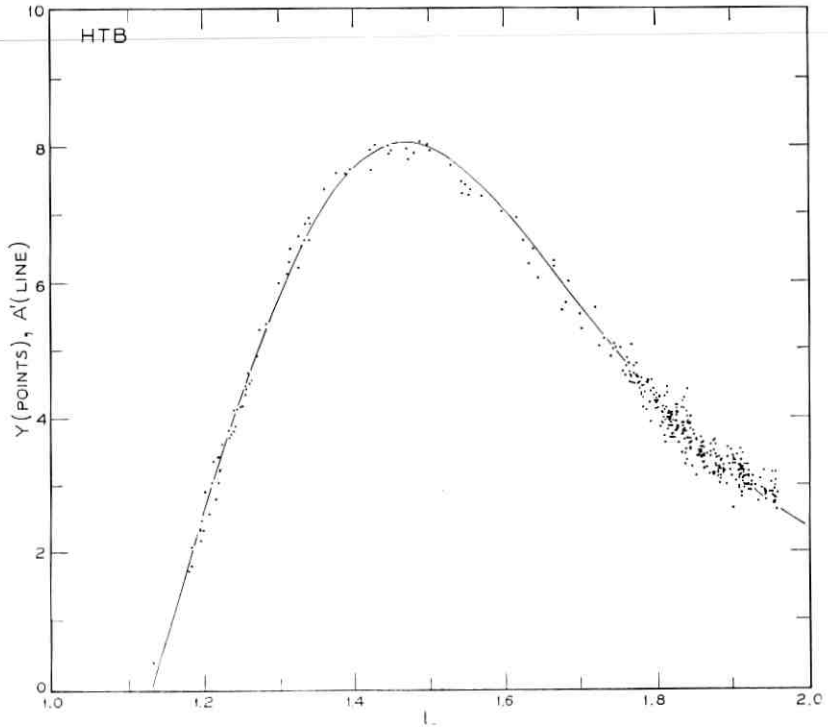


Fig. 21 — All the HTB data for $x < 0.037$ (i.e., within 1° of the magnetic invariant equator) and the equatorial value estimated from the HTB coefficients plotted against L. A' and Y are in units of $(\text{counts/sec})^{1/2}$.

TABLE IV—FIT OF MODEL I TO 960-POINT HTB SAMPLE. (HTB COEFFICIENTS.)

Analysis of variance

Due to	d.f.*	Sum of squares	Mean square
Total	960	5374.7320	
Mean	1	2121.1760	
Corrected total	959	3253.5560	
Model	9	5340.0645	593.3405
Error	951	34.66751	0.03645

Coefficient estimates

	a_1	L_0	a_2	a_3	η	r_1	r_2	r_3	S
Estimate	12.0702	1.1300	0.3006	0.7131	5.6190	0.2600	−0.4937	0.3536	0.3221
Standard error	5.1178	0.0010	0.1205	0.0765	0.3798	0.0090	0.0245	0.0190	0.0048

α values

ρ values	a_1	L_0	a_2	a_3	η	r_1	r_2	r_3	S
a_1		0.13	0.97	−0.98	0.90	0.11	0.09	−0.08	−0.00
L_0	0.4969		0.12	−0.13	0.11	−0.47	0.27	−0.15	−0.00
a_2	0.9995	0.4783		−0.96	0.92	−0.10	0.09	−0.08	−0.00
a_3	−0.9998	−0.4998	−0.9990		−0.90	0.11	−0.09	0.08	0.00
η	0.9948	0.4705	0.9966	−0.9941		−0.10	0.10	−0.09	−0.00
r_1	0.4624	−0.8541	−0.4505	0.4617	−0.4490		−0.64	0.43	0.00
r_2	0.4261	0.6819	0.4199	0.4243	0.4268	−0.9378		−0.72	0.00
r_3	−0.3951	−0.5355	−0.3926	0.3929	−0.4066	0.8219	−0.9589		−0.00
S	−0.0571	−0.0270	−0.0616	0.0657	−0.0680	−0.1166	0.0653	−0.0587	

* degrees of freedom

In Fig. 8 the solid curve, which is $A'(L)$ calculated from the HTB coefficients, may be compared with the dashed curve, which is $A'(L)$ calculated from the CB coefficients. The HTB fit gives higher equatorial values for y' when L is less than ≈ 1.9, as might be expected from the fact, displayed in Figs. 16(a) and (b) and discussed in Section 6.8, that the HTB data select the higher values of Y for $1.4 < L < 1.6$. For L greater than ≈ 1.9, the equatorial values of the HTB fit are somewhat lower than those of the CB fit; however, there is no equatorial data for $L > 1.95$, and the comparison of the fits is not meaningful in this region. The points in Fig. 8 are estimates based on CB, *not* HTB, data and are *not* immediately pertinent to the solid curve.

An estimate of the standard error of the fitted equatorial function $A'(L)$, based on the HTB sample, is plotted as a function of L in Fig. 22(a) (see Section VIII for details). The standard error of $A'(L)$ is typically less than one percent in the range of L ($1.15 < L < 1.95$) over which equatorial data are available. Error bars of this size would hardly be visible in Fig. 21. For the same values of L, the standard errors of $A'(L)$ derived from the HTB fit are substantially smaller than those from the L-slice fits listed in Table II. As might be anticipated, the percent standard error of $A'(L)$ increases as the minimum x values of available data increases with increasing L beyond $L = 2$. This increase to a value of 10 percent at $L = 3$ reflects increasing uncertainty in the extrapolation of the fit. Note that the curves in Fig. 8, which represent the equatorial values of CB and HTB fits, differ, in general, by substantially more than two standard errors and the difference is certainly "statistically significant."

7.5 *Evaluation of Fit at Cutoff*

Figs. 23(b) and (d) show the positions, in x,L and R,λ coordinates, at which zero counts were observed during an 11-second counting interval. Figs. 23(a) and (c) are corresponding plots for one count (one, two, or three counts for $L < 1.5$) per counting interval. Only HTB data are plotted, and the density of points at high L has been reduced to improve the clarity of the display.

Judgments regarding the quality of the fit are made, once again, with reference to the well-defined band of one count, rather than in terms of the more nebulous cutoff. The solid lines in Figs. 23(a) and (c) are the loci of $y'(x, L) = \sqrt{1 \text{ count}/11 \text{ sec}}$, using the HTB coefficients

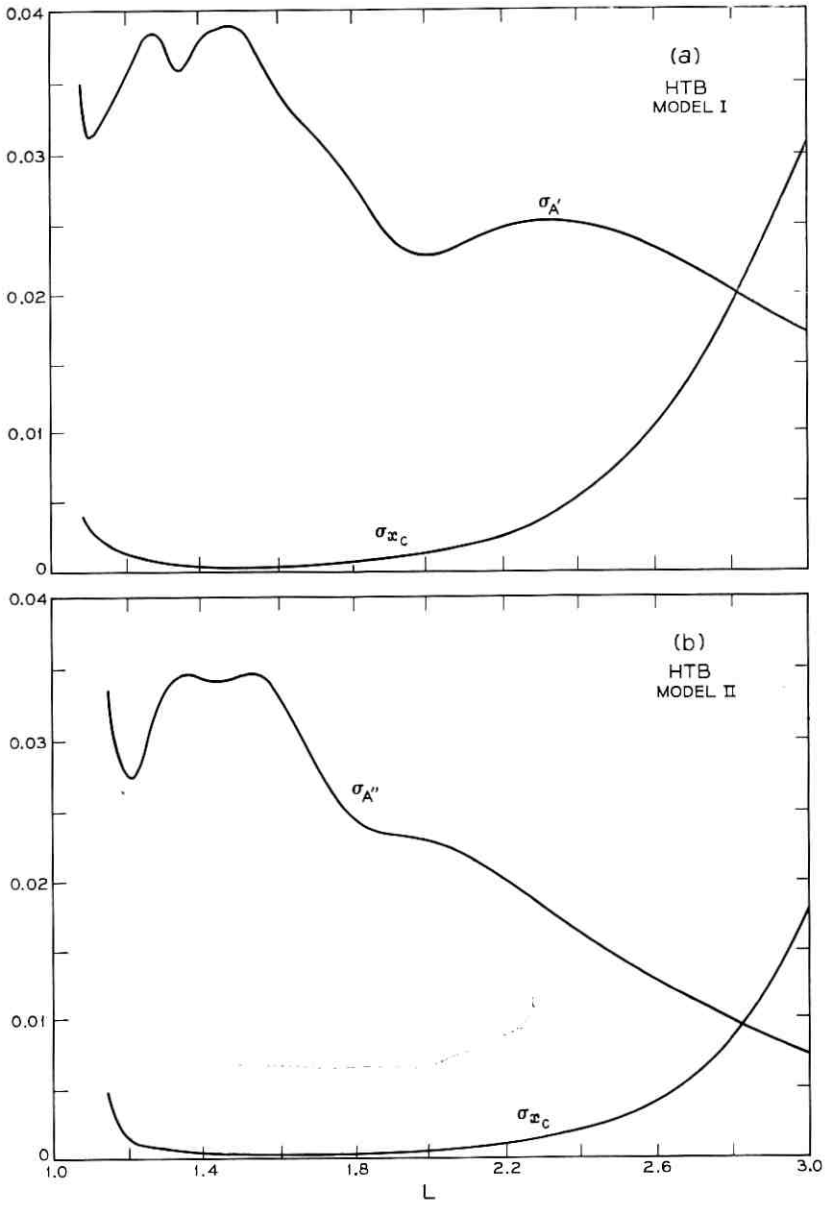


Fig. 22 — The standard deviation of A, σ_A, and the standard deviation of x_c, σ_{x_c}, as functions of L. Units of σ_A and σ_{x_c} are the same as the units of A and x_c, respectively. (a) Model I. (b) Model II.

in the model. These lines represent the data well. Although the fit appears uniformly good in the x,L representation, a slight weakness near the "corner" at $\lambda \approx 40°$ is displayed sensitively in the R,λ plot (see also Fig. 12).

The dashed lines in Figs. 23(a) and (c) show the locus of the fitted cutoff function, $x_c(L)$, calculated from the HTB coefficients. Error bars indicating excursions of one standard error in $x_c(L)$ are shown at two places on Figs. 23(a) and (c). The standard deviation of $x_c(L)$ as a function of L has been estimated (see Section VIII), and is plotted in Fig. 22(a). This standard error is smaller than those produced by the L-slice fits at corresponding values of L (see Table II).

The values of $x_c(L)$ for the HTB and CB coefficients are plotted in Fig. 9. Although there is no discernible difference between the two curves in the figure for $L < 2$, the difference between the tabulated values exceeds twice the standard error (which is very small) over much of the range of L. The two sets of coefficients thus lead to results which differ in a "statistically significant" manner. For L less than ≈ 2, the significance of the standard error is more readily understood when it is interpreted in terms of the altitude of the cutoff. This is done in Section XI.

Beyond $L \approx 2$, the values of x_c for the CB and HTB coefficients diverge noticeably, compare Figs. 12(a) and (c) with Figs. 23(a) and (c), respectively. The magnitude of this divergence is quite sensitive to the method used in selecting the samples to be fitted. As has been discussed, the concept of a cutoff is not well defined in the context of these measurements for $L > 2$. The uncertainty is reflected in the rapid rise in the value of the standard error of $x_c(L)$ [see Fig. 22(a)] as L approaches 3. The significance of this rise may be more readily appreciated by referring once more to the error bars associated with $x_c(L)$ in Figs. 23(a) and (c).

The partitioning of the data on the basis of electrical bias and temperature, and the procedure chosen for selecting the sample to the fitted, introduce statistically significant differences between the values of $x_c(L)$ obtained from the HTB and CB fits, as well as the more readily anticipated significant differences in the values of $A'(L)$.

7.6 *Standard Error of Fitted Value*

The standard error for $y'(x, L)$ is relatively constant, ranging between 0.01 and 0.04, except close to $x_c(L)$. It should be understood that this standard error is based on the fit to the HTB sample, and

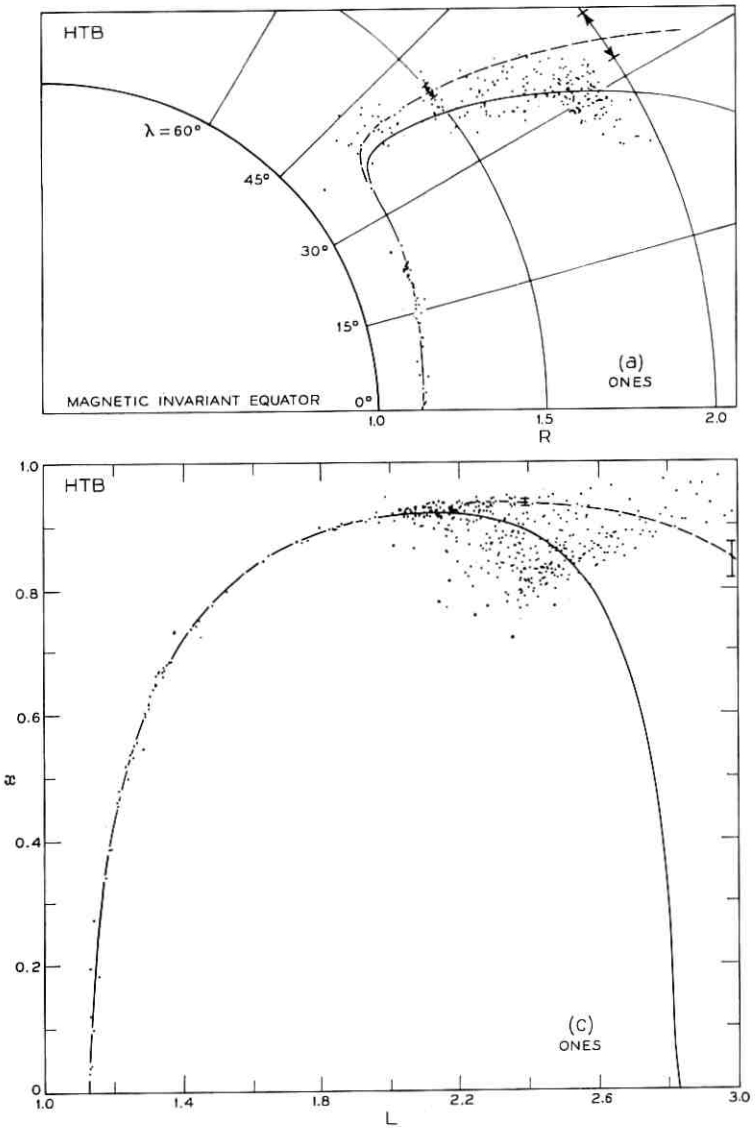


Fig. 23 — All positions for the HTB data in R, λ space (a) and x, L space (c) at which one count (one, two, and three counts for $L < 1.5$) was observed in an 11-second counting interval, and all positions in R, λ space (b) and x, L space (d) at which zero counts were observed in an 11-second counting interval. The solid lines are the loci of positions at which the HTB coefficients estimate one count in 11 seconds. The trace $R = 2.0\ R_e$, which explains the absence of data in the lower right-hand corner of the x, L plots, appears in part (d). The dashed

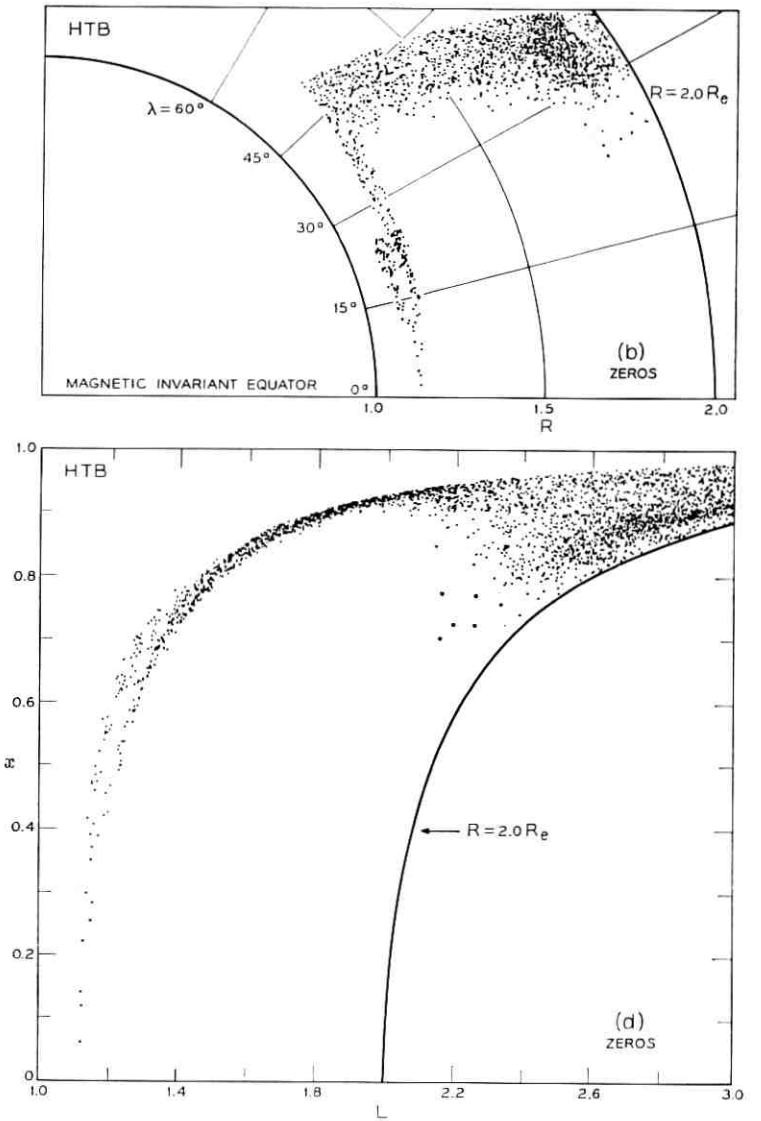


Fig. 23 — (continued)

lines are the loci of the cutoff function $x_c(L)$ or $R_c(L)$ calculated from the HTB coefficients. The cluster of points near $R = 1.1\ R_e$ and $\lambda = 20°$ in part (b) of the figure is data acquired by the telemetry station at Woomera, Australia. They represent observations made near perigee when the satellite was below the bottom edge of the proton belt, which is high over the western Pacific Ocean.

thus applies to the estimate of the *average* value of y and does *not* give the standard deviation of a single predicted observation. The latter would be in the neighborhood of $\sqrt{0.04} = 0.2$ (where 0.04 is approximately the MSE, see Table IV).

Contours of constant percent standard error in the *counting rate*, y^2, are shown by the curves in Fig. 24(a). For $L < 2$ the standard error is less than 2 percent except close to the cutoff, where the value of y^2 is falling fast. (Near the cutoff, the standard error in x_c is more informative.) In the absence of a fitted function, it would be necessary to average between about 30 and 300 observations to achieve a

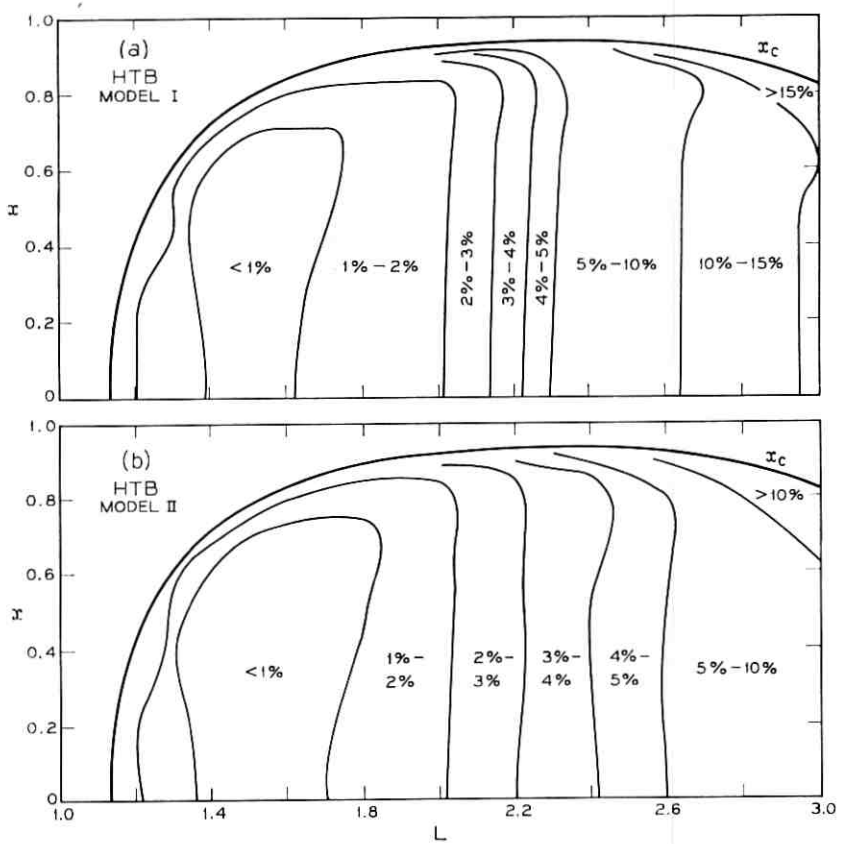


Fig. 24 — Contours of constant percent standard deviation in the *counting rate*, y^2, calculated from the fits to the HTB sample and plotted in x, L space. (a) Model I. (b) Model II.

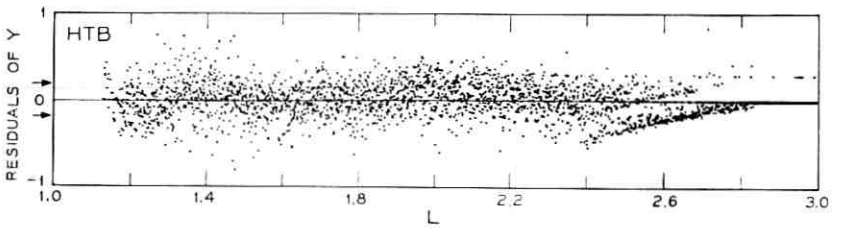


Fig. 25—HTB residuals of Y (i.e., $Y - y$ calculated from the HTB coefficients) plotted against L. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

standard deviation as small as 2 percent. As discussed in Appendix B.4, the estimates of the standard deviation based on the HTB sample are conservative and (if there were no biases in the model) the values that apply to the 40,000 HTB points might be smaller than those in Fig. 24(a) by a factor as large as 6.

The values in Fig. 24(a) are for relative counting rates (or fluxes) and do not include the uncertainty in the absolute calibration of the instrument noted at the end of Appendix A. Other discussion is given in Sections 9.4 and 12.2 and Appendix B.4.

7.7 *Behavior of the Fit on Several L-Slices*

Using the HTB coefficients, values of $y_L(x)$ were calculated for $L_m = 1.35$, 1.805, 2.0215, and 1.79. The results are plotted as the heavy solid lines in Figs. 4 to 7. Recall that the points in these figures are *not* all HTB points. In general, the HTB points are those with the higher values of Y, although this may not be the case at $L \approx 2.2$ because of the temporal effects discussed in Section X. The four figures also allow further appreciation of the difference in results between CB fit and the HTB fit produced by the partitioning of the data and the refinement of the procedure by which the sample was selected.

7.8 *Residuals in x,L Space*

The residuals, $Y - y$, were computed for all the HTB data using the HTB coefficients. Fig. 25 is a plot of residuals against L, and Figs. 26 and 27 are plots of residuals against x, in the indicated L-ranges. These plots are analogous to Figs. 13 to 15, and as they display properties similar to the earlier figures, the discussion of

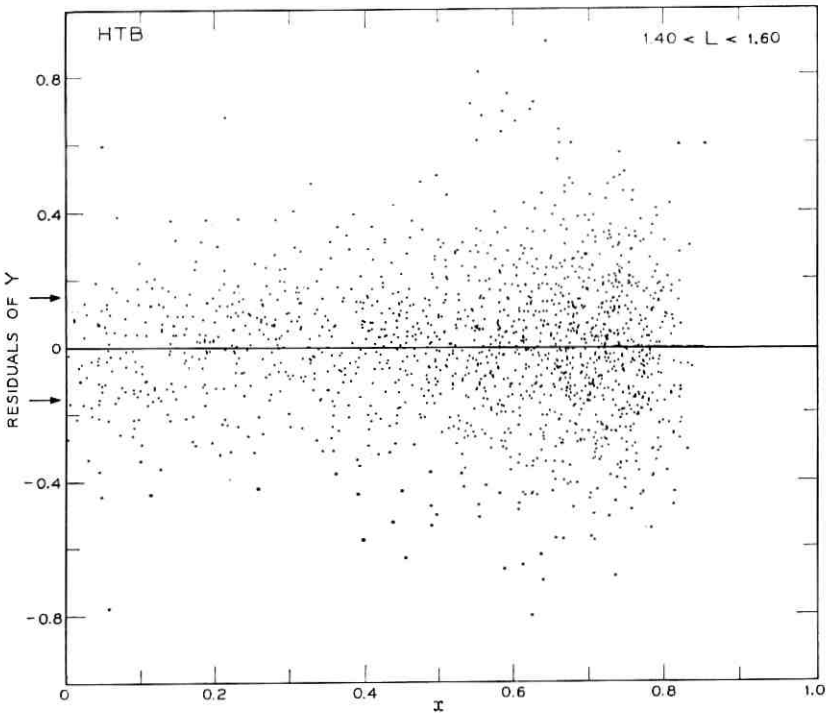


Fig. 26 — HTB residuals of Y (i.e., $Y - y$ calculated from the HTB coefficients) plotted against x for $1.40 < L < 1.60$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

Section 6.6 applies. In particular, there is little indication of a dependence of the residuals on the magnetic coordinates. Moreover, the residuals in Figs. 25 to 27 are more closely clustered about zero than those in Figs. 13 to 15, confirming the fact that there is less scatter in the HTB data. This reduction in the scatter is especially marked in the neighborhood of the peak of the radiation belt (near $x = 0$ between $L = 1.4$ and $L = 1.6$, Fig. 26).

7.9 *Mean Square Residuals in x,L Space*

A breakdown of the mean square residuals (MSR) by L-ranges for the fit to the HTB data is given in Table III. This analysis is analogous to that presented in Section 6.7 for the CB fit. For the Group I data the MSR for the overall fit ($1.1 < L < 3.0$) is about $(1.5)(0.023) = 0.036$ and the largest entry under HTB Group I is

0.059. The anomalous trend of the MSR near $L = 1.4$ evidenced in the fit to the unrestricted data (see Section 6.7) has been largely eliminated. The overall MSR for the Group I data has been reduced by 15 percent.

The breakdown of the MSR by L-ranges is not a particularly refined test of the quality of the fit. This index is based on essentially all the HTB data and, because the averaging procedure is blind to the distribution of data within L-ranges, favors results that fit best where the density of data is high. As the HTB sample was selected using criteria dependent on the preliminary fit to the data and does not necessarily favor x,L regions in which large quantities of data were acquired, the results of fitting this sample does *not* produce the lowest obtainable value of MSR for all of the HTB data. Examina-

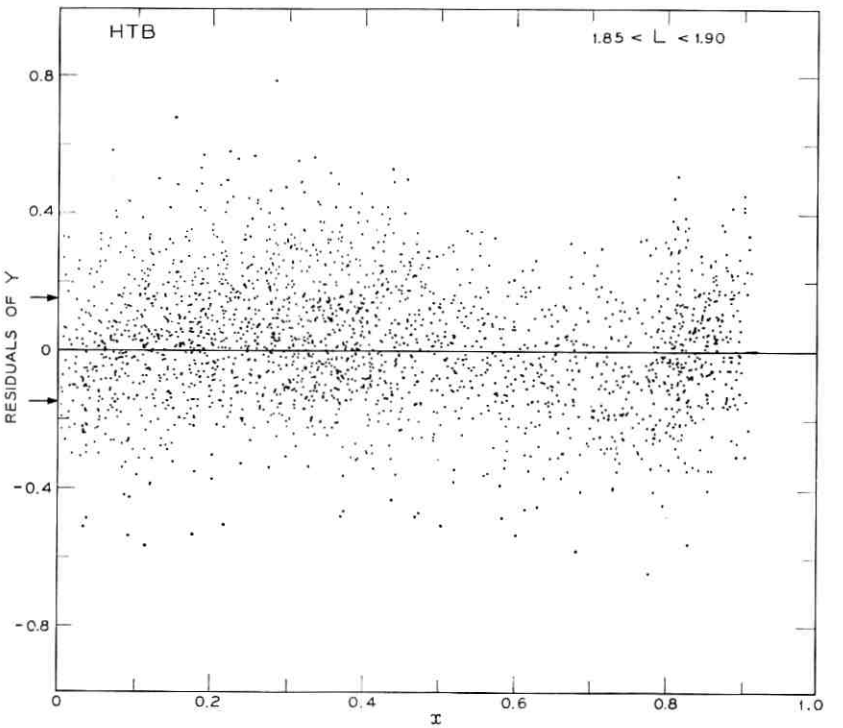


Fig. 27—HTB residuals of Y (i.e., $Y - y$ calculated from the HTB coefficients) plotted against x for $1.85 < L < 1.90$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

tion of the MSR in x,L cells shows the effect of the sample selection procedure on the MSR in L-ranges. Appendix C contains further information and analysis of MSR in x,L cells.

Model I with the HTB coefficients, provides a summary of the HTB data that, in the light of the many sources of variability and measurement errors, reasonably approaches the limit set by expected statistical fluctuations.

7.10 *Sources of Variability in the Data*

The residuals for the HTB data are now examined to see whether further identifiable sources of variability may be associated with them. Possible sources are: instrumental effects, errors in the ephemeris of the satellite, errors in the description of the magnetic field, telemetry errors, fluctuations in the length of the counting interval, deficiencies in the model, and temporal variations. While all these must make some contribution to the MSR, the interrelationships among the coordinates discussed in Section 6.8 and the small size of the individual contributions, make positive identifications very difficult. We have not attempted to examine in detail the large number of small, apparently systematic, deviations discernible on the residual plots, although some of these may be "statistically significant." Instead we have restricted our study to effects which are readily apparent on the residual plots. Where the observations are dense, an effect would be glaringly apparent if it introduced a shift of ≈ 0.05 in the local mean of the residuals. (This corresponds to a change of about 1.2 percent in flux at the peak of the proton intensity, and about 12 percent when the flux is a hundredth of its peak value.)

Instrumental effects are associated with temperature, bias voltage, radiation damage, and imperfections in the omnidirectional characteristics of the detector. Restricting the range of temperature and bias voltage removed the major fraction of the instrumental effects associated with these variables. Directional effects in the detector might show up when the residuals are plotted against γ, the angle between the spin axis and the local magnetic field vector. However, no dependence was observed, indicating that the detector is effectively omnidirectional. Radiation damage, though technically an instrumental effect, is more logically treated with temporal variations.

Examination of plots of residuals versus various geographic coordinates did not reveal any systematic dependencies. In view of the small excess of the MSR over expectation for a random Poisson

process, and the existence of other sources of error, it seems reasonable to conclude that the ephemerides were computed with sufficient accuracy for this analysis.

The plots of residuals against the geographic coordinates as well as against x and L values were used to judge the quality of the coefficients used to calculate the magnetic coordinates L and x. No systematic effects that can be attributed to flaws in the coefficients of the magnetic field were discerned. Nor is there any indication, in the form of excessive scatter of the residuals, that L is an imperfect coordinate in any part of the region of space covered by these data.

Gross telemetry errors and those that occur in conjunction with noise bursts are easily identified and have been discarded. There remain telemetry errors that are indistinguishable from good data on a point-by-point basis, and these erroneous data must make some contribution to the scatter. As noted in Section 8.1, the distribution of the residuals has been looked into and they are found to be very well-behaved. However, it is not possible to make any quantitative estimates of the contribution of the remaining telemetry errors to the MSR.

Temporal variations are an important source of variability, and Section X is devoted to their analysis.

VIII. STATISTICAL CRITIQUE OF MODEL I.

This section presents further information on statistical evaluation of the Model I fit. (Some background concerning relevant statistical techniques is given in Appendix B.) While confirming the very satisfactory performance of Model I in fitting the data, as presented in Section VII, some unsatisfactory aspects are uncovered and several defects of the model are pinpointed. The rectification of these defects is effected by use of Model II, discussed in Section IX.

8.1 *Fit of Model I to the 960-point HTB Sample*

The analysis of variance for the fit of Model I to the 960-point HTB sample is shown in Table IV. This gives various partitionings of the total sum of squares (about 0) of the 960 observations (on the square root of counting rate scale). Table IV indicates the relevance of the model to the data in terms of its statistical effectiveness. Fitting the nine coefficients of the model accounts for more than 99.3 percent of the total sum of squares of the observations, leaving less than 0.7 percent associated with "error" or lack of fit. On a per degree-of-freedom-

basis, the ratio of mean square for "fitted model" with 9 degrees of freedom to mean square for "error" is over 16,000.

Of course, simply fitting the mean of all the data accounts for a sum of squares of 2121.2 of the total of 5374.7. Of the remaining "corrected" total sum of squares about the mean of 3253.6, the part of the model "orthogonal" to the mean accounts for 3218.9, i.e., approximately 98.9 percent (so that the squared multiple correlation coefficient, R^2, is 0.989). The corresponding ratio, mean square for the model with $(9 - 1) = 8$ degrees of freedom to mean square for error, is over 11,000.

It is worth emphasizing that the sample selection process which was used (see Section 7.1) is such that fitting the sample is, on a per observation basis, a more challenging problem than it would be for the entire body of data (see Appendix B.3).

A summary graphical indication of the appropriateness of the fit is given in Fig. 20 which shows the fitted value plotted against the observed value. A perfect fit (essentially impossible here with any model based on x,L coordinates because different integral values of Y are observed near the same x,L point) would be the diagonal straight line shown. Deviations from fit should be gauged as horizontal spread about the line, since the observed quantities are plotted as abscissa, and are seen to be reasonably uniform throughout.

Incisive indication of the quality of fit was provided by various plots of residuals (against L, x, y, time, etc.). Some representative plots over all the HTB data are shown in Figs. 25 to 27 and Figs. 41 to 43.

As a further examination of the adequacy of the fit to the selected HTB data, normal and half-normal probability plots (see Appendix B.8) were prepared for the 745 residuals comprising the subset of the 960-point HTB sample for which $x < x_c(L)$. These plots are shown in Figs. 28 and 29.

Fig. 28 does display a generally good linear configuration indicating that the residuals may reasonably be regarded as a sample from a normal distribution. There is no suggestion of general asymmetry or other distributional peculiarities. There are perhaps three values which are statistically "too large," but not wildly so. Indeed, the plot is remarkably well-behaved and reassuring.

From some points of view, it is useful to consider the statistical behavior of the residuals without regard to their sign. Fig. 29 is a plot of the ordered absolute residuals against standard half-normal (folded standard normal) quantiles. This presentation is more focussed and sensitive to a statistical overabundance of large absolute residuals. The

plot is also very well-behaved, with indication of the same three overly large values.

The reason for omitting from these plots all residuals from points for which $x > x_c(L)$ is that, for those, the predicted value y is 0 and, in the great majority, the observed Y was 0; hence, the residual is 0. Since it was exactly this information which determined the estimate $x_c(L)$ and since one could hardly expect a collection which includes about $\frac{1}{5}$ zeros to behave like a normal sample, these points were omitted

From either Figs. 28 or 29 one can estimate a slope of about 0.21, which is an estimate of the standard deviation of the (counting rate)$^{\frac{1}{2}}$ observations, clear of the confounding influence of the nonvariance-stabilized very low counting rate observations, since observations for $x > x_c(L)$ have been omitted. The corresponding variance estimate, 0.044, clearly exceeds that from the Poisson approximation, 0.023, and also is greater than the pooled value for the MSD(Y), 0.039,

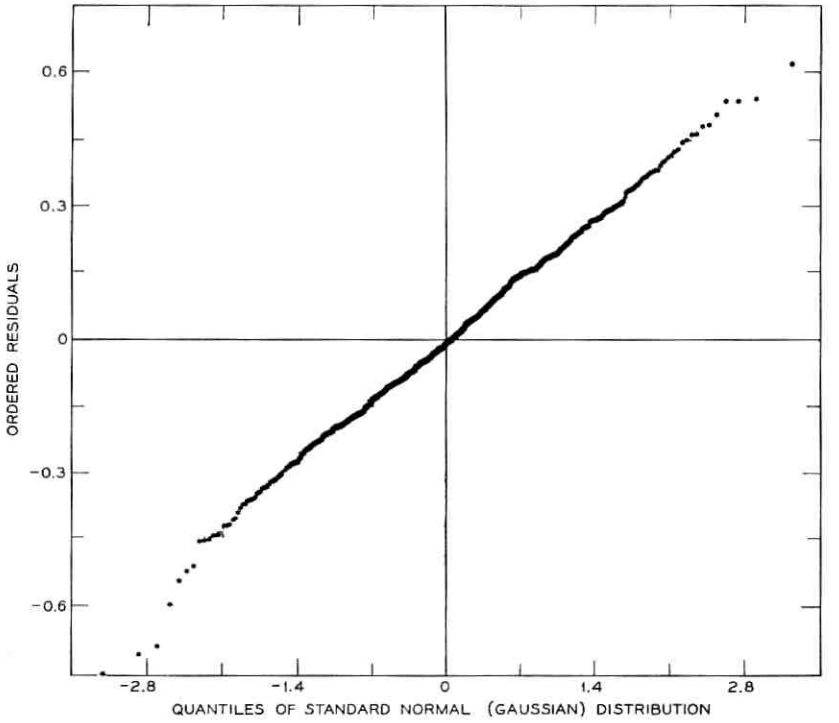


Fig. 28 — Normal probability plot of residuals from fit of the model to the HTB sample.

(Appendix C) the overall HTB data MSR(Y), 0.038, (Appendix C) as well as the MSE(Y) from the fit to the 960 points, 0.036, (Table IV). This is as one would expect, since the variance estimate from the slope of Figs. 28 and 29 is *not* downward biased by the zero (and $\sqrt{1/11}$) residuals from the very low counting rate observations for $x > x_c(L)$, while the other quantities *are* so biased.

The excess of the variance estimate of 0.044 over the Poisson value of 0.023 may be due to any or all of several factors, including: (*i*) the noncorrectness of the Poisson assumption, (*ii*) temporal variations in the radiation belts or the detection equipment, (*iii*) measurement errors or computational biases in time record, ephemeris or magnetic coordinates, etc. (*iv*) noise bursts—the outlandish values were detected and discarded, but the general effect must be an upward bias on variation, and (*v*) inadequacies in the model, including analytic form and coordinates employed.

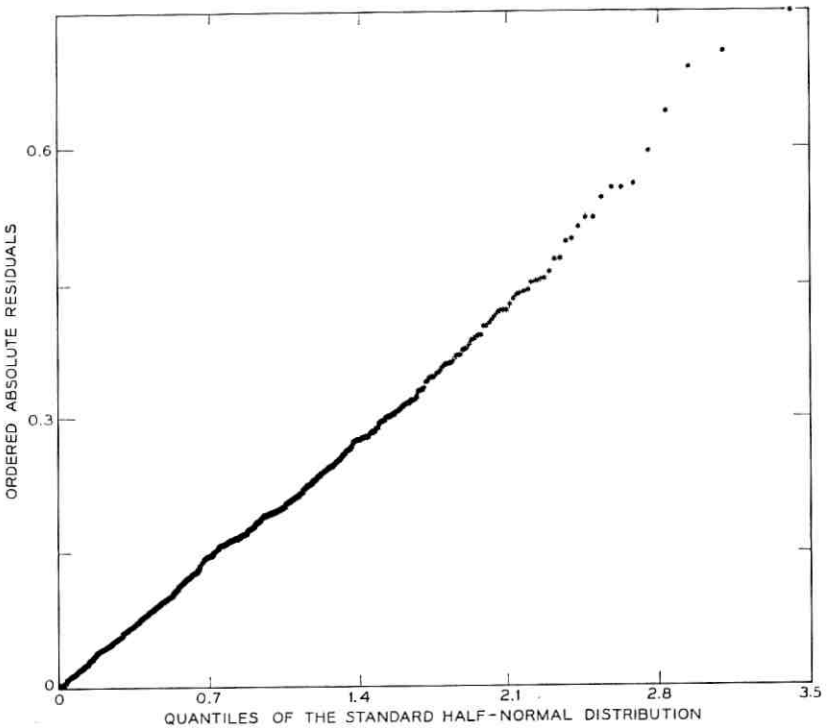


Fig. 29 — Half-normal probability plot of absolute residuals from fit of the model to the HTB sample.

8.2 *Statistical Measures Over All the HTB Data.*

An extensive presentation and comparison of various functions of the residuals over all the HTB data is given in Appendix C. Those results provide (*i*) an empirical justification for the use of the square root transformation; (*ii*) a strong indication that the fit attained by Model I cannot be improved very much in the least squares sense over all the HTB data; (*iii*) information on the extent of "unevenness" of the cell-construction process by which the 960-point HTB sample was selected; and (*iv*) some indication of differential effectiveness of fit of Model I to the data for different x,L regions.

8.3 *Statistical Properties of Estimates of the Coefficients and Coefficient Functions.*

The least squares estimates of the nine coefficients of Model I fitted to the 960-point HTB sample are given in Table IV, with their approximate standard errors and pairwise correlations.* These provide the information needed to obtain estimates and standard errors for functions of the coefficients; e.g., $y'(x,L)$, or $A'(L)$, or the value of the maximum counting rate, or the position in space at which the intensity of high energy protons is maximum, etc. (See Appendix B for the necessary formulae.)

Some of the pairwise correlations in Table IV are exceedingly high. This may be due, in general, either to an unfortunate "design" (i.e., the array of positions of observations in x,L space in this application) or to some inherent "coefficient redundancy" in the model, or to both such blemishes. Occurrence of such near-singularities can lead to practical difficulty with the iterative fitting computation and/or make the individual coefficient estimates poorly determined.

In the present model, only the coefficient L_0 has a direct physical interpretation. Its estimate has a very small standard error and an entirely bearable correlation with the remaining coefficient estimates (all values of $|\alpha| < 0.5$). Otherwise, physical interest centers mainly on the coefficient functions $A'(L)$, $x_c(L)$, and $y'(x,L)$ whose estimation is considered in Sections 7.2, 7.4, 7.5, 7.6, and 8.4.

For a given model and specified coefficient values, the matrix of approximate correlations depends only on the array of data positions in x,L space. Thus, to check on whether the correlational problems might

* A rescaling of the values of ρ, namely as the quantity α defined and motivated in Appendix B.5, is also given in Table IV. The coefficient of dependence α has more nearly the behavior of a "linear utility function."

be due to inadequacy of the practically available (selected) array, a correlation matrix was computed using an 'ideal' x,L array, namely the 1034 values of (x,L) corresponding to the division of x,L space described in Section 7.1 and Appendix B.3. While some minor improvements in some of the correlations were noted, the changes were small. Thus, it would appear that the main reason for the high correlations is in fact some "coefficient redundancy" in the model.

Inspection of Table IV indicates that the very large correlations are associated with some of the parameters of the $A'(L)$ function, namely a_1, a_2, a_3, and η for all pairs of which $|\rho| > 0.99$ (i.e., $|\alpha| > 0.90$). Moreover, it will be seen in Section 8.5 below, that the present parameterization of the model leads to a markedly large indication of nonlinearity and there is reason for believing that this is largely due to the same subset of coefficients. The combination of both defects stimulated development of Model II which overcame them (see Section IX).

8.4 *Estimates of Functions of the Coefficients*

The estimates of the coefficient functions $A'(L)$ and $x_c(L)$ have been discussed in Sections 7.4 and 7.5 and summarized in Figs. 10 and 11. Their estimated standard deviations, on a "pointwise" basis, are graphed in Fig. 22(a), while the approximate correlations of the estimates of $A'(L)$, $x_c(L)$, and S, as functions of L, are shown in Fig. 30(a).

Despite the near-singularities (i.e., $|\rho|$ near 1) in the estimates of some of the individual coefficients of $A'(L)$, it is seen that the estimate of the square root of the equatorial counting rate provided by $A'(L)$ is well-determined over the entire L range. The standard error varies between approximate limits of 0.018 and 0.040, nonmonotonically, and these values are typically less, sometimes by a factor of 5 or more, than the standard errors from the corresponding L-slice estimates (see Table II) reflecting in part the statistical gain from the simultaneous two-dimensional fit.

For $x_c(L)$, the standard error is less than 1 percent over much of the range of L, rising to 3 percent for large L values where the data are statistically inadequate.

The three correlation functions $\rho_{A,x_c}(L)$, $\rho_{A,S}(L)$, and $\rho_{S,x_c}(L)$, for the estimated coefficient functions $A'(L)$, $x_c(L)$, and S, are plotted in Fig. 30(a) (see Appendix B.4 for formulae). In general, these correlations are small ($|\rho| < 0.5$, $|\alpha| < 0.12$). The statement applies to the correlations involving $A'(L)$ despite the very high correlations among individual coefficients. The generally low correlation between $A'(L)$ and $x_c(L)$ is as intuitively expected since $A'(L)$ is influenced mainly

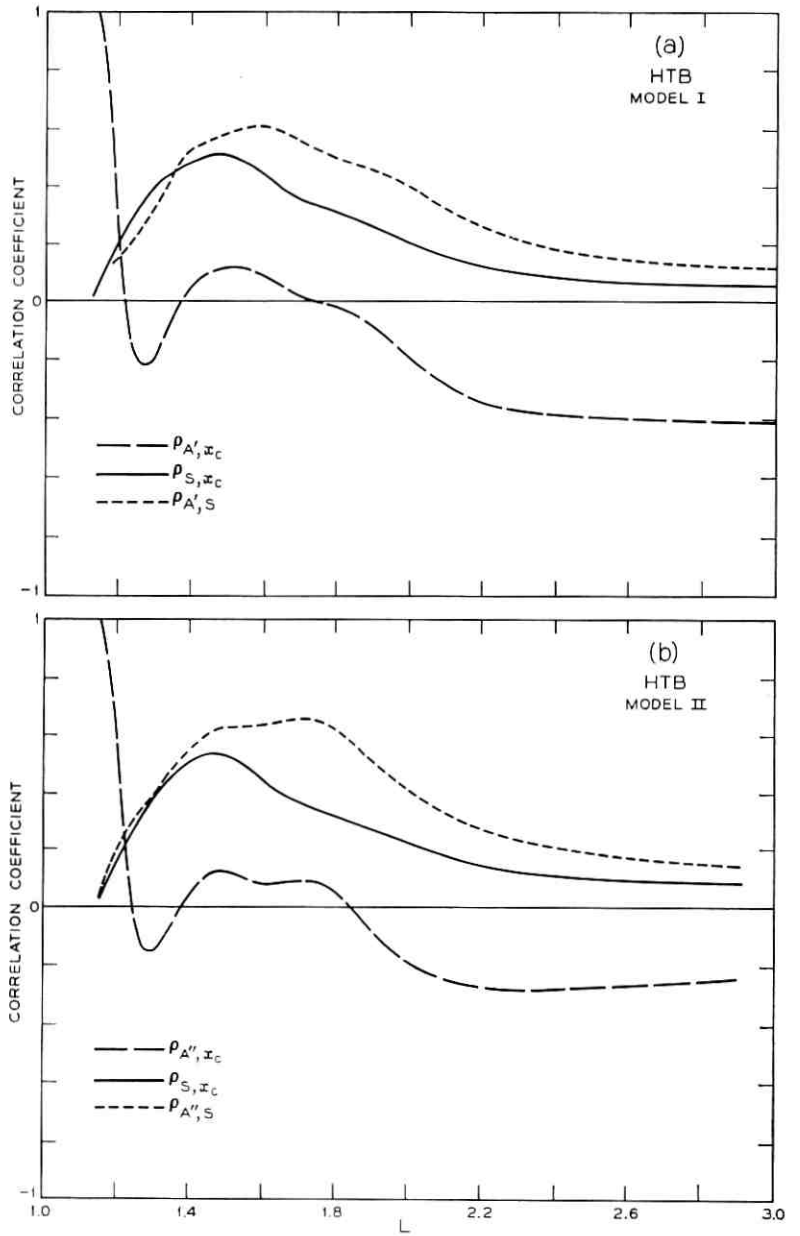


Fig. 30 — Correlation coefficients of A with S, S with x_e, and A with x_e, calculated from the fits to the HTB sample and plotted as functions of L. (a) Model I. (b) Model II.

by observations at small x while $x_c(L)$ is determined mainly by those at large x. The exception is near $L = L_0$, where $\rho_{A,x_c}(L_0)$ approaches 1 as a result of the fact that the coefficient L_0 is common to both functions and that the forms of $A'(L)$ and $x_c(L)$ [see (4), (5), and (6)] require that both functions be zero when $L = L_0$.

The statistical correlation between the fitted A and x_c for the L-slice fits was always positive (see Table II), which is not the case for $\rho_{A,x_c}(L)$. This change in sign gives some indication of basic differences in behavior between the results of the two-dimensional fit and the outcome of the *collection* of one-dimensional L-slice fits.

The (A, S) and (S, x_c) correlations have the same signs in all cases. The magnitude of the correlations among A, x_c, and S is larger for the L-slice fits (see Table II) than for the HTB fit at corresponding values of L [see Fig. 30(a)]. This is very noticeable for L greater than $\approx$1.7. particularly for the large correlation between S and x_c. It is these large correlations which make it difficult to obtain reliable L-slice estimates of x_c or S when $L_m > 2$ (see Fig. 6) or when the distribution of the data within an L-slice is poor (see Fig. 7).

8.5 *Nonlinearity Indices and Dependence of Estimates*

Appendix B.5 discusses the use of the sum of squares function (i.e., sum of squares of differences between observed value and "fitted" value, as a function of proposed coefficients) as an indicator of the joint dependence and behaviour of the coefficient estimates and the fact that the extent to which the contours of the sum of squares function are approximated by a certain family of ellipsoids provides a measure of linearity of the model.

Fig. 31 shows 4 of the 36 pairwise projections of the 9-dimensional ellipsoid, whose size would correspond to a "0.99 joint confidence coefficient" as discussed in Appendix B.5. The axes are scaled in each case according to the standard error of the coefficient. The orientation and shape of the ellipse corresponds directly to the sign and magnitude of the correlation, ρ, or its transform, α, for the pair of coefficients. Thus, for example, Fig. 31(a) shows the projection onto the a_1-a_3 plane. The resulting very narrow positively inclined ellipse corresponds to a very high positive correlation of a_1, a_3 ($\rho = 0.9995$, $\alpha = 0.97$). (The 45° inclination of the graphed ellipses is a result of scaling the axes by their standard errors.) Part (b) of the figure shows a narrow negatively inclined ellipse for the case of rather large negative correlation between a_3 and η estimates. Parts (c) and (d) illustrate results for

small and negligible correlations between L_0, r_2 and r_3, S, respectively.

At various positions on these ellipses there appear numbers which are ratios of the actual sum of squares at that "point" to the minimum sum of squares. The computation of the actual sum of squares is done for the coefficient values corresponding to the point on the 9-dimensional ellipsoid which projects into the point on the plotted ellipse.

If, in fact, the coefficients occurred linearly, all of these numbers on all of the pairwise ellipses would be constant and in the present case would have the value 1.023 corresponding to a sum of squares of residuals of about 35.47. As a basis for judging the actual values and their variability, the following table gives values which this ratio would have, if the coefficients did occur linearly, for various joint (9-dimensional) "confidence coefficients:"

Conf. Coeff.	*Contour Ratio*
0.90	1.015
0.95	1.018
0.99	1.023
0.999	1.029

In view of the variability of the actual ratios in Fig. 31, and of the extent to which some depart from the values in the above table, it is clear that in the present form of the model the coefficients behave jointly in a markedly nonlinear fashion even in a relatively small neighborhood around the least squares estimate.

Inspection of the entire set of $(9)(8)/2 = 36$ pairwise plots strongly suggests that a major part of this nonlinear behavior derives from the coefficients a_1, a_2, a_3, and η of the $A'(L)$ part of the model. These also are the coefficients whose estimates exhibit the undesirably high correlations which have been shown above to be due mainly to a "coefficient redundancy" in the model.

Direct interpretation of the ellipses in Fig. 31, as indicating interdependence of the coefficient estimates, depends heavily on the appropriateness of the linear approximation in the neighborhood of the least squares estimate. Since the nonlinearity index is in fact distressingly large one must be cautious in interpreting the ellipses or their associated correlation or dependence coefficients.

8.6 *Summary Statistical Criticisms of Model I.*

Model I, with coefficients determined by fitting to the 960-point HTB sample, has been shown to provide a very good fit both to the

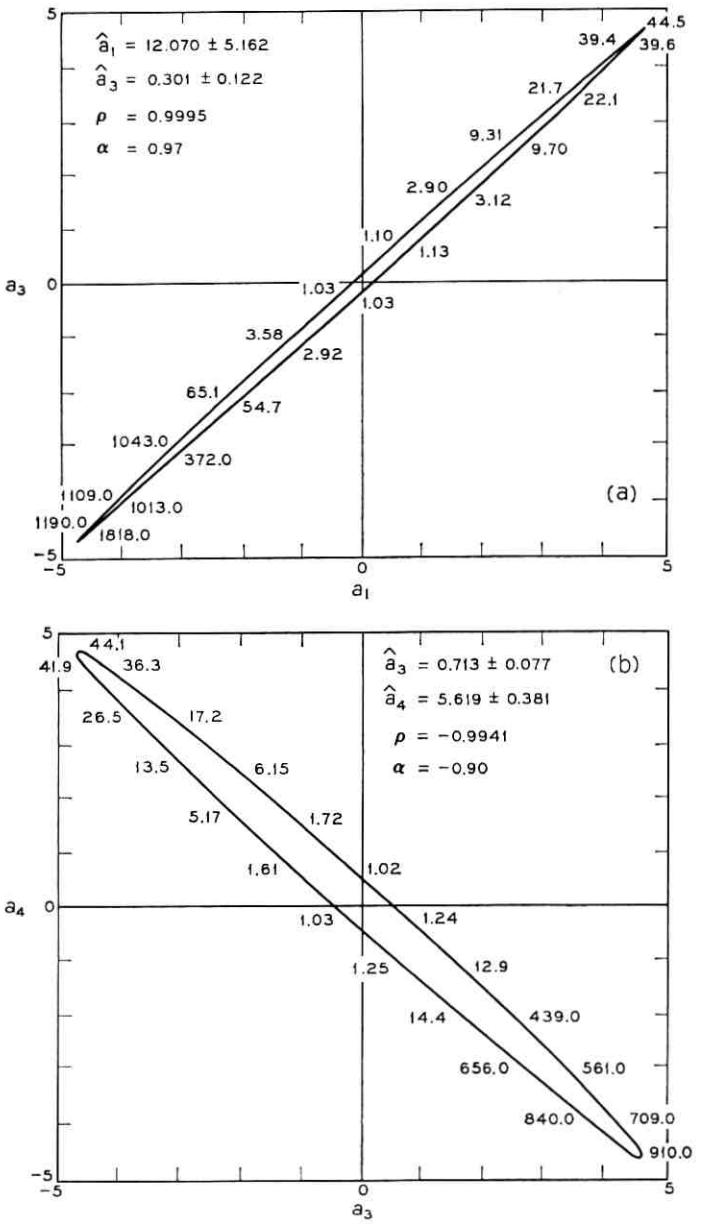


Fig. 31 — Examples of projections of the approximate "0.99 joint confidence region" for the estimates of Model I. (Axes are scaled by standard errors.)

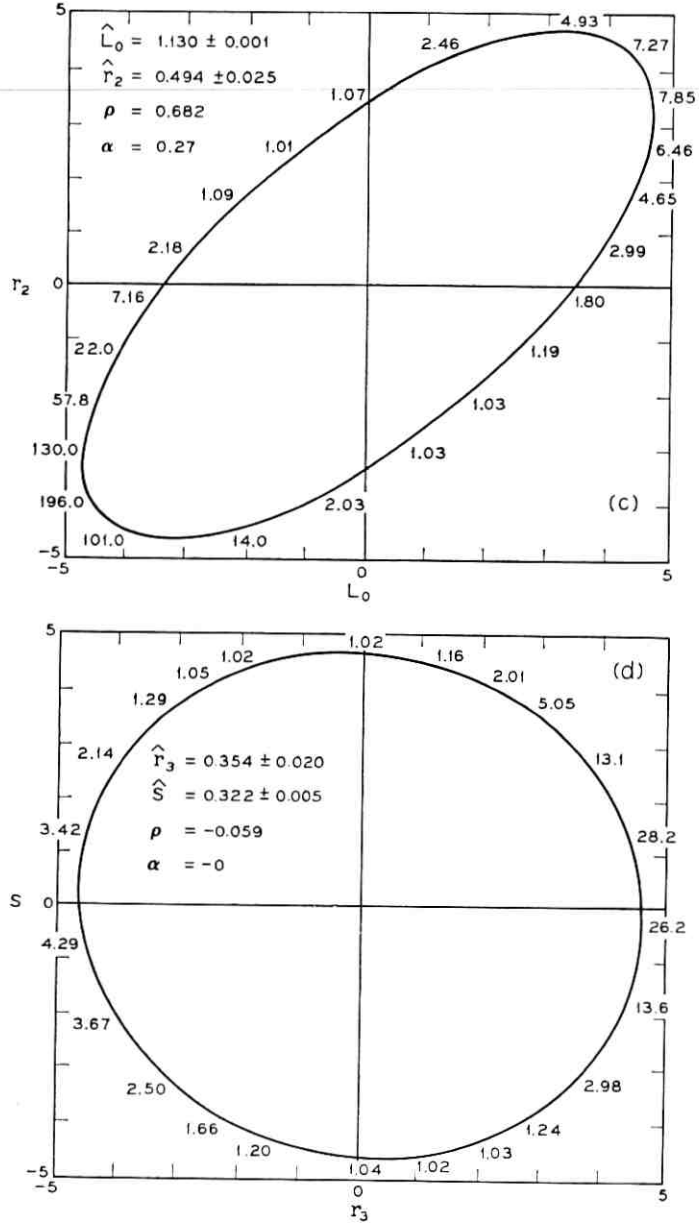


Fig. 31 — (continued)

sample and to the entire body of some 41,000 HTB observations. Moreover, the interesting coefficient functions $y'(x,L)$, $A'(L)$, and $x_c(L)$ have stable statistical properties as has the physically interpretable coefficient L_0.

However, the model has two statistical defects: Firstly, although the model gives an extremely good fit to the data, the parameters a_1, a_2, a_3, and η of the $A'(L)$ part of the model have exceedingly high mutual correlations (see Table IV), and these were shown *not* to be due to an obviously defective design. Secondly, the model coefficients exhibited distressingly high nonlinearity of behavior even within rather close neighborhoods of their least squares estimates, with grounds to suspect that this was caused by the a_1, a_2, a_3, η group of coefficients. In addition, most of the coefficients of Model I do not have any directly meaningful physical interpretation.

The modifications which led to Model II, as discussed in the following Section IX, overcome these defects of Model I while retaining all its virtues.

IX. THE MODEL II FIT TO THE HTB DATA

This section presents the statistical analysis of the HTB data using Model II, a modified version of Model I. The emphasis in the presentation is on comparisons of Models I and II. Since it is shown how very closely the fit of Model II approximates that of Model I, such aspects as the direct presentation of Model II residuals overall the data are unnecessary, and hence omittted.

9.1 *Model II*

The definition of Model II has been given in Section 4.6, together with a discussion of the physical interpretation of its coefficients and its mathematical relation to Model I. Specifically, the 8-coefficient Model II constitutes a specialization and reparameterization of the 9-coefficient Model I. Thus, it follows that the minimum sum of squares in fitting Model II to any body of data can not be less than that from fitting Model I, though this may not be true of the mean square error.

The evolution of Model II from Model I did not arise from any simply described systematic process, as is indeed true in other aspects of this study. Once the basic achievements of Model I were established it was then opportune to focus on major remaining defects. The character of these defects strongly urged elimination of one or more

coefficients in conjunction with a nonlinear reparameterization of the coefficients. The solution achieved was arrived at by empiricism, persistence and good luck.

The remainder of this section documents the assertion that Model II retains all the virtues of Model I while overcoming its defects.

9.2 *The Fit of Model II to the 960-point HTB Sample.*

The analysis of variance from fitting the 960-point HTB sample by means of the 8-coefficient Model II is given in Table V. As expected, the residual sum of squares, 34.7126, of Table V exceeds that of Table IV, namely 34.6675. This difference is associated with the one-degree-of-freedom nonlinear constraint defined in (13). Thus, we see that the sum of squares associated with the one-degree-of-freedom non-linear constraint is (34.7126-34.6675) = 0.0451 and this gives a ratio of less than 1.24 in relation to the mean square error of 0.03645. The value 1.24 corresponds to the upper tail 27 percent point of the chi-squared-with-one-degree-of-freedom distribution. The proportionate increases in the sum of squares for error is about 0.13 percent and the increase in the mean square error is less than one part in 3000. Multiple $R^2 = 0.989$ is effectively unchanged.

For the models of both Tables IV and V, the coefficient S is treated as constant with L. If Model II is modified so that $S(L) = s_0 + s_1 L$, then, fitting this 9-parameter version of Model II yields a sum of squares for error of 34.520. Thus, we would have a sum of squares of (34.713-34.520) = 0.193 associated with the "hypothesis" $s_1 = 0$. The main point of quoting this result is to indicate that these minor differences in the sums of squares for error are judged as unimportant in this context, even if under some highly formalized assumptions the distinctions are "statistically significant."

Of greater interest and sensitivity are the following considerations: (*i*) the behavior of the residuals from Model II as functions of x,L and y; (*ii*) the behavior of the differences between Models I and II; (*iii*) comparisons of the estimates of $A'(L)$ of Model I and $A''(L)$ of Model II [see (6) and (11)]; (*iv*) comparisons of the estimates of $x_c(L)$ from the two models; (*v*) the pattern of correlations of the estimates of the eight Model II coefficients; and (*vi*) the indices of nonlinearity for the coefficients of Model II.

9.3 *Residuals of Model II Fit and Differences Between Models I and II.*

Figs. 32, 33, and 34 are plots of the residuals of the 960-point HTB sample from the fitted values of Model II against L, x and Y, re-

TABLE V—FIT OF MODEL II TO 960-POINT HTB SAMPLE.

Analysis of variance

Due to	d.f.*	Sum of squares	Mean square
Total	960	5374.7321	
Model	8	5340.0195	667.5024
Error	952	34.7126	0.0365

Coefficient estimates

	A_p	L_0	L_p	η	r_1	r_2	r_3	S
Estimate	8.0762	1.1293	1.4644	5.2187	0.2658	−0.5082	0.3638	0.3225
standard error	0.0342	0.0009	0.0016	0.0474	0.0083	0.0241	0.0196	0.0047

α values

ρ values	A_p	L_0	L_p	η	r_1	r_2	r_3	S
A_p		0.01	0.00	0.02	−0.02	0.01	−0.02	0.21
L_0	0.140		−0.05	−0.11	−0.41	0.19	−0.09	0.00
L_p	0.046	−0.304		0.39	0.00	0.00	−0.01	0.01
η	0.203	−0.451	0.794		0.02	−0.00	−0.01	−0.00
r_1	−0.181	−0.806	0.045	0.214		−0.62	0.39	−0.01
r_2	0.162	0.589	0.090	−0.021	−0.923		−0.71	0.00
r_3	−0.171	−0.422	−0.151	−0.107	0.793	−0.956		−0.00
S	0.609	0.015	0.099	−0.010	−0.168	0.098	−0.085	

* degrees of freedom

spectively. These plots show no systematic structure and are quite similar to analogous plots for Model I. Furthermore, Fig. 35, showing the observed Y versus fitted y'' for Model II, is as well-behaved as the corresponding Fig. 20 for Model I.

Figs. 36, 37, and 38 show the deviations between the fitted Models I and II plotted against L, x, and Y, respectively. Of course these figures show a systematic structure since one is plotting the difference of two smooth functions. However, the actual differences are totally insignificant in the light of the data. (Note that the scale for Figs. 36, 37, and 38 differs from that of Figs. 32, 33, and 34 by a factor of 10.)

Thus, on the basis of one less coefficient, Model II fits the data essentially as well as Model I, to which indeed it is a very excellent approximation. It has the merit that the physically arbitrary coef-

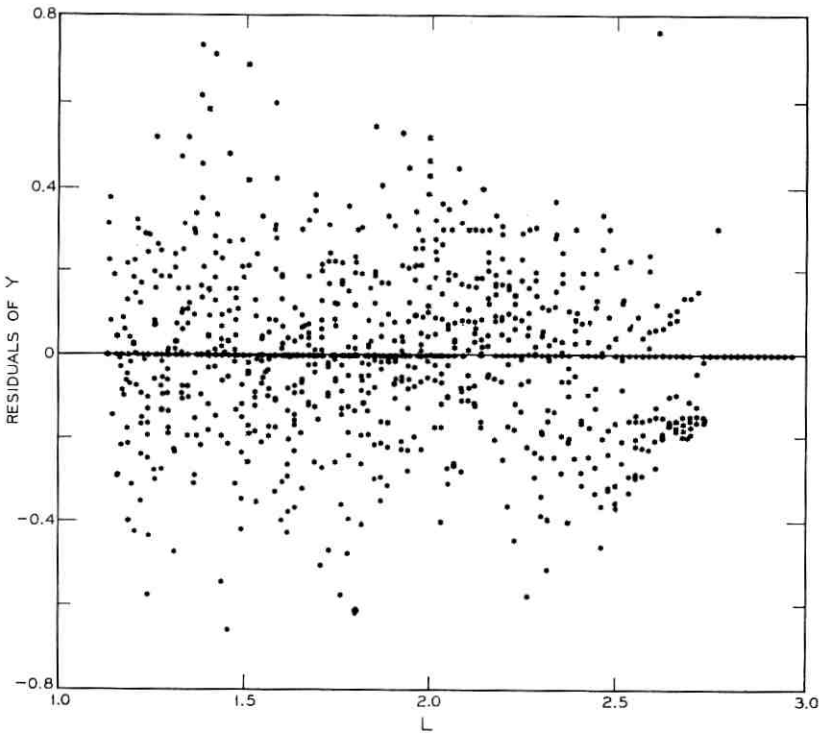


Fig. 32 — Residuals ($Y - y$) from the fit of Model II to the 960-point HTB sample vs L.

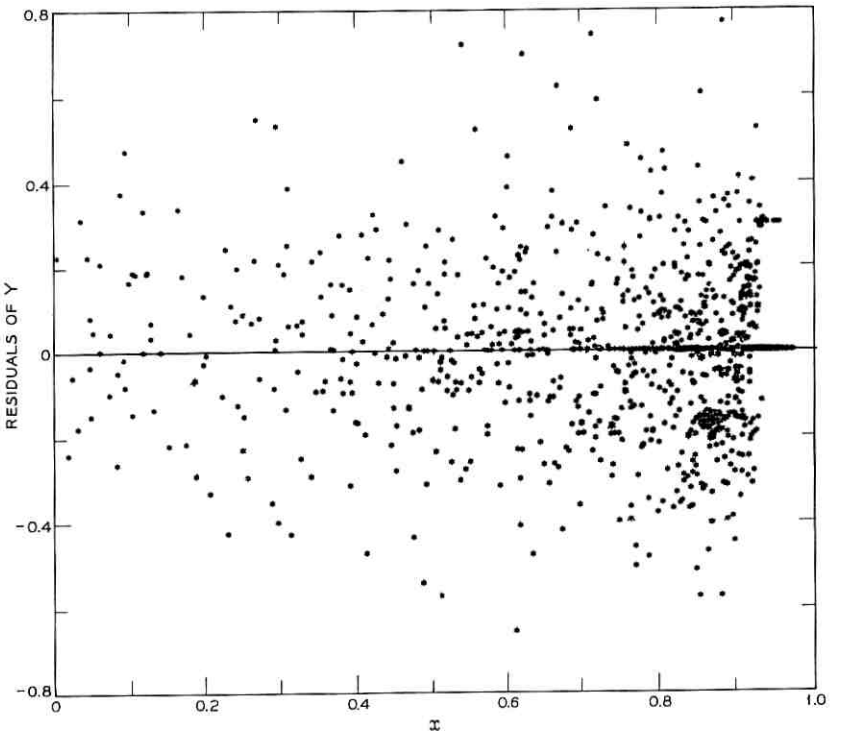


Fig. 33 — Residuals ($Y - y$) from the fit of Model II to the 960-point HTB sample vs. x.

ficients a_1, a_2, and a_3 of Model I have been replaced by A_p and L_p which do have direct physical interpretations. As will be detailed in the next subsection, Model II also has additional attractive statistical attributes.

9.4 *Coefficient Estimates*

Table V gives the least squares estimates of the eight coefficients of Model II together with their approximate standard errors, correlations and α values. The estimates are seen to be extremely well-determined. In particular, for the physically meaningful quantities A_p, L_0, and L_p the standard errors are about 0.4, 0.1, and 0.15 percent, respectively, while for the shape coefficients η and S they are about 1 and 1.5 percent, respectively.

Comparison with Table IV shows that the standard error has de-

creased for *every* coefficient which is common to the models. The most dramatic change is for η for which the standard error diminished by a factor of about 8.

The estimates of $A'(L)$ and $A''(L)$ are in very close correspondence as implied by Fig. 36. The comparison of Fig. 22(b) with Fig. 22(a) indicates that the standard error of $A''(L)$ is uniformly lower than (but in general agreement with) that of $A'(L)$.

Entirely similar remarks apply to comparison of estimates of $x_e(L)$ from Models I and II, as also documented by Figs. 22(a) and 22(b).

It has already been shown that the fitted values of $y'(x, L)$ and $y''(x, L)$ are in very close agreement. The pattern of contours of the percent standard errors of $[y''(x, L)]^{\frac{1}{2}}$, in Fig. 24(b), shows that the standard error is everywhere smaller than the corresponding results for Model I, in Fig. 24(a).

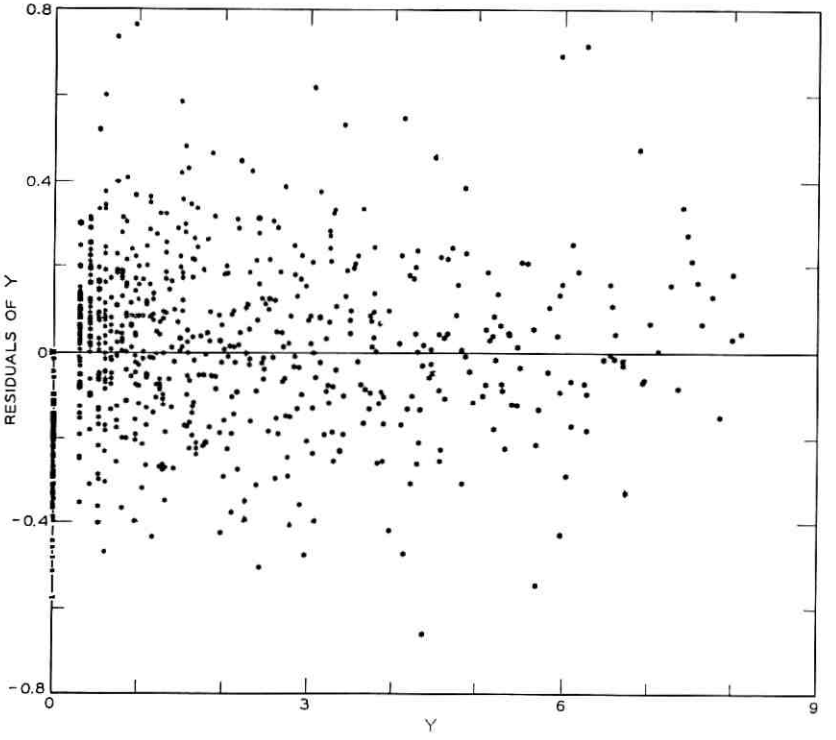


Fig. 34 — Residuals $(Y - y)$ from the fit of Model II to the 960-point HTB sample vs Y.

One of the most dramatic changes between Models I and II is indicated by comparison of the correlations in Tables IV and V. The very large correlations ($|\rho| > 0.99$, $|\alpha| > 0.9$) among the $A'(L)$ coefficients of Model I do not occur for Model II. Only the (r_1, r_2) and (r_2, r_3) coefficient pairs of Model II have $|\alpha|$ values above 0.5. This is inconsequential since these are physically arbitrary coefficients of a cubic polynomial.

The correlations of $A''(L)$, $x_c(L)$, and S from Model II remain much like the corresponding results for Model I, as shown in Fig. 30.

9.5 *Nonlinearity Indices*

The further virtuosity of Model II is indicated by the behavior of the nonlinearity index shown for the examples of "confidence regions"

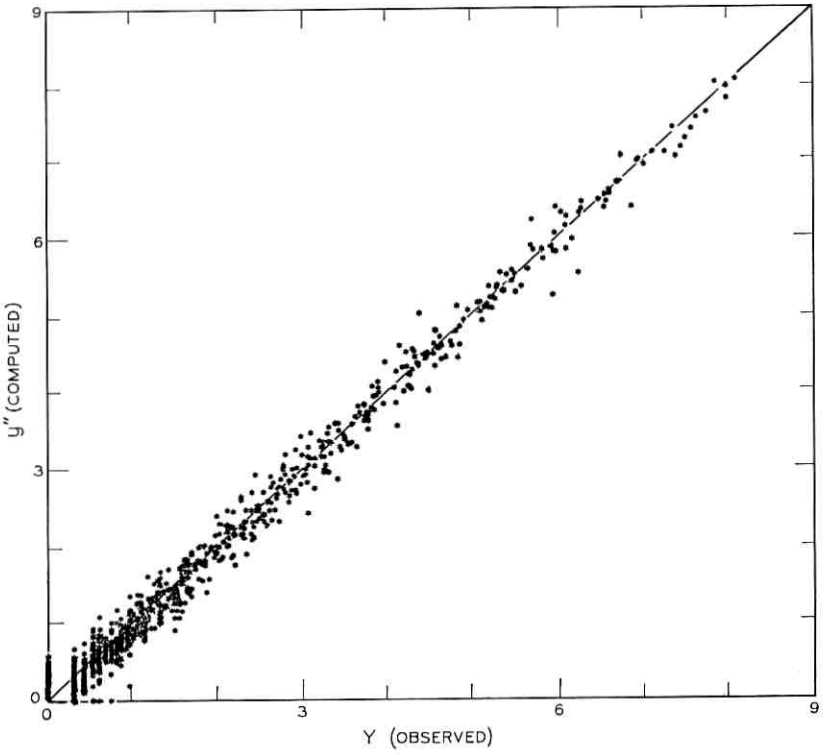


Fig. 35 — The value of y'' computed from the fit of Model II vs the observed value, Y, for the 960-point HTB sample.

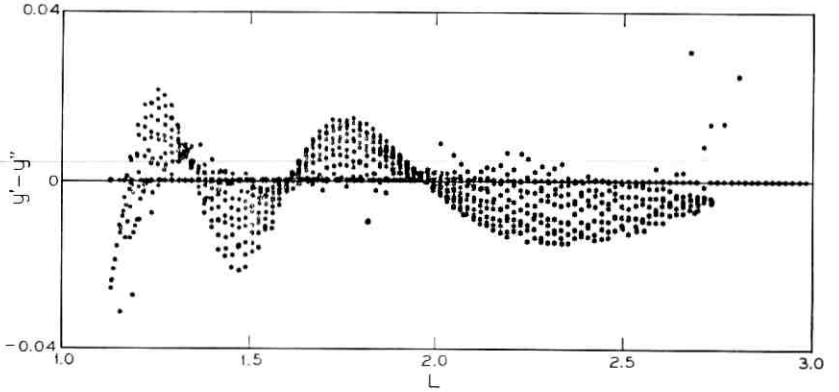


Fig. 36 — Deviations between the Model-I fit, y', and the Model-II fit, y'', vs L, for the 960-point HTB sample.

in Fig. 39. (See Appendix B for general discussion and definition.) Specifically, it is seen that the numbers on the ellipses vary very little and this is true for all 28 of these ellipses. These numbers would be constant and all equal to 1.023 if the model were linear in the fitted coefficients. Comparatively, Model II does indeed behave in a reassuringly linear fashion. For sharp contrast, we may compare Fig. 39 with Fig. 31, for Model I, in which the values range up to 1000 around the 9-dimensional ellipsoid.

The nonlinear behavior of Model I in relation to the linear be-

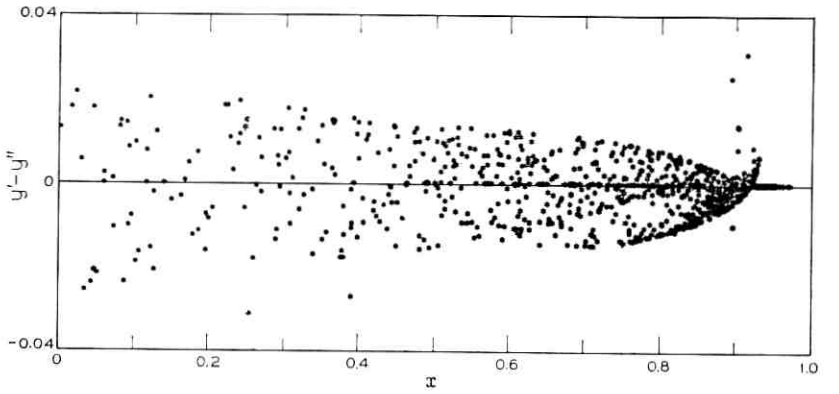


Fig. 37 — Deviations between the Model-I fit, y', and the Model-II fit, y'', vs x, for the 960-point HTB sample.

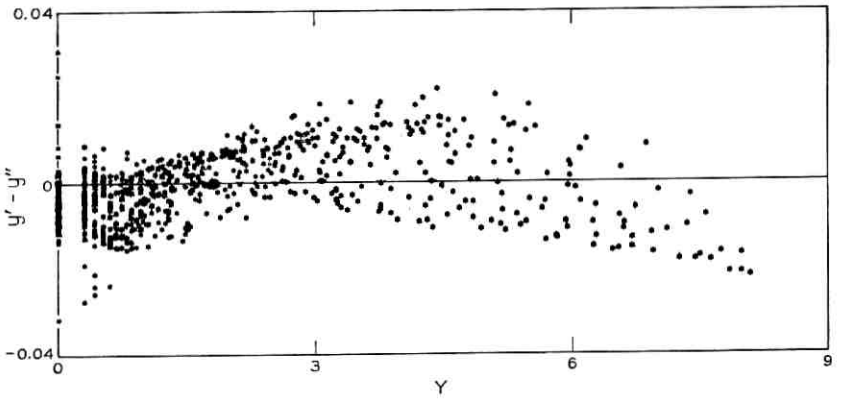


Fig. 38 — Deviations between the Model-I fit, y', and the Model-II fit, y'', vs Y, for the 960-point HTB sample.

havior of its specialized reparameterized version, Model II, is indicative of the reason for the high nonlinearity indices for Model I. Effectively, a p-coefficient model defines a constraining "surface" of p dimensions (p is 9 and 8 for Models I and II, respectively) in the n-dimensional space of the observations (n is 960 in the present case). In a small neighborhood of the least squares estimate, this p-dimensional surface may or may not be planar. If the latter, one will obtain high indices of nonlinearity. If the former, then one will or will not obtain high nonlinearity indices according to whether the individual coefficient *coordinates* within the p-dimensional surface are or are not linearly behaved.

It is likely that the 9-dimensional surface defined by Model I is indeed reasonably planar, but the coordinate system defined by the coefficients is highly nonlinear.

The correlation and nonlinearity effects, it should be noted, are not in principle related. One can have very high correlations with linear models and very low correlations with very nonlinear ones.

9.6 *Summary Comments*

Model II has been presented and validated as an evolution of Model I. Though Model II represents the current recommended fit from this study, several aspects of its justification, and of other comparisons in this paper, are based on the Model I fit. For example, the statistical study of residuals over all the HTB data, discussed in various places

including Appendix C, is based on Model I. This hybrid attitude is entirely sound, since the range of deviation between Models I and II is small compared to the range of residuals from the fitted sample.

Thus Model II provides a fit to the HTB in which the 8 estimated coefficients provide a "good description" of about 41,000 observations. The deviations of the fit from the data are within reasonable statistical fluctuations—variation in telemetered counting rates, orbital errors, observational errors, mapping-to-magnetic-coordinate uncertainties, etc. (See Appendix C.3). A number of the coefficients have physical interpretations and these are statistically well-determined and relatively uncorrelated. Model II, though nonlinear in the coefficients, behaves in a very linear fashion in the neighborhood of the least squares estimates.

X. TEMPORAL VARIATIONS

This section and the two to follow are devoted to discussion of some specific physical results of the analysis.

Temporal variations are considered in three classes: diurnal (day-night), secular, and short term. Residual plots were used to study these effects.

10.1 *Diurnal Effects*

The HTB residuals were plotted against local time for various x,L regions. The HTB data are not well-distributed in local time near the magnetic dipole equator, making it difficult to draw firm conclusions. However, no evidence of a diurnal variation was found.

Specifically, to produce a change of about two percent in the average value of Y on the equator ($x = 0$) would require a diurnal shift in the radial position of the magnetic field line of about 0.01 R_e at $L = 1.35$, and a shift of about 0.02 R_e at $L = 1.55$, if there were no other effects. At these two positions, the value of y is large ($y \approx 8$) and $\partial y/\partial L$ is large, and a two-percent change in y would correspond to a shift in the mean of the residuals of ≈ 0.16 between noon and midnight local time. An effect of this magnitude would be readily observable on the residual plots.

Thus, it is unlikely that displacements larger than 70 km and 140 km, at equatorial L's of 1.35 and 1.55, respectively, would escape detection, and these distances are offered as upper limits to the day-night changes of the magnetic field at the two positions. As both of these displacements are equivalent to a change in field strength of

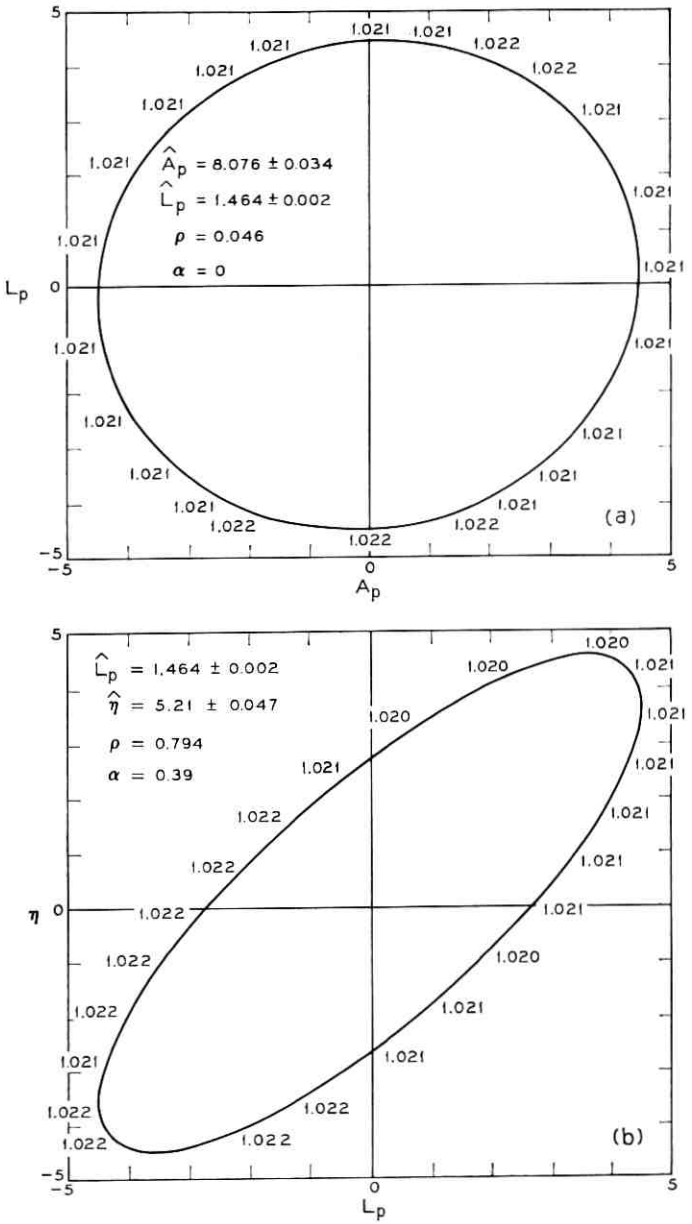


Fig. 39 — Examples of projections of the approximate "0.99 joint confidence region" for the estimates of Model II. (Axes are scaled by standard errors.)

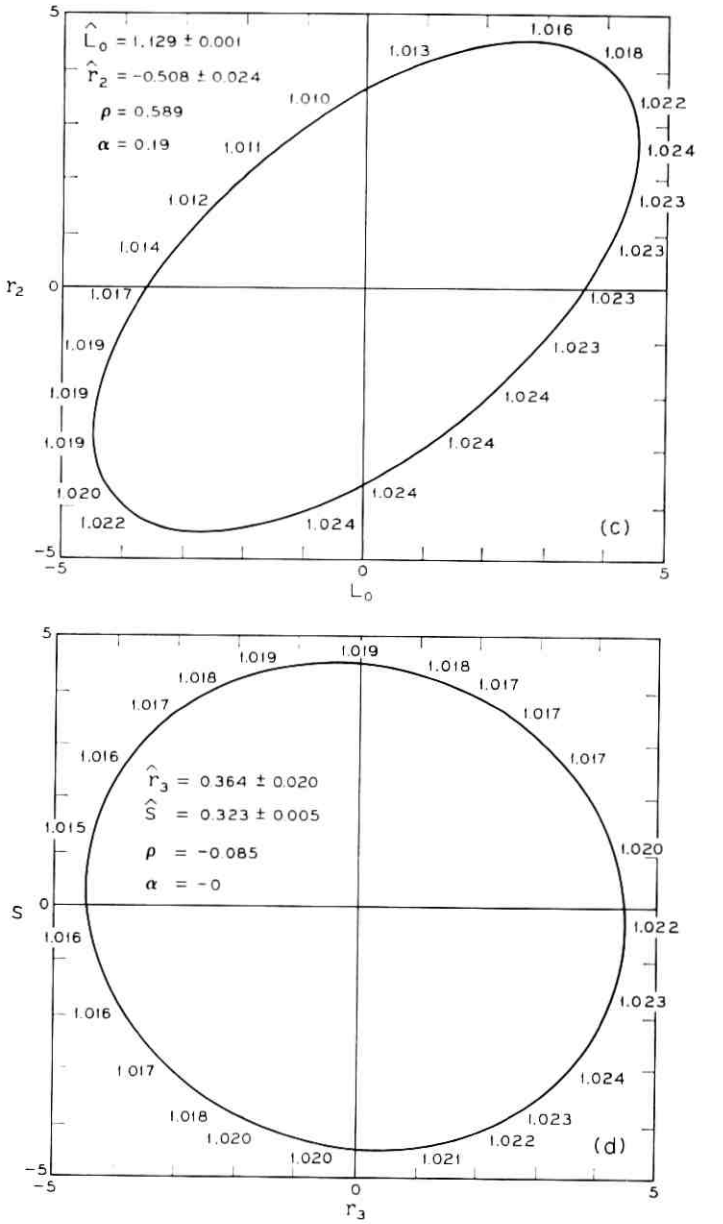


Fig. 39 — (continued)

about 300 gamma (0.003 gauss), this particle experiment does not qualify as a sensitive indicator of adiabatic changes in the earth's magnetic field.

10.2 *Secular Effects*

The HTB residuals are plotted against elapsed time, in days, for $1.85 < L < 1.90$, in Fig. 40. It would appear that the average value of Y decreased between days 191 and 255. This decrease is exhibited in all parts of the belt where we have measurements during this interval. Between days 191 and 225, the orbit of the *Telstar*® 1 satellite did not take it into the central region of the belt $\{1.3 \leqq L \leqq 1.8, \lambda \leqq 10°\}$. In other regions the decrease in the average value of Y over this period is about ten percent. The extremes are two percent and 20 percent, but it

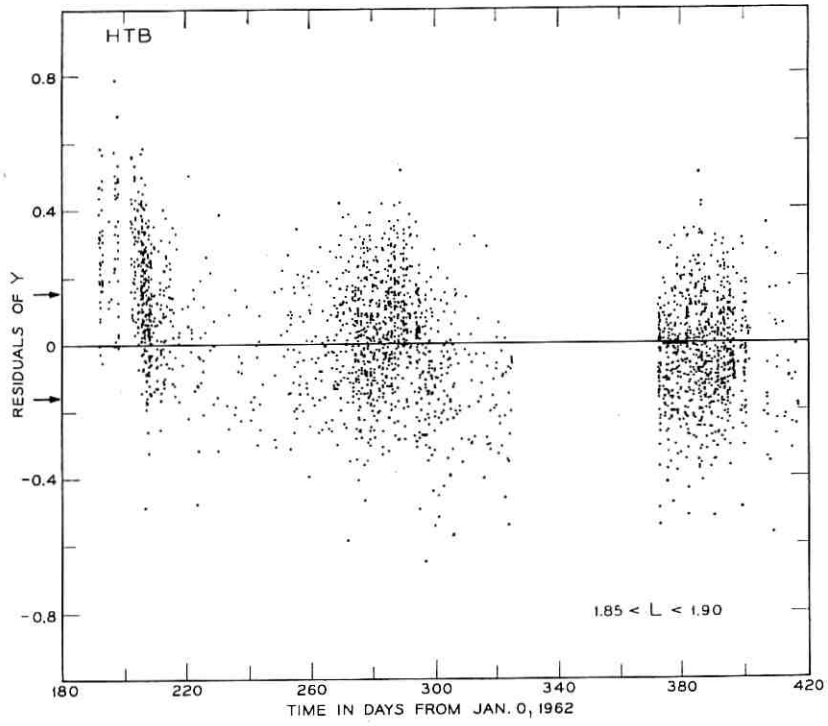


Fig. 40—HTB residuals of Y (i.e., $Y - y$ calculated from the HTB coefficients) plotted against time for $1.85 < L < 1.90$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

is not possible to separate out other variables which may be influencing the results.

From the magnitude of this effect, it is clear that it must be contributing substantially to the MSR. A decrease of ten percent in the average value of Y corresponds to a decrease of about 20 percent in the flux. A fractional change in the flux which is independent of x and L cannot be distinguished from a change in the characteristics of the instrument. Among other possibilities, radiation damage or the decay of protons which might have been associated with the Starfish high-altitude nuclear test of July 9, (day 190) 1962 might have produced the observed effects. Because of this ambiguity, we are unable to offer any well-founded interpretation of the time dependence of the data before day 225. For reasons to be noted shortly, ambiguities are also encountered when interpretation of the temporal behavior of data acquired after day 400 is attempted. In the intermediate period, the time dependence does vary with x and L. By using Fig. 40, which shows comparatively little fluctuation during this intermediate period, as a standard we are able to measure relative changes in the belt. The stretches of sparse data near days 240 and 320 in Fig. 40 are a result of the orbital configuration, there being less opportunity to acquire "high-temperature" data during these periods. The absence of HTB data between day 325 and 373 was caused, as noted in Section 6.9, by the low bias condition that existed during that time. However, an examination of residuals from the CB fit between days 325 and 373 reveals nothing that vitiates the conclusions drawn from the HTB data in what follows.

Residuals versus time-in-days have also been plotted for x,L cells of size 0.1 in L by 0.2 in x. Below $L = 1.9$ we find only one change with time within the sensitivity of our measurements, namely, a secular decrease between days 225 and 400 which occurs only near the ends of the field lines ($x \geqq x_c - 0.2$). We are unable to quantify this effect because, in order to see the droop above the noise, we need to collect residuals from a fairly sizable region of space. The term "sizable" means a region over which y changes so much that an average value of y in the region is not sufficiently representative to be used as a basis for computing a percent change in the flux. Fig. 41 gives an example of an x,L cell near the cutoff where this decrease may be seen. However, in the adjacent lower-x region, Fig. 42, where the ability to discriminate absolute changes in the average value of Y is the same and the ability to discriminate percent change in the average value of Y is much greater

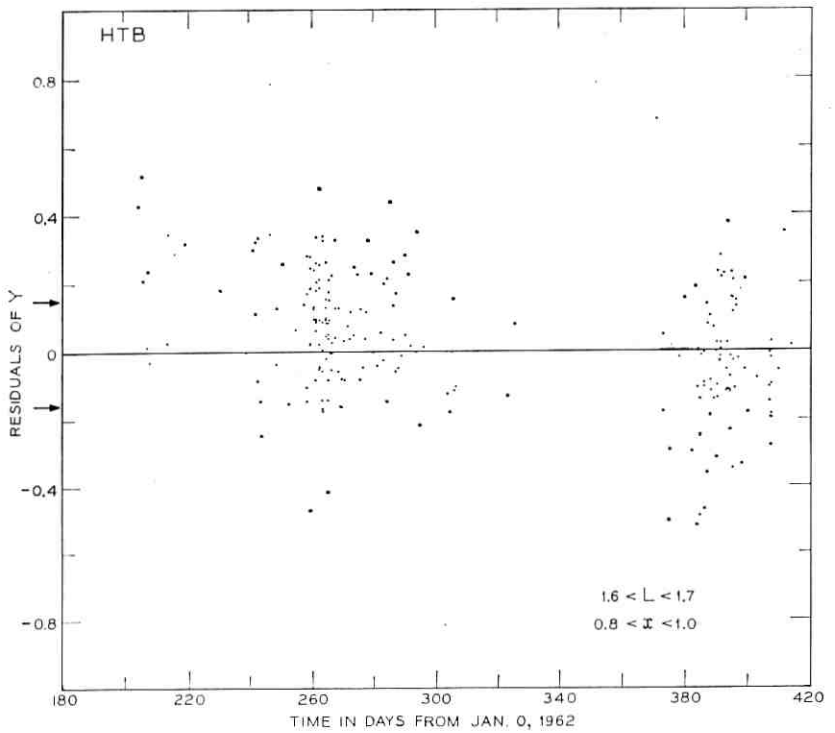


Fig. 41 — HTB residuals of Y (i.e., $Y - y$ calculated from the HTB coefficients) plotted against time for $1.6 < L < 1.7$ and $0.8 < x < 1.0$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

than for the region of Fig. 41, no corresponding secular decrease between days 225 and 400 is evident.

The droop in the residuals after day 400, which is noticeable in Fig. 42, is characteristic of many of the plots of residuals versus time-in-days. The widespread occurrence of this effect confuses instrumental and "real" variations and introduces unresolvable ambiguities when attempts are made to identify the source of the droop.

The observation of the general downward slope in Fig. 41 might be explained by a small decrease in x_c, which corresponds to a small increase in the altitude of the cutoff, between August 1962 and January 1963 on L-shells below 1.9.[23] Alternatively, one might be observing the decay of the 55 MeV protons whose perturbation by the Starfish high-altitude nuclear test of July 10, (day 190) 1962 and

subsequent behavior have been measured by Filz[24] near the bottom of the trapped proton belt. There are too few data for us to attempt further interpretation of this qualitative observation concerning the secular behavior of x_e. The number of points affected and the magnitude of the shift are too small for this effect to contribute interestingly to the MSR.

10.3 *Short-Term Effect*

The plots of the residuals versus time-in-days, for x,L regions, show a short-term fluctuation which is sufficiently singular to be referred to as an event. This event is an increase in the average value of Y over the 30-day period which starts about day 280. It can be seen clearly in Fig. 43. The increase is discernible only for $L > 1.9$.

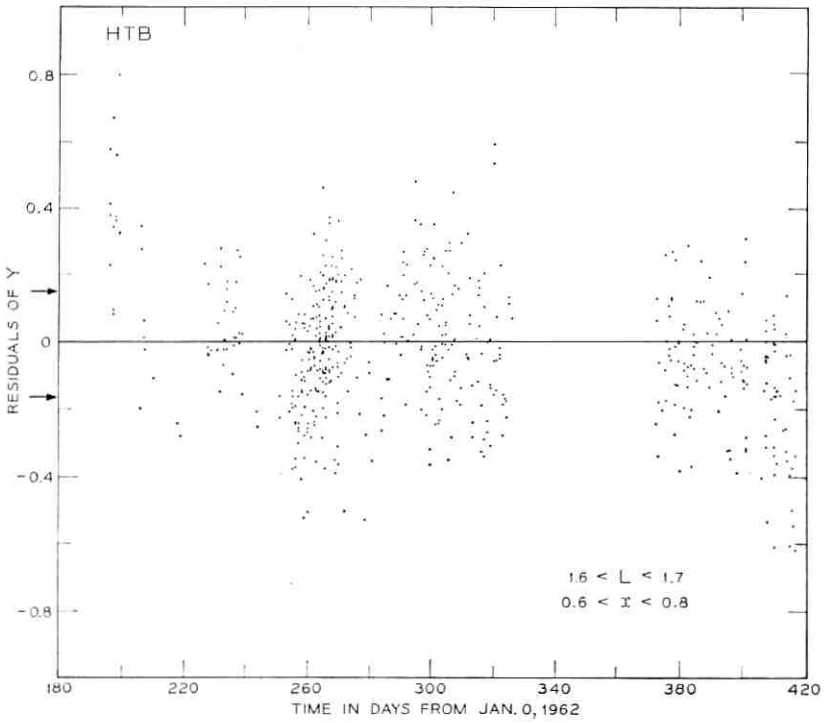


Fig. 42—HTB residuals of Y (i.e., $Y - y$ calculated from the HTB coefficients) plotted against time for $1.6 < L < 1.7$ and $0.6 < x < 0.8$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

TABLE VI—FRACTIONAL INCREASE IN FLUX BETWEEN DAYS 280 AND 310, 1962.

L \ x	0.1	0.3	0.5	0.7	0.9
<1.9	—	—	—	—	—
1.95	0.05	0.07	0.12	0.20	0.70
2.05	—	—	0.37	0.46	0.90
2.15	—	—	0.28	0.33	—

Table VI gives the fractional increase in the average counting rate (Y^2) during this period at various values of x and L. By $L = 2.25$ the change is barely observable and for $L > 2.3$ it has disappeared. The data acquired between days 325 and 373, which are not included among the HTB data because the bias voltage was low, were ex-

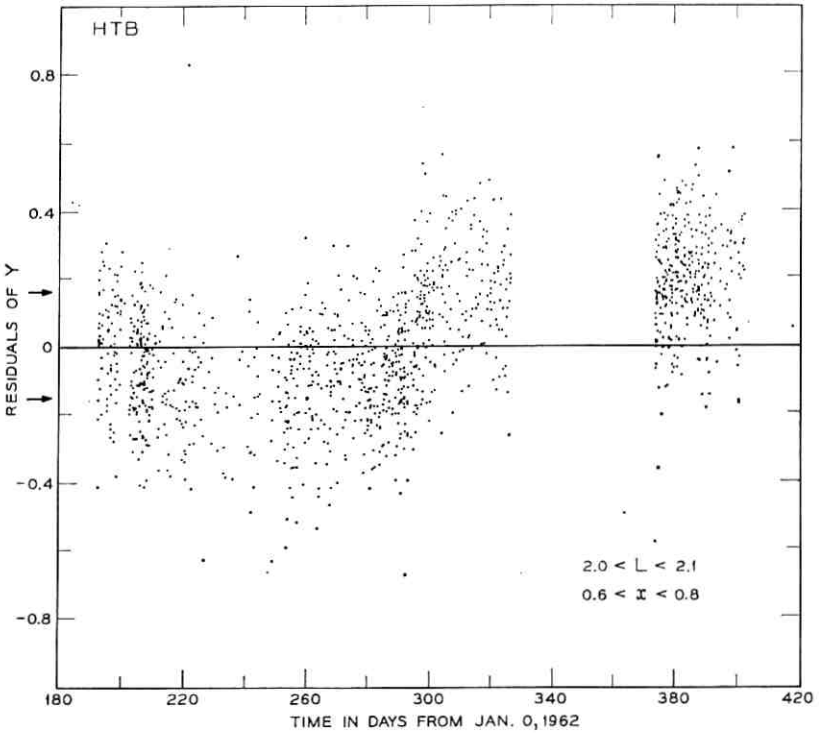


Fig. 43—HTB residuals of Y (i.e., $Y - y$ calculated from the HTB coefficients) plotted against time for $2.0 < L < 2.1$ and $0.6 < x < 0.8$. The arrows indicate $\pm$ the approximate standard deviation if Y^2 were Poisson distributed.

amined; and there appears no reason to believe that there were any changes in the intensity of the >50 MeV protons for $L > 1.9$ during these 48 days.

While it is not possible to be quite sure that we are observing a "true" temporal effect, it is difficult to contrive any alternate explanation. This event can be compared with the changes produced in the high energy proton distribution by the magnetic storm of September 22, 1963, and observed with Relay 1[25] and the *Telstar*® 2 satellite.[5] In both cases only L shells with values above 1.9 were affected, and the effect is more pronounced at higher x's. However, the storm produced a decrease in flux whereas an increase was observed in 1962; the effects of the storm were more severe at larger L's, whereas in this event, a maximum fractional change was observed near $L = 2.05$; and the effect of the storm was sudden, i.e., the flux decrease took place within 24 hours, while the increase observed in 1962 was gradual and required a month to complete. Increases in flux having some of the features described here were observed with Explorer 7.[26] However, it is difficult to be certain that those increases were caused by protons with energies above 18 MeV, rather than electrons with energies greater than 1.1 MeV.

The high-energy protons appear very stable over the seven months covered by our data. In particular, no effects associated with the USSR high-altitude nuclear tests of October 22, October 28, and November 1, 1962, or the large magnetic storm of December 18, 1962 have been observed.

In summary, changes through time in the observed values of the flux are generally less than 20 percent, although they may be larger in some regions of space. We have not been able to detect a diurnal effect. Often, secular changes are not separable from other variables, an exception being an apparent change in the position of the cutoff. An event which appears to comprise a measurable redistribution of the proton flux over an appreciable volume of space and period of time has been noted. We do not know whether the redistribution is in energy or space, and find no indication of the mechanism in the data.

XI. THE CUTOFF

As discussed in Sections V, 6.3, and 7.5, the cutoff function, $x_c(L)$, is defined in terms of our instrument, model and fitting procedure. For $L < 2$, the value of $x_c(L)$ corresponds to the position on the given L

shell at which the omnidirectional flux is of the order of 1 proton/cm² sec, more than three orders of magnitude below the highest flux in the belt. However, because the flux is falling so fast with x, this position is almost certainly very close to the place at which the flux becomes 0. The last statement is not true for $L > 2$. Here, although the value of $x_c(L)$ (the place at which $y = 0$) still corresponds to the point at which the limit of sensitivity of our instrument is reached, the position of x_c is not so well-defined by the fit. In addition, one has only to examine Fig. 23 to realize that x_c may be significantly removed from the value of x at which the flux falls to zero.

The Model-I HTB coefficients of Table IV define the cutoff function, and we have made use of a modification of R. H. Pennington's mirror trace program* to calculate the minimum altitude corresponding to $x_c(L)$ for $L < 2.2$. This inversion was accomplished using the Jensen and Cain magnitude field coefficients for 1960,[13] the same set used to calculate x and L (see Table I). (Other sets of coefficients are available.[27] However, using the GSFC (7/65) coefficients[28] does not produce significantly different altitudes.)

The minimum altitude is smallest in the Southern Hemisphere over the Atlantic Ocean. Fig. 44 shows the results in graphical form. The minimum altitude is $\approx$270 km near the equator ($L = L_0 \approx 1.13$), decreases to a minimum of $\approx$160 km at $L = 1.6$, and increases very rapidly thereafter. For L less than 1.5, the standard error in altitude, derived from the standard error in x_c (see Fig. 22), is about 10 km, which is roughly the accuracy of the inversion procedure as we used it. The standard error in altitude for $L > 1.5$ is indicated by the dashed lines in Fig. 44. At $L = 2$, where the cutoff mechanism is only partially atmospheric, the standard error is nearly 50 km.

The minimum near $L = 1.6$ in the altitude curve of Fig. 44 appears to reflect the existence of the South American magnetic anomaly. Although $R_c(L)$ [see (5)] increases monotonically with L for $L > 1$, the increase is apparently not fast enough to override the influence of the anomaly. This result is true for all the sets of coefficients produced in many trial fits as well as for the HTB coefficients in Table IV. We have not yet carried out the obvious next step of averaging the atmospheric density over the orbital path of the protons to see whether or not the shape of Fig. 44 can be explained on the basis of present models of the atmosphere.

Although the shape of the minimum altitude curve remains the

* Kindly communicated to us by D. J. Williams.

same, the value of the altitude is sensitive to the method of selecting the sample (see Section 7.1). For example, the minimum value of altitude calculated from the CB coefficients is 100 km (again at $L = 1.6$), 60 km lower than the 160 km calculated from the HTB coefficients. The weighting of the HTB sample emphasizes the high x data and gives better representation, and therefore a better expectation of fitting well, near the cutoff. However, the *Telstar*® 1 satellite, with its eccentric orbit and relatively high (950 km) perigee, could not give detailed information about particles near the top of the atmosphere, and this is reflected in the results of the analysis.

In conclusion, the curve of Fig. 44 probably represents the quali-

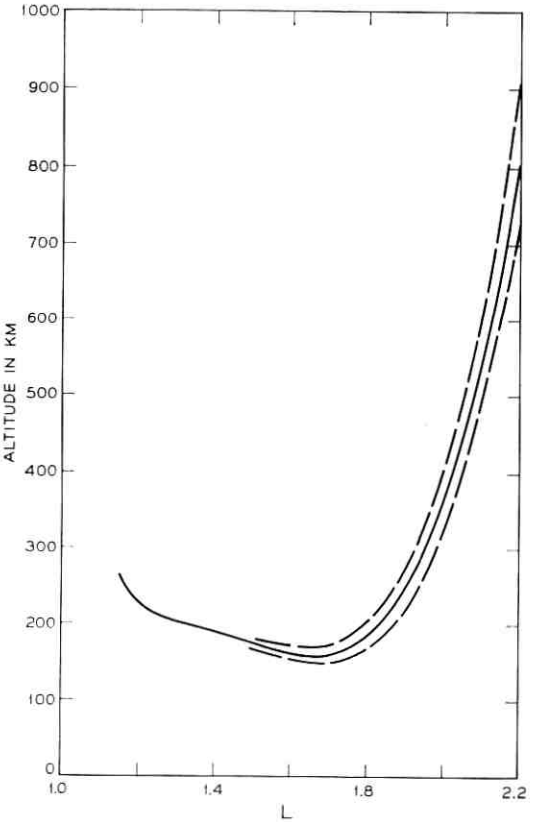


Fig. 44 — The minimum altitude reached by > 50 MeV protons as a function of L. This altitude is determined in geographic coordinates from the transform of $x_c(L)$. The dashed curves are ± one standard error.

tative behavior of the minimum altitude of the cutoff reasonably well, but the uncertainty in the value of the altitude is larger than a simple examination of the standard error plotted in the figure would lead one to believe. The implications of these results for the details of the cutoff mechanism have not been examined in detail; however, it is clear from the sudden upturn of the curve in Fig. 44 that the mechanism is principally atmospheric for L less than about 1.9 and principally nonatmospheric on higher L shells.

XII. COMPARISON WITH OTHER WORK

12.1 *Introduction*

When making comparisons among the various high-energy proton measurements it is desirable that the results be extensive in time and space, reported in terms of omnidirectional fluxes at various positions, and that these positions be expressed in magnetic coordinates derivable from the B,L set. A list of some experiments which meet these desiderata is given in Table VII.

Following a presentation of flux maps, comparisons among these experiments are made with respect to the following features: the absolute intensity at one point in the belt, as close to the maximum of intensity as is practical; the intensity vs L in the equatorial plane; the behavior of the intensity on selected L shells; the flux near the top of the atmosphere, and the equatorial pitch angle distribution. Comparisons covering a larger range of proton energies have also been made by Vette[29] and Fillius.[20]

One of the difficulties encountered in making comparisons among the various bodies of data is that most of the results have been published in graphical form, rendering it necessary to scale numerical values from small plots, an inaccurate procedure at best. A welcome exception is the Explorer 15 data, which McIlwain[18] has made available by means of a series of interpolation functions in the form of a FORTRAN computer program.

12.2 *Telstar® 1 Flux Maps*

For this discussion, the *Telstar®* 1 HTB results have been converted to omnidirectional flux, J, where $J = 4\pi y^2/\bar{g}$. (Note that the value of $\bar{g}$ derives from the assumptions of Appendix A regarding the energy spectrum.) This procedure provides an estimate of the flux of protons with energies between 50 and 130 MeV at positions, (x, L), in mag-

TABLE VII—SOME SATELLITE MEASUREMENTS OF THE HIGH-ENERGY TRAPPED PROTONS.

Satellite	Approx. period covered in reference	Orbit perigee, R_e apogee, R_e incl, deg	Instrument	Approx. energy range	References
Explorer 4	7/26/58 to	1.041	Anton 302 Geiger	> 43 MeV	9
1958 ϵ1	*	1.347	tube (shielded)		30
		50			
Injun 1	7/61 to	1.14	Anton 213 Geiger	> 40 MeV	31.
1961 o2	12/61	1.16	tube (shielded, SpB)		
		67			
1961 $\alpha\delta$1	10/21/61 to	1.59	scintillator	> 59 MeV	32.
(H2)	**	1.55			
		96			
1962 κ1	4/9/62 to	1.54	scintillator	> 59 MeV	32.
(H3)	**	1.44			
		87			
Telstar® 1	7/10/62 to	1.15	solid-state	50–130 MeV	
1962 $\alpha\epsilon$1	2/21/63	1.90	detector		
		45			
Explorer 15	10/27/62 to	1.049	scintillator	40–110 MeV	18
1962 $\beta\lambda$1	1/27/63	3.72			
		18			
Relay 1	5/1/64 to	1.21	scintillator	> 35 MeV	33.
1962 $\beta\nu$1	9/22/64	2.14			
		48			
Injun 3	12/24/62 to	1.037	scintillator	40–110 MeV	19
1962 $\beta\tau$2	9/28/63	1.44			
		70			

* Not stated. Re-entered atmosphere 10/23/59.
** Not stated.

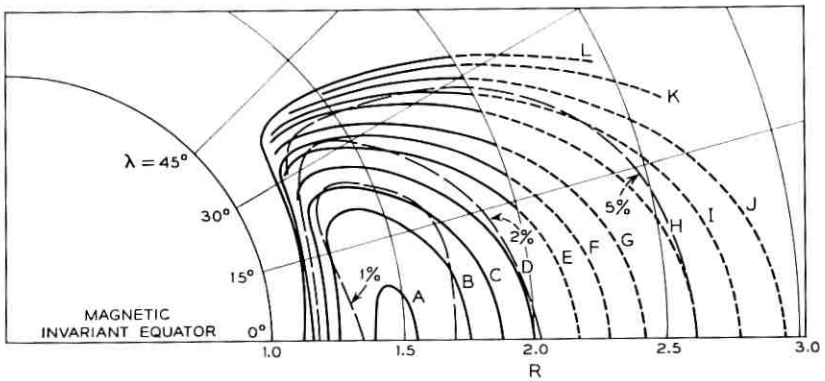


Fig. 45 — Omnidirectional isoflux contours derived from the HTB coefficients and plotted in R,λ space. Dashes indicate extrapolation beyond the region in which data were acquired. Long dashes form contours of constant percent standard deviation.

Label	A	B	C	D	E	....	L
Omnidirectional flux (J)	5×10^3	2×10^3	1×10^3	5×10^2	2×10^2		1×10^0 protons/cm² sec

netic space on the basis of the presently provided model and fit to the HTB data.

For ease of reference, *Telstar*® 1 HTB flux maps are presented in three commonly used forms: Fig. 45 shows contours of constant flux in R,λ coordinates; Fig. 46, contours of constant flux in B,L coordinates; and Fig. 47, log flux vs log B curves for various values of L. These three graphs give an overall picture of the particle distribution. In these figures, dashed lines are used to indicate the extrapolation of fitted values to regions not penetrated by the satellite. Note the way the geometry of the coordinate transformations affects the extrapolated regions. In particular, the functional extrapolation in B,L coordinates gives much more curvature to the contours than might be anticipated. The difference between the functional and straight line extrapolation in B,L can be as large as a factor of 2 in the flux (a shift of 0.2 in L) at $L = 3$. Except for the region of the secondary local maximum in the flux near $L = 2.2$, this functional extrapolation compares surprisingly well with the measurements made on higher altitude satellites.[5, 18]

In the altitude range covered by the data, a single maximum is observed. This maximum in the omnidirectional flux of $\approx 6 \times 10^3$ protons/cm² sec is located on the magnetic equator at $R = L = 1.46$.

The intensity falls abruptly near the bottom of the belt (the top of the atmosphere) and decreases more gradually toward the sides and top of the belt. On a given L shell, the intensity is a maximum at the magnetic equator, and deceases monotonically as the distance from the equator increases.

Neglecting the uncertainties in the calibration of the instrument (−25 to +50 percent), which are discussed in Appendix A and are mentioned in the next subsection, the estimated standard deviation of the estimate of J is less than 2 percent of J over much of the region of space discussed in this section. Smoothed contours of 1 percent, 2 percent and 5 percent standard error are plotted as the dotted lines in Fig. 45. Near the cutoff, where the counting rate is falling to zero, the standard deviation in x_c (see Figs. 44 and 22) is a useful indication of uncertainty in the flux. Other information concerning

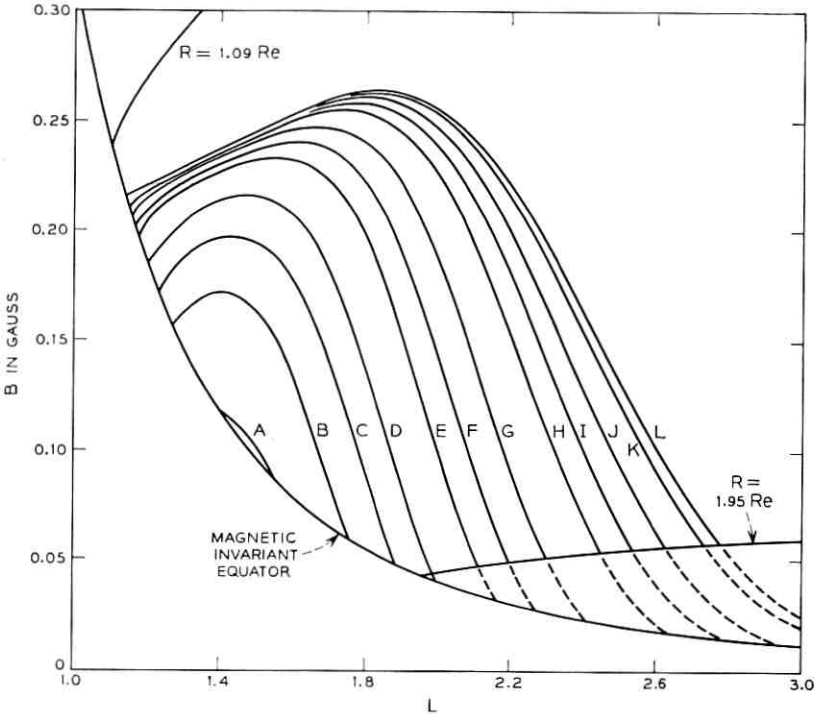


Fig. 46 — Omnidirectional isoflux contours derived from the HTB coefficients and plotted in B,L space. Dashes indicate extrapolation beyond the region in which data were acquired. Labeling is given in Fig. 45.

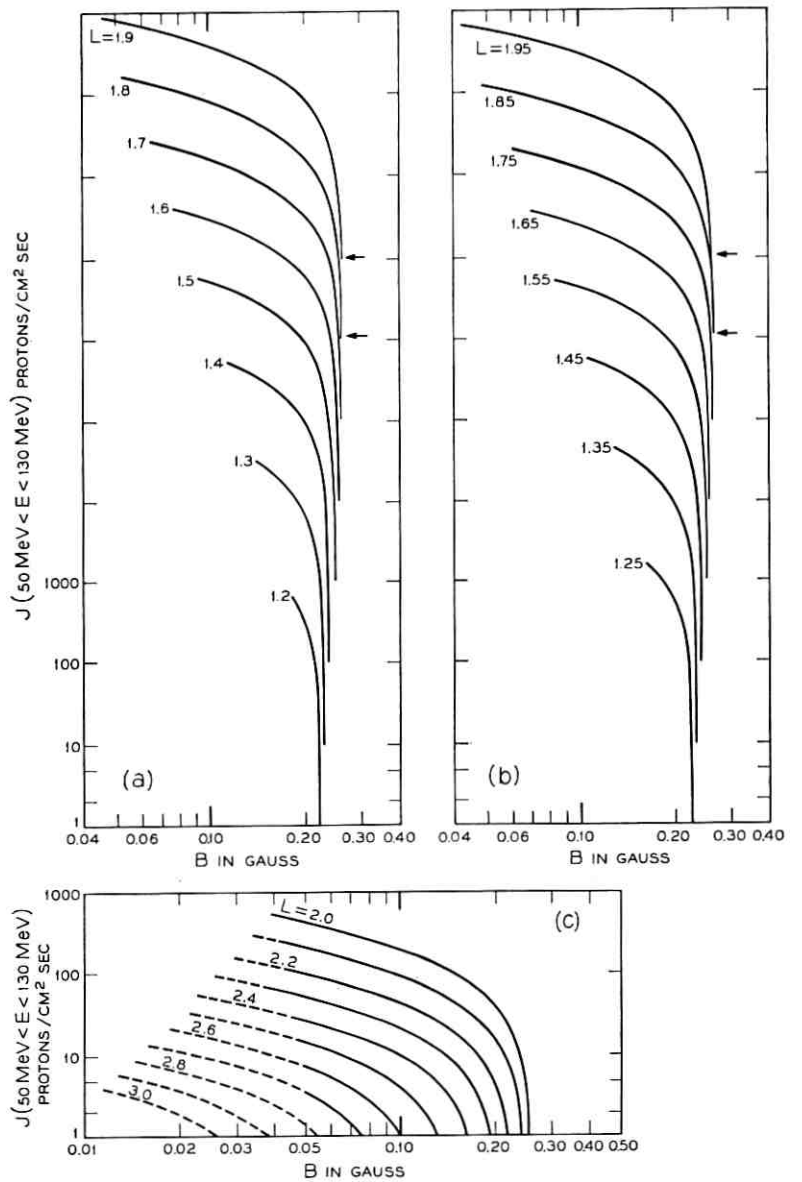


Fig. 47 — Omnidirectional flux on several L-shells derived from the HTB coefficients and plotted against log B. Adjacent curves in parts (a) and (b) are slipped one decade in J. All curves rise from $J = 1$ proton/cm² sec. Dashes indicate extrapolation beyond the region in which data were acquired.

standard deviations may be found in Sections 7.4 to 7.6 and 8.4, Figs. 22, 23(a), 23(c), and 24.

The equations defining Model II (see Section 4.6) and coefficients of Table V, together with the transformation equations among various magnetic coordinate systems, allow accurate relative flux values to be easily calculated in any coordinate system.

12.3 *Comparison of Absolute Intensities*

The solid curve in Fig. 48 is the fitted omnidirectional equatorial flux of 50–130 MeV protons measured by the *Telstar*® 1 satellite. The points are fluxes observed on other satellites (Table VII) at the mag-

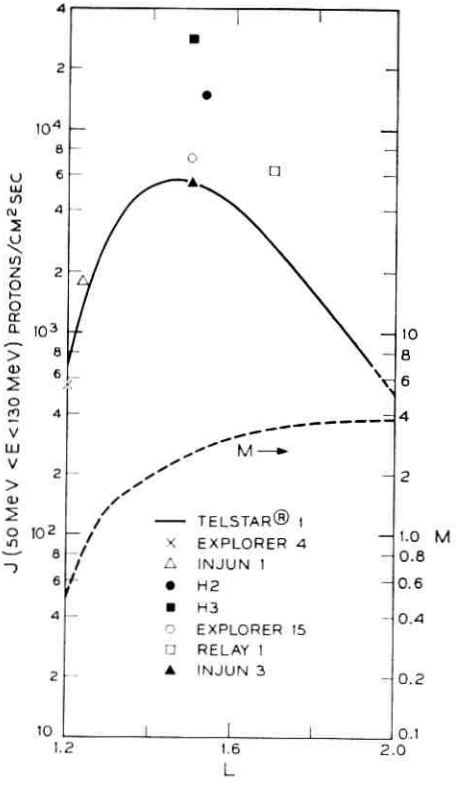


Fig. 48 — Values of equatorial omnidirectional flux, for the satellites indicated in the legend, corrected to the energy range 50–130 MeV and plotted at the appropriate value of L. An integral power-law energy spectrum [see (17)] of exponent $-M$, where M is given is a function of L by the dashed curve, was used in making the corrections. References are given in Table VII.

netic equator and corrected to 50–130 MeV by using a single-component integral energy spectrum of the form

$$N(>E) \propto E^{-M} . \tag{17}$$

The values of M at the magnetic equator are plotted as the dashed line in Fig. 48. These values were taken from Gabbe and Brown,[5] and are consistent with those of Brown, Gabbe, and Rosenzweig,[3] and also those of Fillius and McIlwain,[34] and Freden et al,[35] where the data overlap. Because of uncertainties in the geometric factors of the detectors (see Appendix A) and changes in the belt with time (see Section X), one might expect agreement only within a factor of about 2. On this basis the agreement in absolute intensity is quite reason-

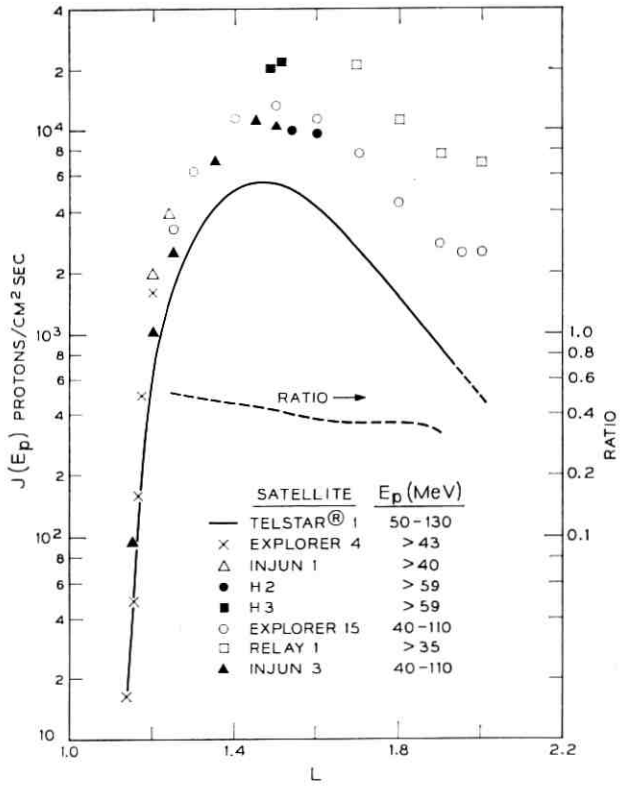


Fig. 49 — Values of equatorial omnidirectional flux, for the satellites and energy ranges indicated in the legend, plotted against L. The dashed curve is the ratio of *Telstar*® 1 to Explorer 15 measurements. References are given in Table VII.

able. However, the *Telstar*® measurements are somewhat on the low side, and those of Imhof and Smith[32] (H_2 and H_3 on Fig. 48) are much higher than the other observations.

The points represent measurements taken before, after, and during the *Telstar*® 1 experiment so it is unlikely that changes in the flux with time explain these differences. It is difficult to account for the discrepancies in absolute flux in terms of the spectral correction, unless more complex spectral forms than those of Appendix A are considered, because the comparisons are among results of detectors whose threshold energies are close to 50 MeV. The most likely sources of the differences are errors in absolute calibration. It follows that a good deal of caution should be exercised in drawing conclusions about temporal effects and energy spectra from measurements made with *different* instruments.

12.4 *Intensity vs L in the Equatorial Plane*

Fig. 49 is a plot of the omnidirectional equatorial flux for each of the satellites listed in the legend of the figure. The data are from detectors having several different energy ranges and no spectral corrections have been made. The general features of the data in these energy ranges have been noted previously in the literature. The flux increases rapidly with L, goes through a maximum near $L = 1.5$ and then decreases. The decrease is not as rapid as the initial rise and in this energy range the flux generally does not decrease monotonically[18, 20] for $L > 2$. Excepting the measurements of Imhof and Smith,[32] the flux decreases with increasing energy, indicating a falling energy spectrum.

The dashed line in Fig. 49 is the ratio of the 50–130 MeV proton flux measured with *Telstar*® 1 to the 40–110 MeV proton flux measured with Explorer 15. This ratio is a good qualitative index of the energy spectrum near 45 MeV, and in these circumstances the change in this index is independent of the absolute calibrations of the instruments. The ratio is seen to decrease monotonically as L goes from 1.25 to 1.9, indicating, in agreement with the references cited in the previous subsection, a softer spectrum* at higher L.

12.5 *Intensity vs B on L Shells*

In Fig. 50 [parts (a), (b), and (c)] measurements from various satellites are compared on the three L shells, 1.3, 1.5, and 1.8. The

* A softer spectrum contains a larger fraction of low-energy particles.

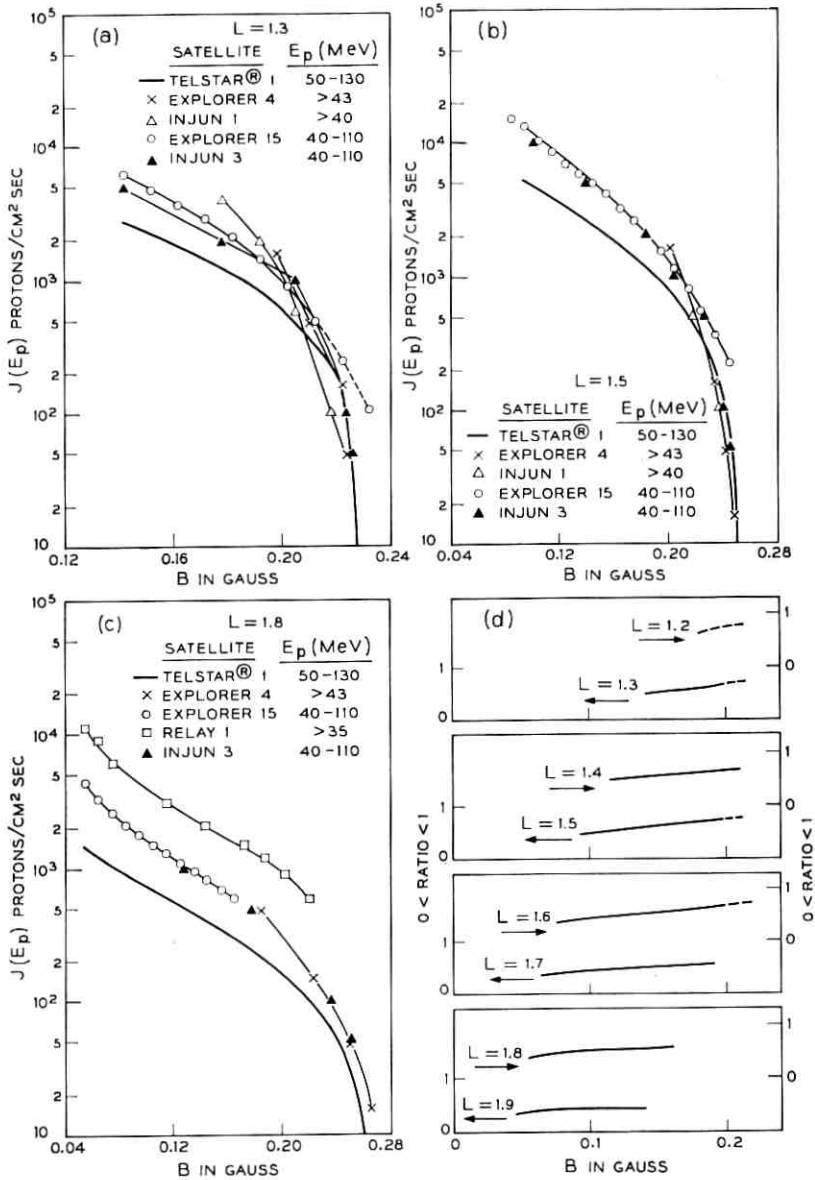


Fig. 50 — Values of omnidirectional flux on various specified L-shells, for the satellites and energy ranges indicated in the legend, are plotted vs B, in parts (a), (b), and (c). Ratios of *Telstar*® 1 to Explorer 15 measurements are shown in part (d). References are given in Table VII.

Explorer 15 and Injun 3 measurements have been compared in more detail by Valerio.[19] Observe that J decreases monotonically with B on all the L shells and the shape of J vs B is very similar for all the measurements on the same shell except for the lowest L shell where the dependence on the energy response of the detector is most important. Information concerning the energy spectrum near 45 MeV is contained in the changes in the ratios of the measurements, and in these circumstances the changes are independent of the absolute calibrations of the instrument.

To cast more light on the qualitative behavior of the energy spectrum, the ratio of the 50–130 MeV proton flux measured with the *Telstar*® 1 satellite to the 40–110 MeV proton flux measured with Explorer 15 has been calculated as a function of B for fixed L. The results are plotted in Fig. 50(d). All the ratios increase with increasing B for L from 1.2 to 1.9 inclusive. The values of B in the plot cover the range from the magnetic equator to a magnetic dipole latitude (λ) of about 30°. The increase in the ratio indicates a spectrum that hardens with increasing B in the neighborhood of 45 MeV. At $L = 1.8$ Freden et al[35] find a spectrum that hardens with increasing B for proton energies between 10 and 35 MeV, but softens with increasing B for proton energies above about 55 MeV. Our results suggest that this change in behavior cannot have occurred below 50 MeV.

12.6 *The Intensity Near the Top of the Atmsophere*

The position of the 8-protons/cm² sec flux contour from the *Telstar*® 1 satellite is plotted in B,L coordinates in Fig. 51 (a), together with our own extrapolation of the published Injun 3 data[19] to a flux of about 10 protons/cm² sec,* and the 16-proton/cm² sec flux contour from Explorer 4. The purpose of this figure is to test whether or not the altitude dependence of contours of constant counting rate at low altitudes is consistent with other data. The qualitative agreement of the results plotted in Fig. 51(a) is quite good, especially for $L < 1.8$, where the atmosphere is controlling. A number of effects may contribute to the divergence of the results for $L > 1.8$. Among them are: temporal effects, this region of the belt is shown to be subject to temporal variations in Section X; instrumental effects, the instruments are near their threshold sensitivities in a region of magnetic space in which the energy spectrum may be anomalous; and biases in the fitting procedure,

* Valerio[19] states that his fits (and therefore his Fig. 8) are not intended to represent the data accurately at low altitude.

examination of residuals give some indication of a slight bias in the fitted function in this region.

It is difficult to get direct insight into the altitude dependence from a B,L plot, so the values of B have been transformed into mini-

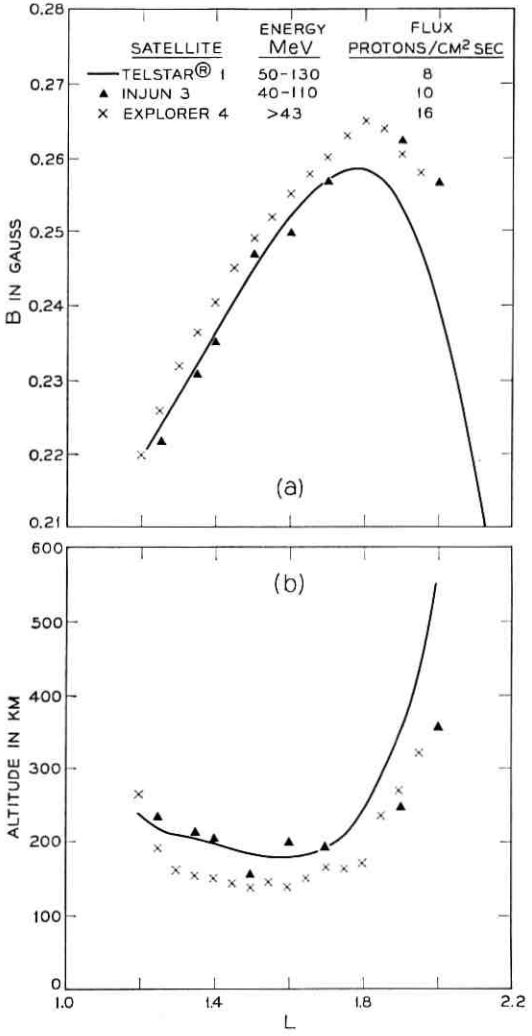


Fig. 51 — Comparison of isoflux contours obtained from three satellites near the top of the atmosphere. Part (a) B, L coordinates, part (b) minimum altitude (near the South American magnetic anomoly). References are given in Table VII.

mum altitude by using the procedure mentioned in Section XI. The minimum altitudes are plotted against L in Fig. 51(b). It is characteristic of all three bodies of data that the minimum in the minimum altitude curve does not occur at minimum L.

It is tempting to consider whether the lower altitude of the Explorer 4 points, coupled with the lower low-energy threshold and high flux associated with the Explorer 4 measurements, might imply that the exosphere was less dense when the Explorer 4 measurements were made. However, the uncertainty in the position of the *Telstar*® contour (see Section XI) is so large that the use of this figure to refute the hypothesis that the atmosphere contracted[23] between 1958 ($\approx$ solar maximum), when the Explorer 4 measurements were made, and 1962, when the *Telstar*® data were taken is precluded, even if one were prepared to overlook the possibility that the energy spectrum at these low altitudes is anomalous[36] and consequently that the calculated geometric factors of the instruments may be in substantial error near the cutoff.

12.7 *Equatorial Pitch Angle Distribution*

The solid curves in Fig. 52(a) represent the equatorial pitch angle distributions, at various values of L, calculated from (8) and the coefficients in Table V. When these are compared with the equatorial pitch angle distributions obtained from the Injun 3 data,[19] which have been replotted as the dashed curves in Fig. 52(a), they are found to be very similar in shape, although the *Telstar*® curves are a trifle flatter. This would be anticipated from the previous discussion of the tendency of the energy spectrum of protons with energies near 45 MeV to harden at high values of B. The shape of the distributions are, however, appreciably different from those derived by Lenchek and Singer[37] from consideration of possible injection and loss mechanisms. This may be seen in Fig. 52(b) which contains the present results as the solid lines, and the results of Lenchek and Singer[37] as the dashed lines.

12.8 *Other Bodies of Data*

A portion of the considerable body of relevent high-energy proton data, some of which does not meet the requirements for inclusion in Table VII, is noted here. The earliest measurements of proton intensities were made on Explorers 1 and 3 by Van Allen.[38] His historic estimate of $\approx 2 \times 10^4$ protons/cm^2 sec with energies >40 MeV at the heart of the inner belt ($x = 0$, $L \approx 1.56$) has been substantiated by all the measurements reported to date. In particular, the high-energy proton

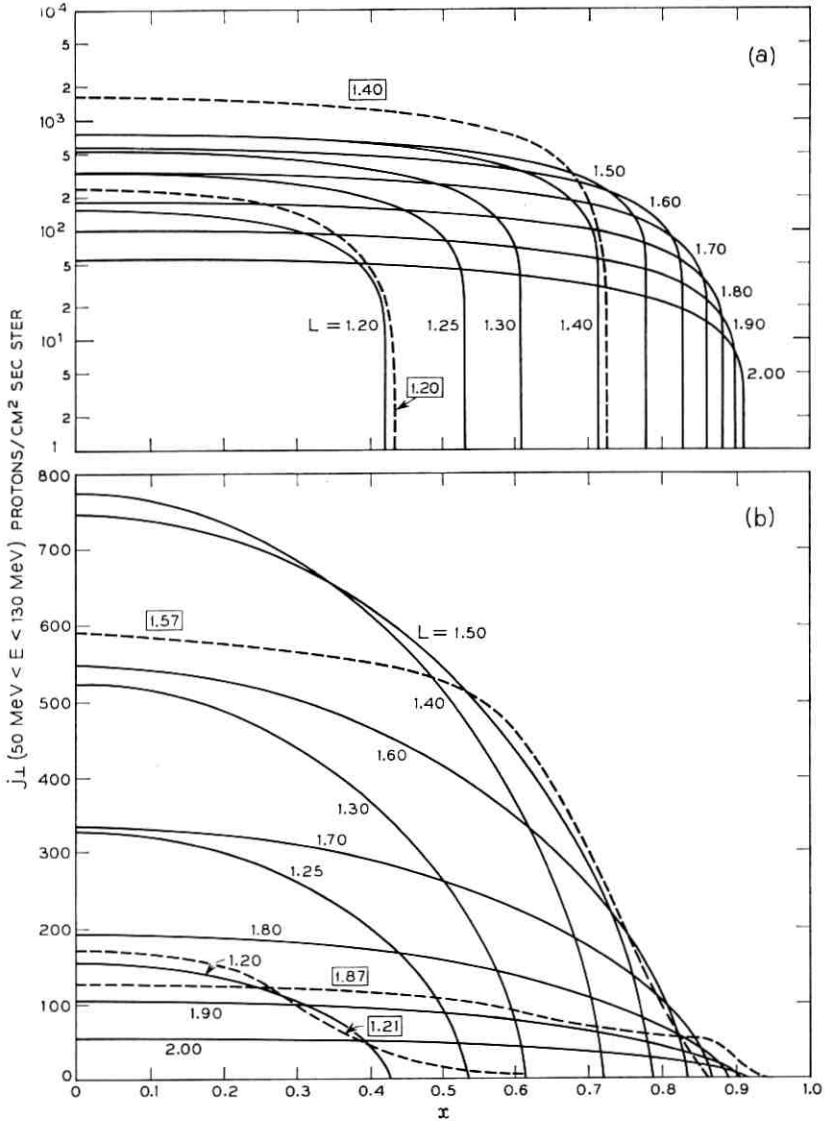


Fig. 52 — Unidirectional flux vs x on various L-shells. The solid curves are derived from the HTB coefficients using (8). The dashed curves in part (a) are Injun 3 results (from Valerio[19] Fig. 8). The dashed curves in part (b) are the results of the theoretical calculations of Lenchek and Singer,[37] taken from their Fig. 10 and arbitrarily normalized to reasonable values of j at $x = 0$.

measurements made in the inner belt by Explorers 6, 12, and 14 and Pioneers 3 and 4, have been noted by Frank et al[39] to agree with each other and with those on Explorers 1 and 3. Reference to some measurements made with ballistic probes may be found in the article by Freden et al.[35]

XIII. QUO VADIS

The mathematical model which has evolved along the lines summarized in Section IV has provided a very satisfactory representation of the high-energy proton data from the *Telstar*® 1 satellite, as discussed in both statistical and physical terms in Sections VI through XII. It is appropriate to consider how this work might be extended.

13.1 *Further Improvements within the Present Scheme*

The final fit of Model II has a mean square error which is less than twice the variance to be expected on the assumption of a Poisson distribution of the count data. Some of this excess is surely due to "experimental error." However, one might seek some additional improvement by the addition of more parameters to the fitting function as indicated in Model III of Section 4.7. Such fits, carried out on an approximately 1000-point selected data set, will almost surely lead to a reduction in the mean square residuals because of the increased freedom the additional parameters provide. However, as noted in Section 4.7, preliminary work with Model III has not led to a really substantial improvement, either statistically or aesthetically as judged by plots of the residuals.

Additionally, one might try to improve further on the representativeness of the sample by simple iteration. Using the HTB fit to Model II to determine new x,L cells, another sample might be selected and fitted. The very small differences between the Model-I CB fit and the Model-I (or II) HTB fit do not suggest that this would be fruitful in the present case. If the preliminary fit used for determining the x boundaries of the cells were a poorer fit, iteration would clearly be worthwhile.

A further extension of the procedure for designating representative cells would involve the development of a two-dimensional version of the basic idea and procedure outlined in Section 7.1. Specifically, one would try to define approximately 1000 x,L cells within each of which the preliminary fit to $y(x, L)$ has the same range. In the present case, the anticipated gain from this refinement did not seem to justify the practical difficulties. However, a practical, well-defined algorithm for

such a process in several dimensions simultaneously might prove very useful.

13.2 *Another Approach to the Model*

All the models presented so far are of the form

$$y(x, L) = A(L) \cdot b(x; e_i(L)), \tag{18}$$

where $A(L)$ represents the variation in intensity along the magnetic equator and $b(x; e_i(L))$ represents the variation with x on an L-shell. The $\{e_i(L)\}$ adjust the nature of the dependence on x, as a function of L. This approach arises from the L-shell orientation of the adiabatic theory of trapped particle motion.

Alternatively, one might focus attention on the shape of y as a function of L at constant x, rather than on y as a function of x at constant L. It is shown in Fig. 19(a) and discussed in Section 7.2 that $y(x, L)$ as a function of L for fixed x forms a simple nesting set of curves at successive values of x. This is a consequence of the monotonic decrease of y with x at any fixed L. With this orientation, a model might be expressed as:

$$y(x, L) = F(L; p_i(x)), \tag{19}$$

where $F(L)$ is the shape of a constant-x section, whose parameters, the $\{p_i\}$, are expressed as functions of x. Although this approach would not contain the L-shell orientation of the particle motion explicitly, it seems to offer very significant practical possibilities.

13.3 *Full Data Utilization*

In the two-dimensional fits that were carried out, only a selected set of data were used, either chosen at random within a set of narrow, contiguous L-slices, as in the fit of Section VI, or chosen on the basis of a preliminary fit to the data as in Section VII. All the data were examined by residual plots and mean square residual measures of the fits, but only a small part of the data were actually used in determining the values of the fitting parameters. With this procedure, information is clearly being lost that could be used to "better" determine the function.

Several methods have been applied in the past to allow all of an existing body of satellite data to influence the mathematical description of that data. The most direct method uses interpolation or smoothing functions. It is often the case that consecutive satellite observations from a particular detector are closely enough spaced to determine

the local spatial variation. Under these conditions a sequence of data points can be averaged or fitted to a local smoothing function. A number of points in a sequence may thus be replaced and represented by a single point which is determined by them all. The replacement may also be made at some particularly convenient coordinate location, for example, at one of a fixed set of L or x values on which functional fitting may subsequently be carried out. This method has been used on the data of Explorer 15, portions of which have been described by McIlwain,[18] Roberts[10] and Brown.[41]

In the context of the high-energy proton data from the *Telstar*® 1 satellite, a different but analogous procedure could be used. Rather than selecting at random one data point within each of approximately 1000 x,L cells, all points within a given cell could be used to determine a value which would represent the observable at the central point of the cell. This might be done by simply averaging the points within the cell, but the cell size is large enough so that the x and L dependence within the cell generally cannot be neglected. A more representative procedure would be to fit the points within an x,L cell with a local smoothing function. This function can be the same function with which the finally selected data values would be fitted across the complete range of x,L space (see Appendix B.7). Although in the present case the average number of points per cell is about 40, in many cells the number of points is fewer than the number of coefficients of the Model II function, and some coefficient constraint would be required. This is not a substantial objection, however, since the function is only being used for smoothing and does not need to be capable of elaborate variation over an x,L cell.

A procedure of this kind greatly reduces the chance that members of a final 1000-point set will be nonrepresentative and acknowledges the experimental weight of adjacent observations in fixing the values of the set. Accordingly, one would expect a reduction in the mean square residuals overall the data, from a fit to such a smoothed sample.

The procedure of smoothing within a cell could be used with larger x, L cells (with more points per cell) to define a point set smaller than 1000. It can of course also be used with much larger bodies of data up to a maximum of 1000 points per cell with the existing computer program.

13.4 *Extension to Other Cases*

There are very evident values in being able to communicate the essence of a large body of data in terms of a mathematical model with

a small number of coefficients. This is very effectively accomplished by the present empirical representation of the *Telstar®* 1 high-energy protons, but the model is very specialized. As previously noted, including a wider range of space such as that explored by the *Telstar®* 2 satellite requires modification of the function. Characterizing the proton distribution for substantially lower energy protons may well require functions outside the generality of even Model III. Treating electrons in almost any region of space requires treating time as well as position variables because a complete set of measurements of the spatial distribution of the particles cannot readily be obtained in a time short compared with significant time variations.

No single formulation yet exists which is capable of coping in a useful way with the range of measurements of particles trapped in the magnetic field of the earth. However, the success of the present formulation as it has been evolved and the general methods that have been developed gives us confidence that other and more complicated cases can be treated.

XIV. SUMMARY AND CONCLUSIONS

This section provides a summary, with references, for the entire document including the appendices.

14.1 *General Accomplishment*

The main accomplishment is the development of a relatively simple (empirical) mathematical model which gives a statistically accurate representation of the spatial distribution of high-energy protons measured with the *Telstar®* 1 satellite.

14.2 *The Data*

14.2.1 *Space and Time Coverage* (Sections I and II)

The data were acquired between July 1962 and February 1963 within the region of space bounded by $1.09\,R_e \leqq R \leqq 1.95\,R_e$ and $0 \leqq \lambda \leqq 58°$. Inside these boundaries good temporal and spatial coverage were achieved.

14.2.2 *Energy Range and Instrumental Sensitivity* (Appendix A)

The nominal energy interval of the detector is $50 < E < 130$ MeV and its nominal geometric factor is $0.143^{+0.071}_{-0.036}$ cm^2 ster. The instrument is effectively omnidirectional and the lower threshold of sensitivity is $\approx$1 proton/cm^2 sec.

14.2.3 *Telemetry* (Section II)

Each observation consisted of the number of counts registered in 11 seconds. With this was associated the time at which the telemetry was received, and auxilliary information.

14.3 *The Models*

14.3.1 *Coordinate System* (Section III)

Each model relates the omnidirectional intensity of high-energy protons to a two-dimensional magnetic space whose coordinates, x,L, derive from a mapping of the earth's main magnetic field onto an axially symmetric dipole field through the adiabatic invariants of the particle motion.

14.3.2 *General Form and Properties* (Section IV)

The models have the form of a product, $A(L) \cdot G(x,L)$, in which the first term expresses the equatorial intensity as a function of L, and the second term describes the diminishment of intensity, as a function of increasing x, for fixed L. The functional expressions for G (excluding G''') transform in closed form to equivalent pitch angle distributions.

14.3.3 *Specializations* (Sections IV and IX)

Retrospectively, all the models may be considered to be specializations of Model III, but historically the two-dimensional models evolved from a series of one-dimensional fits on L-slices. These fits led to the L-slice model which was then generalized empirically to the two-dimensional Model I. Model I was in turn specialized to Model II to overcome some statistical (nonlinearities and high correlations) and interpretive difficulties encountered with Model I.

14.4 *Fitting*

14.4.1 *Criterion* (Section III and Appendix B)

The least squares criterion was used in deriving estimates of the 8 (or 9 or 10) coefficients required by the models to fit the data.

14.4.2 *Scale* (Section III and Appendix B)

To stabilize the variance of the observations, the models have been fitted to the square root of the observed counting rate.

14.4.3 *Sampling* (Sections 6.1, 6.9, 7.1 and Appendix B.3)

Coefficients of Models I and II were estimated by fitting samples containing about 1000 of the nearly 80,000 available observations. Sampling is necessary to avoid exaggerating the importance of regions of x,L space where data are abundant, and also for compatability with existing computer programs. A method of sample selection based on a preliminary fit has been developed to provide a good overall representation of the data. Before selecting the sample, the data were partitioned to remove instrumental effects and outliers identified by studying residuals from preliminary fits.

14.5 *Quality of Fit*

14.5.1 *Criteria of Judgment* (Sections VI to IX and Appendices B and C)

Judgments regarding the quality of fit were largely based on graphical studies of residuals, the behavior of the fit at the boundaries of the radiation belt and various statistical measures. Residuals (equal to observed minus fitted), on the square root scale, were particularly useful as sensitive indicators of the quality and nature of the fit.

14.5.2 *Comparisons Among Models* (Sections V and IX)

The L-slice fits give good one-dimensional representations of very limited regions of data. Both the standard errors of the coefficients and the correlations among coefficients are high compared to the corresponding measures derived from the two-dimensional fits. The fits of Models I and II to the 960-point HTB sample are practically equivalent. However, Model II is superior in the following respects: one less coefficient is required, standard errors are uniformly smaller, correlations among the coefficients are uniformly smaller, the index of nonlinearity is very much smaller, and more of its coefficients have a physical meaning.

14.5.3 *Coordinates* (Sections VI and VII)

Plots of residuals vs x, L, time, etc. indicate the general adequacy of x,L coordinates for the organization of the data.

14.5.4 *Quantitative Measures* (Sections VII, VIII, IX, and Appendices B and C)

Typically, the fits account for nearly 99 percent of the variability about the data mean. The mean square error of fit is about $1\frac{1}{2}$ times

as large as would be anticipated on the basis of assumed Poisson statistics. Even in the worst of quite small spatial regions, the mean square residual does not exceed $2\frac{1}{2}$ times the Poisson-based prediction. Probability plotting procedures indicate that the residuals are closely normally distributed and lead to an estimate of the variance which is about twice the Poisson-based prediction.

14.5.5 *General Limitations* (Appendix C)

Statistical examination of all the data, categorized in x,L cells defined from a preliminary fit, indicates that it is unlikely that the fit given by the present model could be significantly improved by any simple modification based on x,L coordinates alone.

14.6 *Numerical Values of Fitted Coefficients, Standard Errors, etc.*

14.6.1 *L-Slices* (Section V)

Coefficient values and other statistics for four L-slices appear in Table II, and values of coefficients for a large number of L-slices are shown in Figs. 8 to 10.

14.6.2 *Models I and II* (Sections VI to IX, also Sections V, XI, and XII)

Model II is the preferred model. Coefficients, standard errors, correlations, and other summary analysis-of-variance statistics appear in Table IV for Model I and Table V for Model II. The coefficient functions: (*i*) square root of average counting rate, $y(x,L)$; (*ii*) square root of average equatorial counting rate, $A(L)$; and (*iii*) position of cutoff, $x_c(L)$; are well-determined and applicable values, standard errors, and correlations appear in Figs. 19 and 24 for $y(x, L)$ (and Figs. 45 to 47 for the flux); Figs. 8, 11, 21, 22, and 30 for $A(L)$; and Figs. 9, 12, 22, 23, and 30 for $x_c(L)$ (and Fig. 44 for altitude).

14.7 *Some Physical Results*

14.7.1 *Flux Maps* (Section XII)

Flux maps are given in B,L and R,λ coordinates and as J,B contours for constant L, based on the fitted model and using a calibration of the detector assuming certain single-component energy spectra. Neglecting uncertainties of calibration, the *relative* fluxes have a standard error of about 2 percent. The value of the maximum flux is $(5.7^{+1.4}_{-2.8}) \times 10^3$ protons/cm^2 sec at $L = 1.46$ on the magnetic equator.

14.7.2 *The Cutoff* (Section XI)

The minimum geographical altitude corresponding to the fitted cutoff function was computed. This altitude varies as a function of L and has a value of about 270 km at the magnetic equator at $L = L_0 = 1.13$ and a minimum of about 160 km at $L = 1.6$. The shape of this L dependence suggests that the interaction between the protons and the residual atmosphere is of major importance in determining the cutoff for values of L less than 1.9. For larger L values, the loss mechanism determining the cutoff is of different origin.

14.7.3 *Temporal Effects* (Section X)

The general spatial distribution of high-energy protons is very stable in time over the period covered by the present data; however, using residuals as a sensitive indicator, we find two temporal effects that are distinguishable from instrumental effects. Firstly, there appears to be an increase in the flux in the $1.9 < L < 2.2$ region during the 30-day period starting about day 280, 1962. This increase varied from about 5 to 90 percent depending on both x and L. Secondly, there is an indication of a qualitative increase in the altitude of the cutoff over the period of the observations. The present results indicate that any diurnal variability of the earth's magnetic field would have an upper limit of 0.003 Gauss at $L \approx 1.5$.

14.7.4 *Comparison with Other Experiments and Theory* (Section XII)

The absolute fluxes measured in this experiment agree well (within a factor of two) with other extensive experimental measurements, but the present values are in general slightly lower. Spatial distribution of the flux agrees very well with other measurements but differs appreciably from published theoretical calculations.

14.8 *Extensions* (Sections XIII, IV, and Appendix B)

The methods developed in this work have lead to a very satisfactory representation of the high-energy proton data from the *Telstar*® 1 satellite.

With the better methods of utilizing data and selecting samples noted in this paper, and with more general functional forms (some approaches to which have been indicated), it should be possible to represent the radiation intensity for other more extensive and less "well-behaved" bodies of data than the one treated here. Most aspects of the statistical methods developed are generally applicable to problems of modeling data mathematically.

XV. ACKNOWLEDGMENT

The authors are indebted to the many people who have made substantial contributions to various phases of this work. In particular, we thank C. S. Roberts for enlightening discussions regarding various aspects of the model; Mrs. M. Becker, Miss E. A. Blake, Mrs. N. L. Graham, Mrs. A. S. Michaels, and Mrs. M. F. Robbins, for their contributions to the data analysis; D. R. Cox for the suggestion for distinguishing 'coefficient redundancy' from poor 'design'; and J. W. Tukey for the suggestion to examine the mean square deviation of residuals in the x,L cells and for editorial advice.

APPENDIX A.

The Instrument

Energetic electrons and protons were measured on the *Telstar*® 1 satellite by a group of detectors in all of which the sensitive element was a phosphorous-diffused silicon diode specially developed for such particle measurements.[7] The active volume of the device is the disk-shaped space-charge region of the diode under reverse bias. For the detector measuring protons with energies above 50 MeV, the reverse bias was approximately 100 volts, the space-charge region was approximately 2.8 mm in diameter and 0.39 mm thick, and the diode was shielded by about 12 mm of aluminum over a solid angle of 2π and somewhat more than 12 mm of aluminum equivalent over the remaining hemisphere (see Fig. 53).

The thickness of the space-charge region of the detector was measured with protons from a cyclotron. A calculation of the path-length distribution for unscattered particles in the space-charge region and in the surrounding shielding materials has been made. These calculated results have been combined with range-energy information, and the properties of the associated electronic circuits, to give the geometric factor of the instrument, $g(E)$, as a function of the energy, E, of protons incident on the spacecraft. The geometric factor varies with the reverse bias voltage and the temperature of the detector, both of which affect the effective thickness of the active volume of the diode. Fig. 54 is a graph of $g(E)$ vs E for a bias voltage of -97.5 volts and a temperature of 20°C, the nominal operating conditions of the instrument. Note that protons with energies below 50 MeV were not detected.

The geometry of the detector and shield is only approximately omni-

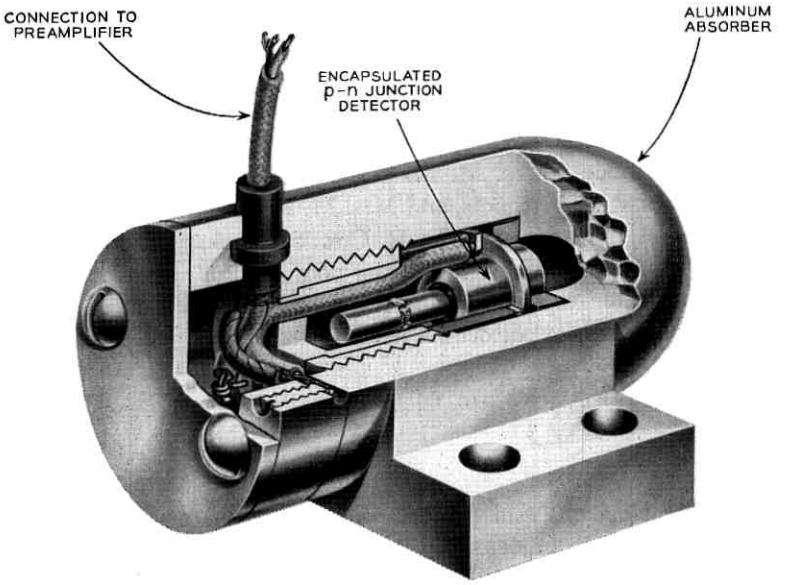


Fig. 53 — The instrument.

directional. However, the satellite was spin stabilized, the symmetry axis of the detector was nearly perpendicular to the spin axis of the satellite, and the telemetered counting rate was an average over at least 15 revolutions of the satellite. This averaging process tends to remove any directionality inherent in the detector geometry. A sensitive analysis noted in Section 7.10 failed to show any directional dependence in the data.

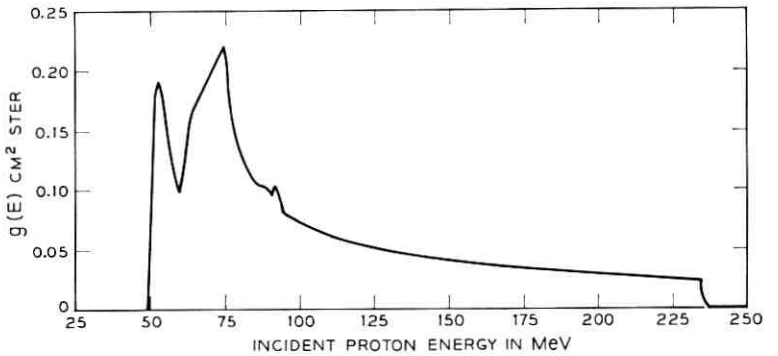


Fig. 54 — Dependence of geometric factor on energy of protons incident on the shielding.

For a differential energy spectrum $N(E)$, where $N(E)dE$ is the number of protons with energies between E and $E + dE$, the average geometric factor, $\bar{g}(E_1, E_2)$, of the detector for particles with energies between E_1 and E_2 is defined by

$$\bar{g}(E_1, E_2) = \frac{\int_0^\infty g(E)N(E)\,dE}{\int_{E_1}^{E_2} N(E)\,dE}. \tag{20}$$

The function $\bar{g}(50\text{ MeV}, E_2)$ has been evaluated numerically for various values of E_2 and forms of $N(E)$. The values of $\bar{g}(50\text{ MeV}, 130\text{ MeV})$ are plotted in Fig. 55 as a function of n for the single-component power-law spectrum $N(E) \propto E^{-n}$, and also as a function of E_0 for the single-component exponential spectrum $N(E) \propto \exp(-E/E_0)$. It may be seen from the figure that $\bar{g}(50\text{ MeV}, 130\text{ MeV})$ varies by less than 6 percent from 0.143 cm^2 ster for $0 < n < 7.5$ and 10 MeV $< E_0 <$ 90 MeV. These ranges of n and E_0 include most experimentally determined values by a comfortable margin.[3,5,29,34,35] The omnidirectional flux, $J(E_1, E_2)$, of protons with energies between E_1 and E_2 is given by

$$J(E_1, E_2) = \frac{4\pi Y^2}{\bar{g}(E_1, E_2)}, \tag{21}$$

where Y^2 is the counting rate of the detector. In the body of this paper, the values $E_1 = 50$ MeV, $E_2 = 130$ MeV and

$$\bar{g} \equiv \bar{g}(50\text{ MeV}, 130\text{ MeV}) = 0.143\text{ cm}^2\text{ ster} \tag{22}$$

are used. The flux $J(50\text{ MeV}, 130\text{ MeV})$ is designated simply by J,

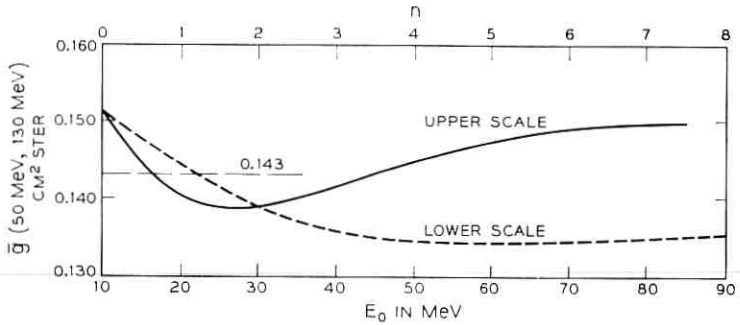


Fig. 55 — Dependence of average geometric factor on the exponent of a differential power-law energy spectrum and the e-folding energy of an exponential energy spectrum.

and the counting rate to flux conversion is considered to be independent of the proton energy spectrum.

While the relative value of $\bar{g}$ shows a variation of less than 6 percent for the wide range of single-component energy spectra noted above, the absolute value of $\bar{g}$ is less well specified. Variations in the ambient temperature and reverse bias voltage may change the effective geometric factor by as much as 25 percent. The difficulty of dealing with the complexities in shielding geometry, caused by embedding the instrument in the spacecraft, introduces additional uncertainties in the absolute value of $\bar{g}$. These uncertainties are in the range of -25 to $+50$ percent.

No provision was made for recalibrating the detector once the satellite was in orbit. However, the evidence, which is discussed in Section X, concerning the temporal variations of the proton distribution is that neither the detector nor the associated circuit elements were substantially affected by the space environment. Instrumental (e.g., temperature and bias voltage) effects are often quite different in character from temporal changes in the proton belts and may be separated from them in many circumstances. It is, of course, possible to postulate instrumental effects that will be inextricably confounded with certain secular changes that might take place in the proton distribution.

APPENDIX B.

Some Statistical Details

B.1 *Introduction*

This appendix presents, heuristically, some facts and formulae concerning the statistical analysis of the data. While a variety of statistical principles, precepts and procedures were employed as guides, the main judgments came from empiricism, scientific intuition and common sense. Various kinds of plots of residuals, used informally, have been of key importance, both for evaluation and for suggestion.

Simply stated, the objective was to produce a statistically accurate analytical description of the intensity distribution of high-energy protons in space surrounding the earth. The process of analysis involved the empirical evolution of a mathematical model, in interaction with the application of fitting and evaluative techniques. The data source and processing have been described in Sections II and III. The iterative and interactive processes of the final stages of model development, fitting, data partitioning and data sampling are described in Sections IV to IX.

Appendix B.2 deals with the basis for use of the square root transformation, Y, of the counting rate data, Y^2. Appendix B.3 discusses the selection of a sub-sample used in fitting. The use of the method of least squares in nonlinear model fitting, to estimate unknown coefficients, or functions of the coefficients, and their standard errors and correlations is reviewed briefly in Appendix B.4. Some remarks on construction of sums of squares contours, often referred to as confidence regions, and of indices of local nonlinearity of the model are given in Appendix B.5. Appendix B.6 discusses several issues relevant to the interpretation of the analysis of variance results. Appendix B.7 describes a mode of "smoothing" data within cells, which could have been used in conjunction with the sub-sampling procedure. Appendix B.8 concerns the technique of probability plotting.

B.2 *The Square Root Transformation*

It appears a reasonable assumption (supported by some empirical evidence) that, in the absence of geophysical disturbances, at a fixed point in space relative to the earth, the number of counts Z, recorded in the detector in 11 seconds, will vary in time according to a Poisson distribution, i.e.,

$$\text{Probability } \{Z = z\} = \frac{e^{-\nu}\nu^z}{z!}, \qquad z = 0, 1, 2, 3, \cdots, \tag{23}$$

where the parameter of the distribution, ν, is the mean value of Z.

With this statistical model, the average intensity of radiation in the region of space measured by the detector is proportional to ν, where the proportionality factor depends on the counter geometry and efficiency. The objective is to develop a function which describes how ν varies in space, based on observations of the quantity Z at different positions in the satellite orbit.

For the Poisson distribution, the variance of Z is also ν, i.e., the average of the squared deviations, $(Z - \nu)^2$, is ν. Thus, as the value of ν changes, the variance of the associated random variable Z also changes. Hence, the scatter of Z about its average value will be different in different regions of space as the average intensity fluctuates.

Working with the experimental data on the scale of Z has two drawbacks. Firstly, if one fitted a mathematical model to the data using a least squares criterion, the different observations would have variable weight, which would require appropriate, troublesome, allowance in the fitting procedure. Secondly, graphical judgment of the adequacy of

any particular fit would be difficult because of the variable scatter of the data about a fitted function in different regions.

Thus, the square root transformation, Y, of the counting rate ($Y^2 = Z/11$ counts per sec) was used to "stabilize" the variance and the model-fitting procedure employed unweighted nonlinear least squares on Y (but with some data weighting as discussed in Section 7.1 and Appendix B.3).

Heuristically, consider the linear Taylor's expansion of Z about ν

$$\sqrt{Z} \doteq \sqrt{\nu} + \frac{(Z - \nu)}{2\sqrt{\nu}} \cdots . \tag{24}$$

Then, the variance of $\sqrt{Z}$ is approximately

$$\text{Var}\,(\sqrt{Z}) \doteq \left(\frac{1}{2\sqrt{\nu}}\right)^2 \text{Var}\,(Z - \nu) + \cdots . \tag{25}$$

If

$$\text{Var}\,(Z - \nu) \propto \nu, \tag{26}$$

then

$$\text{Var}\,(\sqrt{Z}) \doteq \frac{1}{4}\frac{\nu}{\nu} = \frac{1}{4}, \tag{27}$$

that is, Var $(\sqrt{Z})$ would be approximately a constant.

Discussions of this transformation are given by Bartlett[42] and Anscombe.[43] If the distribution is in fact Poisson, then Anscombe shows that the average value of $\sqrt{Z}$ is approximately

$$\sqrt{\nu} - \frac{1}{8\sqrt{\nu}} - \frac{7}{128\nu^{\frac{3}{2}}},$$

while the variance of $\sqrt{Z}$ is, asymptotically,

$$\frac{1}{4}\left\{1 + \frac{3}{8\nu} + \frac{17}{32\nu^2} + \cdots\right\}.$$

Again for the Poisson distribution, Bartlett gives exact values of the dependence of the variance of $\sqrt{Z}$ on ν, summarized in the following:

ν:	0	0.5	1	2	3	4	6	9	15
Var $\sqrt{Z}$:	0	0.310	0.402	0.390	0.340	0.306	0.276	0.263	0.256

For a Poisson distribution, a transformation of the form $\sqrt{Z + 1/2}$ or $\sqrt{Z + 3/8}$ or $(\sqrt{Z} + \sqrt{Z + 1} - 1)$ will improve the variance

stabilization at smaller ν values. In the present application, such a modification would have appeared physically artificial and inconvenient. Moreover, the actual variance of the observations exceeds the Poisson variance (see Appendix C) and the "correction" was thus felt to be unwarranted. Some response to the (empirically defined) variance instability remaining after the square root transformation was made in the form of some weighting in the data selection (described in Section 7.1 and Appendix B.3).

Of course, if one wished to adopt the assumption of a Poisson distribution as an absolute basis for procedure, instead of as a guide, then one might choose to use maximum likelihood to estimate the coefficients of the model. This would mean developing a procedure for determining values of the coefficients [of the function $\nu(x,L)$] which would maximize

$$\prod_{\text{observations}} e^{-\nu(x,L)}[\nu(x, L)]^z/z!.$$

In the present case, a general program for nonlinear least squares was available while a procedure for Poisson likelihood maximization would need to be evolved. Apart from this practical consideration, however, it seemed more robust to use the Poisson assumption as a guide to developing an appropriate transformation preliminary to fitting by least squares. The point is that the square root transformation will effect an approximate variance stabilization not only when the variance is *equal* to the mean (as in the case of the Poisson distribution) but also when, more generally, the variance is *proportional* to the mean. Empirical vindication of this caution is given in Appendix C. Moreover, the least squares approach enables the approximate statistical interpretation of results using familiar procedures from linear multiple regression methods.

The present analysis is based on the quantity Y, where Y^2 = counting rate = $Z/11$ counts per sec. Thus, if in fact Z were a Poisson variable,

$$\text{Var}\,(Y) \doteq \left(\frac{1}{11}\right)\left(\frac{1}{4}\right) = 0.023, \tag{28}$$

as a reasonable approximation. When the average counting rate exceeds 1/11, this value of 0.023 is a lower bound on the variance of Y, even with the Poisson assumption. Moreover, there are many other possible sources of intrinsic variability and experimental error in this situation.

A further benefit which one might expect from the square root transformation in this circumstance is that the distribution of residuals would tend to be more symmetric and more nearly normal (Gaussian).

Some empirical properties of this square root transformation in the present body of data are given in Appendix C.

B.3 *Sample Selection*

As a practical requirement, the available multivariable, multicoefficient, nonlinear least squares fitting program could operate with a maximum of 1000 data points. Hence, the 41,135 HTB observations needed to be sampled or condensed at a 1 in 40 ratio.

As in all real sampling or experimental design situations, many competing criteria and practical difficulties were relevant. Perhaps the overriding point, explicitly understood here (and probably true in most actual model fitting problems), is that the model which was being developed was not the "truth" but was really just a smoothing function which one wanted to fit well over a wide region of space. Thus, it was not appropriate to think of estimating the model coefficients, say, so as to optimize their apparent (indicated) statistical reliability, nor would it be appropriate to try to use all the available data in an equally weighted manner, since accidents of orbital position and instrumental behavior would have too great an effect on the distribution of data points.

The procedure developed for the present use is outlined in Section 7.1, with pertinent remarks also in Section 13.3 and Appendix B.7.

The method of Section 7.1 yielded 960 observations to which the model was then fitted using unweighted least squares. The 960 sampled observations were selected so as to be roughly speaking, "widely spaced," the metric being change in average counting rate. Thus, the challenge of fitting the 960-point sample, as measured by sum of squares of residuals, is *greater,* on a per-observation basis, then would be that of fitting the entire body of 41,135 HTB observations, very many of which are quite close together. The "model bias" difficulties of the entire body of data are concentrated in the sample. The statistical fluctuation would be approximately the same, on a per observation basis, in the sample as in the whole body of data.

B.4 *Estimation Procedure*

The unspecified coefficients of the models defined in Section IV were estimated so as to minimize the sum of squares of deviations between

the observed Y and fitted y, for the sample array of data. The iterative, multivariable, multicoefficient, nonlinear least squares fitting was executed using a computer program due to Huyett and Wilk,[44] based on a procedure outlined by Wilk[45] (see also Lundberg, Wilk and Huyett).[46, 47]

The classical statistical properties of least squares estimation, namely unbiased estimates with minimum variance, apply in the case of statistically uncorrelated observations having equal variances and with the coefficients to be estimated occurring linearly in the model (see, for example, Wilks[16]). In the present case, even with the square root transformation, the observations do not have equal variances but, for practical purposes, the weighting implied by the selection procedure (see Section 7.1) compensates adequately. The model is, however, quite nonlinear in the coefficients. Still, one hopes that the attractive statistical properties of linear least squares carry over approximately to the nonlinear case because, in small enough neighborhoods, nonlinear functions can be linearly approximated. (An index for measuring model nonlinearity is described in Appendix B.5.) In any case, the least squares criterion is geometrically appealing and primitively meaningful.

Among the by-products of the fitting procedure, applied to the particular array of data in x,L space, are approximate values for the standard errors of the estimated coefficients, a matrix of approximate pairwise correlation coefficients for the estimated coefficients, an analysis-of-variance table giving the sum of squares accounted for and not accounted for by the fitted model, a list of residuals (equal to observed minus fitted), and various plots.

The least squares estimates of single-valued functions of the coefficients, such as $A(L)$, $x_c(L)$, or $y(x, L)$ are simply the same functions of the estimates of the coefficients (since least squares is an invariant process). Approximate variances and correlations of functions of the coefficients may be derived as follows: If $\theta' = (\theta_1, \cdots, \theta_p)$ denotes the coefficients of the model, and $\hat{\theta}$ their estimates, then the approximate covariance of the estimates $g(\hat{\theta})$ and $h(\hat{\theta})$ of the functions $g(\theta)$ and $h(\theta)$ is

$$\begin{aligned}
&\text{Covariance } (g(\hat{\theta}), h(\hat{\theta}) \\
&\quad = \text{Cov } (g(\hat{\theta}), h(\hat{\theta})) \\
&\quad = \text{Statistical average of } \{(g(\hat{\theta}) - g(\theta))(h(\hat{\theta}) - h(\theta))\}
\end{aligned}$$

$$\cong \sum_i \sum_j \left[\frac{\partial g(\theta)}{\partial \theta_i}\right]_{\hat{\theta}} \left[\frac{\partial h(\theta)}{\partial \theta_j}\right]_{\hat{\theta}} \text{Cov } (\hat{\theta}_i, \hat{\theta}_j)$$

$$= \sum_i \left[\frac{\partial g(\theta)}{\partial \theta_i}\right]_{\hat{\theta}} \left[\frac{\partial h(\theta)}{\partial \theta_i}\right]_{\hat{\theta}} \text{Var } (\hat{\theta}_i)$$

$$+ 2 \sum_{i<j} \left[\frac{\partial g(\theta)}{\partial \theta_i}\right]_{\hat{\theta}} \left[\frac{\partial h(\theta)}{\partial \theta_j}\right]_{\hat{\theta}} [\text{Var } (\hat{\theta}_i) \text{ Var } (\hat{\theta}_j)]^{\frac{1}{2}} \rho_{ij} , \qquad (29)$$

where ρ_{ij} is the correlation of $\hat{\theta}_i$ and $\hat{\theta}_j$. The formula for the approximate variance of $g(\hat{\theta})$ is then just a specialization of the above, putting $g = h$.

Some associated facts and issues are worth mentioning here. First, the approximate statistical correlations ρ_{ij} of the estimated coefficients of the model, or of functions of these, depend on (*i*) the distribution of the sample in x,L space, (*ii*) the values of the coefficients and (*iii*) the nature of the mathematical model; but do *not* depend on the actual adequacy or appropriateness of the fit. Similarly, the approximate standard errors of estimates are each made up as a product of which one term depends upon the square root of the mean square of the residuals of fit and the other depends only on the same factors as do the ρ_{ij}. Second, the various statistical measures, such as standard errors of estimated coefficients which are obtained from the fit to the 960-point HTB sample are, in a narrow statistical sense, conservative because they refer to the *sample* only and do not make allowance for the fact that the fitted model does indeed fit very well to the *entire body* of 41,135 HTB data. Thus, if statistical fluctuations were the only factor in the uncertainty of the estimates, one might further reduce this uncertainty by some factor, roughly approximated by $6 \approx \sqrt{41{,}135/960}$. This view of statistical uncertainty does not, however, give appropriate weight to the "model bias", which will not be eliminated by *any* number of observations. Third, all the summary statistical measures, which are referred to as standard errors, correlations, confidence regions, etc., should be used and interpreted in a data analytic way, i.e., as indicating facets of the body of data and the adequacy of its description by the model and analysis—rather than in terms of some supposedly "true" model or hypothesis which one is trying to evaluate in probabilistic terms.

B.5 *Sums of Squares Contours, "Confidence Regions" and Nonlinearity Indices*

The models of Section IV are defined up to the values of the unspecified coefficients. Any set of values for these coefficients may be

said to provide a "fit" to the 960-point sample of data. Thence one can define a sum of squares function of the set of coefficients as

$$SS \text{ (coefficients)} = \sum \text{(observed} - \text{"fitted")}^2, \tag{30}$$

which will take on various (positive) values as one varies the values of the coefficients. In the space of the coefficients there exist then, in principle, contours of this "sum of squares" function.

While standard errors provide information on reasonable allowances for the estimate of a single parameter in the light of the fit of the model to the actual body of data, they do not carry any information on the *joint* statistical properties of the estimates. A reasonable (robust and primitive) indication of joint statistical behavior is provided by these "sum of squares" contours in coefficient space.

In the case of models in which the unknown coefficients occur *linearly,* these contours are a family of ellipsoids defined by certain simple quadratic functions of the coefficients. The orientation and shape of this family of ellipsoids indicate the interdependence of the estimates of the coefficients in the light of the data, and show which coefficients are well-determined and which poorly. However, the interpretive value depends heavily on geometrical appreciation and, for more than a few coefficients, high-dimensional representation cannot be achieved directly.

The ellipsoid (even in the linear case) is *not* defined, in general, by its one-dimensional projections. (The standard error of a coefficient estimate is half the length of the projection of the unit ellipsoid of the family onto the coefficient axis.) But, as a matter of simple geometrical fact, *all pairs* of two-dimensional projections *do* uniquely define the ellipsoid. Thus, one practical means of a complete graphical representation of the high-dimensional ellipsoid is in terms of all possible pairwise planar projections.

For the case of linear models, on the basis of a series of assumptions —namely that the differences between the model and the observations are due to statistical fluctuations which are normally and independently distributed all with zero mean and the same variance—some may choose the abstract probabilistic interpretation of these ellipsoids as "confidence regions" (see, for example, Wilks[16]). If this interpretation is used, it is necessary that the distinctions and relationships between the joint, pairwise and marginal confidence coefficients and regions or intervals be understood. Details will not be provided here. Briefly, if a nine-dimensional ellipsoid were specified to have a confidence coefficient of β_9, then any two-dimensional projection would have a con-

fidence coefficient of β_2, interpreted marginally. The relation between β_9 and β_2 is indicated by the following:

β_9	β_2
0.13	0.90
0.25	0.95
0.50	0.984
0.75	0.997
0.90	0.9994
0.95	0.99995

In the present case, the model is nonlinear and the fluctuations are not normal. Contours of the sums of squares function as a function of the coefficients can, in principle, be obtained for a given body of data and will *not* be ellipsoids. In practice, however, obtaining these contours is so laborious as to be virtually impossible.

However, one may consider a linear (planar) approximation to the nonlinear model in the neighborhood of the least squares estimates of the coefficients and thence obtain expressions for a family of ellipsoids which *may* be reasonably good approximations to contours of the sums of squares function. An index of the effective nonlinearity of the model is the nonconstancy of the sums of squares of residuals on these ellipsoids and this can be normalized by division by the value of the minimum sum of squares. Such measures are presented and discussed in Sections VIII and IX.

Given that the linear approximation is adequate, the nonnormality of the observations should not deter those who seek (and who believe in) the general probabilistic confidence interpretation since the statistical process is likely very robust.

Sections VIII and IX contain specific examples of some of the pairwise projections of these "approximations to sum of squares contours." Specifically, the size of the 9-dimensional ellipsoid was such that, if all the statistical assumptions applied, a joint 0.99 confidence coefficient could be attached. Since a complete set of pairwise projections for nine coefficients involves 36 ellipses only a few are shown. As a summary indicator of the nature and behavior of these ellipses the quantity

$$\alpha = (\text{sign of } \rho) \cdot (1 - \sqrt{1 - \rho^2}) \tag{31}$$

is tabulated (in Tables IV and V), where ρ is the correlation of the pair of coefficients involved. The value of $1 - |\alpha|$ is the ratio of the area of the actual ellipse to that of the largest ellipse which could be inscribed in the rectangle formed by the horizontal and vertical tangents to the actual ellipse (see Wilk[48]). The range of α is $-1 \leqq \alpha \leqq 1$ and large values of $|\alpha|$ (say above 0.75) corresponds to narrow ellipses with major axis oblique to the coordinate axes, and represent situations of high interdependence of the coefficient estimates.

B.6 *The Analysis of Variance*

The analysis of variance provides a summary description of the apportionment of the "variability" of a body of data in the light of the model employed for analysis, where variability is defined in terms of sum of squares.

Given n observations, one may visualize an n-dimensional observation space, whose coordinates represent the possible values of each of the n observations. The data are then represented by a fixed point in this space.

The model, having p unspecified coefficients, implies certain functional relationships amongst the coordinates of the observation space. Thus the model effectively defines a constraining "surface" of p dimensions, and each point on this surface corresponds to some set of values of the unspecified coefficients of the model. The least squares estimate of the coefficients corresponds to that point on the constraining surface which is closest to the actual data point. If the coefficients in the model occur linearly then the constraining surface is a hyperplane which ordinarily, by definition of the observations, contains the origin, and, if the model includes a constant term, also contains the equiangular line (corresponding to the mean).

The squared distance of the data point to the origin is then the total sum of squares, $\sum Y_i^2$, while its shortest squared distance to the constraining surface is the error or residual sum of squares, associated with lack of fit. The difference between these may be termed the model sum of squares and, for linear models, this is actually the squared distance from the least squares estimates point to the origin.* If a constant term is included in a linear model, then the model sum of squares may be further decomposed additively in terms of the squared

* In the linear case, the model sum of squares is easily computed directly as the squared length of the projection onto the hyperplane of the line joining the data point and the origin. This fact is used in the present iterative computer program in checking convergence.

perpendicular distance (call it D_1^2) of the least squares estimate point to the equiangular line and the squared distance (call it D_2^2) along the equiangular line, from the foot of that perpendicular, to the origin. This latter quantity D_2^2 is usually termed the sum of squares due to the mean. The squared distance of the above-defined point on the equiangular line to the data point is called the corrected total sum of squares, $\sum (Y_i - \bar{Y})^2$ and is just $\sum Y_i^2 - D_2^2$. The ratio of the squared length D_1^2 to the corrected total sum of squares is defined as the squared multiple correlation, R^2, and often used as a measure of accomplishment of a model. It is easy to show that R^2 defined above is equal to

$$1 - \frac{\text{sum of squares for error}}{\text{total corrected sum of squares}}.$$

This latter quantity is computable even when the model is nonlinear and/or does not contain a constant term.

One may define contours of sums of squares of residuals in the constraining surface as the loci of the intersections with the surface of given radii from the observation point. In the event that the constraining surface was a hyperplane, which would be true if the unspecified coefficients in the model occur linearly, then these loci (or contours) would be a family of p-dimensional spheres. For nonlinear models, this will be approximately true for a sufficiently small neighborhood of the least squares point.

The particular form of the model, in regard to the unspecified coefficients, defines a coordinate system within the constraining surface. Three cases are worth distinguishing. First, the constraining surface is a hyperplane and the coefficients are linear. Second, the surface is a hyperplane but its coordinates are nonlinear. The second case may be reduced to the first by appropriately transforming the coefficient coordinate system. Third, the surface is nonlinear. In this case one can approximate the surface by a hyperplane in a small neighborhood. Thus, in a sufficiently small neighborhood, the situation can be regarded as linear.

The approximately or exactly linear coordinates implied by the model will in general be nonorthogonal. Thus, the representation of the spherical (exact or approximate) contours in an orthogonal coordinate system for the coefficients yields a family of ellipsoids. In the sense of measuring lack of fit by sums of squares between fitted and observed values, these contours in coefficient space constitute sets whose members are "equidistant" from the data point.

B.7 *A Procedure for Smoothing in Cells*

In Sections 7.1, 13.3 and Appendix B.3, discussion of why and how to sample and possibly "smooth" the data has been given. One specific practical possibility is now described.

Suppose one has a preliminary fit of the model, represented by $g(w_i\ ;\ \hat{\theta})$, where $\hat{\theta}' = (\hat{\theta}_1\ ,\ \hat{\theta}_2\ ,\ \cdots\ ,\ \hat{\theta}_p)$ are the fitted coefficient values and w_i denotes the independent variables. Suppose this preliminary fit is used to partition the space of the independent variables (here x and L) into some approximation of equirange cells, as described earlier. As argued in Section XIII, it may be profitable to "smooth" the data in each cell so as to yield a value generally representative of all the observations in that cell, instead of using a random selection from the cell.

A sensible smoothing function for each cell is, clearly, the model $g(w;\ \theta)$. A simple procedure is, for each cell separately, to carry out one stage of linear adjustment, doing the linear least squares regression of $\{Y_i - g(w_i\ ;\ \hat{\theta})\}$ on $\partial g/\partial\theta_1\,|_{\hat{\theta}}\ ,\ \cdots\ ,\ \partial g/\partial\theta_p\,|_{\hat{\theta}}$, to obtain the regression coefficients $\tilde{\delta}' = (\tilde{\delta}_1\ ,\ \cdots\ ,\ \tilde{\delta}_p)$, for that particular cell. Then the smoothing function for that cell would be $g(w;\ \tilde{\theta})$ where $\tilde{\theta} = \hat{\theta} + \tilde{\delta}$. A representative "smoothed observation" for that cell might then be the quantity $g(\bar{w};\ \tilde{\theta})$, where $\bar{w}$ is, say, the mid-point of the cell.

This process permits each cell, overall, to determine a single value to represent it in the entire fitting process and diminishes the chance that a random selection from a cell may be unnecessarily nonrepresentative of that cell behavior.

If one had wished to fit to *all* the available data, then the smoothed cell values would be weighted in proportion to the number of data in the cell. In the present case, this was deliberately *not* done.

The goodness of fit of a model to smoothed cell values, not differentially weighted, cannot be statistically judged directly from the analysis of variance since the residuals are no longer individually statistically comparable and the mean square residual is not an estimate of the error variance of the observations. However, the fitted model can be assessed by functions of its residuals from the original data (or a sample thereof).

B.8 *Probability Plotting*

The techniques of probability plotting are useful for data analysis in a wide variety of circumstances. (See Wilk and Gnanadesikan[17] for a general discussion of probability plotting techniques.) For instance,

in the present work, plots of residuals against various variables have provided invaluable guidance, but one is also interested in the distributional behavior *per se* of the collection of residuals. As presented in Section 8.1, normal and half-normal probability plots have been used for this purpose.

The rationale for such probability plots is roughly as follows: If one draws a random sample of size n from a population which is normally distributed with mean μ and variance σ^2 then the ordered observations would be expected to approximate, roughly, to a linear function, $\mu + \sigma z_i(n)$, of appropriate "representative" values $z_i(n)$ from a standard normal ($\mu = 0$, $\sigma^2 = 1$) distribution. Thence a plot of the ordered observations against the $z_i(n)$ would tend to be linear, with intercept approximately μ and slope approximately σ. For the representative value, $z_i(n)$, corresponding to the ith ordered observation, one can use the standard normal quantile for the proportion $(i-\frac{1}{2})/n$.

This plotting technique displays the individual observations in a sample graphically and does so against a backdrop such that the existence of outliers and asymmetry, as well as other distributional properties, are sensitively indicated. Of course such plots are usually profitably supplemented by others that order or partition the data according to information extraneous to the responses themselves.

We expect the mean of the residuals, $Y - y$, in the present study (see Section 8.1) to lie near 0. Also we expect that their variances will be approximately the same, since that is the purpose of the square root transformation. As a further benefit of the square root transformation we expect that the distribution of the residuals will tend to be symmetric and to approach normality; thence the present application of *normal* probability plotting of the residuals. The fact that these residuals are not entirely statistically independent—since they derive from a commonly estimated fitted function—is a minor issue since the number of observations is so much larger than the number of fitted coefficients.

Half-normal probability plotting employs the ordered *absolute* residuals plotted against standard half-normal (standard normal folded at 0) distribution quantiles. Such a plot eliminates any symmetry-type information but provides an incisive focus in bringing together on the plot the largest departures from fit.

Probability plots can provide very sensitive indications of distributional peculiarities especially in regard to "overly" large values. Sometimes the indications are of little practical interest, such as minor

lumps which one can see in Fig. 29, but in other regards, such as in estimating an "intrinsic" error standard deviation, the plots may permit a good judgment on how to discount or correct for apparently aberrant values which might otherwise have an undue influence, say, on mean square error.

Error standard deviations may be estimated from normal or half-normal probability plots as the "slope of the linear configuration." Typically, it will not be relevant to make a great show of objectivity in this process since the declared purpose is to permit an informal discounting of unexpected distributional peculiarities. Thus, in Fig. 29, one takes the slope as defined essentially by the bulk of residuals, ignoring the few largest.

APPENDIX C

Statistical Measures Over All the HTB Data

This appendix presents various statistical measures over all the 41,135 HTB data. These measures concern the fit of Models I and II and the partition of the x,L space (as described in Section 7.1 and Appendix B.3) into 1034 cells of which 813 were nonempty of observations. The partition is such that the range of y within cells is relatively small. For each cell, two functions are used: (*i*) The mean square deviation (MSD) defined as

$$\text{MSD}\,(u) = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - \bar{u})^2, \tag{32}$$

where the cell has n observations and u_i denotes some function of a cell observation, e.g., Y_i or Y_i^2, and $\bar{u}$ is the mean of the u_i in the cell; (*ii*) The mean square residual (MSR) defined as

$$\text{MSR}\,(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - y_i)^2, \tag{33}$$

where y_i is the fitted value (from Model I or II) corresponding to the observed Y_i.

C.1 *Empirical Justification of Square Root Transformation*

Figs. 56 and 57 show plots of MSD(Y^2) versus the cell mean of Y^2 and MSD(Y) versus the cell mean of Y, respectively. It is seen that MSD(Y^2) shows a distinct and major dependence on the average value of the counting rate, Y^2, while MSD(Y) does not show syste-

matic increase relative to the average value of Y, except, as expected, in the close neighborhood of zero counting rate.

A more detailed analsis of the results of Fig. 56 indicates that the dependence of MSD(Y^2) on cell mean of Y^2 is somewhat curvilinear having larger slope for larger Y^2 values. This curvilinearity is very likely mainly due to the mode of definition of the x,L cells. The procedure used tends to produce cells which are "too large" in regions where the counting rate is also large, thus leading to an apparent *extra* increase in MSD(Y^2) with Y^2. At all values, however, the dependence of MSD(Y^2) on Y^2 is greater than the slope 0.09 $(=1/11)$ which would be associated with the Poisson distribution. The empirically observed slope varies from about 0.15, based on small values, to 0.3, based on large values of the MSD(Y^2).

These results suggest that one cannot hope to achieve, by means of

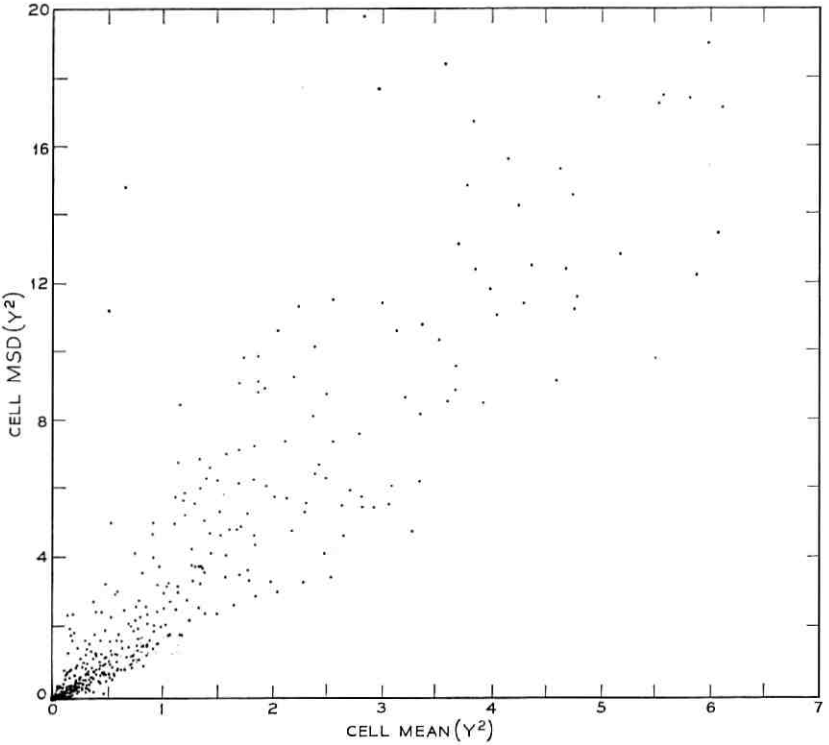


Fig. 56 — Cell MSD (Y^2) vs cell mean of Y^2 for the x, L cells defined in Section 7.1 and Appendix B.

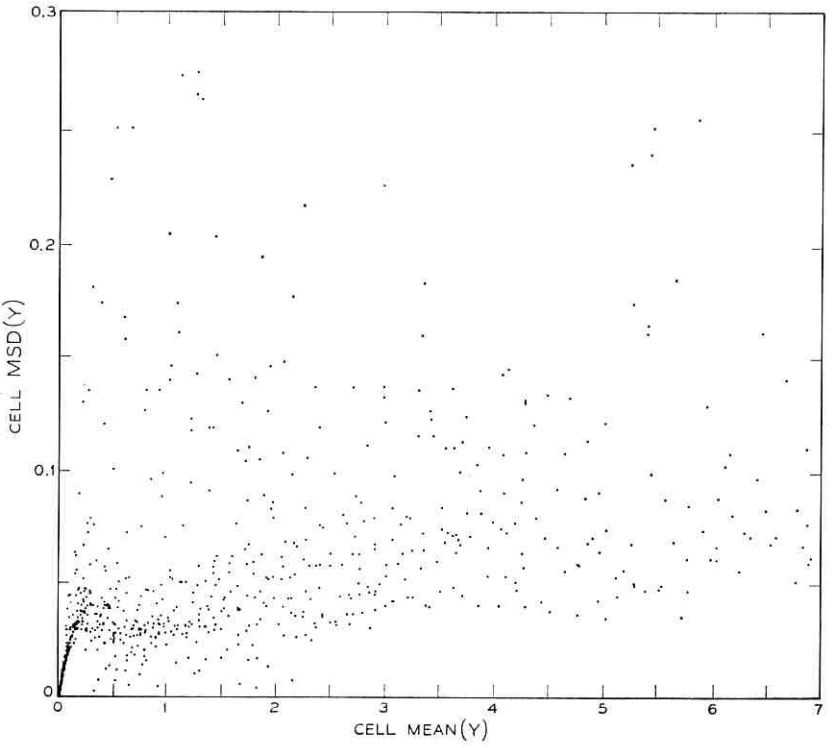


Fig. 57 — Cell MSD (Y) vs cell mean of Y for the x, L cells defined in Section 7.1 and Appendix B.

any fitted model based on x,L coordinates, on the scale of Y, a mean square residual (error) as low as 0.023 which is associated with the Poisson assumption.

Although the Poisson assumption provided a useful stimulus toward a profitable transformation of the data, these results confirm that it would have been unwise to have tied oneself too closely to the assumption as a complete basis for analysis, as for instance in basing the fit on maximization of the Poisson likelihood function (see Appendix B.2). Possible sources of variability and error in the data, beyond Poisson fluctuations in counts, have been discussed elsewhere in this paper.

C.2 *Determination of Weights*

The sample selection procedure involved "weighting" the 813 non-empty cells by selecting 2, 3/2, or 1 observation per cell. The observed

MSD (Y) were classified into three groups defined by: $0 \leqq \text{MSD} \leqq 0.013$; $0.013 < \text{MSD} \leqq 0.02$; $0.02 < \text{MSD}$. The x,L coordinates of the midpoints of cells so identified are shown in Fig. 58. The actual assignment of weights was based on applying contiguity considerations to this plot.

C.3 *Analysis of Variance Over All the HTB Data*

Table VIII summarizes the analysis of variance over all the 41,135 HTB data. The table covers the fit of Models I and II to all the data, using the estimated coefficients (see Tables IV and V) from the fit to the 960-point sample. Also, one can regard the collection of averages of the Y values in each of the 813 nonempty cells as providing a fit depending on 813 fitted quantities. The corresponding "error" (cell deviations) is the pooled cell MSD(Y). Finally, the residuals of the fit of Model I (or II) can be "fitted" by 813 cell

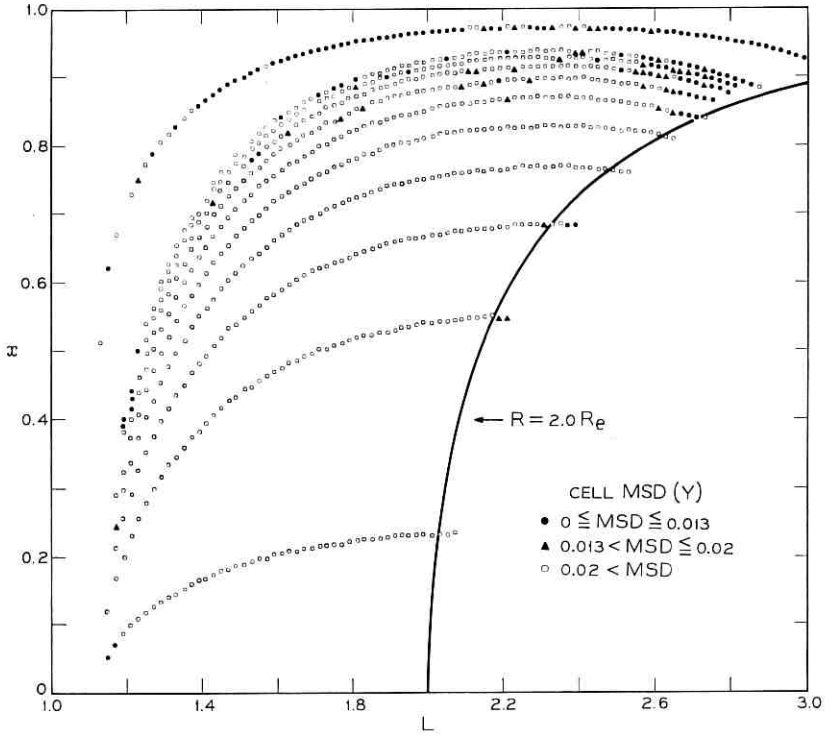


Fig. 58 — Positions of centers of regions in x, L space having certain ranges of cell MSD (Y). The ranges are indicated in the legend.

TABLE VIII—ANALYSIS OF VARIANCE OVER ALL THE HTB DATA (41,135 POINTS MINUS 226 OUTLIERS).

Due to	d.f.	Sum of squares	Mean square
Total (41,135–226)	40,909	230,267.45	
Mean	1	115,755.39	
Corrected total	40,908	114,512.07	
Model I residuals	40,900	1,411.3	0.0345
Model II residuals	40,901	1,419.6	0.0347
Cell deviations	40,096	1,541.4	0.0384
Cell dev. of Model I res.	40,085	1,167.0	0.0291
Cell dev. of Model II res.	40,086	1,166.9	0.0291
	Multiple R^2 value		
Model I	0.988		
Model II	0.988		

averages of the residuals, leading to an "error" which is the pooled cell MSD($Y - y$), i.e., due to the cell deviations of the Model I (or II) residuals.

The fits to all the data provided by Models I and II are equally good, as was true for the 960-point sample. The mean square residuals over all the data (about 0.035) is *lower* than the value (about 0.036) obtained for the sample even though the fit of the model was determined by the sample. This bears out the expectation (see Appendix B.3) that the mode of selection of the sample is such that the sample was *harder* to fit on a per-observation basis than the entire body of data.

The cell means provide overall a poorer actual fit than Model I or II, and allowing for the number of fitted coefficients, the mean square for cell deviations exceeds that for the models by about 12 percent.

Fitting cell means to the model residuals yields an additional substantial reduction in the sum of squares of the model residuals and a mean square of about 0.029, which is some 17 percent lower than the value for the models. If in fact the models gave an "unbiased" fit everywhere, then one would expect that the values of pooled MSR(Y) and pooled MSD($Y - y$) would be nearly the same. The excess of the former is due mainly to systematic inadequacies of the fit (see Appendix C.4).

The value 0.029 represents virtually a lower bound on the achievable 'mean square error' for this body of data. Despite its downward bias from the substantial number of 'zero counting rate' observations, it exceeds the 'Poisson' variance of 0.023 by about 25 percent. This

excess is probably due to a combination of factors, including incomplete elimination of temperature and bias voltage instrumental effects, as discussed further below.

The 'improvement' of the MSD$(Y - y)$ over MSR *cannot* be taken to mean that some smooth "simple" adjustment of the model based on x,L coordinates might be found so as to yield similar improvement. Some of the bias apparently associated with x,L coordinates in different regions may be due to artifactual association with temporal, instrumental, or other small effects and such corrections could not be made overall in terms of a "simple" x,L dependence.

C.4 *Analysis of the Excess Variation*

A study of plots of cell MSD(Y) against each of y, x, and L indicates that large MSD values occur mainly in the $1.2 < L < 1.7$ region, at high average counting rates. This excess is due largely to

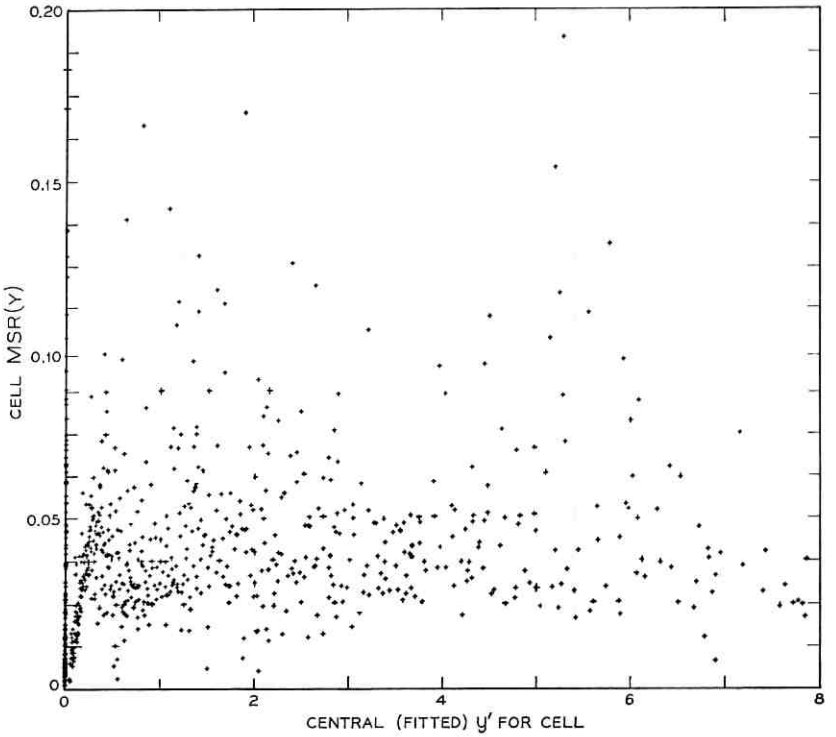


Fig. 59 — Cell MSR (Y) vs central value of y' for cell.

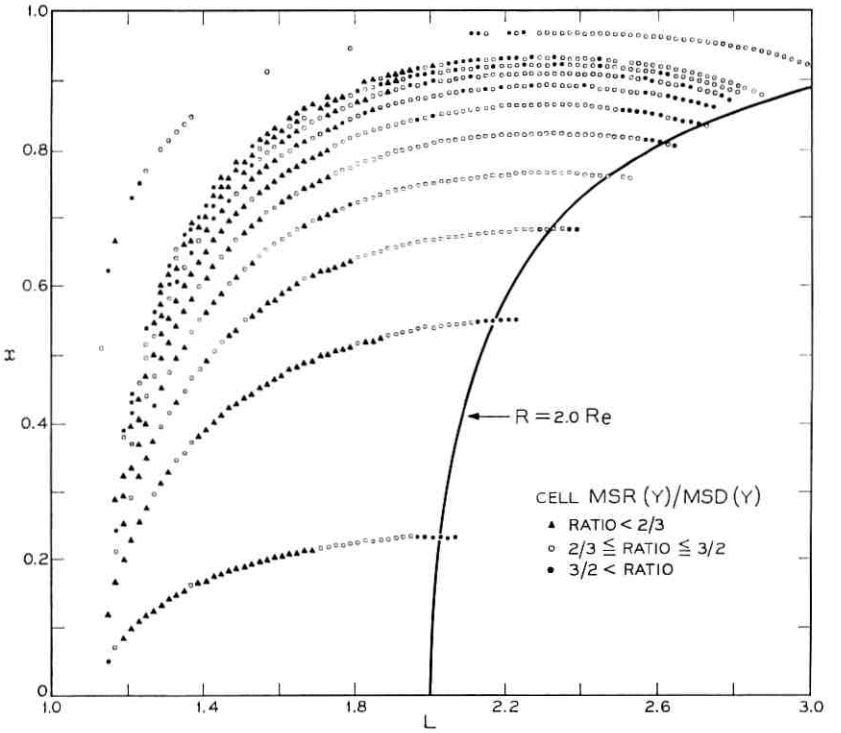


Fig. 60 — Positions of centers of regions in x, L space having certain ranges of cell MSR(Y)/MSD(Y). The ranges are indicated in the legend. (Plotted points are mid-points for the cells. Points appearing to the right of the boundary $R = 2.0\ R_e$ represent cells which have data only in one corner.)

the hybrid mode of x,L cell formation, in which the L-slices were equal length intervals, while within each L-slice, the x segments were chosen to have equal increments of y. Thus, at L values where y is large, the x,L cells will tend to have a larger y range.

The tendency of MSD to rise with cell average counting rate is not mirrored by cell MSR behaviour. As shown by Fig. 59, the level of MSR is not dependent on y except, as expected, for those cells where the counting rate is near zero. Roughly speaking, the average level of cell MSR for y values away from zero is about 0.04, in agreement with the probability plot estimate of Section 8.1. Of course, Fig. 59 shows both smaller ordinate values and less dependence on the abscissa values than the comparable plot of Fig. 57.

The relation of cell MSR to cell MSD is partially indicated in Fig.

60, showing positions in x,L space of various ranges of the value of the ratio MSR/MSD. One sees that MSR tends to exceed MSD along the "outside" of the data region. The excess along the $R = 2\ R_c$ boundary is due mainly to model bias or inadequacy. The excess at high L-high x is probably associated with temporal effects. The large ratios along the left edge of the data, which is the cutoff region, is likely a reflection of deficiency of the function. The excess of MSR over MSD is associated in the main with small y values.

Fig. 61 shows cell mean square deviations of residuals, MSD$(Y - y)$, plotted against y. This plot shows less scatter (most noticably for MSD$(Y - y) > 0.075$) than that of Fig. 59, and a lower average level of MSD$(Y - y)$ for $y > 0$, as expected. The high values of MSD$(Y - y)$ are not related to y as such but rather, as other plots show, with "extra fluctuations" in the $1.2 < L < 1.7$ region. This is probably associated with the coarse HTB data partition which does not com-

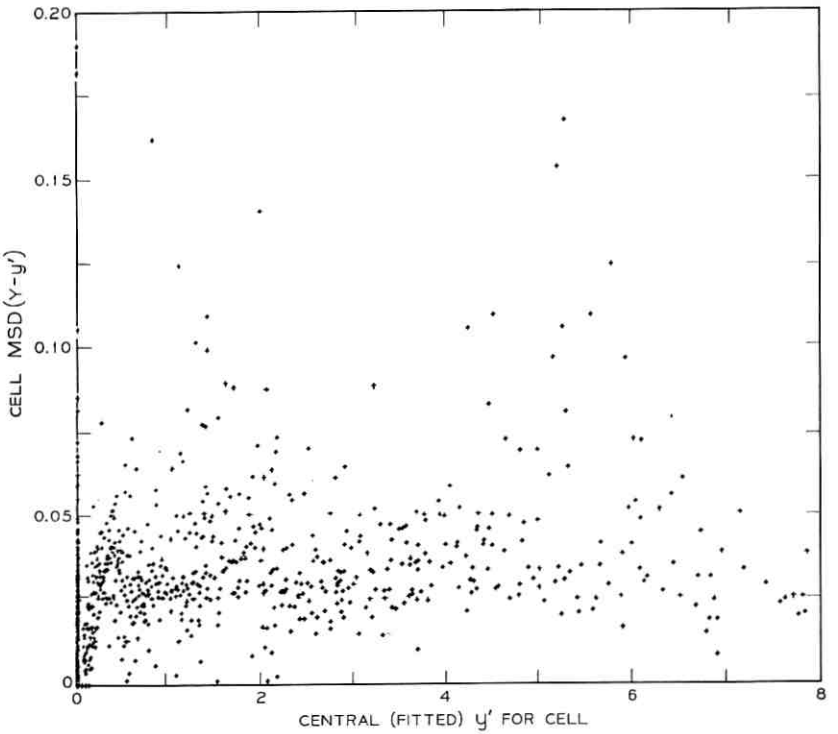


Fig. 61 — Cell MSD $(Y - y')$ vs central value of y' for cell.

pletely take care of the temperature and bias voltage instrumental effects.

REFERENCES

1. The *Telstar* Experiment, B.S.T.J., *42*, July, 1963, pp. 739–1940.
2. Brown, W. L., Buck, T. M., Medford, L. V., Thomas, E. W., Gummel, H. K., Miller, G. L., and Smits, F. M., The Spacecraft Radiation Experiments, B.S.T.J., *42*, July 1963, pp. 899–942.
3. Brown, W. L., Gabbe, J. D., and Rosenzweig, W., Results of the Telstar Radiation Experiments, B.S.T.J., *42*, July, 1963, pp. 1505–1560.
4. Brown, W. L. and Gabbe, J. D., The Electron Distribution in the Earth's Radiation Belts During July, 1962, as Measured by Telstar, J. Geophys. Res., *68*, 1963, pp. 607–618.
5. Gabbe, J. D. and Brown, W. L., Some Observations of the Distribution of Energetic Protons in the Earth's Radiation Belts between 1962 and 1964, *Radiation Trapped in the Earth's Magnetic Field*, B. M. McCormac, ed., D. Ridel Publishing Co., Dordrecht, Netherlands, 1966, pp. 165–184.
6. Chapman, S. and Bartels, J., *Geomagnetism*, Oxford University Press, London, 1940.
7. Buck, T. M., Wheatley, G. H., and Rodgers, J. W., Silicon *p-n* Junction Radiation Detectors for the Telstar Satellite, IEEE Trans. Nuc. Sci., *11*, 1964, pp. 294–301.
8. Courtney-Pratt, J. S., Hett, J. H., and McLaughlin, J. W., Optical Measurements on Telstar to Determine the Orientation of the Spin Axis, and the Spin Rate, J. Soc. Motion Pict. Telev. Eng., *72*, 1963, pp. 462–484.
9. McIlwain, C. E., Coordinates for Mapping the Distribution of Magnetically Trapped Particles, J. Geophys. Res., *66*, 1961, pp. 3681–3692.
10. Roberts, C. S., On the Relationship between the Unidirectional and Omnidirectional Flux of Trapped Particles on a Magnetic Line of Force, J. Geophys. Res., *70*, 1965, pp. 2517–2528.
11. Chamberlain, J. W., *Motion of Charged Particles in the Earth's Magnetic Field*, Gordon and Breach Publishing Co., New York, New York, 1964.
12. McIlwain, C. E., Program INVAR, private communication, 1962.
13. Jensen, D. C. and Cain, J. C., An Interim Geomagnetic Field (abstract), J. Geophys. Res., *67*, 1962, pp. 3568–3569.
14. Heiskanan, W. A., Geodetic Data, *American Institute of Physics Handbook*, D. E. Gray, ed., McGraw-Hill Book Company, New York, 1957.
15. *The American Ephemeris and Nautical Almanac* for the Year 1962 (and for the Year 1963), United States Government Printing Office, Washington, D. C., 1960 (and 1961).
16. Wilks, S. S., *Mathematical Statistics*, John Wiley, New York, 1962.
17. Wilk, M. B., and Gnanadesikan, R., Probability Plotting Methods for the Analysis of Data. Submitted to Biometrika, 1966.
18. McIlwain, C. E., The Radiation Belts, Natural and Artificial, Sci., *142*, 1963, pp. 355–361.
19. Valerio, J., Protons from 40–100 MeV Observed on Injun 3, J. Geophys. Res., *69*, 1964, pp. 4949–4958.
20. Fillius, R. W., Trapped Protons of the Inner Radiation Belt, J. Geophys. Res., *71*, 1966, pp. 97–123.
21. Gabbe, J. D., Michaels, A. S., and Wilk, M. B., unpublished work, 1967.
22. McIlwain, C. E., Long-Term Changes in the Distribution of the 40- to 110-MeV Trapped Protons, Trans., AGU, *46*, 1965 (abstract), p. 141.
23. Blanchard, R. C., and Hess, W. N., Solar Cycle Changes in Inner-Zone Protons, J. Geophys. Res., *69*, 1964, pp. 3927–3938.
24. Filz, R. C., The Low-Altitude Inner-Zone Trapped Proton Flux during 1965 Measured at 55 MeV, Trans., AGU, *47*, 1966 (abstract), p. 129.
25. McIlwain, C. E., Redistribution of Trapped Protons during a Magnetic Storm, Department of Physics, Univ. of Calif. at San Diego, April, 1964.

26. Pizzella, G., McIlwain, C. E., and Van Allen, J. A., Time Variation of Intensity of the Earth's Inner Radiation Zone, October, 1959, through December, 1960, J. Geophys. Res., *67,* 1962, pp. 1235–1253.
27. Cain, J. C., Daniels, W. E., Hendricks, S. J., and Jensen, D. C., An Evaluation of the Main Geomagnetic Field, 1940–1962, J. Geophys. Res., *70,* 1965, pp. 3647–3674.
28. Cain, J. C., private communication, September, 1965.
29. Vette, J. I., Models of the Trapped Radiation Environment, *Volume I: Inner Zone Protons and Electrons,* National Aeronautics and Space Administration, Washington, D. C., 1966 (NASA SP-3024).
30. McIlwain, C. E. and Pizzella, G., On the Energy Spectrum of Protons Trapped in the Earth's Inner Van Allen Zone, J. Geophys. Res., *68,* 1963, pp. 1811–1823.
31. Bostrom, C. O., Zmuda, A. J., and Pieper, G. F., Trapped Protons in the South Atlantic Magnetic Anomaly, July through December, 1961; (2) Comparisons with Nerv and Relay 1 and Discussion of the Energy Spectrum, J. Geophys. Res., *70,* 1965, pp. 2035–2043.
32. Imhof, W. L. and Smith, R. V., Proton Intensities and Energy Spectrums in the Inner Van Allen Belt, J. Geophys. Res., *69,* 1964, pp. 91–100.
33. McIlwain, C. E., Fillius, R. W., Valerio, J., and Dave, A., Relay 1 Trapped Radiation Measurements, Department of Physics, University of California at San Diego, March, 1964.
34. Fillius, R. W., and McIlwain, C. E., Anomalous Energy Spectrum of Protons in the Earth's Radiation Belt, Phys. Rev. Letters, *12,* 1964, pp. 609–612.
35. Freden, S. C., Blake, J. B., and Paulikas, G. A., Spatial Variation of the Inner Zone Trapped Proton Spectrum, J. Geophys. Res., *70,* 1965, pp. 3113–3116.
36. Knecht, D. J., The Energetic-Proton Spectrum at the Lower Edge of the Inner Belt, Trans., AGU, *47,* 1966, (abstract), p. 128.
37. Lenchek, A. M. and Singer, S. F., Geomagnetically Trapped Protons from Cosmic-Ray Albedo Neutrons, J. Geophys. Res., *67,* 1962, pp. 1263–1287.
38. Van Allen, J. A., The Geomagnetically Trapped Corpuscular Radiation, J. Geophys. Res., *64,* 1959, pp. 1683–1689.
39. Frank, L. A., Van Allen, J. A., Whelpley, W. A., and Craven, J. D., Absolute Intensities of Geomagnetically Trapped Particles with Explorer 14, J. Geophys. Res., *68,* 1963, pp. 1573–1579.
40. Roberts, C. S., Electron Loss from the Van Allen Zones due to Pitch Angle Scattering by Electromagnetic Disturbances, *Radiation Trapped in the Earth's Magnetic Field,* B. M. McCormac, ed., D. Ridel Publishing Co., Dordrecht, Netherlands, 1966, pp. 403–421.
41. Brown, W. L., Observations of Transient Behavior of Electrons in the Artificial Radiation Belts, *Radiation Trapped in the Earth's Magnetic Field,* B. M. McCormac, ed., D. Ridel Publishing Co., Dordrecht, Netherlands, 1966, pp. 610–633.
42. Bartlett, M. S., The Square Transformation in Analysis of Variance, J. Roy. Stat. Soc. Suppl. *3,* 1936, pp. 68–78.
43. Anscombe, F. J., The Transformation of Poisson, Binomial and Negative Binomial Data. Biometrika, *35,* 1948, pp. 246–254.
44. Huyett, Marilyn J. and Wilk, M. B., A General Fortran II Program for Nonlinear Least Squares Regression, unpublished manuscript, 1960.
45. Wilk, M. B., An Identity of Use in Nonlinear Least Squares (abstract), Ann. Math. Statist., *29,* 1958, p. 618.
46. Lundberg, J. L., Wilk, M. B., and Huyett, Miss M. J., Estimation of Diffusivities and Solubilities from Sorption Studies, J. Polymer Sci., *57,* 1962, pp. 275–299.
47. Lundberg, J. L., Wilk, M. B., and Huyett, Miss M. J., Sorption Studies using Automation and Computation. Ind. Eng. Chem. Fundamentals, *2,* 1963, pp. 37–43.
48. Wilk, M. B., Gauging the Magnitude of Correlation. Submitted to J. Amer. Stat. Assoc., 1966.

Gold Doped Silicon Compandor Diodes For N2 and N3 Carrier Systems

By K. R. GARDNER and T. R. ROBILLARD

(Manuscript received May 2, 1967)

Companding has proven to be a valuable technique for improving the signal-to-noise ratio of voice transmission at baseband frequencies. A compandor consists of a compressor element which reduces the dynamic range of a transmitted signal in a predetermined manner and an expandor element which restores the signal range at the receiver. Practical Bell System applications to date have used electron tubes, germanium point-contact semiconductor diodes and unpassivated silicon mesa diodes. Each of these variolosser elements had serious shortcomings. Two new diode pairs have been designed which eliminate the problems of impedance range control and linearity, diode noise and electrical stability. The new design utilizes heavy gold doping of a planar oxide-passivated wafer design to produce a bulk controlled device capable of unusually high manufacturing yields.

I. INTRODUCTION

Compandors are a special application for a diode because the diodes are used as variolossers and the electrical parameter which must be controlled is the small signal forward impedance as a function of bias current. Furthermore, control of impedance is required over two orders of magnitude. Other requirements are low noise and good stability. The 484/489A and 484/489B diode pairs, which are electrically identical and differ only in mechanical outline, are silicon "planar" diodes which were designed for use in this application. The new devices were designed to replace two pairs of troublesome unpassivated "mesa" type devices in both the N2 and N3 carrier systems.

A comprehensive diode design was not previously available for this application. Diodes were obtained by selection from available types at low yield. This paper discusses the theoretical and empirical design and the fabrication of the new diodes.

II. CIRCUIT FUNCTION OF COMPANDORS

Noise is an important problem associated with long distance telephone transmission. Elimination or reduction of this undesirable effect is a consideration in the design of all transmission equipment. Noise occurs in transmission from many sources such as thermal noise, external interferences, and crosstalk. The compandor circuit, which was first introduced into the Bell System in the transatlantic radio circuit in 1932,[1] is one of the methods used to reduce noise. While the first compandor circuits used vacuum triodes[1] as the variolosser units, later compandor circuits used semiconductor diodes when they become available.

A compandor[2] is composed of two-parts, compressor and expandor, one at each end of a transmission path. The compressor circuit compresses the dynamic range of the transmitted signal power by taking the square root of the signal (although other functions could be used). If the maximum signal levels are transmitted at the same power with compression as they would be without compression, then the minimum signals will be transmitted at relatively higher power with compression than without. Therefore, a higher signal-to-noise ratio results for the minimum signals on the transmission path if compression is used. At the receiving end of the transmission line the expandor circuit squares (expands) the signal to its original dynamic range.

The N2 and N3 carrier system compandors[3] compress a 60-dB signal range into a 30-dB range for transmission. Therefore, 30-dB higher noise may be potentially tolerated in the transmission path. At the receiving terminal the expandor portion of the compandor expands the signal range to its original value of 60 dB and restores the signal to its original form. Since a compandor is an interchangable plug-in unit and the compressor and the expandor in the same unit do not work together, it is necessary that all compressor circuits track closely with all expandor circuits.

The core of the compressor and expandor circuits is a pair of variolosser diodes. The stringent requirements on the compandor circuits are reflected in stringent requirements on the variolosser diodes. This paper reports the development of two diode pairs which meet the unique requirements of these circuits.

III. THE DIODES

3.1 *General Description*

The first semiconductor diodes used as the compandor variolossers were selected from available types. While the New York-London long

wave radiotelephone circuit (1932) used vacuum triodes,[1] the N1 carrier system used germanium point-contact diodes;[4] the P1 and O carrier systems used silicon alloy diodes;[5] and the N2 carrier system originally used diffused silicon, mesa diodes.[3] Several problems arose with the use of these state-of-the-art-diodes although careful selection and circuit adjustment could correct most of them. The major problems of high device cost, high noise, and periodic supply shortages arose directly from a lack of understanding of the physical mechanism controlling the forward impedance characteristic.

It was possible, by designing a new diode, to overcome all of the problems and at a much lower cost. The new design uses silicon planar techniques coupled with controlled gold doping and heat treatment to produce the desired diode characteristics.

The following parameters are used to characterize the diodes for the compandor applications:

(*i*) The small-signal forward impedance,* Z_f, at a specified mid-range dc dias current.

(*ii*) The ratio, R_1, of the small-signal forward impedance at a specified lower current to the impedance at the above mid-range current.

(*iii*) The ratio, R_2, of the impedance at the mid-range current to the impedance at a specified higher current.

(*iv*) The impedance difference between the diodes of a pair measured separately at the idling current.

(*v*) Noise generated by the diode over the current and frequency ranges of interest.

Table I shows the limits for these parameters for both the mesa and the planar type diodes.

3.2 *Design Theory*

The primary parameter to be controlled was the small-signal forward impedance, Z_f. The theoretical forward impedance of the semiconductor diode may be obtained by differentiation of the current-voltage equation. For semiconductor diodes the relationship is:

$$I_F = I_s(\exp qV/nkT - 1), \tag{1}$$

where

I_s = saturation current,

*For simplicity the expression 'small-signal forward impedance' will often be shortened to 'forward impedance' or 'impedance' in this paper.

q = electronic charge,
V = applied voltage,
n = experimentally determined constant commonly between 1 and 2,
k = Boltzmann's constant,
T = absolute temperature.

Therefore,

$$Z_f \equiv \frac{\partial V}{\partial I_F} = \frac{nkT}{qI_s} \exp - qV/nkT.$$

For forward bias, V, greater than a few nkT/q ($kT/q = 0.026$ volts

TABLE I — SALIENT CHARACTERISTICS

Parameter	Planar compressor and expandor	Mesa compressor and expandor	Units
Z_f, Forward impedance, at 50 μA dc bias			
For single diode of pair	900 ± 35	1045 ± 125	ohm
For diode pair in series	1800 ± 70	2070 ± 70*	ohm
R_1, Impedance ratio = $\frac{Z_f \text{ at } 10\ \mu\text{A dc}}{Z_f \text{ at } 50\ \mu\text{A dc}}$			
For single diode of pair	5.0 ± 0.2	4.9 ± 0.4	—
For diode pair in series	5.0 ± 0.2	4.9 ± 0.2*	—
R_2, Impedance ratio = $\frac{Z_f \text{ at } 50\ \mu\text{A dc}}{Z_f \text{ at } 300\ \mu\text{A dc}}$			
For single diode of pair	6.0 ± 0.2	6.2 ± 0.5	—
For diode pair in series	6.0 ± 0.2	6.15 ± 0.2*	—
Parameter	**Compressor only**	**Compressor only**	**Units**
ΔZ_f = Difference in impedance of diodes of pair measured separately:			
At 2 μA dc bias	2000 max	2000 max	ohms
At 10 μA dc bias	500 max	500 max	ohms
v_n = Noise voltage of single diode or pair at 2.5μA dc bias. Bandwidth 200–3500Hz. Parallel resistance 17,000 ohm.	20 max	— —	μVrms

* Computer selected.

at room temperature) one has $\exp qV/nkT \gg 1$ and to a good approximation*

$$Z_f = \frac{nkT}{q}\frac{1}{I_F}. \tag{2}$$

It is this Z_f-I_F functional relation which is used in the design of the N2 and N3 compandor circuits.

There are five sources of current in a forward biased p-n junction; diffusion, bulk recombination, surface recombination, channel and tunneling currents. The total diode current is given by

$$\begin{aligned} I_F &= I_D \text{ (diffusion)} + I_{BR} \text{ (bulk recombination)} \\ &\quad + I_{SR} \text{ (surface recombination)} + I_{CL} \text{ (channel)} \\ &\quad + I_T \text{ (tunneling)}. \end{aligned} \tag{3}$$

The diffusion current[6] at small bias is given by

$$I_D = I_d(\exp qV/nkT - 1), \tag{4}$$

where

$$I_d = qA[p_n(D_p/\tau_p)^{\frac{1}{2}} + n_p(D_n/\tau_n)^{\frac{1}{2}}]$$

and

τ_p = lifetime of holes on n side of junction,
τ_n = lifetime of electrons on p side of junction,
p_n = hole concentration in n-region,
n_p = electron concentration in p-region,
D_p = diffusion constant for holes,
D_n = diffusion constant for electrons,
A = area.

At high forward bias (injection) the current will be given by

$$I_D = I_d(\exp qV/2kT - 1) \tag{5}$$

and in the intermediate range the current equation will be similar to (1).

The bulk recombination current[7,8] for bias voltages, V, greater than several kT/q is given by

$$I_{BR} = I_{r_0} \exp qV/2kT, \tag{6}$$

where

$$I_{r_0} = \frac{\pi}{2}\left(\frac{kT}{qE}\right)\frac{qn_i}{\tau_0}A,$$

n_i = intrinsic carrier concentration,

* The error is less than 1 percent for voltages greater than 0.2 volts and less than 0.1 percent for voltages greater than 0.3 volts.

E = electric field at junction,

and τ_0 = lifetime.

The exact voltage dependence of (6) depends in a complicated way on the physical parameters of the junction, but in a given range may be described by

$$I_{BR} = I_{br} \exp qV/n_{br}kT, \tag{7}$$

where $n_{br} \leqq 2$ and accounts for the voltage dependence of I_{re}, while I_{br} is the voltage independent factor.

Surface recombination current may be described by a similar equation;[8]

$$I_{SR} = I_{sr} \exp qV/n_{sr}kT, \tag{8}$$

where $n_{sr} > 1$.

Channel current[8] at $V > kT/q$ may be described by

$$I_{CL} = I_{cl} \exp qV/n_{cl}kT, \tag{9}$$

where $n_{cl} = 1$ up to 4 or 5 or more for poorly stabilized surfaces.

By considering the five currents in parallel one may calculate a small-signal impedance for each current, and the forward impedance of the diode may be expressed as five impedances in parallel.

$$\frac{1}{Z_f} = \frac{1}{\frac{kT}{qI_D}} + \frac{1}{\frac{n_{br}kT}{qI_{BR}}} + \frac{1}{\frac{n_{sr}kT}{qI_{SR}}} + \frac{1}{\frac{n_{cl}kT}{qI_{CL}}} + \frac{1}{Z_T} \tag{9}$$

or

$$\frac{1}{Z_f} = \frac{q}{kT}\left(I_D + \frac{I_{BR}}{n_{br}} + \frac{I_{SR}}{n_{sr}} + \frac{I_{CL}}{n_{cl}} + \frac{kT}{qZ_T}\right). \tag{10}$$

Diffusion current cannot be made dominant over recombination current in silicon except at high current densities where the value of the multiplier, n, may be modified by carrier injection. In both mesa and planar diodes the diffusion current was reduced by using heavily doped starting material (approximately 0.005 ohm-cm p-type silicon). The use of such low resistivity material had the further advantage of reducing the series resistance of the bulk silicon to about 0.04 ohms. At all currents of interest the 0.04 ohms made a negligible contribution to the diode impedance.

The bulk recombination current was greatly enhanced by introducing trapping centers by heavily gold doping the diodes. The effect on the diode parameters of various gold doping levels was investigated by

varying temperature and time of gold diffusion and subsequent heat treatments. These effects are discussed in detail in Section 3.4.3.

From the previous equations and measurements on existing diodes the relative values of the five currents and their contributions to the total impedance may be compared. As an example, values are calculated for a bias of 0.4 volts and diode parameters which approximate those of the actual diodes.

The diffusion current is calculated from (4) as

$$I_D = 10^{-9}A = 0.001\mu A,$$

where the following values are assumed:

$\tau_n = \tau_p = 10^{-10}$ sec (Ref. 9)
$N_n = N_p = 2 \times 10^{19}$ cm^{-1} (0.005 ohm-cm)
$\mu_p = 30$ cm^2/volt sec (Ref. 10)
$\mu_n = 75$ cm^2/volt sec (Ref. 10).

The bulk recombination current is calculated from (6) by using the approximation

$$E = (\psi_0 - V)/W,$$

where ψ_0 is the built-in voltage and W is the space-charge width. The bulk recombination current is

$$I_{BR} = 30\mu A \gg I_D = 0.001\mu A.$$

Estimates of surface recombination and channel currents were made from measurements on planar type diodes. These estimated values were much smaller (by several orders of magnitude) than the bulk recombination current. Likewise, the observed magnitude of the forward tunneling current is negligible since doping levels are relatively light and the junction is graded.

Hence, bulk recombination current is dominant and the junction impedance becomes

$$Z_j = \frac{nkT}{q}\frac{1}{I_F} \cong \frac{n_{br}kT}{q}\frac{1}{I_{BR}}. \tag{11}$$

By way of contrast the impedance of the mesa diode was primarily dependent on surface damage introduced during mechanical formation of the active diode wafer. High surface recombination (mechanically damaged) wafer edges were created when the diffused slices were diamond sawed into wafers. A portion of this damage was then removed

by chemical etching to set the diode multiplication factor, n, and hence forward impedance to the nominal value.

3.3 *Mesa Diode Deficiencies*

System and manufacturing experience pointed out three major shortcomings of the unpassivated mesa design. The first of these shortcomings was lack of good control of the nominal impedance value and range. Manufacturing problems were experienced until the planar redesign efforts delineated the physical mechanisms controlling forward impedance. Even with this understanding manufacturing control could not be improved sufficiently to obtain narrow distributions of impedance; a computer selection of individual diode pairs was necessary for reasonable yields.

A second electrical characteristic which could not be controlled in the manufacturing operation was the noise voltage produced by the device in the 200-3500 Hz band. The N2 and N3 systems require that the noise voltage be less than 20 microvolts for the compressor pair and 40 microvolts for the expandor pair when operating at a direct current of 2.5 microamperes. This characteristic was checked on a nonparametric basis at the equipment assembly location; and, quite frequently, shipments of diode pairs would be found which exhibited excessively high noise.

Finally, the short-term stability objectives of the systems could never be achieved with the unpassivated device.

As shown in this paper, the redesigned device readily meets all noise and stability objectives and permits manufacture of diodes at very high yields without the need for computer matching.

3.4 *Planar Diode Design*

3.4.1 *Structural Features*

While the primary compandor diode design effort was directed toward understanding and controlling the physical variables associated with the active semiconductor chip, the encapsulating structure was also changed to provide an assembly more suited to printed circuit board mounting. As shown in Fig. 1, each diode pair of the mesa type was composed of two metal package diodes molded in epoxy and glued together with an epoxy cement. This arrangement is costly and results in a double ended structure whose leads must be trimmed and formed for mounting. The redesign diode structures are simply two TO-18 en-

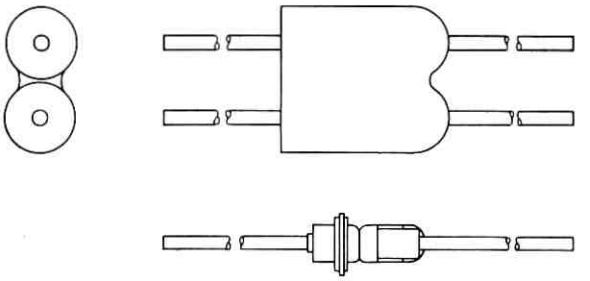


Fig. 1 — Outline of mesa diode pairs and package outline.

capsulations which are snap fitted into an acytal copolymer plastic block. The lead arrangement shown in Fig. 2(a) was used for immediate production and field replacements; the straight-through lead arrangement shown in Fig. 2(b) is being used in new equipment which incorporates modified circuit boards. The latter structure requires neither lead trimming nor forming for insertion. Code and date markings are molded into the plastic carrier which eliminates the need for coding individual finished devices. The plastic carriers are bullet-shaped to identify polarity and are color coded to provide positive differentiation of compressor and expandor pairs in the equipment assembly areas. The leads of the device are solder coated to facilitate wave soldering to printed circuit boards.

The essential features of both mesa and planar type wafers are depicted in Fig. 3. Fig 3(a) shows the mesa structure used in the earlier diode. In this case, a p-n junction is formed approximately

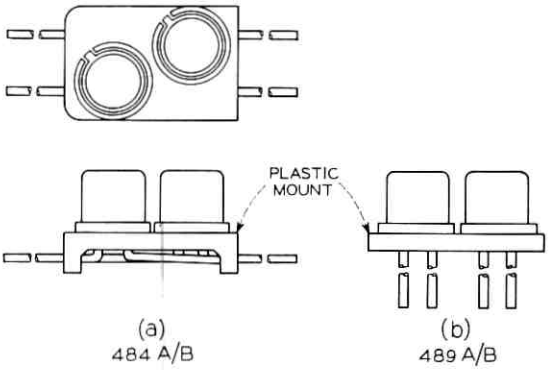


Fig. 2 — Outlines of 484A/B and 489A/B diode pairs.

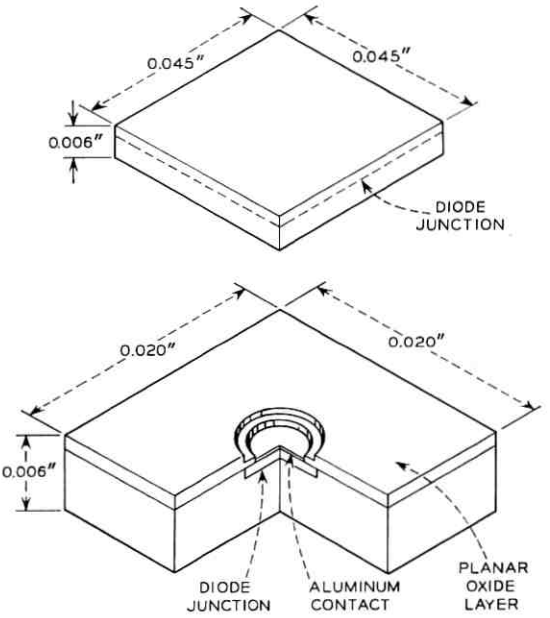


Fig. 3 — (a) Mesa structure of component diodes. (b) Planar structure of component diodes.

0.0002 inch below both faces of a one-inch diameter silicon wafer by gaseous diffusion producing a p-n-p structure. One of the p-n junctions is then removed from the wafer by mechanical lapping. The lapped slices are next plated with nickel and gold to form ohmic contacts. The wafers are then cut into 0.045-inch square chips by a diamond sawing operation to produce the final chip. This chip is subsequently eutectic-bonded to the package mounting stud, etched to remove a controlled amount of sawing damage (thus adjusting the impedance to the nominal value), and finally spring contacted during final encapsulation to complete the device. A cut-away view of this structure is shown in Fig. 4.

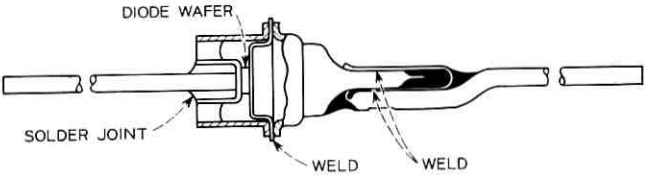


Fig. 4 — Cut-away view of mesa diode package.

Fig. 3(b) depicts the mechanical features of the planar wafer. In this case, a p-n junction is formed in the p-type silicon by diffusing phosphorus through a 0.008-inch hole cut into the protective layer of silicon dioxide. Since the starting crystal is very heavily doped with boron (approximately 2×10^{19} atoms/cc), it is difficult to overdope this material and produce a deep junction; in this device, the junction lies 0.0003 inch below the initial surface of the silicon. As explained elsewhere, the silicon is also very heavily gold doped by a high temperature diffusion to control the recombination-generation current and thus the diode multiplication factor which in turn controls diode impedance. An aluminum contact is evaporated and alloyed selectively into the hole in the oxide to complete wafer fabrication. The planar wafer is next eutectic-bonded to a gold-plated TO-18 header and a thermocompression wire bond is made between the metal button and the header lead. Final closure of the device is accomplished by resistance welding a Kovar can to the gold-plated Kovar header. A barium oxide impregnated porous nickel cylinder is brazed to the top of the can and serves as a moisture getter. A cut-away view of the individual planar device is shown in Fig. 5.

3.4.2 *Fabrication Process*

Many of the basic processes used to fabricate these diodes are common to other planar silicon devices and have been presented elsewhere.[11, 12] This section, then, will deal mainly with those processes which determine the forward impedance, noise and stability aspects of the device. A basic flow chart of the major assembly operations is presented in Fig. 6. In this chart, the header assembly operations, getter fabrication, activation and assembly operations and the semiconductor crystal growing operations are not shown.

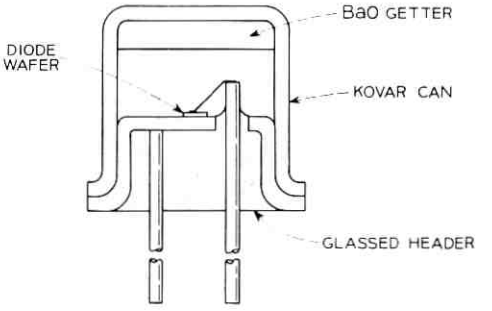


Fig. 5 — Cross-section view of planar component diode.

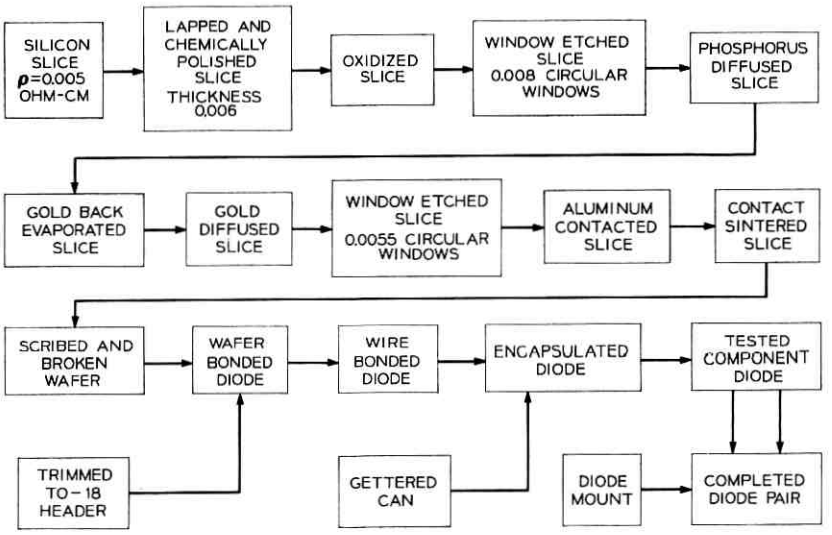


Fig. 6 — Flow chart for fabrication of planar diodes. Only fundamental operations are shown.

The first fundamental design choice involves the selection of resistivity type and doping level. The choice of p-type silicon allows use of a junction diffusant (phosphorus) which can be easily cleaned off of the surface of the wafer contact area. The choice of very heavily doped starting material is also of paramount importance in producing a stable device. It has been demonstrated[13] that alkali ions, a universal source of contamination, can electrolize through a protective silicon dioxide layer at high temperature under reverse bias and invert the conductivity type of the p-type material surrounding the junction. This inverted area can cause high channel currents to flow and also drastically increase the capacitance of a device when operated under reverse bias.

When operated in the forward direction, a "channeled" device will exhibit a multiplication factor of typically 2–4 and occasionally up to 10. Obviously, such changes would drastically shift the impedance levels of the device. However, with starting material doped to a level of 2×10^{19} atoms/cm^3, it is estimated from the curves of Ref. 14 that 10^{13} surface charges/cm^2 would be necessary to invert the material. Contamination levels of this magnitude are not encountered if minimal care is exercised in the oxide growing, diffusion and contact evap-

oration steps. Hence, as discussed in Section 3.2, the choice of heavily doped starting material essentially eliminates the contribution of channel recombination-generation current to the total diode current when compared to the bulk recombination-generation current. Also, as calculated in Section 3.2, the diffusion current contribution to the total diode current for this (or any practical) starting resistivity is also negligible compared to the bulk recombination-generation current.

The next pair of design choices, heavy gold doping and planar oxide-passivated technology, combine to produce a very large bulk controlled recombination-generation current and a negligibly small surface current contribution. As indicated in Fig. 6, the polished slice is first oxide passivated and then is selectivly etched to open 0.008-inch circular holes in the oxide using photolithographic techniques. Kodak Thin Film Resist (KTFR) is used as the emulsion in the photo-shaping operation. After junction diffusion, the junction assumes the shape shown in Fig. 3(b). The junction diffuses laterally as well as vertically. Lateral diffusion under the oxide layer provides a p-n junction which terminates at the semiconductor surface at a low surface charge location (under the passivating oxide). The low surface charge results in low surface recombination current. Thus, the resultant low resistivity, planar, oxide-protected, heavily-gold-doped combination results in a device which is completely bulk controlled and capable of being predictably controlled in manufacture.

The gold doping level must next be selected to provide the desired value of diode multiplication factor and hence forward impedance. Since many mesa devices are currently in field service, and since both the N2 and N3 were designed to accommodate this device, it was desirable to attempt to set the impedance level at a value of 1035 ohms at 50 microamperes or a multiplication factor of approximately 2. Since values of the multiplication factor at room temperature from gold doping as high as 1.85 had been reported in the literature,[8] this approach appeared to offer promise of successfully achieving the desired objective. As shown in Fig. 7, the impedance level of the device is a very strong function of the gold diffusion temperature. As can be seen from this combined plot of impedance and maximum solid solubility[15] of gold in silicon as a function of temperature, the impedance level is directly related to only the bulk properties of the device as calculated in Section 3.2 and discussed previously in this section. The impedance values presented in this plot were achieved with other impedance controlling variables held constant. In particular, the time and tempera-

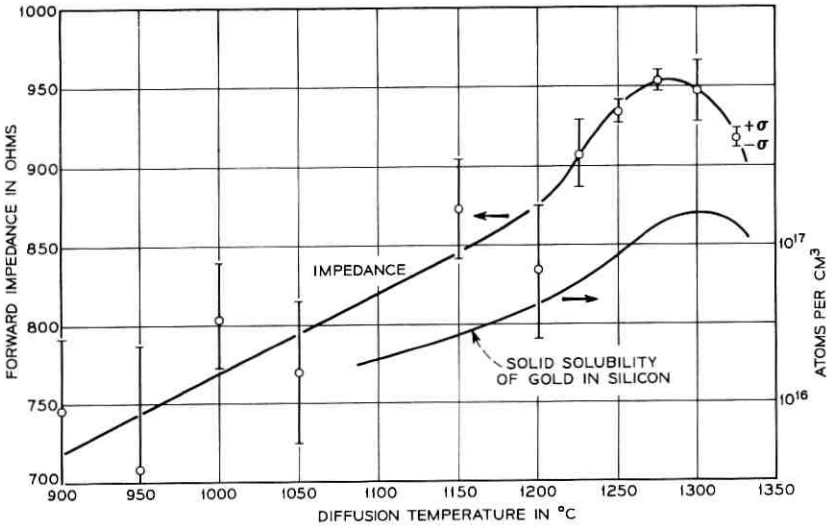


Fig. 7 — Forward impedance at 50μA bias current as a function of gold diffusion temperature. Diffusion time 10 minutes. Contact sinter time 3 minutes. Solid solubility after Trumbore.

ture used for contact sintering to provide contact adherence were held at 3 minutes and 625°C, respectively. Fig. 16 presents the variation of impedance level with contact sintering time at 625°C. It can be seen that maximum impedance results when the contact is not sintered. Heat treatment of the gold-loaded slice (even without metal contacts present) results in lowered impedance probably through an oxide-gettering or precipitation mechanism. With minimum contact sintering, average values of impedance as high as 990 ohms at a forward current of 50 microamperes have been achieved for a gold diffusion temperature of 1300°C. The corresponding multiplication factor for these experimental conditions is 1.91. For good mechanical adherence of the contact it was desirable to sinter the contact at about 625°C for 9.5 minutes (a standard process); hence, the impedance level of the redesigned device was set at 900 $\pm$ 35 ohms for a gold diffusion temperature of 1300°C. This shift in impedance nominal from the mesa component diode nominal value of 1045 $\pm$ 125 ohms necessitated a change of a few resistor values in the compandors.

As discussed in Section 3.3, noise in the low audio frequency range was a serious problem with the mesa diode. While a detailed study of the physical noise mechanisms in silicon was not undertaken in this

development, design information was obtained which clearly indicates methods of controlling this important parameter. Control of the $1/f$ low-frequency noise results as a by-product of heavy gold doping for impedance control. As illustrated in Fig. 10, the noise voltage of the device in the 200-3500 Hz band when biased in the forward direction at 2.5 microamperes is independent of the gold level up to about 1200°C then drops sharply and begins to level out beyond 1300°C. Since the mesa device saw no high temperature gold diffusion, no beneficial effect of the gold was realized.

As seen from Fig. 10, at the specified diffusion temperature of 1300°C the bulk of the planar component diodes are approaching the test set lower limit of 2.4 microvolts and no devices are approaching the compressor or expandor limits of 20 and 40 microvolts, respectively. This parameter is now easily controlled in manufacture; hence, both compressor as well as expandor limits have been set at 20 microvolts.

After oxidation, diffusion and contacting, the slices are simply diamond scribed, cracked apart, eutectic (gold-silicon) wafer bonded and thermocompression wire bonded to the T0-18 header. Finally, the metal can containing an activated moisture getter is resistance welded to the assembled header. The excellent device stability which will be presented in a later section is attributable to the use of very low resistivity semiconductor material, to extremely high gold doping and to the use of oxide passivation techniques.

The design factors discussed in this section combine to produce a device with a very narrow range of impedance, a low noise voltage, extremely stable electrical characteristics and which can be produced with good manufacturing control.

3.4.3 *Design Variables*

The diffusion of gold into silicon is a complex process involving interstitial-substitutional equilibrium.[16] In addition, both diffusion constant and solid solubility are partially dependent on the concentration of other impurities such as boron and phosphorus.[17, 18] Because complete data were not available on the entire ranges of interest of diffusion temperature or boron and phosphorus concentration, and because data were not available on the effect of annealing which would necessarily occur during the contact sintering, the effects of gold diffusion temperature and time and contact sintering time were determined empirically. A matrix experiment was performed where one parameter was varied, and then another etc., holding the other parameters constant.

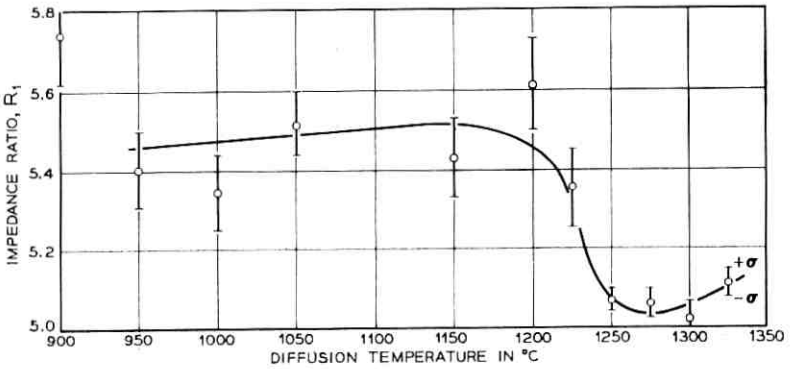


Fig. 8 — Impedance ratio, $R_1 = Z(10\mu A)/Z(50\mu A)$, as a function of gold diffusion temperature. Diffusion time 10 minutes. Contact sinter time 3 minutes.

Selected results of the experiments are shown in Figs. 7 through 16. Each point represents the average of about 40 diodes. Unless indicated otherwise, the gold diffusion temperature was 1300°C, the gold diffusion time was 10 minutes, and the sintering time was 3 minutes.

The effect of gold diffusion temperature on $Z(50\mu A)$, R_1, R_2, noise voltage, capacitance and forward voltage is shown in Figs. 7 through 12. Below about 1200°C the gold diffusion has little effect, but at higher temperatures the diode parameters depend mainly on the gold solubility. Above 1200°C the spread of measurements was also much smaller, which indicates that the bulk rather than the surface prop-

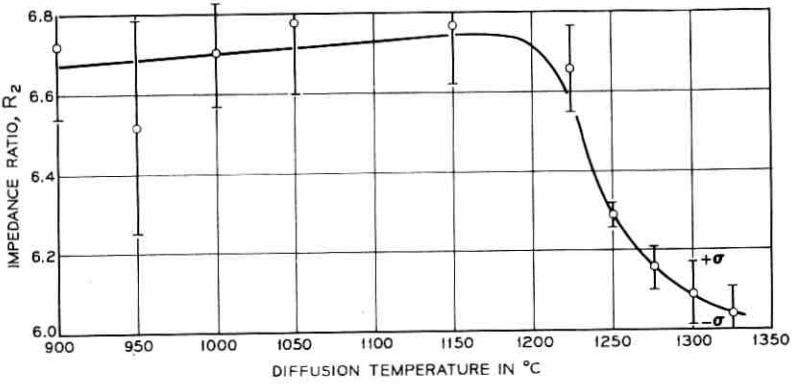


Fig. 9 — Impedance ratio, $R_2 = Z(50\mu A)/Z(300\mu A)$, as a function of gold diffusion temperature. Diffusion time 10 minutes. Contact sinter time 3 minutes.

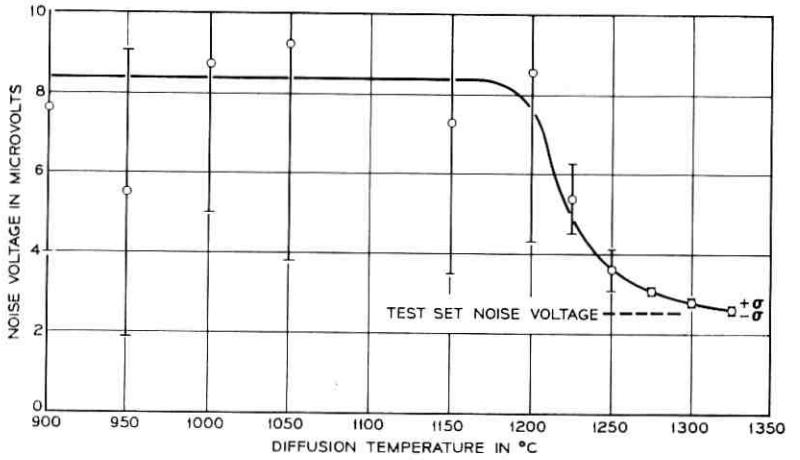


Fig. 10 — Noise voltage (200-3500 Hz) at 2.5μA forward bias current as a function of gold diffusion temperature. Diffusion time 10 minutes. Contact sinter time 3 minutes.

erties were dominant. This information resulted in the choice of a high gold diffusion temperature of 1300°C.

The effect of gold diffusion time on $Z(50\mu A)$, forward voltage and noise voltage is shown in Figs. 13 through 15. At times greater than 10 minutes the forward voltage and noise did not change with time. However, the diode impedance and hence the multiplier, n, did change which means that an equilibrium condition was not reached. Since

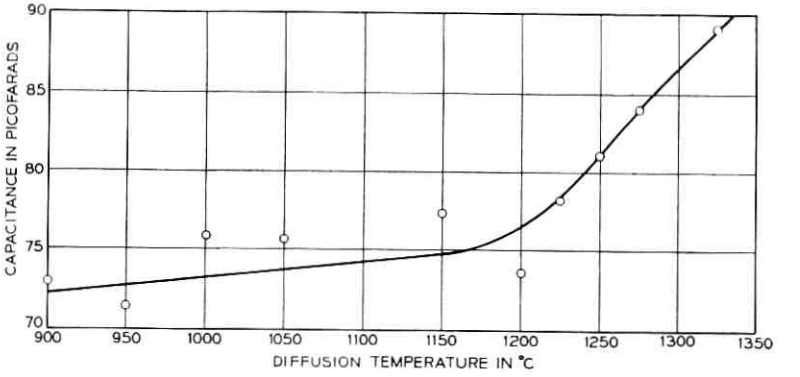


Fig. 11 — Capacitance at 1 MHz and zero bias as a function of gold diffusion temperature. Diffusion time 10 minutes. Contact sinter time 3 minutes.

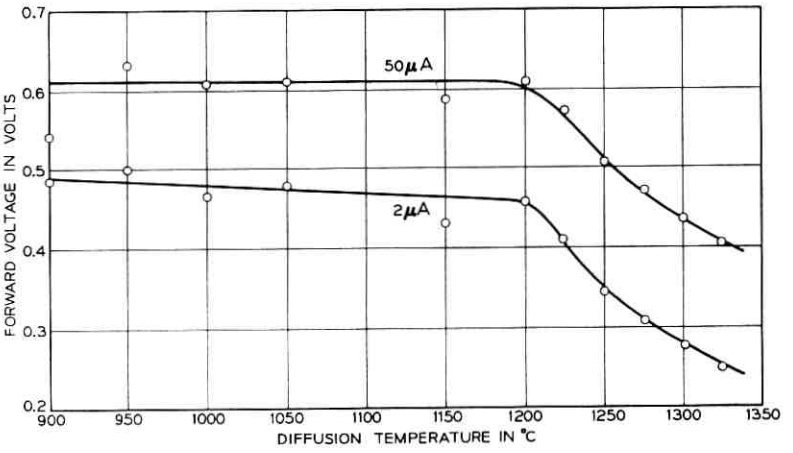


Fig. 12 — Forward voltage at 2μA and 50μA bias currents as a function of gold diffusion temperature. Diffusion time 10 minutes. Contact sinter time 3 minutes.

gold is known to precipitate or collect in phosphorus doped silicon dioxide[19] and at dislocations, as well as to form a complex with phosphorus, equilibrium would not be expected only on the basis that solid solubility had been reached. Ten minutes was chosen for the diffusion time.

The importance of contact sintering time can be seen in Fig. 16 which shows forward impedance, $Z(50\mu A)$, as a function of sintering time. A sintering time of 9.5 minutes was chosen because it corresponds to a standard transistor process which results in good contact adher-

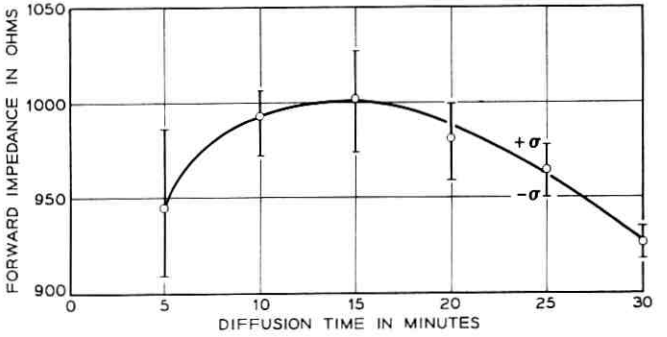


Fig. 13 — Forward impedance at 50μA bias current as a function of gold diffusion time. Diffusion temperature 1300°C. Contact sinter time 3 minutes.

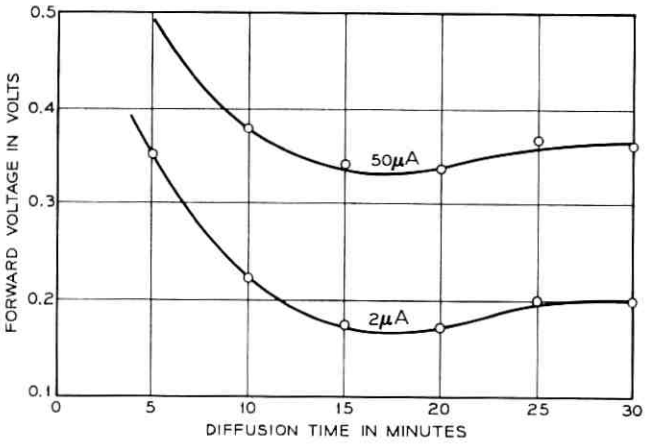


Fig. 14 — Forward voltage at 2μA and 50μA bias current as a function of gold diffusion time. Diffusion temperature 1300°C. Contact sinter time 3 minutes.

ence and because the slope of impedance versus sinter time is low at that time.

Studies were carried out in which the diffusion depth was varied from 0.3 to 0.8 mils while holding the gold diffusion to the standard conditions. There was no effect on diode parameters.

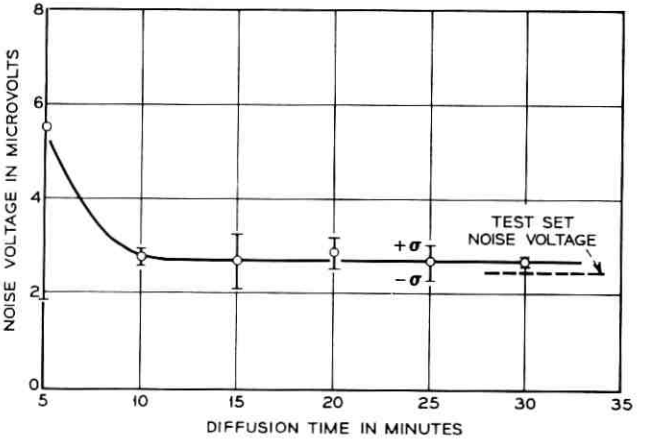


Fig. 15 — Noise voltage (200–3500 Hz) at 2.5μA bias current as a function of gold diffusion time. Diffusion temperature 1300°C. Contact sinter time 10 minutes.

3.4.4 *Electrical Characteristics*

The salient electrical characteristics of the planar diodes namely: forward impedance, impedance ratios and impedance differences are summarized in Table I. The impedance of a typical unit is shown in Fig. 17 in the range of 1μA to 10mA bias. Impedance measurements are made at a frequency of 1000 Hz. In the frequency range of interest, the capacitance has negligible effect on the impedance. In the worst case, at the highest frequency of interest (3500 Hz), and at a forward current bias of 10μA, the capacitive reactance is more than two orders of magnitude greater than the resistive component. Therefore, the difference between the total magnitude of impedance and the resistive component is less than 0.01 percent. The dependence of forward impedance with temperature is shown in Fig. 18 for $Z(50\mu A)$. The temperature coefficient of 0.85 ohms/°C is less than would be predicted directly from (2) and implies a temperature dependence of the multiplier, n, which has been noted elsewhere.[8]

Although no requirements are placed on forward voltage, a plot of forward voltage versus forward current for a typical component diode is shown in Fig. 19 for completeness. It is, of course, the linearity of this semilogarithmic plot which results in the excellent impedance control of the new diodes with current.

The stability requirement placed on the diodes is that the impedance value, $Z(50\mu A)$, should not drift with time; in particular it should not drift in the first few minutes of application of bias. The short term drift, as has been noted, was a problem with the mesa diodes. No short term drift has been detected in the planar diodes by a test system

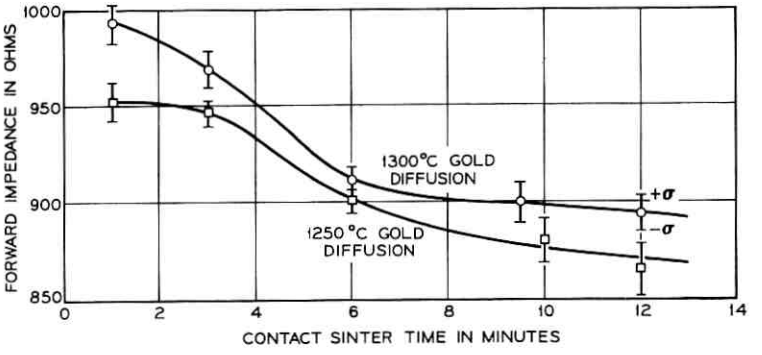


Fig. 16 — Small signal forward impedance at 50μA bias as a function of contact sinter time. Nominal sintering temperature was 625°C.

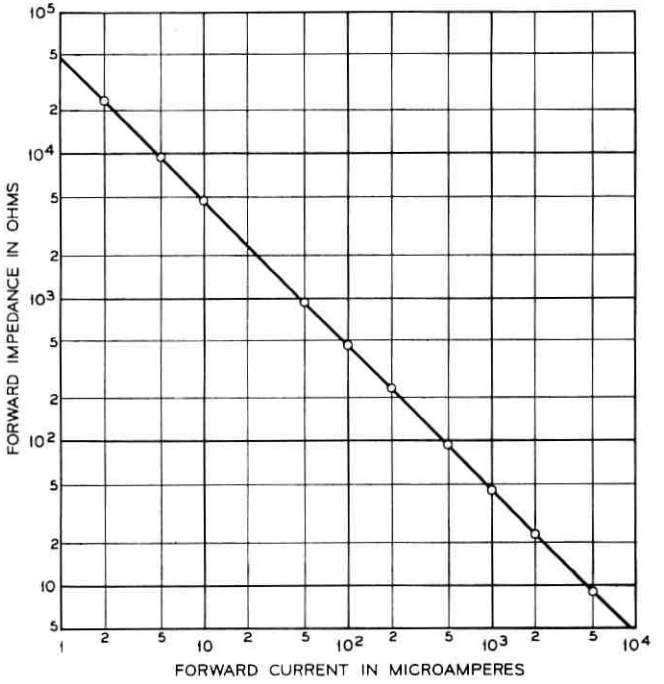


Fig. 17 — Typical forward impedance versus bias current at 25°C for component diode.

capable of detecting a drift of 1 ohm (0.1 percent change). Likewise, no long term drift has been detected either. In one life study a sample of 24 component diodes showed a drift of less than 0.5 percent (which was test set limit) after 4000 hours of greatly accelerated switched power aging (I_0 = 50 mA, V_R (peak) = 5V) at an ambient of 150°C. Another important characteristic is the noise generated by the diodes in the frequency range 200 Hz to 3500 Hz (C message weighting). The noise appears as a hissing sound to the listener when no voice signal is present. Measurements are made with 17,000 ohms in parallel with the diode or diode pair which is what appears in the actual circuit. Since the diode impedance at 2.5 μA is comparable to 17,000 ohms, and the noise voltages add as the square root of the sum of the squares, the measured noise voltage of two diodes in series is actually less than the noise voltage of either diode singly. The circuit requirement was less than 20 μV rms for a compressor pair and 40 μV rms for an expandor pair. Therefore, by requiring a single diode to have less than

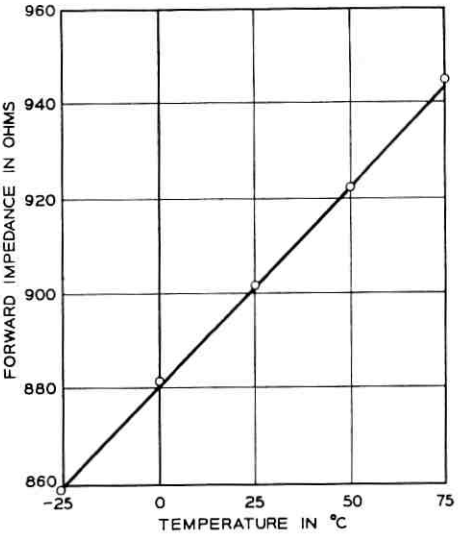


Fig. 18 — Typical forward impedance at 50μA bias current as a function of temperature.

20-μV rms noise, the pairs are guaranteed to meet the 20-μV rms limit. Most diodes had noise voltages less than or comparable to the test set limit of 2.4 μV rms. A check of 862 diodes produced during the development showed only 1 device to fail the noise limit.

Measurements made on noisy mesa diodes and other diodes produced

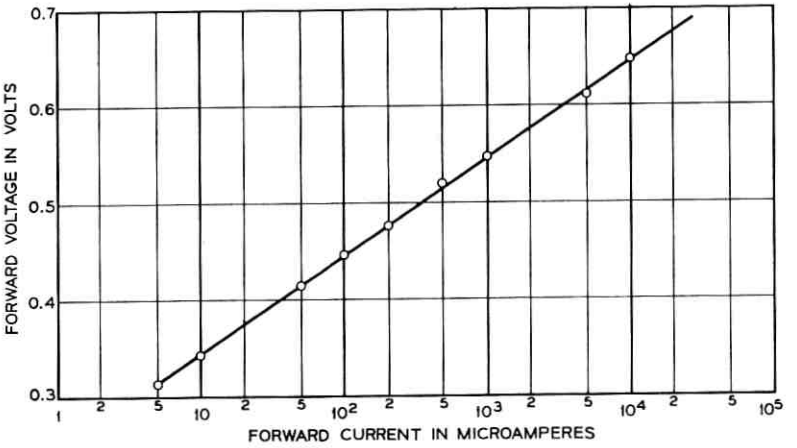


Fig. 19 — Forward current versus forward voltage for typical component diode.

during the development of the planar diode and the reasons for the noise improvement are discussed in the next section.

3.4.5 *Noise Discussion*

The decrease of noise voltage with increasing gold doping can be seen in Fig. 10.* Based on measurements made on units with measur-

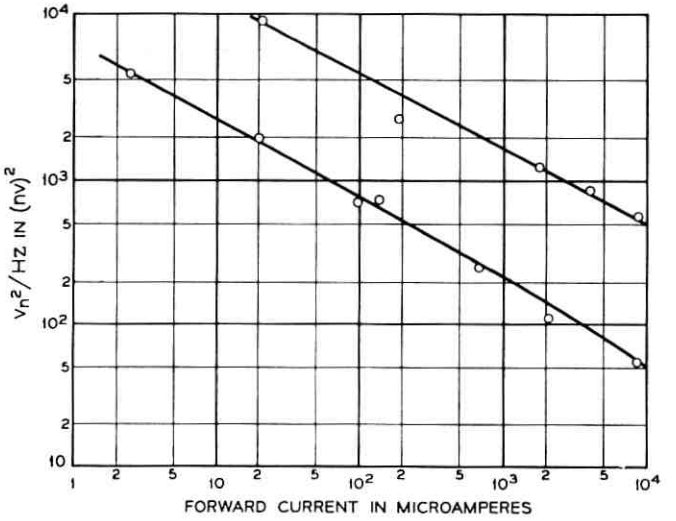


Fig. 20 — Noise voltage squared as a function of forward bias current, I_F, for two noisy development diodes.

able noise, the noise is $1/f$ noise over the audio frequency range, i.e.;

$$\Delta v_n^2 = (\text{const}/f)\Delta f$$

or

$$\Delta i_n^2 = (\text{const}/f)\Delta f,$$

where v_n = noise voltage, i_n = noise current and Δf = small frequency range.

Measurement as in Fig. 20 shows that the dependence of noise voltage on total dc current, I_F, is

$$v_n^2 = (\text{const})/I_F^{0.6}.$$

* Except for the noise voltages plotted in Fig. 10, which are measured as described in the last section, all noise voltages are equivalent open circuit voltages. All noise currents are equivalent short circuit currents.

Because of this dependence of noise voltage on forward current, the noise limit is specified at the low current of 2.5 μA.

Note that because $i_n = v_n/Z_f$ and $Z_f = nkT/qI_F$,

$$i_n^2 = (\text{const})I_F^{1.4}. \tag{12}$$

The decrease of noise at a given forward bias current, I_F, with increasing gold doping may be explained in the following manner. As recombination-generation centers are increased, the forward voltage, V_F, required to attain a given forward current decreases. If there is a secondary current (much less in magnitude than the recombination-generation current), which is the noise generating current, and if the noise due to this current increases with increasing forward bias voltage, then adding gold decreases the forward voltage for a specified current and the decreased forward voltage results in lower noise. This secondary current is quite likely associated with surface, bulk or channel leakage components or excess tunneling current derived from anomalous intermediate energy states.

If the above explanation is correct, the noise current measured at a specified forward voltage should be the same for various gold doping levels. The noise current is compared for a forward voltage of 0.463 volts for several groups from Fig. 10. The reason for comparing noise currents rather than noise voltages will be made clear shortly. Average noise voltages from Fig. 10 were corrected for test set noise and for the parallel 17,000-ohm resistor and converted to noise current.

$$v_n^2 \text{ (corrected)} = v_n^2 \text{ (measured)} - v_n^2 \text{ (set) where } v_n^2 \text{ (set)} = 2.4\mu V$$

$$v_n = v_n \text{ (open circuit)} = v_n \text{ (corrected)} \ (1.7 \times 10^4 + Z_f)/1.7 \times 10^4$$

$$i_n = v_n/Z_f .$$

The V-I characteristics of each group gave the bias current, I_F, for $V_F = 0.463$ volts for that group. The empirical equation (12) was used to find the noise current at this new current, since the constant in (12) could be found from the measurement at 2.5 μA above. Calculations were made for gold diffusion temperatures of 1150, 1200, 1225, and 1250°C where the greatest change in noise appeared to take place. The results, in Table II, are in rather good agreement with the hypothesis that the noise current depends only on the voltage V_F.

The fact that the noise must be described as a current generator (rather than a voltage generator) follows logically from a circuit analysis of the physical diode. An equivalent circuit for the diode is given

TABLE II — NOISE CURRENT AND DC CURRENT AT $V_F = 0.463$ VOLTS.

Gold diffusion temperature	1150°C	1200°C	1225°C	1250°C
I_F	3.75μA	2.5μA	6.0μA	21μA
i_n (short circuit)	1.05nA	0.95nA	1.03nA	1.36nA

in Fig. 21(a), where I_d is the diode current calculated earlier and I_x is an excess current.

An equivalent ac circuit including noise sources is given in Fig. 21(b), where x refers to excess current quantities and d to the dominant diode current quantities. If it is assumed that $I_x \ll I_d$, $Z_x \gg Z_d$ and $i_{nx} \gg i_{nd}$, the circuit of Fig. 21(c) results. The Thevenin equivalent of Fig. 21(c) is shown in Fig. 21(d). The noise voltage, v_n, is dependent on quantities related to two independent currents. The noise voltage measured at a specified voltage, changes with gold doping because Z_f (which equals nkT/I_F) changes with gold doping, while i_{nx} remains constant. Thus, the noise current is directly related to the noise mechanism, while the noise voltage is indirectly related.

As previously mentioned, there are several candidates for the current which produces the excess noise. It does not appear to be associated with bulk recombination current because gold doping does not change it. It could be associated with surface, channel or bulk leakage

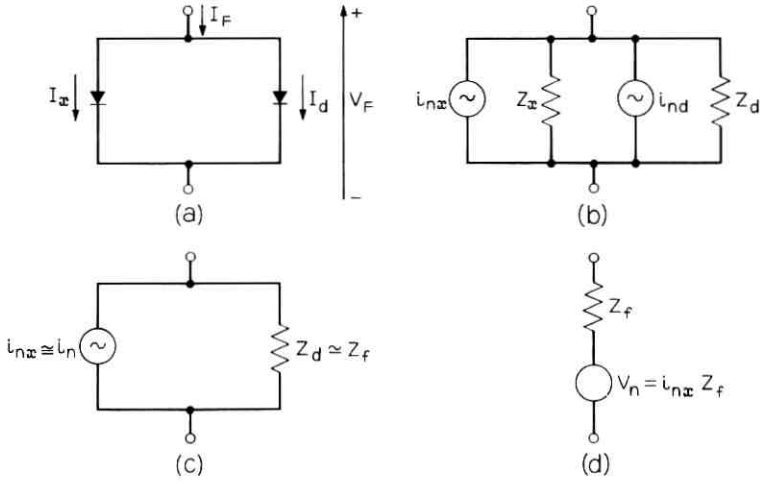


Fig. 21 — Equivalent circuits of diode with noise sources.

currents since $1/f$ noise has been widely reported for these components. It could also be the excess current associated with tunneling[20] since $1/f$ noise has been reported for this current in germanium.[21] While these diodes do not exhibit measurable tunnel current, they are near the tunnel diode doping levels.

IV. CONCLUSIONS

A new semiconductor compandor diode has been developed in which the critical small signal forward impedance characteristics are controlled by bulk material properties. The heavy gold doping employed in this design forces bulk recombination-generation currents to dominate over all surface, channel and diffusion currents and results in a low-noise device with well-controlled electrical characteristics. Oxide passivation and very low resistivity semiconductor material combine to produce an extremely stable device capable of being manufactured at yields governed almost exclusively by assembly workmanship. These devices were initially designed for use in the N2 and N3 Carrier Systems and have also been incorporated into the 3A Echo Suppressor System as a variolosser element.

V. ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of H. W. Hoffman, including the design and construction of the impedance and noise test sets and the drift and aging studies.

REFERENCES

1. Mathes, R. C. and Wright, S. R., The Compandor — An Aid Against Static in Radio Telephony, B.S.T.J., *13,* July, 1934, pp. 315–332.
2. Carter, R. O., Theory of Syllabic Compandors, Proc. IEEE, *111,* March, 1964, pp. 503–513.
3. Lundry, W. R. and Willey, L. F., The N2 Carrier Terminal Circuit Design, B.S.T.J., *44,* May-June, 1965, pp. 761–785.
4. Kahl, W. E. and Pedersen, L., Some Design Features of the N1 Carrier Telephone System, B.S.T.J., *30,* April, 1951, p. 418.
5. Body, L. C., Howard, J. D., and Pedersen, L., A New Carrier System for Rural Service, B.S.T.J., *36,* March, 1957, p. 357.
6. W. Shockley, The Theory of P-N Junctions in Semiconductors and P-N Junction Transistors, B.S.T.J., *28,* July, 1949, pp. 435-489.
7. Sah, C. T., Noyce, R. N., and Shockley, W., Carrier Generation and Recombination in P-N Junctions and P-N Junction Characteristics, Proc. IRE, *45,* September, 1957, pp. 1228–1243.
8. Sah, C. T., Effect of Surface Recombination and Channel on P-N Junction and Transistor Characteristics, IRE Trans. Electron Devices, *ED-9,* 1962, p. 94.
9. Bakanowski, A. E. and Forster, J. H., Electrical Properties of Gold Doped Diffused Silicon Computer Diode, B.S.T.J., *39,* January, 1960, pp. 87–103.

10. Phillips, A. B., *Transistor Engineering,* McGraw-Hill Book Company, Inc., New York, N.Y., 1962, p. 68.
11. Tanenbaum, M. and Thomas, D. E., Diffused Emitter and Base Transistors, B.S.T.J., *35,* January, 1956, pp. 1–22.
12. Attala, M. M., Tannenbaum, E., and Scheibner, E. J., Stabilization of Silicon Surfaces by Thermally Grown Oxides, B.S.T.J., *38,* May, 1959, pp. 749–783.
13. Mathews, J. R., Griffin, W. A., and Olson, K. H., Inversion of Oxidized Silicon Surfaces by Alkali Metals, J. Electrochem. Soc., *112,* 1965, p. 899.
14. Young, C. E., Extended Curves of the Space Charge, Electric Field and Free Carrier Concentration at the Surface of a Semiconductor and Curves of Electrostatic Potential Inside a Semiconductor, J. Appl. Phys., *32,* March, 1961, pp. 329–332.
15. Trumbore, F. A., Solid Solubilities of Impurity Elements in Germanium and Silicon, B.S.T.J., *39,* January, 1960, pp. 205–233.
16. Wilcox, W. R. and LaChapelle, T. J., Mechanism of Gold Diffusion in Silicon, J. Appl. Phys., *35,* January, 1964, pp. 240–246.
17. Joshi, M. L. and Dash, S., Distribution and Precipitation of Gold in Phosphorus-Diffused Silicon, J. Appl. Phys., *37,* May, 1966, pp. 2453–2457.
18. Wilcox, W. R., LaChapelle, T. J., and Forbes, D. H., Gold in Silicon: Effect on Resistivity and Diffusion in Heavily-Doped Layers, J. Electrochem. Soc., *111,* pp. 1377–1380.
19. Goetzberger, A. and Shockley, W., Metal Precipitates in Silicon p-n Junctions, J. Appl. Phys., *31,* October, 1960, pp. 1821–1824.
20. Chynoweth, A. G., Feldmann, W. L., and Logan, R. A., Excess Tunnel Current in Silicon Esaki Junctions, Phys. Rev., *121,* February, 1961, pp. 684–694.
21. Yajima, T. and Esaki, L., Excess Noise in Narrow Germanium p-n Junctions, J. Phys. Soc. Japan, *13,* November, 1958, pp. 1281–1287.

Noise-Like Structure in the Image of Diffusely Reflecting Objects in Coherent Illumination

By L. H. ENLOE

(Manuscript received April 13, 1967)

Holographic and other imaging systems utilizing coherent light introduce a speckled or noise-like pattern in the image of a diffuse object which severely degrades image quality. It is desirable to understand this effect quantitatively. Intelligent design in many cases requires knowledge of the mean-square value, spatial power spectral density, and autocorrelation junction of the noise-like fluctuations. These quantities have been determined for the image of a uniform diffuse object. Major results are:

(i) The mean-square value of the fluctuation in the image intensity is equal to the square of the mean intensity.

(ii) One can decrease the relative magnitude of the noise-like fluctuations at the cost of a corresponding increase in the aperture required of the optical system (or hologram) over that required to resolve the desired image in a spatial frequency sense. In a holographic facsimile or TV system, this calls for a corresponding increase in electrical bandwidth.

(iii) The improvement in (ii) is not possible for direct viewing with the human eye, since the resolution of a healthy eye is known to be limited by diffraction at the iris.

I. INTRODUCTION

Holographic and other imaging systems using coherent light have been receiving considerable attention lately.[1, 2, 3, 4] Most analyses on this subject assume that the object reflects specularly, or transmits specularly if the object is a transparency, i.e., the reflectivity or transmissivity of the object varies smoothly. Most objects, however, are more nearly diffuse reflectors. When the image of a diffusely reflecting object is formed it will be covered with a noise or grain-like structure[5, 6, 7] which is the speckle pattern which one sees when laser light is used to illuminate an object.

In this paper we investigate the noise-like or speckled nature of the image of a uniform diffuse surface. It should be emphasized that we are interested in the properties of the image in contradistinction to the direct backscattered field studied by Goldfisher.[8] We show that the intensity consists of two parts. The first is the mean or ensemble average intensity and is proportional to the intensity which would be obtained if incoherent light were used for illumination. This is the desired component of the image and might be likened to a signal. The second part of the image is the speckled or noise-like component which tends to obscure the average intensity. This noise-like component occurs because of the random phase angles associated with the scattering centers comprising the microstructure of the diffuse surface. The spatial autocorrelation function and power spectral density of the speckle pattern in the image are found, and are shown to be dependent upon the size of the aperture stop. It is shown that the variance of the intensity fluctuation is equal to the square of the mean intensity. The fluctuation may be reduced, however, if one is willing to sacrifice resolution by recording the image on film whose resolution is much poorer than that set by the aperture of the optics. Unfortunately, this alternative is not available when viewing with the human eye, since the resolution of a healthy eye is known to be determined by the diffraction limit of the iris.[9] This seems to place definite limitations upon the use of coherent light in visual systems.

II. ARBITRARY APERTURE

The model which we shall use for a diffuse object is shown in Fig. 1. Although the object is shown to be a granular transparency, it could equally well have been shown as a reflector without loss of generality. The essential point is that a monochromatic coherent light wave of unit intensity is assumed to be scattered by a random set of point scatterers. Each scatterer is assumed to be a unit scatterer which is many wavelengths in depth from its neighbor. The relative phase of the wave scattered from each scatterer may be assumed to be a random variable which is statistically independent of the phase of the waves scattered from other scatterers. Any phase change between 0 and 2π is equally likely. Multiple scattering will be neglected.

The scattered field just to the right of the granular transparency can be expressed by the equation

$$F_0(x, y) = \sum_{i=1}^{K} \delta(x - x_i, y - y_i)\epsilon^{j\theta_i}, \tag{1}$$

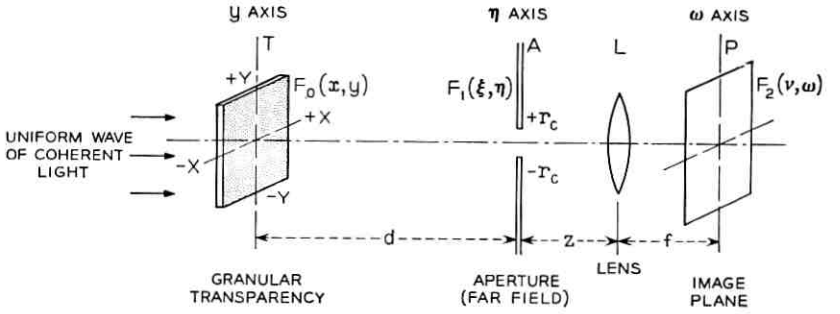


Fig. 1 — A uniform wave of coherent light is incident on a transparency composed of randomly distributed unit point scatterers. Light collected by the aperture A, placed in the far-field, is imaged by lens L on plane P.

where θ_i is the relative phase of the wave scattered from the scatterer located at $x = x_i$, $y = y_i$. θ_i, x_i and y_i are assumed to be random variables uniformly distributed in the intervals $(0,2\pi)$, $(-X,+X)$ and $(-Y,+Y)$, respectively. Notice that because of our assumptions, the statistics of the scattered field are independent of any deterministic variation in the phase of the illuminating field.

A Fourier transform relationship exists between the scattered field given by (1) and its far-field. The far-field is given by

$$\begin{aligned} F_1(\xi, \eta) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F_0(x, y)\epsilon^{j(2\pi/\lambda d)(x\xi+y\eta)}\, dx\, dy \\ &= \sum_{i=1}^{K} \epsilon^{j\theta_i + j(2\pi/\lambda d)(x_i\xi+y_i\eta)}, \end{aligned} \tag{2}$$

where we have suppressed the time factor $\epsilon^{+j\Omega t}$. Notice that each scatterer has produced a plane wave, and that the slope of the phase front of each wave with respect to the ξ, η axes is determined by the position (x_i, y_i) of the random scatterer.

Let the far-field $F_1(\xi, \eta)$ be passed through an aperture having an amplitude transmission function $H(\xi, \eta)$, and then through a lens which is placed a distance z behind the aperture. Since the field at the back focal plane of a lens is a Fourier transform like function of the field in front of the lens, an image of the granular transparency, as modified by the aperture, will be formed in the back focal plane, and is given by[10]

$$\begin{aligned} F_2(\nu, \omega) &= \epsilon^{jc(\nu^2+\omega^2)} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} H(\xi, \eta)F_1(\xi, \eta)\epsilon^{j(2\pi/\lambda f)(\xi\nu+\eta\omega)}\, d\xi\, d\eta \\ &= \epsilon^{jc(\nu^2+\omega^2)} \sum_{i=1}^{K} h\left(\frac{\nu}{\lambda f} + \frac{x_i}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_i}{\lambda d}\right)\epsilon^{j\theta_i}, \end{aligned} \tag{3a}$$

where $c = \pi(z - f)/\lambda f^2$, and where $h(t, u)$ and $H(\xi, \eta)$ are Fourier transform pairs in the sense

$$h(t, u) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} H(\xi, \eta) \epsilon^{j2\pi(\xi t + \eta u)} \, d\xi \, d\eta. \tag{3b}$$

Notice that except for the unimportant phase factor $\epsilon^{jc(\nu^2+\omega^2)}$, (3a) differs from (1) for the field at the granular transparency only in that $h(\ )$ functions have replaced the delta functions. That is to say, the delta function of light from the scatterer at (x_i, y_i) is reproduced as a broadened $h(\ )$ function located at $\nu = -(f/d)x_i$, $\omega = -(f/d)y_i$. The image is reversed, and magnified by the factor $m = f/d$. Notice that because of the random phase θ_i of each, the impulse functions will add vectorially in a random fashion when they overlap one another.

The situation is analogous to passing shot noise impulses through a low-pass filter having an impulse response $h(\ )$. The impulses are broadened into $h(\ )$ pulses whose width depends inversely upon the filter bandwidth. In the coherent light case, however, the process is two dimensional and the applied impulses have random phase angles distributed uniformly between 0 and 2π, rather than being constrained to be positive impulse functions as is the case for shot noise.

The quantity of greatest interest to us is the intensity of the image, which is found by multiplying the image field by its conjugate.

$$\begin{aligned} I(\nu, \omega) &= F_2(\nu, \omega) F_2^*(\nu, \omega) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{K} h\left(\frac{\nu}{\lambda f} + \frac{x_k}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_k}{\lambda d}\right) \cdot h^*\left(\frac{\nu}{\lambda f} + \frac{x_i}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_i}{\lambda d}\right) \epsilon^{j(\theta_k - \theta_i)}. \end{aligned} \tag{4}$$

The uniform diffuse object is assumed to exist in the region $-X \leqq x \leqq +X$, $-Y \leqq y \leqq +Y$. The number K of point scatterers in this region is a random variable, as are their positions (x_i, y_i) and their relative phase angles θ_i. We may, therefore, obtain the ensemble average of the image intensity I by averaging (4) with respect to the $2K + 1$ random variables consisting of the K positions (x_i, y_i), K phase angles θ_i, and K itself:

$$\begin{aligned} \bar{I} = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} & \left[\sum_{k=1}^{K} h\left(\frac{\nu}{\lambda f} + \frac{x_k}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_k}{\lambda d}\right) \epsilon^{j\theta_k}\right] \\ & \cdot \left[\sum_{i=1}^{K} h^*\left(\frac{\nu}{\lambda f} + \frac{x_i}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_i}{\lambda d}\right) \epsilon^{-j\theta_i}\right] \\ & \cdot W(x_1, y_1; x_2, y_2; \cdots x_K, y_K; \theta_1; \cdots \theta_K; K) \\ & \cdot dx_1 \, dy_1 \cdots dx_K \, dy_K \, d\theta_1; \cdots d\theta_K \, dK, \end{aligned} \tag{5}$$

where $W(\ )$ is the multi-dimensional probability density function.

Now the positions (x_i, y_i) are considered to be statistically independent variables, as are the relative phase angles θ_i. They are also independent of K, so we may simplify (5) to obtain

$$\bar{I} = \int_{-\infty}^{+\infty} W(K)\, dK \cdot \left[\sum_{k=1}^{K} \sum_{i=1}^{K} \int_{-X}^{+X} \frac{dx_1}{2X} \cdots \int_{-X}^{+X} \frac{dx_K}{2X} \int_{-Y}^{+Y} \frac{dy_K}{2Y} \int_{0}^{2\pi} \frac{d\theta_1}{2\pi} \cdots \int_{0}^{2\pi} \frac{d\theta_K}{2\pi} \cdot \left\{ \epsilon^{j(\theta_k - \theta_i)} h\left(\frac{\nu}{\lambda f} + \frac{x_k}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_k}{\lambda d} \right) \cdot h^*\left(\frac{\nu}{\lambda f} + \frac{x_i}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_i}{\lambda d} \right) \right\} \right]. \tag{6}$$

We see that the above expression vanishes unless $\theta_i = \theta_k$, i.e., $i = k$. Further, all of the $h(\ )$ functions have the same shape so that if the size of a resolution element in the image is small compared to the field of view, i.e., the extent of $h(\nu/\lambda f, \omega/\lambda f)$ is small compared to X and Y, then we may replace the limits of integration $\pm X$ and $\pm Y$ by $\pm \infty$ to obtain

$$\bar{I} = \frac{d^2 \lambda^2 \rho_1(0, 0)}{4XY} \int_{-\infty}^{+\infty} KW(K)\, dK. \tag{7}$$

$\rho_1\ (u, v)$ is the autocorrelation function of the aperture impulse function $h(\xi, \eta)$, i.e.,

$$\begin{aligned} \rho_1(u, v) &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} h^*(t, \tau) h(t + u, \tau + v)\, dt\, d\tau \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} H(\xi, \eta) H^*(\xi, \eta) \epsilon^{j2\pi(\xi u + \eta v)}\, d\xi\, d\eta. \end{aligned} \tag{8}$$

If we now assume that the number of scatterers per unit area of the transparency has a Poisson distribution of mean $\bar{N}$, then the mean intensity is

$$\bar{I} = d^2 \lambda^2 \bar{N} \rho_1(0, 0). \tag{9}$$

Next we wish to determine an expression for the autocorrelation function of the intensity, from which we may determine the spatial power spectral density and variance of the noise-like fluctuations. The autocorrelation function of the intensity as given by (4) is

$$\begin{aligned} R(\tau, t) &= \overline{I_2(\nu, \omega) I_2(\nu + \tau, \omega + t)} \\ &= \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left[\sum_{k=1}^{K} h\left(\frac{\nu}{\lambda f} + \frac{x_k}{\lambda d}, \frac{\omega}{\lambda f} + \frac{y_k}{\lambda d} \right) \epsilon^{j\theta_k} \right] \end{aligned} \tag{10}$$

$$\cdot\left[\sum_{i=1}^{K} h^*\left(\frac{\nu}{\lambda f}+\frac{x_i}{\lambda d},\frac{\omega}{\lambda f}+\frac{y_i}{\lambda d}\right)\epsilon^{-j\theta_i}\right]$$
$$\cdot\left[\sum_{m=1}^{K} h\left(\frac{\nu+\tau}{\lambda f}+\frac{x_m}{\lambda d},\frac{\omega+t}{\lambda f}+\frac{y_m}{\lambda d}\right)\epsilon^{j\theta_m}\right]$$
$$\cdot\left[\sum_{n=1}^{K} h^*\left(\frac{\nu+\tau}{\lambda f}+\frac{x_n}{\lambda d},\frac{\omega+t}{\lambda f}+\frac{y_n}{\lambda d}\right)\epsilon^{-j\theta_n}\right]$$
$$\cdot W(x_1, y_1; x_2, y_2; \cdots x_K y_K; \theta_1; \cdots \theta_K, K)\, dx_1 \cdots dK.$$

Because of the statistical independence of the phase angles θ_i, positions (x_i, y_i) and K, and because of the assumed uniform distribution, we may simplify (10) to

$$R(\tau, t) = \int_{-\infty}^{+\infty} W(K)\, dK$$
$$\cdot\sum_{k=1}^{K}\sum_{i=1}^{K}\sum_{m=1}^{K}\sum_{n=1}^{K}\int_{-\infty}^{+\infty}\frac{dx_1}{2X}\int_{-\infty}^{+\infty}\frac{dy_1}{2Y}\cdots\int_{-\infty}^{+\infty}\frac{dx_K}{2X}\int_{-\infty}^{+\infty}\frac{dy_K}{2Y}$$
$$\cdot\int_0^{2\pi}\frac{d\theta_1}{2\pi}\cdots\int_0^{2\pi}\frac{d\theta_K}{2\pi}\left[\epsilon^{j(\theta_k-\theta_i+\theta_m-\theta_n)}h\left(\frac{\nu}{\lambda f}+\frac{x_k}{\lambda d},\frac{\omega}{\lambda f}+\frac{y_k}{\lambda d}\right)\right.$$
$$\cdot h^*\left(\frac{\nu}{\lambda f}+\frac{x_i}{\lambda d},\frac{\omega}{\lambda f}+\frac{y_i}{\lambda d}\right)h\left(\frac{\nu+\tau}{\lambda f}+\frac{x_m}{\lambda d},\frac{\omega+t}{\lambda f}+\frac{y_m}{\lambda d}\right)$$
$$\left.\cdot h^*\left(\frac{\nu+\tau}{\lambda f}+\frac{x_n}{\lambda d},\frac{\omega+t}{\lambda f}+\frac{y_n}{\lambda d}\right)\right]. \tag{11}$$

We see that the integral vanishes unless

$$i = k \quad \text{and} \quad n = m$$

or

$$n = k \quad \text{and} \quad i = m,$$

which gives

$$R(\tau, t) = \int_{-\infty}^{+\infty} W(K)\, dK$$
$$\cdot\sum_{k=1}^{K}\sum_{m=1}^{K}\int_{-\infty}^{+\infty}\frac{dx_1}{2X}\int_{-\infty}^{+\infty}\frac{dy_1}{2Y}\cdots\int_{-\infty}^{+\infty}\frac{dx_K}{2X}\int_{-\infty}^{+\infty}\frac{dy_K}{2Y}$$
$$\cdot\left|h\left(\frac{\nu}{\lambda f}+\frac{x_k}{\lambda d},\frac{\omega}{\lambda f}+\frac{y_k}{\lambda d}\right)\right|^2\left|h\left(\frac{u+\tau}{\lambda f}+\frac{x_m}{\lambda d},\frac{\omega+t}{\lambda f}+\frac{y_m}{\lambda d}\right)\right|^2$$
$$+h\left(\frac{\nu}{\lambda f}+\frac{x_k}{\lambda d},\frac{\omega}{\lambda f}+\frac{y_k}{\lambda d}\right)\cdot h^*\left(\frac{\nu+\tau}{\lambda f}+\frac{x_k}{\lambda d},\frac{\omega+t}{\lambda f}+\frac{y_k}{\lambda d}\right)$$
$$\cdot h^*\left(\frac{\nu}{\lambda f}+\frac{x_m}{\lambda d},\frac{\omega}{\lambda f}+\frac{y_m}{\lambda d}\right)h\left(\frac{\nu+\tau}{\lambda f}+\frac{x_m}{\lambda d},\frac{\omega+t}{\lambda f}+\frac{y_m}{\lambda d}\right). \tag{12}$$

Now, we have two subcases here. There are $K(K-1)$ terms for which $k \neq m$, and there are K terms for which $k = m$.

$$
\begin{aligned}
R(\tau, t) = & \int_{-\infty}^{+\infty} K(K-1)W(K)\, dK \\
& \cdot\left\{\left[\frac{1}{4XY}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left| h\left(\frac{\nu}{\lambda f}+\frac{x}{\lambda d}, \frac{\omega}{\lambda f}+\frac{y}{\lambda d}\right)\right|^2 dx\, dy\right]\right. \\
& \quad + \left|\frac{1}{4XY}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h^*\left(\frac{\nu}{\lambda f}+\frac{x}{\lambda d}, \frac{\omega}{\lambda f}+\frac{y}{\lambda d}\right)\right. \\
& \qquad \left.\left. \cdot h\left(\frac{\nu+\tau}{\lambda f}+\frac{x}{\lambda d}, \frac{\omega+t}{\lambda f}+\frac{y}{\lambda d}\right) dy\, dx\right|^2\right\} \\
& + 2\int_{-\infty}^{+\infty} KW(K)\, dK \\
& \cdot\left[\frac{1}{4XY}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left| h\left(\frac{\nu}{\lambda f}+\frac{x}{\lambda d}, \frac{\omega}{\lambda f}+\frac{y}{\lambda d}\right)\right|^2\right. \\
& \qquad \left.\cdot\left| h\left(\frac{\nu+\tau}{\lambda f}+\frac{x}{\lambda d}, \frac{\omega+t}{\lambda f}+\frac{y}{\lambda d}\right)\right|^2 dy\, dx\right]. \qquad (13)
\end{aligned}
$$

Assuming that the distribution of scatterers $W(K)$ is Poisson and using the definition of h ( ) given in (3b), straightforward evaluation of the integrals in (13) yields

$$
R(\tau, t) = \bar{I}^2\left[1 + \frac{|\rho_1(\tau/f\lambda, t/f\lambda)|^2}{\rho_1(0,0)^2}\right] + 2\frac{\bar{I}}{\rho_1(0,0)}\rho_2(\tau/f\lambda, t/f\lambda), \quad (14)
$$

where $\rho_1(u, v)$ is defined in (8) and

$$
\begin{aligned}
\rho_2(u, v) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |h(\tau, t)|^2 \, |h(\tau+u, t+v)|^2 \, d\tau\, dt \\
&= \text{autocorrelation function of the magnitude squared of the aperture impulse function.}
\end{aligned}
$$

The spatial power spectral density is found by taking the Fourier transform of (14). After simplification we obtain

$$
\begin{aligned}
S(q, p) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} R(\tau, t)\epsilon^{-j2\pi(\tau q + tp)}\, d\tau\, dt \\
&= \bar{I}^2\left\{\delta(q, p) + \frac{(f\lambda)^2}{\rho_1(0,0)^2}|H(\lambda f q, \lambda f p)|^2 \circledast |H(\lambda f q, \lambda f p)|^2\right. \\
&\qquad \left. + \frac{2(f\lambda)^2}{\rho_1(0,0)\bar{I}}|H(\lambda f q, \lambda f p) \circledast H(\lambda f q, \lambda f p)|^2\right\}, \qquad (15)
\end{aligned}
$$

where $\circledast$ stands for convolution. In particular we define

$$F(\lambda fq, \lambda fp) \circledast G(\lambda fq, \lambda fp) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F^*(x, y) G(\lambda fq + x, \lambda fp + y)\, dy\, dx. \tag{16}$$

Equation (9), which gives the mean intensity of the image, and (15), which gives the power spectral density of the intensity fluctuations, are the major results of this section.

III. CIRCULAR APERTURE

Now consider the special case of a circular aperture of radius r_c, and let it be located on axis so that

$$H(\xi, \eta) = \begin{cases} 1, & r \leqq r_c, \\ 0 & r > r_c, \end{cases} \tag{17}$$

where

$$r = +\sqrt{\xi^2 + \eta^2}.$$

The average intensity in the image plane is given by (9) and is

$$\bar{I} = d^2\lambda^2 \bar{N} \rho_1(0, 0) = \pi \bar{N} (\lambda d r_c)^2, \tag{18}$$

where $\rho_1(0,0)$ was evaluated from the integral

$$\rho_1(0, 0) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |\, H(\xi, \eta)\, |^2\, d\xi\, d\eta = \pi r_c^2\,. \tag{19}$$

Evaluation of the integrals in (15) gives the power spectral density

$$S(q, p) = \bar{I}^2 \Bigg[\delta(q, p) + \frac{1}{\pi s_c^2} \Bigg\{ 1 - \frac{2}{\pi} \sin^{-1}\left(\frac{s}{2s_c}\right) - \frac{2}{\pi}\left(\frac{s}{2s_c}\right)\sqrt{1 - \left(\frac{s}{2s_c}\right)^2} + \frac{1}{F}\left(1 - \frac{2}{\pi}\sin^{-1}\left(\frac{s}{2s_c}\right) - \frac{2}{\pi}\left(\frac{s}{2s_c}\right)\sqrt{1 - \left(\frac{s}{2s_c}\right)^2}\right)\Bigg\}\Bigg], \tag{20}$$

where

q, p = image plane spatial frequencies in rectangular coordinates,

$s = +\sqrt{q^2 + p^2}$

$s_c = r_c/f\lambda$ = cutoff frequency produced by diffraction at the circular aperture.

$F = \dfrac{d^2\lambda^2\bar{N}}{2\pi r_c^2}$

= overlap factor.

The overlap factor F warrants some discussion. Basically it is equal to the average number of point scatterer image centers contained within an area equal to that occupied by the image of a single scatterer. That is, a single point scatterer located at (0,0) in the object plane would produce a point in the image plane at (0,0) having intensity

$$I = \left| h\left(\frac{\nu}{\lambda f}, \frac{\omega}{\lambda f}\right) \right|^2 = I_0 \left[\frac{2J_1\left(\frac{2\pi r_c}{f\lambda}\sqrt{\nu^2+\omega^2}\right)}{\frac{2\pi r_c}{f\lambda}\sqrt{\nu^2+\omega^2}} \right]^2.$$

The intensity is down[11] approximately 50 percent at $(2\pi r_c/f\lambda)\sqrt{\nu_1^2+\omega_1^2} = \sqrt{2}$, and the area covered by the image of the point scatterer at this 50 percent value is $A = \pi(\nu_1^2+\omega_1^2) = f^2\lambda^2/2\pi r_c^2$. For a diffuse object, the average number of imaged scattering centers per unit area in the image plane is $\bar{n} = (d/f)^2\bar{N}$. If we define the overlap factor F as the average number of scatterer image centers falling in the area of one of these images we have

$$F = \bar{n}A = \frac{d^2\lambda^2\bar{N}}{2\pi r_c^2}.$$

For a truly diffuse surface, the overlap factor $F \gg 1$ so that (20) reduces to

$$S(q,p) = \bar{I}^2\left[\delta(q,p) + \frac{1}{\pi s_c^2}\left\{1 - \frac{2}{\pi}\sin^{-1}\left(\frac{s}{2s_c}\right) - \frac{2}{\pi}\left(\frac{s}{2s_c}\right)\sqrt{1-\left(\frac{s}{2s_c}\right)^2}\right\}\right] \quad (21)$$

which is plotted in Fig. 2. Note that it is symmetrical about the vertical axis. For very small spatial frequencies, (21) can be approximated by

$$S(q,p) = \bar{I}^2\left[\delta(q,p) + \frac{1}{\pi s_c^2}\right]. \quad (22)$$

The total fluctuation or noise power occurring in spatial frequencies less than some frequency s_1 is

$$P = \bar{I}^2\left(\frac{1}{\pi s_c^2}\right)(\pi s_1^2) = \bar{I}^2\left(\frac{s_1}{s_c}\right)^2. \quad (23)$$

IV. CONCLUSIONS

We have found that the image of a uniform diffuse object illuminated with monochromatic coherent light consists of two parts. The first is the mean or ensemble average given by (18), for a circular aperture, and is proportional to the intensity which would be obtained if noncoherent light were used for illumination. This is the desired component and might be likened to the signal component of the image. The square of this term appears as the first term in (20), (21), and (22), and as the delta function in Fig. 2. The second part of the image is a grainy or noise-like component which tends to obscure the mean intensity or signal. This noise-like component occurs because of the random phase angles associated with the point scatterers comprising the microstructure of the diffuse object. This component is shown as the second term in (20), (21), and (22), and as the continuous part of the power spectrum in Fig. 2. Integration of (21) shows that the variance of the noise-like fluctuations in the intensity is equal to the square of the mean intensity (or to the signal power). This is fortunate to the extent that when the signal is small, the noise is likewise small. However, while our analysis was for the particular case of a uniform diffuse surface, we can safely predict that for nonuniform diffuse objects fine detail in the image will be largely obscured by the noise-like fluctuations if resolution is limited by diffraction.

The noise-like fluctuations in the image can be reduced if one *records* the image on film whose modulation transfer function has a bandwidth which is much smaller than the diffraction limit of the optical system. The high-frequency noise in Fig. 2 will not be resolved in this case. For instance, if one requires the "signal-to-noise" ratio to be increased from unity to 10^3 (30 dB), then from (23) we see that

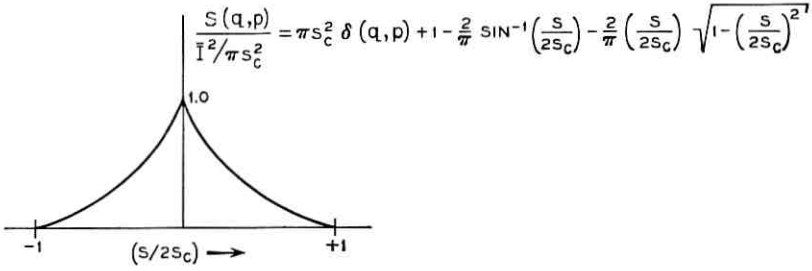


Fig. 2 — Section of the spatial power spectral density for a uniform diffuse surface imaged through a circular aperture. The complete two dimensional spectrum is obtained by rotating the above curve about the vertical axis.

the diffraction bandwidth s_e of the improved optical system must be $10^{3/2} = 31.6$ times the bandwidth s_1 resolvable by the film, and therefore by the whole system. (Since most transducers produce a signal which is proportional to the intensity of the incident light, it seems appropriate to consider the square of the mean intensity as signal power and the variance of the intensity fluctuations as noise power.)

Although we have analyzed the very special optical system shown in Fig. 1, our results are not critically dependent upon the placement of the aperture. The aperture could be the lens aperture rather than an independent physical device, or it could be the aperture defined by the finite size of a hologram, for instance. Our results should also hold approximately for the human eye, since the resolution of a healthy eye is known to be determined by the diffraction limit of the iris.[9] The predicted value of unity for the signal-to-noise ratio is the right order of magnitude for what one observes when laser light is used for illumination if one is careful to hold the eye stationary and hence not average the noise out as a function of time. Although moving the eye tends to average out the noise, the residual noisiness remains objectionable. This places definite limitations upon the use of coherent light in visual systems.

The author wishes to thank Messrs. C. B. Rubinstein and A. B. Larsen for helpful discussions.

REFERENCES

1. Cutrona, L. J., Leith, E. N., Palermo, C. J., and Porcello, L. J., Optical Data Processing and Fittering Systems, IRE Trans. Inform. Theory, June, 1960, pp. 386–400.
2. Leith, E. N. and Uptatieks, J., Reconstructed Wavefronts and Communication Theory, J. Opt. Soc. Amer. *52*, No. 10, October, 1962, pp. 1123–1130.
3. Stroke, G. W. and Falconer, D. G., Lensless Fourier-Transform Method for Optical Holography, Appl. Phys. Lett, *6*, No. 10, May 15, 1965, pp. 201–203.
4. Gabor, D., Microscopy by Reconstructed Wavefronts, Proc. Roy. Soc., *A197*, 1949, pp. 454–487.
5. Considine, P. S., Effects of Coherence on Imaging Systems, J. Opt. Soc. Amer. *56*, No. 8, August, 1966, pp. 1001–1008.
6. Rigden, J. D. and Gordon, E. I., The Granularity of Scattered Optical Maser Light, Proc. IRE, *50*, No. 11, November, 1962, pp. 2367–2368.
7. Cutler, C. C., Coherent Light, Int. Sci. Tech., September, 1963, pp. 54–63.
8. Goldfisher, L. I., Autocorrelation Function and Power Spectral Density of Laser-Produced Speckle Patterns, J. Opt. Soc. Amer. *55*, No. 3, March, 1965, pp. 247–253.
9. Graham, C. H. (Editor), *Vision and Visual Perception*, John Wiley and Sons, Inc., New York, 1965, Chapter 11, pp. 321–345.
10. Leith, E. N. and Palermo, C. J., *Introduction to Optical Data Processing*, University of Michigan Engineering Summer Conferences, May 24-June 4, 1965, pp. 2–12 to 2–20.
11. Born, M. and Wolf, E., *Principles of Optics*, Pergamon Press, New York, 1959, p. 396.

The Excitation of Planar Dielectric Waveguides at p-n Junctions, I

By J. McKENNA

(Manuscript received April 26, 1967)

The fields excited within a planar dielectric waveguide by an externally incident electromagnetic field are studied in this paper. The dielectric waveguide fills the half space $z > 0$, while the half space $z < 0$ is air. The waveguide is formed by a nonuniform, anisotropic, nonabsorbing, dielectric medium. Different choices of the dielectric tensor for this medium yield different waveguides. Certain models which are particularly relevant to electro-optic diode waveguides and laser diode amplifiers are studied in some detail. An arbitary incident field will, in general, excite not only a finite number of propagating modes, but also a background of continuum modes. Integral representations of the total transmitted field within the waveguide as well as of the reflected field are obtained. The representation of the total transmitted field can be decomposed into a finite sum of discrete propagating modes, a continuum propagating field, and an evanescent field. Explicit evaluation of the fields depends on the solution of a pair of integral equations. In practice, the dielectric tensor of the waveguide differs but little from the dielectric constant of the surrounding material. An approximate solution is found for this case, and numerical results will appear in a following paper.

1. INTRODUCTION

Recently there has been great interest in the guiding of light by the p-n junction region in certain piezoelectric semiconductors, for it has been noted that the Pockels effect due to the electric field within the p-n junction can be used to modulate light which propagates parallel to the junction plane.[1-4] This effect was first observed, and has been most intensively studied, with visible light in GaP junctions,[1] but it has also been observed with infrared light in GaAs junctions.[1, 4]

All treatments of the effect so far have assumed that the p-n junction region, which has a higher dielectric constant than the surround-

ing, normal GaP, behaves like a dielectric waveguide.[1-5] A detailed analysis of this waveguide would require a knowledge of the optical properties in the neighborhood of the junction. However, since these properties change significantly in a fraction of a wavelength, it is extremely difficult to investigate them individually by experimental means. In order to get around this difficulty it has been necessary to adopt an indirect approach based on analyzing a number of different mathematical models and comparing their predictions with experiment.

As part of this program Nelson and McKenna[6] have investigated the possible discrete modes which can propagate in a number of different models and have studied in considerable detail the properties of the lowest-order mode of each polarization. Recent experimental work has made it increasingly clear, however, that a knowledge of the discrete modes alone is not enough to provide an understanding of these p-n junction dielectric waveguides. This is because a beam of light, when focused on the face of a junction waveguide, excites within the waveguide not only a finite number of discrete modes, but also a background of continuum modes. In many cases this background light is intense enough to mask important features of the discrete propagating modes. Thus, unless an understanding of this background light is available, the task of comparing the predictions of different mathematical models with experiment is almost impossible. An understanding of the electromagnetic boundary value problem involved also has great relevance to understanding what happens when light is introduced into a laser diode amplifier.

The purpose of this paper is to study in some detail a class of mathematical models of the excitation of dielectric waveguides. These models are simple enough so that the mathematical analysis can be performed and the background light can be investigated carefully. At the same time, it is felt that the models are realistic enough so that their predictions can be compared with experiment.

The models can be described as follows. The waveguide consists of the half space $z > 0$, as shown in Fig. 1, while the region $z < 0$ is air. The waveguide itself is assumed to be formed by a nonuniform, anisotropic, nonabsorbing dielectric. The components of the dielectric tensor are functions of the coordinate x only, and for each value of x the dielectric tensor is diagonal in the fixed coordinate system shown in Fig. 1. As an example, for the GaP electro-optic diode modulator studied in *NM* this corresponds to the cases where the junction field is in the [111] or [100] directions. Each such model is determined by its

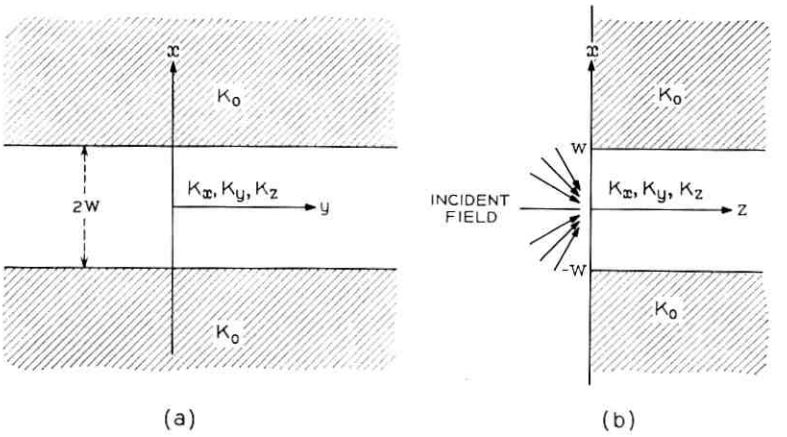


Fig. 1 — Symmetric step model illustrating the coordinate system used in all the models. The dielectric tensor is always diagonal in this fixed coordinate system.

dielectric tensor whose diagonal elements we will denote by $K_n(x)$ $(n = x, y, z)$.

It was shown in *NM* that the amount of absorption encountered in GaP electro-optic diode modulators was too small to affect significantly the shape of the modes. It is, therefore, felt that the study of absorptionless models here is well justified. It was also shown in *NM* that the detailed analytical form of the functions $K_n(x)$ is not important when only the lowest-order discrete mode of each polarization can propagate. The most important features of the discrete modes can be determined by studying models for which the functions $K_n(x)$ are step functions (piece-wise constant). Although it is possible to carry out a good deal of the analysis without specifying the functions $K_n(x)$, the final detailed results naturally depend on the choice of $K_n(x)$. We shall concentrate here on two models, the symmetric step model and the asymmetric step model. The symmetric step model is defined by the equations[6]

$$K_m(x) = K_m , \qquad |x| < w \tag{1}$$

$$= K_0 , \qquad |x| > w \tag{2}$$

and the asymmetric step model is defined by the equations[6]

$$K_m(x) = K_m , \qquad |x| < w \tag{3}$$

$$= K_1 , \qquad x < -w \tag{4}$$

$$= K_2 , \qquad x > w, \tag{5}$$

where $K_2 < K_1$, and $K_m > K_j \geqq 1$, $m = x, y, z$, $j = 0, 1, 2$ (see Fig. 2). In the case of the GaP electro-optic diode modulators there are relations of the form[6]

$$K_m = n^2(1 + \delta_m), \qquad (m = x, y, z) \tag{6}$$

$$K_0 = n^2(1 - \Delta), \qquad K_j = n^2(1 - \Delta_j), \qquad (j = 1, 2). \tag{7}$$

In (6) and (7) n is the index of refraction of normal GaP, the quantities δ_m are linear in the junction field (the linear electro-optic effect), and $0 \leqq |\delta_m| < \Delta \ll 1$.

In Section II we will write down general integral representations for incident waves in the region $z < 0$, as well as integral representations for the resulting reflected and transmitted fields. These integral representations will involve a number of unknown functions. Some of these functions are determined directly from the structure of the waveguide and are independent of the incident field and the boundary condition at $z = 0$. The remaining unknown functions are determined by the incident field and the boundary conditions at $z = 0$. We show that these functions satisfy a set of linear integral equations. The results of Section II are independent of the specific form of $K_m(x)$ and the incident field. In Section III we explicitly calculate the unknown functions which depend only on the structure of the waveguide for the symmetric and asymmetric step models. In Section IV we obtain approximate solutions of the integral equations for a special class of

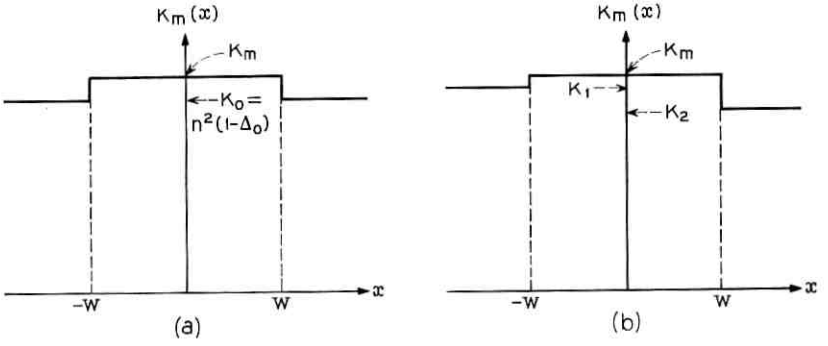


Fig. 2 — (a) The function $K_m(x)$ for the symmetric step model. (b) The function $K_m(x)$ for the asymmetric step model.

waveguide models. The remaining unknown functions are determined for these models in terms of the incident field. In a second paper on this subject we will give asymptotic expansions and numerical results for the fields within the waveguide for the symmetric and asymmetric models when they are excited by a Gaussian incident wave.

II. A GENERAL DESCRIPTION OF THE FIELDS

In this section we study formal solutions of Maxwell's equations which describe an incident wave in the region $z < 0$ moving to the right and striking the waveguide from the left, a reflected wave in the region $z < 0$, and a transmitted wave in the region $z > 0$. The fields are assumed to be monochromatic and independent of the coordinate y. We write for the total electric and magnetic field vectors

$$\mathbf{E}(x, z, t) = \mathrm{Re}\,(\mathbf{e}(x, z)e^{i\omega t}), \qquad \mathbf{H}(x, z, t) = \mathrm{Re}\,(\mathbf{h}(x, z)e^{i\omega t}), \tag{8}$$

and for the total electric displacement and magnetic induction vectors

$$\mathbf{D}(x, z, t) = \mathrm{Re}\,(\mathbf{d}(x, z)e^{i\omega t}), \qquad \mathbf{B}(x, z, t) = \mathrm{Re}\,(\mathbf{b}(x, z)e^{i\omega t}), \tag{9}$$

where Re denotes the real part and $\omega = 2\pi f$ is the angular frequency of the radiation. Then Maxwell's equations are

$$\begin{aligned} \nabla \times \mathbf{e} &= -i\omega\mathbf{b}, & \nabla\cdot\mathbf{d} &= 0, \\ \nabla \times \mathbf{h} &= i\omega\mathbf{d}, & \nabla\cdot\mathbf{b} &= 0. \end{aligned} \tag{10}$$

From our assumptions about the model, the constitutive equations can be written as

$$\mathbf{b} = \mu_0\mathbf{h}, \qquad \mathbf{d} = \epsilon_0\mathbf{K}\cdot\mathbf{e}, \tag{11}$$

where ϵ_0 and μ_0 are, respectively, the permittivity and permeability of free space. The dielectric matrix $\mathsf{K} = \mathsf{K}(x, z)$ is the unit matrix for $z < 0$, and for $z > 0$ it is a diagonal matrix whose diagonal elements, $K_n(x)$ $(n = x, y, z)$, are functions of x only. It is a straightforward matter to show that any solution of Maxwell's equations satisfying the above assumptions can be written as the linear combination of a *TE* solution and a *TM* solution. We consider these solutions separately.

2.1 *TE Fields*

We first look for *TE* solutions having the form

$$\mathbf{e}(x, z) = [0, e_y(x, z), 0], \qquad \mathbf{h}(x, z) = [h_x(x, z), 0, h_z(x, z)]. \tag{12}$$

In the region $z < 0$, e_y must satisfy the Helmholtz equation

$$\frac{\partial^2 e_y}{\partial x^2} + \frac{\partial^2 e_y}{\partial z^2} + k^2 e_y = 0, \tag{13}$$

where the free-space wavenumber k is defined by

$$k = \omega(\epsilon_0 \mu_0)^{\frac{1}{2}} = 2\pi/\lambda$$

and λ is the free-space wavelength. The total field in $z < 0$ is the sum of the incident field $e_y^{(i)}$ and the reflected field $e_y^{(r)}$ and both $e_y^{(i)}$ and $e_y^{(r)}$ are solutions of (13). In the region $z > 0$ there is only the transmitted field which satisfies the equation

$$\frac{\partial^2 e_y}{\partial x^2} + \frac{\partial^2 e_y}{\partial z^2} + k^2 K_y(x) e_y = 0. \tag{14}$$

A solution of (13), which can be found by separation of variables, and which describes a general incident field due to sources in $z < 0$ at a finite distance from the plane $z = 0$, is

$$e_y^{(i)}(x, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{E}_y^{(i)}(l) \exp \{-i\Omega(l)z - ilx\}\, dl, \tag{15}$$

where

$$\begin{aligned} \Omega(l) &= +\sqrt{k^2 - l^2}, \qquad |\, l \,| \leqq k \\ &= -i\sqrt{l^2 - k^2}, \qquad |\, l \,| \geqq k. \end{aligned} \tag{16}$$

The components of the magnetic field vector can be obtained with the aid of Maxwell's equations by differentiating (15). Let $\Sigma(z_0)$ denote the strip $-\infty < x < \infty$, $0 \leqq y \leqq 1$, lying in the plane $z = z_0$. Then the time averaged power incident on $\Sigma(z)$, $z \leqq 0$, is independent of z and is

$$\begin{aligned} P_i &= -\tfrac{1}{2} \operatorname{Re} \int_{-\infty}^{\infty} e_y^{(i)}(x, z) h_x^{(i)}(x, z)^*\, dx \\ &= (4\pi\omega\mu_0)^{-1} \int_{-k}^{k} \sqrt{k^2 - l^2}\, |\, \mathcal{E}_y^{(i)}(l) \,|^2\, dl, \end{aligned} \tag{17}$$

where * denotes complex conjugation. We will assume that

$$\int_{-\infty}^{\infty} |\, \mathcal{E}_y^{(i)}(l) \,|^2\, dl < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} |\, \Omega(l) \,|\, |\, \mathcal{E}_y^{(i)}(l) \,|^2\, dl < \infty.$$

(15) is to describe an incident field due to sources at $z = -\infty$, then it is easy to see that we must have $\mathcal{E}_y^{(i)}(l) = 0$, $|\, l \,| > k$. Since the incident field must be specified, it will always be assumed that $\mathcal{E}_y^{(i)}(l)$ is known.

A solution of (13) describing a general wave reflected from the waveguide surface $z = 0$ is

$$e_y^{(r)}(x, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{E}_y^{(r)}(l) \exp \{i\Omega(l)z - ilx\}\, dl. \tag{18}$$

We will always assume that the source of the incident radiation is perfectly absorbing so that $e_y^{(i)}(x, z) + e_y^{(r)}(x, z)$ is the total field in the region between the source and the surface of the waveguide at $z = 0$. It will be seen that because of the boundary conditions at $z = 0$, $\mathcal{E}_y^{(i)}(l)$ generally does not vanish outside some finite l interval. Because of the factor $\exp \{i\Omega(l)z\}$, that part of the integral in (18) between the limits $-k$ and k, $\int_{-k}^{k} \{\ \}\, dl$, represents a traveling field, while the remainder of the integral represents an evanescent field which damps out very rapidly in the negative z direction. The time averaged power reflected back through the strip $\Sigma(z)$, $z \leqq 0$, is

$$P_r = (4\pi\omega\mu_0)^{-1} \int_{-k}^{k} \sqrt{k^2 - l^2}\, |\, \mathcal{E}_y^{(r)}(l)\, |^2\, dl. \tag{19}$$

We now turn to the transmitted field. We use the method of separation of variables, and we seek transmitted waves which are linear superpositions of solutions of (14) of the form

$$e_y^{(t)}(x, z) \approx e_\nu(x) \exp \{-i\sqrt{-\nu}z\}. \tag{20}$$

In (20) ν is a real separation parameter, and if $\nu > 0$, $\sqrt{-\nu} = -i\sqrt{\nu}$. If (20) is substituted into (14) we get the eigenvalue equation

$$\frac{d^2 e_\nu}{dx^2} + (k^2 K_y(x) + \nu) e_\nu = 0. \tag{21}$$

Equation (21) defines a singular, self-adjoint, second-order boundary value problem on the interval $-\infty < x < \infty$. The theory of this equation is well known, and we refer the reader to Coddington and Levinson[7] for a detailed treatment. We give a summary here of those properties of such equations which we will need.

For all the models under consideration, the functions $K_m(x)$ are positive, bounded functions, which are bounded away from zero, and which are differentiable except for at most a finite number of step discontinuities. Equation (21), therefore, defines a problem which is called limit-point type at both plus and minus infinity. This means that for arbitrary, complex ν, (21) possesses exactly one solution (up to a constant factor) which is square integrable over $0 < x < \infty$, and exactly one solution which is square integrable over $-\infty < x < 0$.

For a given real number ν, let $\varphi_1(x, \nu)$ and $\varphi_2(x, \nu)$ be the two solutions of (21) which satisfy the conditions that $\varphi_j(x, \nu)$ and $\varphi_j'(x, \nu)$ be continuous and which satisfy the initial conditions

$$\varphi_1(0, \nu) = 1, \qquad \varphi_1'(0, \nu) = 0, \tag{22}$$

$$\varphi_2(0, \nu) = 0, \qquad \varphi_2'(0, \nu) = 1, \tag{23}$$

where $' = d/dx$. Equation (21) also determines a 2×2 matrix-valued function $\rho(\nu)$, $-\infty < \nu < \infty$, having the following properties: (*i*) $\rho(\nu)$ is Hermitian ($\rho_{jk}(\nu) = \rho_{kj}^*(\nu)$). (*ii*) $\rho(\nu) - \rho(u)$ is positive semidefinite if $\nu > \mu$. (*iii*) $\rho_{jk}(\nu)$ is of bounded variation on every finite interval. The matrix $\rho(\nu)$ is called the spectral density matrix and its construction is outlined in Section III. Then if $f(x)$ is any square integrable function ($\int_{-\infty}^{\infty} | f(x) |^2 dx < \infty$), we can define two transforms of $f(x)$, $g_j(\nu)$ ($j = 1, 2$), such that

$$\lim_{L\to\infty} \int_{-\infty}^{\infty} \sum_{j,k=1}^{2} \left\{ g_j(\nu) - \int_{-L}^{L} f(x)\varphi_j(x, \nu)\, dx \right\} \cdot \left\{ g_k(\nu) - \int_{-L}^{L} f(x)\varphi_k(x, \nu)\, dx \right\}^* d\rho_{jk}(\nu) = 0. \tag{24a}$$

This is referred to as convergence in the mean with respect to the measure $\rho(\nu)$, and in the manner of Fourier transforms of $\mathfrak{L}^2$ functions, we write

$$g_j(\nu) = \int_{-\infty}^{\infty} f(x)\varphi_j(x, \nu)\, dx \qquad (j = 1, 2). \tag{24b}$$

In terms of these transforms, the Parseval equality

$$\int_{-\infty}^{\infty} | f(x) |^2 dx = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} g_j(\nu)^* g_k(\nu)\, d\rho_{jk}(\nu), \tag{25}$$

and the expansion

$$f(x) = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \varphi_j(x, \nu) g_k(\nu)\, d\rho_{jk}(\nu) \tag{26}$$

are valid. Equation (26) is defined in terms of convergence in the mean. The set of real points ν at which the functions $\rho_{jk}(\nu)$ are nonconstant is the spectrum of (21). The set of points where any $\rho_{jk}(\nu)$ is discontinuous is the point spectrum and for each such value of ν, (21) has exactly one square integrable solution. The continuous spectrum is the set of points of continuity of $\rho(\nu)$ which are in the spectrum. In Section III we will exhibit the spectral density matrices for two important models.

We can now write down a formal expression for the transmitted field:

$$e_y^{(t)}(x, z) = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \exp\{-i\sqrt{-\nu}z\}\varphi_j(x, \nu)g_k(\nu)\, d\rho_{jk}(\nu). \tag{27}$$

The two initial value solutions $\varphi_j(x, \nu)$ $(j = 1, 2)$, as well as the functions $\rho_{jk}(\nu)$ $(j, k = 1, 2)$ are determined, independently of the boundary conditions at $z = 0$, by (21) and we can assume that they are known. The two unknown functions $g_j(\nu)$ $(j = 1, 2)$ in (27) are determined by the field at $z = 0$, since with the aid of (24) we can write

$$g_j(\nu) = \int_{-\infty}^{\infty} e_y^{(t)}(x, 0)\varphi_j(x, \nu)\, dx. \tag{28}$$

It is clear that because of the factor $\exp\{-i\sqrt{-\nu}z\}$, the parts of the integrals $\int_{-\infty}^{0}$ in (27) represent the propagating portion of the transmitted field, while the parts $\int_{0}^{\infty}$ reprent the evanescent portion of the transmitted field. With the aid of the Parseval relation, (25), we can write down an expression for the time averaged power transmitted across any $\Sigma(z)$, $z \geqq 0$,

$$P_t = (2\omega\mu_0)^{-1} \sum_{j,k=1}^{2} \int_{-\infty}^{0} \sqrt{-\nu}g_j(\nu)^* g_k(\nu)\, d\rho_{jk}(\nu). \tag{29}$$

We now make use of the conditions that $e_y(x, z)$ and $h_x(x, z)$ must be continuous at $z = 0$ in order to write down a set of integral equations which determines $\mathcal{E}_y^{(r)}(l)$, $g_1(\nu)$, and $g_2(\nu)$.

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} [\mathcal{E}_y^{(i)}(l) + \mathcal{E}_y^{(r)}(l)]e^{-ilx}\, dl = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \varphi_j(x, \nu)g_k(\nu)\, d\rho_{jk}(\nu), \tag{30}$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \Omega(l)[\mathcal{E}_y^{(i)}(l) - \mathcal{E}_y^{(r)}(l)]e^{-ilx}\, dl$$

$$= \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \sqrt{-\nu}\varphi_j(x, \nu)g_k(\nu)\, d\rho_{jk}(\nu). \tag{31}$$

Although there appear to be only two equations in three unknown functions, because of (24) and (26), (30) and (31) are sufficient to determine the unknown functions. We indicate formally why this is true, although it will be clear from the results of Section IV that this scheme must be modified in specific cases. We do not go into these modifications, because in Section IV we use a different scheme to get approximate solutions. With the aid of (24b), solve (30) and (31) for

$g_j(\nu)$, giving the four equations

$$g_i(\nu) = \int_{-\infty}^{\infty} \varphi_i(x, \nu)\, dx \frac{1}{2\pi} \int_{-\infty}^{\infty} [\mathcal{E}_\nu^{(i)}(l) + \mathcal{E}_\nu^{(r)}(l)]e^{-ilx}\, dl, \tag{32}$$

$$\sqrt{-\nu} g_i(\nu) = \int_{-\infty}^{\infty} \varphi_i(x, \nu)\, dx \frac{1}{2\pi} \int_{-\infty}^{\infty} \Omega(l)[\mathcal{E}_\nu^{(i)}(l) - \mathcal{E}_\nu^{(r)}(l)]e^{-ilx}\, dl, \tag{33}$$

for $j = 1, 2$. On eliminating $g_i(\nu)$ between these equations we get the two equations in the unknown $\mathcal{E}_\nu^{(r)}(l)$.

$$\begin{aligned} &\int_{-\infty}^{\infty} \varphi_i(x, \nu)\, dx \frac{1}{2\pi} \int_{-\infty}^{\infty} (\sqrt{-\nu} + \Omega(l))\mathcal{E}_\nu^{(r)}(l)e^{-ilx}\, dl, \\ &= \int_{-\infty}^{\infty} \varphi_i(x, \nu)\, dx \frac{1}{2\pi} \int_{-\infty}^{\infty} (-\sqrt{-\nu} + \Omega(l))\mathcal{E}_\nu^{(i)}(l)e^{-ilx}\, dl, \end{aligned} \tag{34}$$

for $j = 1, 2$. Now from (26) we can write $f(x) = f_1(x) + f_2(x)$ where

$$f_k(x) = \sum_{j=1}^{2} \int_{-\infty}^{\infty} \varphi_j(x, \nu) g_k(\nu)\, d\rho_{jk}(\nu) \qquad (k = 1, 2), \tag{35}$$

$$\int_{-\infty}^{\infty} f_k(x)\varphi_j(x, \nu)\, dx = \delta_{jk} g_j(\nu) \qquad (j, k = 1, 2), \tag{36}$$

and δ_{jk} is the Kronecker delta function. It is this decomposition of an arbitrary $f(x)$ into components lying in the two subspaces spanned by $\varphi_1(x, \nu)$ and $\varphi_2(x, \nu)$ which is reflected in the two integral equations (34). The solution of (34) with given j yields the component of the reflected field lying in the subspace spanned by the corresponding $\varphi_j(x, \nu)$. Let $\mathcal{E}_{\nu j}^{(r)}(l)$, $j = 1, 2$, denote the two solutions. Then $\mathcal{E}_\nu^{(r)}(l) = \mathcal{E}_{\nu 1}^{(r)}(l) + \mathcal{E}_{\nu 2}^{(r)}(l)$ describes the total reflected field. With this result $g_j(\nu)$ $(j = 1, 2)$ can be obtained from either (32) or (33). We have been unable to obtain exact solutions for the integral equations (30)–(31) for any of the models considered here. However, in Section IV approximate solutions are obtained for certain situations of interest.

2.2 *TM Fields*

We next seek TM solutions of Maxwell's equations of the form

$$\mathbf{e}(x, z) = (e_x(x, z), 0, e_z(x, z)), \qquad \mathbf{h}(x, z) = (0, h_y(x, z), 0). \tag{37}$$

In the region $z < 0$, h_y must satisfy (13). In the region $z > 0$, h_y must satisfy the equation

$$\frac{\partial}{\partial x}\left\{(1/K_z(x))\frac{\partial h_y}{\partial x}\right\} + \frac{\partial}{\partial z}\left\{(1/K_z(x))\frac{\partial h_y}{\partial z}\right\} + k^2 h_y = 0. \tag{38}$$

Just as for the TE fields, a general incident field due to sources in $z < 0$ at a finite distance from the plane $z = 0$ is

$$h_y^{(i)}(x, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathfrak{C}_y^{(i)}(l) \exp \{-i\Omega(l)z - ilx\}\, dl. \tag{39}$$

The time averaged power due to this wave which is incident on $\Sigma(z)$, $z \leqq 0$, is

$$P_i = \tfrac{1}{2} \operatorname{Re} \int_{-\infty}^{\infty} e_x^{(i)}(x, z) h_y^{(i)}(x, z)^*\, dx = (4\pi\omega\epsilon_0)^{-1} \int_{-k}^{k} \Omega(l) \mid \mathfrak{C}_y^{(i)}(l) \mid^2 dl. \tag{40}$$

We assume that $\int_{-\infty}^{\infty} \mid \mathfrak{C}_y^{(i)}(l) \mid^2 dl < \infty$ and $\int_{-\infty}^{\infty} \mid \Omega(l) \mid \mid \mathfrak{C}_y^{(i)}(l) \mid^2 dl < \infty$. As for the *TE* field if the sources of the *TM* field are at $z = -\infty$ then $\mathfrak{C}_y^{(i)}(l) = 0$, $\mid l \mid > k$. Furthermore, it will always be assumed that $\mathfrak{C}_y^{(i)}(l)$ is known.

A solution of (13) describing a general reflected wave is

$$h_y^{(r)}(x, z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathfrak{C}_y^{(r)}(l) \exp \{i\Omega(l)z - ilx\}\, dl. \tag{41}$$

Just as in the case of the *TE* field, $h_y^{(r)}(x, z)$ can be split into a propagating field and an evanescent field. The time averaged power reflected back through the strip $\Sigma(z)$, $z \leqq 0$, is

$$P_r = (4\pi\omega\epsilon_0)^{-1} \int_{-k}^{k} \Omega(l) \mid \mathfrak{C}_y^{(r)}(l) \mid^2 dl. \tag{42}$$

The transmitted field is again treated by separation of variables, and we write

$$h_y^{(t)}(x, z) \approx h_y(x) \exp \{-i\sqrt{-\nu}\, z\}.$$

Then $h_y(x)$ satisfies the eigenvalue equation

$$K_x(x) \frac{d}{dx} \left\{ (1/K_x(x)) \frac{dh_y}{dx} \right\} + (k^2 K_x(x) + \nu) h_y = 0. \tag{43}$$

Equation (43) is not in the canonical form of a self-adjoint boundary value problem. However, if we make the change of variables

$$u = \int_0^x \{K_x(t)\}^{-1}\, dt, \tag{44}$$

(43) is transformed to the equation

$$\frac{d}{du} \left[\{K_x(u) K_x(u)\}^{-1} \frac{dh_y}{du} \right] + (k^2 K_x(u) + \nu) h_y = 0. \tag{45}$$

This equation defines a self-adjoint boundary value problem,[8] and even though the function $\{K_x(u)K_z(u)\}^{-1}$ may have step discontinuities, the techniques of Ref. 7 can be shown to be still valid. Equation (45) is limit-point at $u = \pm\infty$, and so on transforming back to the variable x, the following statements can be made.

For a given real number ν, let $\psi_1(x, \nu)$ and $\psi_2(x, \nu)$ be the two solutions of (43) which satisfy the requirements that

$$\psi_j(x, \nu) \quad \text{and} \quad \{K_z(x)\}^{-1}\psi_j'(x, \nu)$$

be continuous for all x, and which satisfy the initial conditions

$$\psi_1(0, \nu) = 1, \qquad (1/K_z(0))\psi_1'(0, \nu) = 0, \tag{46}$$

$$\psi_2(0, \nu) = 1, \qquad (1/K_z(0))\psi_2'(0, \nu) = 1. \tag{47}$$

Equation (43) determines a 2×2 spectral density matrix $\sigma(\nu)$ whose construction is given in Section III. If $f(x)$ is any square integrable function of x, we define two transforms of $f(x)$,

$$h_j(\nu) = \int_{-\infty}^{\infty} f(x)\psi_j(x, \nu)\{K_z(x)\}^{-1}\, dx \qquad (j = 1, 2), \tag{48}$$

where equality in (48) is defined in terms of convergence in the mean with respect to the measure $\sigma(\nu)$. In terms of these transforms, the Parseval equality

$$\int_{-\infty}^{\infty} |f(x)|^2 \{K_z(x)\}^{-1}\, dx = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} h_j(\nu)h_k(\nu)^*\, d\sigma_{jk}(\nu), \tag{49}$$

and the expansion

$$f(x) = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \psi_j(x, \nu)h_k(\nu)\, d\sigma_{jk}(\nu). \tag{50}$$

are valid. The last equality is again defined in the sense of convergence in the mean.

We can write down a formal expression for the transmitted field

$$h_\nu^{(t)}(x, z) = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \exp\{-i\sqrt{-\nu}\, z\}\psi_j(x, \nu)h_k(\nu)\, d\sigma_{jk}(\nu). \tag{51}$$

The two initial value solutions $\psi_j(x, \nu)$ $(j = 1, 2)$, as well as the functions $\sigma_{jk}(\nu)$ $(j, k = 1, 2)$ are determined, independently of the boundary conditions at $z = 0$, by (43) and we can assume that they are known. The two unknown functions $h_j(\nu)$ $(j = 1, 2)$ in (51) are determined by the field at $z = 0$ since with the aid of (48) we can write

$$h_j(\nu) = \int_{-\infty}^{\infty} h_\nu^{(t)}(x, 0)\psi_j(x, \nu)\{K_x(x)\}^{-1}\, dx. \tag{52}$$

With the aid of the Parseval relation, (49), we can write down an expression for the time averaged power transmitted across any $\Sigma(z)$, $z \geqq 0$:

$$P_t = (2\omega\epsilon_0)^{-1} \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \sqrt{-\nu}\, h_j(\nu)^* h_k(\nu)\, d\sigma_{jk}(\nu). \tag{53}$$

We can now make use of the conditions that $e_x(x, z)$ and $h_y(x, z)$ must be continuous at $z = 0$ in order to write down a set of integral equations from which $\mathfrak{IC}_\nu^{(r)}(l)$, $h_1(\nu)$, and $h_2(\nu)$ can be determined.

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} [\mathfrak{IC}_\nu^{(i)}(l) + \mathfrak{IC}_\nu^{(r)}(l)]e^{-ilx}\, dl = \sum_{j,k=1}^{2} \int_{-\infty}^{\infty} \psi_j(x, \nu)h_k(\nu)\, d\sigma_{jk}(\nu), \tag{54}$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \Omega(l)[\mathfrak{IC}_\nu^{(i)}(l) - \mathfrak{IC}_\nu^{(r)}(l)]e^{-ilx}\, dl$$

$$= \sum_{j,k=1}^{2} \{1/K_x(x)\} \int_{-\infty}^{\infty} \sqrt{-\nu}\, \psi_j(x, \nu)h_k(\nu)\, d\sigma_{jk}(\nu). \tag{55}$$

Just as in the case of the TE field, the solution of (54) and (55) reduces to the solution of the two integral equations

$$\int_{-\infty}^{\infty} \psi_j(x, \nu)\, dx \frac{1}{2\pi}\int_{-\infty}^{\infty} \{\sqrt{-\nu}/K_x(x) + \Omega(l)\}\mathfrak{IC}_\nu^{(r)}(l)e^{-ilx}\, dl$$

$$= \int_{-\infty}^{\infty} \psi_j(x, \nu)\, dx \frac{1}{2\pi}\int_{-\infty}^{\infty} \{-\sqrt{-\nu}/K_x(x) + \Omega(l)\}$$

$$\cdot \mathfrak{IC}_\nu^{(i)}(l)e^{-ilx}\, dl, \qquad (j = 1, 2). \tag{56}$$

III. THE SPECTRAL DENSITY MATRIX FOR SEVERAL MODELS

3.1 *General Outline of the Construction*

In Section II it was shown that the determination of the transmitted field for a given model depended on a knowledge of the initial value solutions $\varphi_j(x, \nu)$ and $\psi_j(x, \nu)$ $(j = 1, 2)$ and the spectral density matrices $\rho(\nu)$ and $\sigma(\nu)$. In this section we study these functions in some detail for two simple but important models, the symmetric step model and the asymmetric step model. These calculations illustrate the technique for treating the whole class of piecewise constant models.

We first outline the general construction of the spectral density matrices.[7] The solutions of (21) have the property that the functions

$\varphi_j(x, \nu)$, $\varphi_j'(x, \nu)$ $(j = 1, 2)$ are entire functions of ν for each fixed x, when ν is a complex variable. The first step is to determine the two functions of ν, $m_\infty(\nu)$ and $m_{-\infty}(\nu)$ such that when Im $\nu > 0$, $\varphi_1(x, \nu) + m_\infty(\nu)\varphi_2(x, \nu)$ is a square integrable function of x over $[0, \infty]$ and $\varphi_1(x, \nu) + m_{-\infty}(\nu)\varphi_2(x, \nu)$ is square integrable over $[-\infty, 0]$. The elements of the spectral density matrix are then given by the formula

$$\rho_{jk}(\nu) - \rho_{jk}(\mu) = \lim_{\epsilon \to +0} \frac{1}{\pi} \int_\mu^\nu \text{Im } M_{jk}(\eta + i\epsilon)\, d\eta \tag{57}$$

where μ and ν are real, Im denotes the imaginary part, and for arbitrary complex ν

$$M_{11}(\nu) = (m_{-\infty}(\nu) - m_\infty(\nu))^{-1}, \tag{58}$$

$$M_{12}(\nu) = M_{21}(\nu) = \tfrac{1}{2}(m_{-\infty}(\nu) + m_\infty(\nu))(m_{-\infty}(\nu) - m_\infty(\nu))^{-1}, \tag{59}$$

$$M_{22}(\nu) = m_{-\infty}(\nu)m_\infty(\nu)(m_{-\infty}(\nu) - m_\infty(\nu))^{-1}. \tag{60}$$

Equation (57) defines $\rho_{jk}(\nu)$ uniquely at points of continuity up to an arbitrary, additive constant. The functions $M_{jk}(\nu)$ $(j, k = 1, 2)$ are meromorphic if Im $\nu \neq 0$ and all their real poles are simple. The point spectrum consists exactly of the points which are real poles of one of the $M_{jk}(\nu)$. There are at most a countable number of such points. Let ν_0 be a real pole of $M_{jk}(\nu)$ and let a_{jk} be the residue there,

$$M_{jk}(\nu) = \frac{a_{jk}}{\nu - \nu_0} + \cdots . \tag{61}$$

Then it follows from (57) and (61) that

$$\rho_{jk}(\nu_0 + 0) - \rho_{jk}(\nu_0 - 0) = -\text{Re } (a_{jk}). \tag{62}$$

If ν_0 is not a pole of any $M_{jk}(\nu)$, and Im $M_{jk}(\nu_0) \neq 0$ for some (j, k), then ν_0 is a point of the continuous spectrum and

$$d\rho_{jk}(\nu_0) = \frac{1}{\pi} \text{Im } M_{jk}(\nu_0). \tag{63}$$

If ν_0 is not a pole of any $M_{jk}(\nu)$ and Im $M_{jk}(\nu) = 0$ for all (j, k) in some neighborhood of ν_0, then ν_0 is not in the spectrum and

$$d\rho_{jk}(\nu) = 0 \qquad (j, k = 1, 2) \tag{64}$$

in a neighborhood of ν_0.

3.2 *TE Fields for Symmetric Step Model*

We now apply these formulas to the symmetric step model for the case of the *TE* field. The functions $K_n(x)$ $(n = x, y, z)$ are defined by

(1 and 2). Equation (21) has constant coefficients in the two regions $|x| < w$ and $|x| > w$. Since $e_y(x, z)$ and $h_z(x, z)$ must be continuous at $x = \pm w$, the desired solution of (21) must be continuous and have a continuous derivative. We have

$$\varphi_1(x, \nu) = \cos(\omega_\nu x), \qquad |x| \leqq w \tag{65}$$

$$= \cos(\omega_\nu w)\cos\{\omega_0(|x| - w)\} - (\omega_\nu/\omega_0)\sin(\omega_\nu w)\sin\{\omega_0(|x| - w)\}, \qquad |x| \geqq w \tag{66}$$

$$\varphi_2(x, \nu) = (1/\omega_\nu)\sin(\omega_\nu x), \qquad |x| \leqq w \tag{67}$$

$$= (1/\omega_\nu)\sin(\omega_\nu w)\cos\{\omega_0(x - w)\} + (1/\omega_0)\cos(\omega_\nu w)\sin\{\omega_0(x - w)\}, \qquad x \geqq w \tag{68}$$

$$\varphi_2(x, \nu) = -\varphi_2(-x, \nu), \qquad x \leqq -w \tag{69}$$

where

$$\omega_n = \{\nu + k^2 K_n\}^{\frac{1}{2}} \qquad (n = 0, x, y). \tag{70}$$

In (70) ω_n is defined as a single-valued function of ν in the complex plane cut along the real axis from $-k^2K_n$ to ∞. That branch is chosen which is positive real on the upper side of the cut. Simple calculations now yield

$$m_\infty(\nu) = -m_{-\infty}(\nu) = \{\omega_\nu \sin(\omega_\nu w) + i\omega_0 \cos(\omega_\nu w)\}\{\cos(\omega_\nu w) - i(\omega_0/\omega_\nu)\sin(\omega_\nu w)\}^{-1}. \tag{71}$$

Therefore,

$$M_{11}(\nu) = -1/\{4M_{22}(\nu)\} = 1/\{2m_{-\infty}(\nu)\}, \tag{72}$$

$$M_{12}(\nu) = M_{21}(\nu) = 0. \tag{73}$$

In order to determine the spectrum, we begin by decomposing the whole real axis into the union of three intervals

$$I_1 = [-\infty, -k^2K_\nu], \quad I_2 = (-k^2K_\nu, -k^2K_0), \quad I_3 = [-k^2K_0, \infty]. \tag{74}$$

From (57) and (73) it is clear that $\rho_{12}(\nu)$ and $\rho_{21}(\nu)$ are constant for all ν, hence

$$d\rho_{12}(\nu) = d\rho_{21}(\nu) = 0, \qquad -\infty \leqq \nu \leqq \infty. \tag{75}$$

It is easily seen that $M_{11}(\nu)$ and $M_{22}(\nu)$ are real and have no poles or zeros in I_1. Therefore, I_1 contains no points of the spectrum, and

$$\rho_{ij}(\nu) = \rho_{ij}(-\infty), \qquad d\rho_{ij}(\nu) = 0 \qquad \nu \varepsilon I_1 \qquad (j = 1, 2). \tag{76}$$

In the interval I_2, $M_{11}(\nu)$ and $M_{22}(\nu)$ can each have a finite number of poles, and from (72) it follows that the poles $M_{11}(\nu)$ are the zeros of $M_{22}(\nu)$ and vice versa. The real poles of $M_{11}(\nu)$ are the real solutions of

$$\omega_\nu \sin(\omega_\nu w) + i\omega_0 \cos(\omega_\nu w) = 0 \tag{77}$$

and the real poles of M_{22} are the real solutions of

$$\cos(\omega_\nu w) - i(\omega_0/\omega_\nu) \sin(\omega_\nu w) = 0. \tag{78}$$

For $\nu \in I_2$, ω_ν is real while ω_0 is purely imaginary. If we let

$$b(\nu) = \omega_\nu(\nu), \qquad p(\nu) = -i\omega_0(\nu) = (-\nu - k^2K_0)^{\frac{1}{2}}, \tag{79}$$

then (77) in the single unknown ν can be replaced by the set of three equations

$$-\nu = k^2K_0 + p^2, \qquad -\nu = k^2K_\nu - b^2, \qquad b \tan bw = p, \tag{80}$$

in the two positive real unknowns b and p and the original unknown ν. Similarly, (78) can be replaced by the set of equations

$$-\nu = k^2K_0 + p^2, \qquad -\nu = k^2K_\nu - b^2, \qquad b \cot bw = -p. \tag{81}$$

These equations are well known and their solutions have been determined.[6, 9] The set of equations (80) has a finite number of real solutions and always has at least one solution for all positive values of the parameters, w, k, $K_\nu - K_0$. These are the even modes of NM. We denote corresponding values of ν by ν_{1j}, $j = 1, 2, \cdots, R_1$. The set of equations (81) also has a most finite number of solutions, although if $(wk)^2 \times (K_\nu - K_0)$ is small enough it has no real solutions. These are the odd modes of NM. We denote the values of ν corresponding to these roots ν_{2j}, $j = 1, 2, \cdots, R_2$. The points ν_{1j}, ν_{2j}, which are all in the interval I_2, comprise the point spectrum of (21). Let

$$\delta\rho(\nu) = \lim_{\epsilon \to +0} \{\rho(\nu + \epsilon) - \rho(\nu - \epsilon)\}. \tag{82}$$

Then with the aid of (62) it is easy to show that

$$\delta\rho_{11}(\nu_{1i}) = p(\nu_{1i})/\{1 + wp(\nu_{1i})\}, \quad \delta\rho_{22}(\nu_{1i}) = 0, \quad j = 1, 2, \cdots, R_1, \tag{83}$$

$$\delta\rho_{11}(\nu_{2j}) = 0, \qquad \delta\rho_{22}(\nu_{2i}) = b^2(\nu_{2i})p(\nu_{2i})/\{1 + wp(\nu_{2i})\}, \quad j = 1, 2, \cdots, R_2. \tag{84}$$

With the aid of (65) through (69) and (77) through (79) it is readily shown that

$$\varphi_1(x, \nu_{1i}) = \cos(b(\nu_{1i})x), \qquad |x| \leqq w \tag{85}$$

$$= \cos(b(\nu_{1i})w) \exp\{p(\nu_{1i})(w - |x|)\}, \qquad |x| \geqq w \tag{86}$$

$$\varphi_2(x, \nu_{2i}) = \{1/b(\nu_{2i})\} \sin(b(\nu_{2i})x), \qquad |x| \leqq w \tag{87}$$

$$= \{1/b(\nu_{2i})\} \sin(b(\nu_{2i})w) \exp\{p(\nu_{2i})(w - x)\}, \qquad x \geqq w \tag{88}$$

$$\varphi_2(x, \nu_{2i}) = -\varphi_2(-x, \nu_{2i}). \qquad x \leqq -w \tag{89}$$

It is also true that

$$\int_{-\infty}^{\infty} \varphi_i(x, \nu_{jk})^2 \, dx = 1/\delta\rho_{ii}(\nu_{jk}), \quad k = 1, 2, \cdots, R_i, \quad j = 1, 2. \tag{90}$$

The remaining points in I_2 are not in the spectrum.

Finally, in the interval I_3 it is readily shown that $M_{11}(\nu)$ and $M_{22}(\nu)$ have no poles. It is shown easily then that the whole interval I_3 is in the continuous spectrum, and in this interval

$$d\rho_{ii}(\nu) = \rho'_{ii}(\nu) \, d\nu \qquad (j = 1, 2), \tag{91}$$

where

$$\rho'_{11}(\nu) = \frac{1}{2\pi} [\omega_\nu^2 \sin^2(\omega_\nu w) + \omega_0^2 \cos^2(\omega_\nu w)]^{-1} \omega_0 , \tag{92}$$

$$\rho'_{22}(\nu) = \frac{1}{2\pi} [\omega_\nu^2 \cos^2(\omega_\nu w) + \omega_0^2 \sin^2(\omega_\nu w)]^{-1} \omega_\nu^2 \omega_0 . \tag{93}$$

In summary, the spectrum of (21) consists of the points ν_{jk}, $k = 1, 2, \cdots, R_j$, $j = 1, 2$, and the interval I_3. Equation (27) for the transmitted field can be written as

$$e_y^{(t)}(x, z) = \sum_{j=1}^{2} \sum_{k=1}^{R_j} \delta\rho_{ii}(\nu_{jk}) g_i(\nu_{jk}) \varphi_i(x, \nu_{jk}) \exp\{-i\sqrt{-\nu_{jk}}\, z\}$$

$$+ \sum_{i=1}^{2} \int_{-k^2 K_0}^{0} \exp\{-i\sqrt{-\nu}\, z\} \varphi_i(x, \nu) g_i(\nu) \rho'_{ii}(\nu) \, d\nu$$

$$+ \sum_{i=1}^{2} \int_{0}^{\infty} \exp\{-\sqrt{\nu}\, z\} \varphi_i(x, \nu) g_i(\nu) \rho'_{ii}(\nu) \, d\nu. \tag{94}$$

The terms in the first, double summation in (94) are just the possible *TE* modes which can be excited in the waveguide. The terms in the second summation represent the propagating continuum field while the terms in the last summation represent the evanescent part of the transmitted field. A useful interpretation of the propagating continuum field can be obtained as follows. Consider within the waveguide in the

region $x < -w$ an incident plane wave of the form

$$e_\nu^{(0)}(x, z, \nu) = \exp\{-i\sqrt{-\nu}\, z - i\omega_0(\nu)x\}, \tag{95}$$

so that if θ is the direction of propagation of this wave (measured clockwise from the positive z axis), then

$$\cos\theta = \sqrt{-\nu}/k\sqrt{K_0}\,, \quad \sin\theta = \omega_0(\nu)/k\sqrt{K_0}\,. \tag{95}$$

On striking the region of higher dielectric constant, $|x| < w$, part of this wave will be reflected and part of it will be transmitted through the region $|x| < w$. Denote by $\chi_+(x, z, \nu)$ this total electromagnetic field set up by the incident wave, (95). Similarly, denote by $\chi_-(x, z, \nu)$ the total electromagnetic field set up by the incident wave in the region $x > w$

$$e_\nu^{(0)}(x, z, \nu) = \exp\{-i\sqrt{-\nu}\, z + i\omega_0(\nu)x\}. \tag{97}$$

In Fig. (3) we give a schematic description of χ_+ and χ_-. Then it can be shown that for $-k^2K_0 \leqq \nu \leqq 0$,

$$\exp\{-i\sqrt{-\nu}\, z\}\varphi_j(x, \nu) = a_j(\nu)\chi_+(x, z, \nu) + b_j(\nu)\chi_-(x, z, \nu) \quad (j = 1, 2). \tag{98}$$

For the above values of ν the directions of propagation of the incident waves for χ_+ and χ_- fill the interval $-\pi/2 \leqq \theta \leqq \pi/2$. Thus, the propagating continuum field is just a wave packet of plane waves appropriate to the medium defined by the dielectric tensor $K_n(x)$.

Similarly, the evanescent part of the field can be interpreted as a superposition of waves bound to the surface $z = 0$ and propagating in

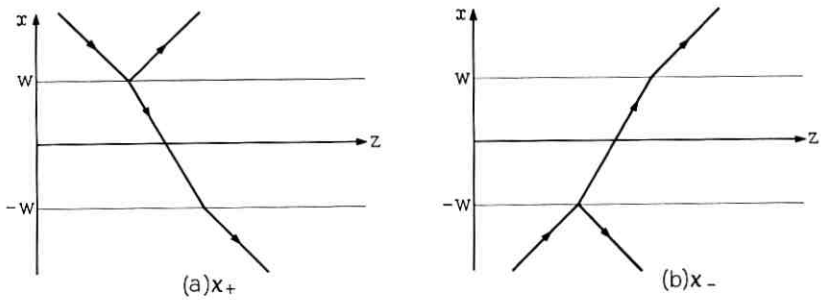


Fig. 3 — A schematic diagram of the plane waves appropriate to the dielectic medium in the symmetric step model. The wave χ_+ is incident on the junction region from the positive x direction, while χ_- is incident from the negative x direction.

the positive and negative x directions. The distinction between the propagating and evanescent parts of the transmitted field is further shown in the expression for the time averaged transmitted power, (29), which for the symmetric step model is

$$P_t = (2\omega\mu_0)^{-1} \sum_{i=1}^{2} \sum_{k=1}^{R_i} \sqrt{-\nu_{ik}} \mid g_i(\nu_{ik}) \mid^2 \delta\rho_{ii}(\nu_{ik}) + (2\omega\mu_0)^{-1} \sum_{i=1}^{2} \int_{-k^2K_0}^{0} \sqrt{-\nu} \mid g_i(\nu) \mid^2 \rho'_{ii}(\nu)\, d\nu. \quad (99)$$

As this expression shows, the evanescent part of the field transmits no energy on the average.

3.3 *TM Fields For Symmetric Step Model*

The *TM* fields of the symmetric step model can be treated similarly. Equation (43) has constant coefficients in the two regions $\mid x \mid < w$ and $\mid x \mid > w$. Since $e_z(x, z)$ and $h_y(x, z)$ must be continuous at $x = \pm w$, the solutions of (43) must be such that $\psi_j(x, \nu)$ and $\{1/K_x(x)\}\psi'_j(x, \nu)$ $(j = 1, 2)$ are continuous. We have

$$\psi_1(x, \nu) = \cos(K_r\omega_x x), \qquad \mid x \mid \leqq w \quad (100)$$

$$= \cos(K_r\omega_x w) \cos\{\omega_0(\mid x \mid - w)\}$$

$$- \{(\omega_x K_0)/(\omega_0 K_g)\} \sin(K_r\omega_x w) \sin\{\omega_0(\mid x \mid - w)\}, \qquad \mid x \mid \geqq w \quad (101)$$

$$\psi_2(x, \nu) = \{K_g/\omega_x\} \sin(K_r\omega_x x), \qquad \mid x \mid \leqq w \quad (102)$$

$$= \{K_g/\omega_x\} \sin(K_r\omega_x w) \cos\{\omega_0(x - w)\}$$

$$+ \{K_0/\omega_0\} \cos(K_r\omega_x w) \sin\{\omega_0(x - w)\}, \qquad x \geqq w \quad (103)$$

$$\psi_2(x, \nu) = -\psi_2(-x, \nu), \qquad x \leqq -w \quad (104)$$

where

$$K_g = (K_x K_z)^{\frac{1}{2}}, \qquad K_r = (K_z/K_x)^{\frac{1}{2}}, \quad (105)$$

and ω_x and ω_0 are defined in (70). Next,

$$m_\infty(\nu) = -m_{-\infty}(\nu) = \{(\omega_x/K_g) \sin(K_r\omega_x w) + i(\omega_0/K_0) \cos(K_r\omega_x w)\} \cdot \{\cos(K_r\omega_x w) - i(K_g\omega_0/K_0\omega_x) \sin(K_r\omega_x w)\}^{-1}. \quad (106)$$

Therefore,

$$M_{11}(\nu) = -1/\{4M_{22}(\nu)\} = 1/\{2m_{-\infty}(\nu)\}, \quad (107)$$

$$M_{12}(\nu) = M_{21}(\nu) = 0, \quad (108)$$

and from (57) and (108) we have

$$d\sigma_{12}(\nu) = d\sigma_{21}(\nu) = 0, \qquad -\infty < \nu < \infty. \tag{109}$$

The spectrum in the case of TM fields is determined in the same way as in the case of TE fields, and we merely state the results. There are no points of the spectrum in the interval $I_1 = [-\infty, -k^2K_x]$,

$$\sigma_{jj}(\nu) = \sigma_{jj}(-\infty), \qquad d\sigma_{jj}(\nu) = 0, \qquad \nu \,\varepsilon\, I_1 \qquad (j = 1, 2). \tag{110}$$

The interval $I_2 = (-k^2K_x, -k^2K_0)$ contains a finite number of points in the point spectrum. The points of discontinuity of $\sigma_{11}(\nu)$ are the real solutions of

$$(\omega_z/K_z) \sin (K_r\omega_x w) + i(\omega_0/K_0) \cos (K_r\omega_x w) = 0, \tag{111}$$

while the points of discontinuity of $\sigma_{22}(\nu)$ are the real solutions of

$$\cos (K_r\omega_x w) - i(K_z\omega_0/K_0\omega_x) \sin (K_r\omega_x w) = 0. \tag{112}$$

If we let

$$b(\nu) = K_r\omega_x(\nu), \qquad p(\nu) = -i\omega_0(\nu) = (-\nu - k^2K_0)^{\frac{1}{2}}, \tag{113}$$

then (111) in the single unknown ν can be replaced by the set of equations

$$-\nu = k^2K_0 + p^2, \qquad -\nu = k^2K_x - K_xb^2/K_z, \qquad bK_0 \tan bw = pK_z, \tag{114}$$

in the two positive real unknowns b and p and the original unknown ν. In the same way, (112) can be replaced by the set of equations

$$-\nu = k^2K_0 + p^2, \qquad -\nu = k^2K_x - K_xb^2/K_z, \qquad bK_0 \cot bw = -pK_z. \tag{115}$$

The set of (114) has a finite number of real solutions and for all positive values of the parameters K_0/K_z, K_x/K_z, w, $k^2(K_x - K_0)$ there is always at least one solution.[6, 9] These are the even modes of NM. The corresponding values of ν are denoted by ν_{1j}, $j = 1, 2, \cdots, S_1$. The set of equations (115) also has at most a finite number of solutions, although if $(wk)^2 (K_x - K_0)$ is small enough it has no real solutions. These are the odd modes of NM. The corresponding values of ν are denoted by ν_{2j}, $j = 1, 2, \cdots, S_2$. The points ν_{1j}, ν_{2j} are the point spectrum of (43) and they all lie in the interval I_2. Furthermore,

$$\delta\sigma_{11}(\nu_{1i}) = \mathrm{S}(p(\nu_{1i})), \qquad \delta\sigma_{22}(\nu_{1i}) = 0, \qquad j = 1, 2, \cdots, S_1, \tag{116}$$

$$\delta\sigma_{11}(\nu_{2i}) = 0, \quad \delta\sigma_{22}(\nu_{2i}) = b(\nu_{2i})^2 \mathrm{S}(p(\nu_{2i}))/K_z^2, \quad j = 1, 2, \cdots, S_2, \tag{117}$$

where

$$s(p) = K_x p\left[wp + \frac{k^2K_0K_x(K_x - K_0)}{(K_xK_z - K_0^2)p^2 + k^2K_0^2(K_x - K_0)}\right]^{-1}. \quad (118)$$

From (100) through (104) and (111) through (113) it follows that

$$\psi_1(x, \nu_{1j}) = \cos(b(\nu_{1j})x), \qquad |x| \leqq w \quad (119)$$

$$= \cos(b(\nu_{1j})w) \exp\{p(\nu_{1j})(w - |x|)\}, \qquad |x| \geqq w \quad (120)$$

$$\psi_2(x, \nu_{2j}) = \{K_z/b(\nu_{2j})\} \sin(b(\nu_{2j})x), \qquad |x| \leqq w \quad (121)$$

$$= \{K_z/b(\nu_{2j})\} \sin(b(\nu_{2j})x) \exp\{p(\nu_{2j})(w - x)\}, \qquad x \geqq w, \quad (122)$$

$$\psi_2(x, \nu_{2j}) = -\psi_2(-x, \nu_{2j}). \qquad x \leqq -w \quad (123)$$

It is also true that

$$\int_{-\infty}^{\infty} \psi_j(x, \nu_{jk})^2\{K_x(x)\}^{-1}\,dx = 1/\delta\sigma_{jj}(\nu_{jk}), \quad k = 1, 2, \cdots, S_j, \quad j = 1, 2. \quad (124)$$

The remaining points in I_2 are not in the spectrum.

The continuous spectrum is the interval $I_3 = [-k^2K_0, \infty]$. For points of the continuous spectrum

$$d\sigma_{jj}(\nu) = \sigma'_{jj}(\nu)\,d\nu \qquad (j = 1, 2), \quad (125)$$

where

$$\sigma'_{11}(\nu) = \frac{1}{2\pi}[K_0^2\omega_x^2 \sin^2(K_r\omega_x w) + K_xK_z\omega_0^2 \cos^2(K_r\omega_x w)]^{-1}K_0K_xK_z\omega_0, \quad (126)$$

$$\sigma'_{22}(\nu) = \frac{1}{2\pi}[K_0^2\omega_x^2 \cos^2(K_r\omega_x w) + K_xK_z\omega_0^2 \sin^2(K_r\omega_x w)]^{-1}K_0\omega_z^2\omega_0. \quad (127)$$

To summarize these results, the spectrum consists of the points ν_{jk}, $k = 1, 2, \cdots, S_j$, $j = 1, 2$ and the interval I_3, and the transmitted field can be written in the form

$$h_y^{(t)}(x, z) = \sum_{j=1}^{2}\sum_{k=1}^{S_j} \sigma_{jj}(\nu_{jk})h_j(\nu_{jk})\psi_j(x, \nu_{jk}) \exp\{-i\sqrt{-\nu_{jk}}\,z\} + \sum_{j=1}^{2}\int_{-k^2K_0}^{0} \exp\{-i\sqrt{-\nu}\,z\}\psi_j(x, \nu)h_j(\nu)\sigma'_{jj}(\nu)\,d\nu + \sum_{j=1}^{2}\int_{0}^{\infty} \exp\{-\sqrt{\nu}\,z\}\psi_j(x, \nu)h_j(\nu)\sigma'_{jj}(\nu)\,d\nu. \quad (128)$$

Just as for the *TE* fields, the terms in the first, double summation in (128) are the possible *TM* modes which can be excited in the waveguide. The terms in the second summation represent the propagating continuum field while the terms in the last summation represent the evanescent part of the transmitted field. Just as for the *TE* fields, the propagating part of the continuum field can be interpreted as a wave packet of reflected and refracted plane waves, and the evanescent part of the field can be interpreted in terms of surface waves at $z = 0$. Equation (53) for the transmitted energy is

$$P_t = (2\omega\epsilon_0)^{-1} \sum_{j=1}^{2} \sum_{k=1}^{S_j} \sqrt{-\nu_{jk}} \mid h_j(\nu_{jk}) \mid^2 \delta\sigma_{jj}(\nu_{jk}) + (2\omega\epsilon_0)^{-1} \sum_{j=1}^{2} \int_{-k^2K_0}^{0} \sqrt{-\nu} \mid h_j(\nu) \mid^2 \sigma'_{jj}(\nu)\, d\nu. \quad (129)$$

3.4 *TE Fields For Asymmetric Step Model*

We now turn to the second of the two models which are studied in detail and examine the *TE* fields for the asymmetric step model. The functions $K_n(x)$ $(n = x, y, z)$ are defined by (3) through (5). Equation (21) has constant coefficients in the regions $|x| < w$, $x > w$, $x < -w$, and we seek solutions which are continuous and have continuous first derivatives. Then

$$\varphi_1(x, \nu) = \cos(\omega_y x), \qquad |x| \leqq w \quad (130)$$

$$= \cos(\omega_y w) \cos\{\omega_2(x - w)\} - (\omega_y/\omega_2) \sin(\omega_y w) \sin\{\omega_2(x - w)\}, \qquad x \geqq w \quad (131)$$

$$= \cos(\omega_y w) \cos\{\omega_1(x + w)\} + (\omega_y/\omega_1) \sin(\omega_y w) \sin\{\omega_1(x + w)\}, \qquad x \leqq -w \quad (132)$$

$$\varphi_2(x, \nu) = (1/\omega_y) \sin(\omega_y w), \qquad |x| \leqq w \quad (133)$$

$$= (1/\omega_y) \sin(\omega_y w) \cos\{\omega_2(x - w)\} + (1/\omega_2) \cos(\omega_y w) \sin\{\omega_2(x - w)\}, \qquad x > w \quad (134)$$

$$= -(1/\omega_y) \sin(\omega_y w) \cos\{\omega_1(x + w)\} + (1/\omega_1) \cos(\omega_y w) \sin\{\omega_1(x + w)\}, \qquad x \leqq -w, \quad (135)$$

where

$$\omega_n(\nu) = (\nu + k^2K_n)^{\frac{1}{2}} \qquad (n = 1, 2, x, y). \quad (136)$$

As before ω_n is defined as a single-valued function of ν in the complex plane cut along the real axis from $-k^2K_n$ to ∞. Then

$$m_\infty(\nu) = \{\omega_\nu \sin(\omega_\nu w) + i\omega_2 \cos(\omega_\nu w)\} \cdot \{\cos(\omega_\nu w) - i(\omega_2/\omega_\nu) \sin(\omega_\nu w)\}^{-1}, \quad (137)$$

$$m_{-\infty}(\nu) = -\{\omega_\nu \sin(\omega_\nu w) + i\omega_1 \cos(\omega_\nu w)\} \cdot \{\cos(\omega_\nu w) - i(\omega_1/\omega_\nu) \sin(\omega_\nu w)\}^{-1}. \quad (138)$$

From (58) through (60) and (137) through (138) we obtain

$$M_{jk}(\nu) = N_{jk}(\nu)/D(\nu) \qquad (j, k = 1, 2), \quad (139)$$

where

$$N_{11}(\nu) = -\tfrac{1}{2}[(1 - \omega_1\omega_2/\omega_\nu^2) + (1 + \omega_1\omega_2/\omega_\nu^2) \cos(2\omega_\nu w) - i\{(\omega_1 + \omega_2)/\omega_\nu\} \sin(2\omega_\nu w)], \quad (140)$$

$$N_{12}(\nu) = N_{21}(\nu) = (i/2)(\omega_1 - \omega_2), \quad (141)$$

$$N_{22}(\nu) = \tfrac{1}{2}\{(\omega_\nu^2 - \omega_1\omega_2) - (\omega_\nu^2 + \omega_1\omega_2) \cos(2\omega_\nu w) + i\omega_\nu(\omega_1 + \omega_2) \sin(2\omega_\nu w)\}, \quad (142)$$

$$D(\nu) = (\omega_\nu + \omega_1\omega_2/\omega_\nu) \sin(2\omega_\nu w) + i(\omega_1 + \omega_2) \cos(2\omega_\nu w). \quad (143)$$

To determine the spectrum we note first that in the interval $I_1 = [-\infty, -k^2K_\nu]$, the functions $M_{jk}(\nu)$ $(j, k = 1, 2)$ are analytic and real. This interval, therefore, contains no points of the spectrum and

$$d\rho_{jk}(\nu) = 0 \qquad (j, k = 1, 2), \qquad \nu \in I_1 . \quad (144)$$

The only real poles of the functions $M_{jk}(\nu)$ are in the interval $I_2 = (-k^2K_\nu, -k^2K_1)$. These poles are the real solutions of $D(\nu) = 0$. In I_2, ω_ν is real while ω_1 and ω_2 are purely imaginary. If we let

$$b(\nu) = \omega_\nu(\nu), \qquad p_n(\nu) = -i\omega_n(\nu) = (-\nu - k^2K_n)^{\frac{1}{2}} \qquad (n = 1, 2), \quad (145)$$

then the equation $D(\nu) = 0$ is equivalent to the set of four equations

$$-\nu = k^2K_1 + p_1^2 , \qquad -\nu = k^2K_2 + p_2^2 , \qquad -\nu = k^2K_\nu - b^2, \quad (146)$$

$$\tan 2bw = \{p_1/b + p_2/b\}/\{1 - (p_1/b)(p_2/b)\},$$

in the three positive real unknowns b, p_1, p_2 and the original unknown ν. These equations and their solutions have also been studied in detail.[5, 6] In order that (146) have a solution, it is necessary and suffi-

cient that

$$K_\nu > K_n \qquad (n = 1, 2), \tag{147}$$
$$2wk(K_\nu - K_1)^{\frac{1}{2}} > \tan^{-1}\{(K_1 - K_2)/(K_\nu - K_1)\}^{\frac{1}{2}}.$$

If conditions (147) are satisfied, $D(\nu) = 0$ has a finite number of real solutions, ν_j, $j = 1, 2, \cdots, R$ which all lie in the interval I_2. This is the first significant difference between the symmetric and asymmetric step models. The symmetric step model always has at least one point in its point spectrum while the asymmetric step model may have no point spectrum.

We can write, assuming that (146) and (147) are satisfied.

$$\delta\rho_{jk}(\nu_l) = -N_{jk}(\nu_l)/D'(\nu_l), \qquad j, k = 1, 2, \qquad l = 1, 2, \cdots, R, \tag{148}$$

where $D'(\nu) = (d/d\nu)\, D(\nu)$. If we make use of (145), it is easy to show that

$$\{\delta\rho_{12}(\nu_l)\}^2 = \delta\rho_{11}(\nu_l)\delta\rho_{22}(\nu_l), \qquad l = 1, 2, \cdots, R. \tag{149}$$

Neither of the functions $\varphi_1(x, \nu_j)$ or $\varphi_2(x, \nu_j)$ is square integrable over $-\infty < x < \infty$ for $j = 1, 2, \cdots, R$. However, because of (149), they appear in (27) for $e_\nu^{(t)}(x, z)$ only in the combination

$$\Phi(x, \nu_j) = \sqrt{\delta\rho_{11}(\nu_j)}\,\varphi_1(x, \nu_j) + \{\delta\rho_{12}(\nu_j)/\sqrt{\delta\rho_{11}(\nu_j)}\}\varphi_2(x, \nu_j), \qquad j = 1, 2, \cdots, R. \tag{150}$$

If we define

$$\Phi_0(x, \nu_j) = \sqrt{\delta\rho_{11}(\nu_j)}\cos(b(\nu_j)x) + \{\delta\rho_{12}(\nu_j)/\sqrt{\delta\rho_{11}(\nu_j)}\, b(\nu_j)\}\sin(b(\nu_j)x), \tag{151}$$

then because of (146)

$$\Phi(x, \nu_j) = \Phi_0(x, \nu_j), \qquad |x| \leqq w \tag{152}$$
$$= \Phi_0(w, \nu_j)\exp\{p_2(\nu_j)(w - x)\}, \qquad x \geqq w \tag{153}$$
$$= \Phi_0(-w, \nu_j)\exp\{p_1(\nu_j)(w + x)\}. \qquad x \leqq -w \tag{154}$$

Thus, the functions $\Phi(x, \nu_j)$ are square integrable, and, as we shall see, are just the possible propagating modes in the wave guide. The remaining points in the interval I_2 are not in the spectrum.

The remainder of the real axis, the interval $-k^2K_1 \leqq \nu \leqq \infty$, forms the continuous spectrum. To show this, consider first the interval $I_3 = [-k^2K_1, -k^2K_2]$. In I_3, ω_ν and ω_1 are real, while ω_2 is purely

imaginary. The functions $M_{jk}(\nu)$ have no poles in I_3 and their imaginary parts are not zero. We introduce the notation

$$\omega_y(\nu) = b(\nu), \qquad \omega_1(\nu) = p_1(\nu), \qquad \omega_2(\nu) = ip_2(\nu), \qquad \nu \in I_3 . \tag{155}$$

Then we can write

$$d\rho_{jk}(\nu) = \frac{1}{\pi} \{p_1(\nu)/\Delta(\nu)\} r_j(\nu) r_k(\nu)\, d\nu \qquad (j, k = 1, 2), \tag{156}$$

where

$$r_1(\nu) = \cos bw + (p_2/b) \sin bw, \tag{157}$$

$$r_2(\nu) = b \sin bw - p_2 \cos bw, \tag{158}$$

$$\Delta(\nu) = \{b \sin 2bw - p_2 \cos 2bw\}^2 + \{(p_1 p_2/b) \sin 2bw + p_1 \cos 2bw\}^2. \tag{159}$$

For $\nu \in I_3$ it is clear from (131), (134), and (155) that $\varphi_1(x, \nu)$ and $\varphi_2(x, \nu)$ both grow exponentially as $x \to +\infty$. However, from (156) we see that in (27) for $e_\nu^{(t)}(x, z)$, the functions $\varphi_j(x, \nu)$ $(j = 1, 2)$ appear only in the combination

$$\Lambda(x, \nu) = r_1(\nu)\varphi_1(x, \nu) + r_2(\nu)\varphi_2(x, \nu) \tag{160}$$

when $\nu \in I_3$. However,

$$\Lambda(x, \nu) = \cos \{b(x - w)\} - (p_2/b) \sin \{b(x - w)\}, \qquad |x| \leqq w \tag{161}$$

$$= \exp \{p_2(w - x)\}, \qquad x \geqq w \tag{162}$$

$$= (\cos 2bw + (p_2/b) \sin 2bw) \cos \{p_1(x + w)\} + (1/p_1)(b \sin 2bw - p_2 \cos 2bw) \cdot \sin \{p_1(x + w)\}, \qquad x \leqq -w. \tag{163}$$

Equations (161) through (163) represent the second important difference between the symmetric and asymmetric step models. In the symmetric model all the components of the continuum field are oscillatory functions of x on both sides of the waveguide while in the asymmetric model some of the components of the continuum field are exponentially damped on one side of the waveguide. The physical interpretation of $\Lambda(x, \nu)$ will be discussed later.

In the remaining interval, $I_4 = [-k^2 K_2, \infty]$, the functions ω_n $(n = 1, 2, y)$ are all real and the functions $M_{jk}(\nu)$ $(j, k = 1, 2)$ have no poles. Therefore,

$$d\rho_{jk}(\nu) = \rho'_{jk}(\nu)\, d\nu, \tag{164}$$

where

$$\rho'_{11}(\nu) = \frac{1}{\pi}(\omega_1 + \omega_2)\{\omega_\nu^2 \cos^2(\omega_\nu w) + \omega_1\omega_2 \sin^2(\omega_\nu w)\}/\mathfrak{D}, \tag{165}$$

$$\rho'_{12}(\nu) = \rho'_{21}(\nu) = \frac{1}{\pi}\omega_\nu(\omega_1 - \omega_2)(\omega_\nu^2 + \omega_1\omega_2)\sin(\omega_{\nu w})\cos(\omega_\nu w)/\mathfrak{D}, \tag{165}$$

$$\rho'_{22}(\nu) = \frac{1}{\pi}\omega_\nu^2(\omega_1 + \omega_2)\{\omega_\nu^2 \sin^2(\omega_\nu w) + \omega_1\omega_2 \cos^2(\omega_\nu w)\}/\mathfrak{D}, \tag{167}$$

$$\mathfrak{D}(\nu) = (\omega_\nu^2 + \omega_1\omega_2)^2 \sin^2(2\omega_\nu w) + \omega_\nu^2(\omega_1 + \omega_2)^2 \cos^2(2\omega_\nu w). \tag{168}$$

The spectrum for the TE fields of the asymmetric model consists of the (possibly empty) set of points ν_j, $j = 1, 2, \cdots, R$ and the interval $-k^2K_1 \leqq \nu \leqq \infty$. The transmitted field can now be written in the following way.

$$\begin{aligned} e_\nu^{(t)}(x, z) = {} & \sum_{j=1}^{R}\left\{\sum_{k=1}^{2}\{\delta\rho_{11}(\nu_j)\}^{-\frac{1}{2}}\delta\rho_{1k}(\nu_j)g_k(\nu_j)\right\}\exp\{-i\sqrt{-\nu_j}\,z\}\Phi(x, \nu_j) \\ & + \frac{1}{\pi}\int_{-k^2K_1}^{-k^2K_2}\exp\{-i\sqrt{-\nu}\,z\}\Lambda(x, \nu)\left\{\sum_{j=1}^{2} r_j(\nu)g_j(\nu)\right\}\{p_1(\nu)/\Delta(\nu)\}\,d\nu \\ & + \sum_{j,k=1}^{2}\int_{-k^2K_2}^{0}\exp\{-i\sqrt{-\nu}\,z\}\varphi_j(x, \nu)g_k(\nu)\rho'_{jk}(\nu)\,d\nu \\ & + \sum_{j,k=1}^{2}\int_{0}^{\infty}\exp\{-\sqrt{\nu}\,z\}\varphi_j(x, \nu)g_k(\nu)\rho'_{jk}(\nu)\,d\nu. \end{aligned} \tag{169}$$

The expression for $e_\nu^{(t)}(x, z)$ has been split up into a sum of parts in order to facilitate its physical interpretation. The first part represents the possible discrete, propagating modes which can be excited in the system. The form of these modes has been studied in detail elsewhere,[5,6] and as pointed out earlier, unless condition (147) is satisfied, no such modes can be excited. In order to interpret the second term, consider within the waveguide in the region $x < -w$ an incident plane wave of the form

$$e_\nu^{(0)}(x, z, \nu) = \exp\{-i\sqrt{-\nu}\,z - i\omega_1(\nu)x\}. \tag{170}$$

At the surface $x = -w$, part of this wave will be reflected and part will be transmitted. However, at the surface $x = w$, the wave will suffer total internal reflection. The total electromagnetic field set up by $e_\nu^{(0)}(x, z, \nu)$ is proportional to $\Lambda(x, \nu)\exp\{-i\sqrt{-\nu}z\}$. The second term is then just a superposition of plane waves which are totally reflected at $x = w$. In Fig. 4 we give a schematic description of these

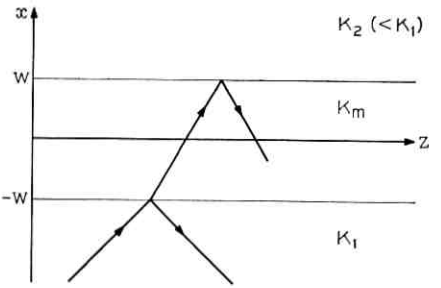


Fig. 4 — A schematic diagram of the totally reflected wave in the asymmetric step model. The wave is incident on the junction region at $x = -w$ where it is partly reflected and partly transmitted. The partly transmitted portion is then totally reflected at $x = w$.

waves. In microscopy the theory of the Becke line is based on just such a superposition of totally reflected plane waves.[10] The third term is a superposition of plane waves which are reflected and refracted at $x = \pm w$. The last term is a superposition of waves bound to the surface $z = 0$ and propagating in the positive and negative x directions.

The time averaged, transmitted power is

$$P_t = (2\omega\mu_0)^{-1} \sum_{l=1}^{R} \sqrt{-\nu_l} \mid \sqrt{\delta\rho_{11}(\nu_l)}\, g_1(\nu_l)$$
$$+ \{\delta\rho_{12}(\nu_l)/\sqrt{\delta\rho_{11}(\nu_l)}\}\, g_2(\nu_l) \mid^2$$
$$+ (2\omega\mu_0\pi)^{-1} \int_{-k^2K_1}^{-k^2K_2} \sqrt{-\nu} \mid r_1(\nu)g_1(\nu) + r_2(\nu)g_2(\nu) \mid^2 \{p_1(\nu)/\Delta(\nu)\}\, d\nu$$
$$+ (2\omega\mu_0)^{-1} \int_{-k^2K_2}^{0} \sqrt{-\nu} \left\{ \sum_{i,k=1}^{2} g_i(\nu)^* g_k(\nu) \rho'_{ik}(\nu) \right\} d\nu. \tag{171}$$

3.5 *TM Fields For Asymmetric Step Model*

The *TM* fields for the asymmetric model present no new features, and we merely record the results. We have

$$\psi_1(x, \nu) = \cos(K_r\omega_x x), \qquad \mid x \mid \leqq w \tag{172}$$
$$= \cos(K_r\omega_x w) \cos\{\omega_2(x - w)\}$$
$$- (\omega_x K_2/\omega_2 K_v) \sin(K_r\omega_x w) \sin\{\omega_2(x - w)\}, \quad x \geqq w \tag{173}$$
$$= \cos(K_r\omega_x w) \cos\{\omega_1(x + w)\}$$
$$+ (\omega_x K_1/\omega_1 K_v) \sin(K_r\omega_x w) \sin\{\omega_1(x + w)\}, \quad x \leqq -w \tag{174}$$

$$\psi_2(x, \nu) = (K_g/\omega_x) \sin (K_r\omega_x x), \qquad |x| \leqq w \tag{175}$$

$$= (K_g/\omega_x) \sin (K_r\omega_x w) \cos \{\omega_2(x - w)\} + (K_2/\omega_2) \cos (K_r\omega_x w) \sin \{\omega_2(x - w)\}, \qquad x \geqq w \tag{176}$$

$$= -(K_g/\omega_x) \sin (K_r\omega_x w) \cos \{\omega_1(x + w)\} + (K_1/\omega_1) \cos (K_r\omega_x w) \sin \{\omega_1(x + w)\}, \qquad x \leqq -w \tag{177}$$

where $\omega_n(\nu)$ $(n = x, 1, 2)$ are defined in (136) and K_g and K_r are defined in (105). Next,

$$m_\infty(\nu) = \{(\omega_x/K_g) \sin (K_r\omega_x w) + i(\omega_2/K_2) \cos (K_r\omega_x w)\} \cdot \{\cos (K_r\omega_x w) - i(\omega_2 K_g/\omega_x K_2) \sin (K_r\omega_x w)\}^{-1}, \tag{178}$$

$$m_{-\infty}(\nu) = -\{(\omega_x/K_g) \sin (K_r\omega_x w) + i(\omega_1/K_1) \cos (K_r\omega_x w)\} \cdot \{\cos (K_r\omega_x w) - i(\omega_1 K_g/\omega_x K_1) \sin (K_r\omega_x w)\}^{-1}. \tag{179}$$

Then from (58) through (60), (178), and (179) we obtain

$$M_{jk}(\nu) = N_{jk}(\nu)/D(\nu) \qquad (j, k = 1, 2), \tag{180}$$

where

$$N_{11}(\nu) = -\tfrac{1}{2}[(1 - \omega_1\omega_2 K_g^2/\omega_x^2 K_1 K_2) + (1 + \omega_1\omega_2 K_g^2/\omega_x^2 K_1 K_2) \cos (2K_r\omega_x w) - i(K_g/\omega_x)(\omega_1/K_1 + \omega_2/K_2) \sin (2K_r\omega_x w)], \tag{181}$$

$$N_{12}(\nu) = N_{21}(\nu) = (i/2)(\omega_1/K_1 - \omega_2/K_2), \tag{182}$$

$$N_{22}(\nu) = \tfrac{1}{2}[(\omega_x^2/K_g^2 - \omega_1\omega_2/K_1 K_2) - (\omega_x^2/K_g^2 + \omega_1\omega_2/K_1 K_2) \cos (2K_r\omega_x w) + i(\omega_x/K_g)(\omega_1/K_1 + \omega_2/K_2) \sin (2K_r\omega_x w)], \tag{183}$$

$$D(\nu) = (\omega_x/K_g + \omega_1\omega_2 K_g/\omega_x K_1 K_2) \sin (2K_r\omega_x w) + i(\omega_1/K_1 + \omega_2/K_2) \cos (2K_r\omega_x w). \tag{184}$$

There are no points of the spectrum in $I_1 = [-\infty, -k^2 K_x]$. The only real poles of the functions $M_{jk}(\nu)$ are in the interval $I_2 = (-k^2 K_x, -k^2 K_1)$. In I_2, ω_x is real while ω_1 and ω_2 are imaginary. If we let

$$b(\nu) = K_r\omega_x(\nu), \qquad p_n(\nu) = -i\omega_n(\nu), \qquad n = 1, 2, \tag{185}$$

then the equation determining the poles, $D(\nu) = 0$ is equivalent to the set of equations

$$-\nu = k^2K_1 + p_1^2, \quad -\nu = k^2K_2 + p_2^2, \quad -\nu = k^2K_x - \frac{K_x}{K_z} b^2,$$
$$\tan 2bw = (p_1K_z/bK_1 + p_2K_z/bK_2)/(1 - p_1p_2K_z^2/b^2K_1K_2). \tag{186}$$

In order that these equations have a solution, it is necessary and sufficient that[5, 6]

$$K_x > K_n \qquad (n = 1, 2),$$
$$2wk\{K_z(K_x - K_1)/K_x\}^{\frac{1}{2}} > \tan^{-1}\{K_xK_z(K_1 - K_2)/K_2^2(K_x - K_1)\}^{\frac{1}{2}}. \tag{187}$$

If conditions (187) are satisfied, $D(\nu) = 0$ has a finite number of real solutions in I_2, ν_j, $j = 1, 2, \cdots, S$.

If (186) and (187) are satisfied, we can write

$$\delta\sigma_{jk}(\nu_l) = -N_{jk}(\nu_l)/D'(\nu_l), \qquad j, k = 1, 2, \qquad l = 1, 2, \cdots, S. \tag{188}$$

Just as for the TE fields, it is true that

$$\{\delta\sigma_{12}(\nu_l)\}^2 = \delta\sigma_{11}(\nu_l)\delta\sigma_{22}(\nu_l), \qquad l = 1, 2, \cdots, S. \tag{189}$$

Because of (189) the functions $\psi_1(x, \nu_l)$ and $\psi_2(x, \nu_l)$ appear in (49) for $h_y^{(l)}(x, z)$ only in the combination

$$\Psi(x, \nu_j) = \sqrt{\delta\sigma_{11}(\nu_j)}\,\psi_1(x, \nu_j) + \{\delta\sigma_{12}(\nu_j)/\sqrt{\delta\sigma_{11}(\nu_j)}\}\psi_2(x, \nu_j), \qquad j = 1, 2, \cdots, S. \tag{190}$$

If we define

$$\Psi_0(x, \nu_j) = \sqrt{\delta\sigma_{11}(\nu_j)}\cos(b(\nu_j)x) + \{K_z\delta\sigma_{12}(\nu_j)/\sqrt{\sigma_{11}(\nu_j)}\, b(\nu_j)\}\sin(b(\nu_j)x), \tag{191}$$

then because of (186)

$$\Psi(x, \nu_j) = \Psi_0(x, \nu_j), \qquad |x| \leqq w \tag{192}$$
$$= \Psi_0(w, \nu_j)\exp\{p_2(\nu_j)(w - x)\}, \qquad x \geqq w \tag{193}$$
$$= \Psi_0(-w, \nu_j)\exp\{p_1(\nu_j)(w + x)\}. \qquad x \leqq -w \tag{194}$$

The remaining points in I_2 are not in the spectrum.

The remainder of the real axis, the interval $-k^2K_1 \leqq \nu \leqq \infty$ forms the continuous spectrum. In the subinterval $I_3 = [-k^2K_1, -k^2K_2)$, ω_x and ω_1 are real while ω_2 is imaginary. If we let

$$K_x\omega_x(\nu) = b(\nu), \qquad \omega_1(\nu) = p_1(\nu), \qquad \omega_2(\nu) = ip_2(\nu), \qquad \nu \in I_3, \tag{195}$$

then we can write

$$d\sigma_{jk}(\nu) = \frac{1}{\pi}\{p_1(\nu)/K_1\Delta(\nu)\}s_j(\nu)s_k(\nu)\,d\nu \qquad (j, k = 1, 2), \tag{196}$$

where

$$s_1(\nu) = \cos bw + (p_2K_x/bK_2)\sin bw, \tag{197}$$

$$s_2(\nu) = -(p_2/K_2)\cos bw + (b/K_x)\sin bw, \tag{198}$$

$$\Delta(\nu) = \{(b/K_x)\sin 2bw - (p_2/K_2)\cos 2bw\}^2 + \{(p_1p_2K_x/bK_1K_2)\sin 2bw + (p_1/K_1)\cos 2bw\}^2. \tag{199}$$

When $\nu \in I_3$, $\psi_1(x, \nu)$ and $\psi_2(x, \nu)$ appear in (51) for $h_\nu^{(t)}(x, z)$ only in the combination

$$\Xi(x, \nu) = s_1(\nu)\psi_1(x, \nu) + s_2(\nu)\psi_2(x, \nu). \tag{200}$$

We have

$$\Xi(x, \nu) = \cos\{b(x - w)\} - (p_2K_x/bK_2)\sin\{b(x - w)\}, \qquad |x| \leqq w \tag{201}$$

$$= \exp\{p_2(w - x)\}, \qquad x \geqq w \tag{202}$$

$$= \{\cos 2bw + (p_2K_x/bK_2)\sin 2bw\}\cos\{p_1(x + w)\} + (1/p_1)\{(bK_1/K_x)\sin 2bw - (p_2K_1/K_2)\cos 2bw\}\sin\{p_1(x + w)\}, \qquad x \leqq -w. \tag{203}$$

In the remaining interval, $I_4 = [-k^2K_2, \infty]$, the functions ω_n $(n = 1, 2, x)$ are all real and we can write

$$d\sigma_{jk}(\nu) = \sigma'_{jk}(\nu)\,d\nu, \tag{204}$$

where

$$\sigma'_{11}(\nu) = \frac{1}{\pi}(\omega_1/K_1 + \omega_2/K_2)\{(\omega_x^2/K_g^2)\cos^2(K_r\omega_x w) + (\omega_1\omega_2/K_1K_2)\sin^2(K_r\omega_x w)\}/\mathfrak{D}, \tag{205}$$

$$\sigma'_{12}(\nu) = \sigma'_{21}(\nu) = \frac{1}{\pi}(\omega_x/K_g)(\omega_1/K_1 - \omega_2/K_2) \cdot \{\omega_x^2/K_g^2 + \omega_1\omega_2/K_1K_2\}\sin(K_r\omega_x w)\cos(K_r\omega_x w)/\mathfrak{D}, \tag{206}$$

$$\sigma'_{22}(\nu) = \frac{1}{\pi}(\omega_x/K_g)^2(\omega_1/K_1 + \omega_2/K_2)\{(\omega_1\omega_2/K_1K_2)\cos^2(K_r\omega_x w) + (\omega_x/K_g)^2\sin^2(K_r\omega_x w)\}/\mathfrak{D}, \tag{207}$$

$$\mathfrak{D} = \{(\omega_x/K_y)^2 + (\omega_1\omega_2/K_1K_2)\}^2 \sin^2 (2K_r\omega_x w) + (\omega_x/K_y)^2(\omega_1/K_1 + \omega_2/K_2)^2 \cos^2 (2K_r\omega_x w). \quad (208)$$

To summarize, the spectrum for the TM waves of the asymmetric model consists of the (possibly empty) set of points ν_l, $l = 1, 2, \cdots, S$, and the interval $-k^2K_1 \leqq \nu \leqq \infty$. The transmitted field can be written as

$$h_y^{(t)}(x, z) = \sum_{j=1}^{S} \sum_{k=1}^{2} \{\delta\sigma_{11}(\nu_j)\}^{-1} \delta\sigma_{1k}(\nu_j) h_k(\nu_j) \exp\{-i\sqrt{-\nu_j}\, z\} \Psi(x, \nu_j)$$
$$+ \frac{1}{\pi} \int_{-k^2K_1}^{-k^2K_2} \exp\{-i\sqrt{-\nu}\, z\} \Xi(x, \nu) \sum_{j=1}^{2} s_j(\nu) h_j(\nu) \{p_1(\nu)/K_1\Delta(\nu)\}\, d\nu$$
$$+ \sum_{j,k=1}^{2} \int_{-k^2K_2}^{0} \exp\{-i\sqrt{-\nu}\, z\} \psi_j(x, \nu) h_k(\nu) \sigma'_{jk}(\nu)\, d\nu$$
$$+ \sum_{j,k=1}^{2} \int_{0}^{\infty} \exp\{-\sqrt{\nu}\, z\} \psi_j(x, \nu) h_k(\nu) \sigma'_{jk}(\nu)\, d\nu. \quad (209)$$

The time averaged, transmitted power is

$$P_t = (2\omega\epsilon_0)^{-1} \sum_{l=1}^{S} \sqrt{-\nu_l}\, |\sqrt{\sigma_{11}(\nu_l)}\, h_1(\nu_l) + \{\sigma_{12}(\nu_l)/\sqrt{\sigma_{11}(\nu_l)}\} h_2(\nu_l)|^2$$
$$+ (2\omega\epsilon_0)^{-1} \int_{-k^2K_1}^{-k^2K_2} \sqrt{-\nu}\, |s_1(\nu)h_1(\nu) + s_2(\nu)h_2(\nu)|^2 \{p_1(\nu)/K_1\Delta(\nu)\}\, d\nu$$
$$+ (2\omega\epsilon_0)^{-1} \int_{-k^2K_2}^{0} \sqrt{-\nu} \sum_{j,k=1}^{2} h_j(\nu)^* h_k(\nu) \sigma'_{jk}(\nu)\, d\nu. \quad (210)$$

IV. APPROXIMATE SOLUTION OF THE INTEGRAL EQUATIONS

In Section II we obtained general expressions for the reflected and transmitted fields for the TE fields in (18) and (27) and for the TM fields in (41) and (51). In (27) and (51) there appear the functions $\varphi_i(x, \nu)$ and $\psi_i(x, \nu)$ and the spectral density matrices $\rho(\nu)$ and $\sigma(\nu)$. A technique for determining these quantities in certain cases was illustrated in Section III by explicitly calculating them for the symmetric and asymmetric step models. In order to complete the determination of the reflected and transmitted fields, the functions $\mathcal{E}_y^{(r)}(l)$, $\mathcal{H}_y^{(r)}(l)$, $g_k(\nu)$, and $h_k(\nu)$ must be calculated. In Section II we showed that these functions were determined by the integral equations (30)–(31) and (54)–(55).

We have been unable to solve these integral equations exactly for the general case. However, there are certain cases of great physical

interest, such as the electro-optic diode modulator, where excellent approximate solutions can be obtained. Let

$$M_n = \max_x K_n(x), \qquad m_n = \min_x K_n(x), \qquad (n = x, y, z) \tag{211}$$

and assume that

$$(M_n - m_n)/m_n \ll 1 \qquad (n = x, y, z). \tag{212}$$

Then the incident field impinges on an essentially uniform, plane dielectric interface, and the reflected field can be calculated as if the region $z > 0$ were a uniform dielectric. Let $\bar{K}_n$ $(n = x, y, z)$ be suitably chosen, constant values for the dielectric tensor for $z > 0$. Then it is readily shown that for the TE fields

$$\mathcal{E}_y^{(r)}(l) = R_e(l)\mathcal{E}_y^{(i)}(l), \tag{213}$$

and for the TM fields

$$\mathcal{H}_y^{(r)}(l) = R_h(l)\mathcal{H}_y^{(i)}(l), \tag{214}$$

where the reflection coefficients are

$$R_e(l) = \{\Omega(l) - k_y\Omega(l/k_y)\}\{\Omega(l) + k_y\Omega(l/k_y)\}^{-1}, \tag{215}$$

$$R_h(l) = \{k_x\Omega(l) - \Omega(l/k_z)\}\{k_x\Omega(l) + \Omega(l/k_z)\}^{-1}, \tag{216}$$

$$k_n = (\bar{K}_n)^{\frac{1}{2}} \qquad (n = x, y, z), \tag{217}$$

and $\Omega(l)$ is defined in (16). In this approximation, the total fields at $z = 0$ for the TE and TM fields are, respectively,

$$e_y(x, 0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} T_e(l)\mathcal{E}_y^{(i)}(l)e^{-ilx}\,dl, \tag{218}$$

$$h_y(x, 0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} T_h(l)\mathcal{H}_y^{(i)}(l)e^{-ilx}\,dl, \tag{219}$$

where the transmission coefficients are

$$T_n(l) = 1 + R_n(l), \qquad n = e, h. \tag{220}$$

Now that $e_y(x, 0)$ and $h_y(x, 0)$ are known, $g_j(\nu)$ $(j = 1, 2)$ can be calculated from (28) and $h_j(\nu)$ $(j = 1, 2)$ can be calculated from (52), since $e_y(x, 0) = e_y^{(t)}(x, 0)$ and $h_y(x, 0) = h_y^{(t)}(x, 0)$.

We illustrate some features of the calculation of $g_k(\nu)$ and $h_k(\nu)$ with the symmetric and asymmetric step models. We first note that if these models are used to study an electro-optic diode modulator, typical values of the parameters defining the dielectric tensors in (1) through (7) are[6] $n = 3.31$, $\Delta \cong 10^{-3}$, $\delta_n \cong 2 \times 10^{-4}$ $(n = x, y, z)$, $\Delta_1 = 0.96\Delta$,

$\Delta_2 = 1.04\ \Delta$. Then $M_n - m_n \cong 1.4 \times 10^{-2}$, $m_n \approx 10.9$. Condition (212) is thus well satisfied.

For the symmetric step model we let $\bar{K}_n = K_0$ $(n = x, y, z)$. If the functions $\mathcal{E}_\nu^{(i)}(l)$ and $\mathcal{H}_\nu^{(i)}(l)$ are sharply peaked about $l = 0$, then (218) and (219) can be further approximated by

$$e_\nu(x, 0) = T_e(0)\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathcal{E}_\nu^{(i)}(l)e^{-ilx}\,dl = T_e(0)e_\nu^{(i)}(x, 0), \tag{221}$$

$$h_\nu(x, 0) = T_h(0)h_\nu^{(i)}(x, 0). \tag{222}$$

The calculation of $g_k(\nu)$ and $h_k(\nu)$ is now reduced to quadratures. If the incident field is not sharply peaked, we define

$$\Phi_j(l, \nu) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi_j(x, \nu)e^{-ilx}\,dx, \tag{223}$$

$$\Psi_j(l, \nu) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi_j(x, \nu)\{k_x(x)\}^{-1}e^{-ilx}\,dx, \tag{224}$$

so that

$$g_j(\nu) = \int_{-\infty}^{\infty} T_e(l)\mathcal{E}_\nu^{(i)}(l)\Phi_j(l, \nu)\,dl, \tag{225}$$

$$h_j(\nu) = \int_{-\infty}^{\infty} T_h(l)\mathcal{H}_\nu^{(i)}(l)\Psi_j(l, \nu)\,dl, \qquad j = 1, 2. \tag{226}$$

If ν is in the continuous spectrum, $\Phi_j(l, \nu)$ and $\Psi_j(l, \nu)$ are distributions which are easily determined with the aid of the relation[11]

$$\int_0^{\infty} e^{ix\sigma}\,dx = 1/(i\sigma) + \pi\delta(\sigma), \tag{227}$$

where $\delta(\sigma)$ is the delta function and when $1/\sigma$ appears under an integral sign, it is assumed that the Cauchy principal value is taken. If ν is in the point spectrum, $\Phi_j(l, \nu)$ and $\Psi_j(l, \nu)$ are ordinary functions.

For the asymmetric step model we let $\bar{K}_n = \frac{1}{2}(K_1 + K_2)$, $(n = x, y, z)$. For this model, a straightforward application of (28) and (52) fails in general if ν is the point spectrum or if $\nu \varepsilon I_3$, because $\varphi_j(x, \nu)$ and $\psi_j(x, \nu)$ now grow exponentially as x tends to either plus infinity or minus infinity. This apparent difficulty is merely a reflection of the manner of convergence of the integrals defining $g_k(\nu)$ and $h_k(\nu)$. For our purposes here, it is enough to note from (169) and (209) that when ν is in the point spectrum, the functions $g_k(\nu)$ and $h_k(\nu)$ do not appear independently, but only in the linear combinations

$$\sum_{k=1}^{2} \{\delta\rho_{11}(\nu_j)\}^{-\frac{1}{2}}\delta\rho_{1k}(\nu_j)g_k(\nu_j) = \int_{-\infty}^{\infty} e_\nu(x, 0)\Phi(x, \nu_j)\,dx, \quad j = 1, 2, \cdots, R, \tag{228}$$

$$\sum_{k=1}^{2} \{\delta\sigma_{11}(\nu_j)\}^{-\frac{1}{2}}\delta\sigma_{1k}(\nu_j)h_k(\nu_j) = \int_{-\infty}^{\infty} h_\nu(x, 0)\Psi(x, \nu_j)\,dx, \quad j = 1, 2, \cdots, S. \tag{229}$$

The integrals on the right of (218) and (219) are now well defined. Similarly, if $\nu \in I_3$, the relevant quantities to calculate are

$$\sum_{k=1}^{2} r_k(\nu)g_k(\nu) = \int_{-\infty}^{\infty} e_\nu(x, 0)\Lambda(x, \nu)\,dx, \tag{230}$$

$$\sum_{k=1}^{2} s_k(\nu)h_k(\nu) = \int_{-\infty}^{\infty} h_\nu(x, 0)\Xi(x, \nu)\,dx. \tag{231}$$

If $\nu \in I_4$, (28) and (52) can be applied directly. Now, all the techniques discussed in the case of the symmetric model can be applied here.

V. SUMMARY

In Section I we have defined a class of dielectric waveguide models. The waveguide is formed by an anisotropic, nonuniform dielectric filling the half space $z > 0$. The dielectric tensor is diagonal in the fixed coordinate system of Fig. 1, and the diagonal matrix elements are functions of x only, $K_n(x)$ $(n = x, y, z)$.

Integral representations for the incident, reflected, and transmitted fields were given in (15), (18), and (27), respectively, for the *TE* fields, and in (39), (41) and (51), respectively, for the *TM* fields. These representations are very general, holding for a large class of functions $K_n(x)$ and incident fields. These integral representations, however, contain the unknown functions $\varphi_j(x, \nu)$, $\psi_j(x, \nu)$, $\rho_{jk}(\nu)$ and $\sigma_{jk}(\nu)$ $(j, k = 1, 2)$, which are determined solely by the dielectric tensor, $K_n(x)$, and the unknown functions $g_k(\nu)$, $h_k(\nu)$, $(k = 1, 2)$, $\mathcal{E}_\nu^{(r)}(l)$, and $\mathcal{H}_\nu^{(r)}(l)$, which also depend on the incident field and the boundary conditions at $z = 0$. It was shown that this latter group of unknown functions are the solutions of two sets of integral equations, (30)–(31) for the *TE* fields and (54)–(55) for the *TM* fields. These equations are very complicated, and we have been unable to solve them exactly for any specific models of interest.

In Section III we gave a detailed calculation of the functions $\varphi_j(x, \nu)$, $\psi_j(x, \nu)$, $\rho_{jk}(\nu)$, and $\sigma_{jk}(\nu)$ $(j, k = 1, 2)$ for both the symmetric and asymmetric step models. These calculations are important in their own

right, since the symmetric and asymmetric step models have been used extensively in the study of the electro-optic diode modulators.[1-6] However, these computations also illustrate the technique for treating the whole class of piecewise constant models. This is important, for it is not yet completely established which is the correct model to use in exploring the behavior of the electro-optic diode modulator, and it is felt that any actual physical situation can be well approximated by a piecewise constant model.

It should be noted that the success of the techniques used in this paper depends on being able to obtain exact analytic solutions of (21) and (43), or at least good analytic approximations to these solutions. There are a number of other models for which the exact solutions of (21) can be obtained, for example the continuous dielectric constant models described in Section III of *NM*. It is, however, much more difficult to find models, other than the piecewise constant models, for which (43) is solvable in terms of known functions. Nevertheless, the possibility remains of investigating the *TE* fields for a fairly wide varity of models.

The calculations of Section III provide a method of determining the discrete modes which is different from the methods used in earlier treatments.[5,6,9] These calculations showed also that the asymmetry of the background light is accentuated in the asymmetric step model by total internal reflection at the junction region boundary.

Finally, in Section IV it was shown that good approximations can be found for the functions $g_k(\nu)$, $h_k(\nu)$, $\mathfrak{IC}_\nu^{(r)}(l)$, and $\mathcal{E}_\nu^{(r)}(l)$ in certain cases of physical interest. In particular, these approximations are valid for the electro-optic diode modulator. These approximations do not depend on a particular choice of the incident field.

The final results of this paper then are integral representations for the fields for both the *TE* and *TM* fields. Of the various functions in the integrands, some have been determined exactly and good approximations have been found for the remainder for a number of important models and for arbitrary incident fields.

These integral representations are complicated in appearance, but when z is large enough, asymptotic expansions of them can be found which lend themselves to numerical analysis. In a subsequent paper asymptotic expansions of the transmitted fields will be presented for the symmetric and asymmetric step models in the case that the incident field is Gaussian and numerical results for cases of experimental interest will be presented.

VI. ACKNOWLEDGMENT

The author wishes to thank D. F. Nelson for numerous helpful conversations and continual encouragement.

REFERENCES

1. Nelson, D. F., and Reinhart, F. K., Appl. Phys. Letters, *5*, 1964, p. 148.
2. Yariv, A., and Leite, R. C. C., Appl. Phys. Letters, *2*, 1963, p. 55.
3. Ashkin, A., and Gershenzon, M., J. Appl. Phys., *34*, 1963, p. 2116.
4. Walters, W. L., J. Appl. Phys., *37*, 1966, p. 916.
5. Anderson, W. W., IEEE J. Quant. Elec., *QE-1*, 1965, p. 228.
6. Nelson, D. F., and McKenna, J., to be published in J. Appl. Phys. This will be referred to hereafter as *NM*.
7. Coddington, E. A., and Levinson, N., *Theory of Ordinary Differential Equations*, McGraw-Hill Book Company, Inc., New York, 1955, Chapter 9.
8. Stone, M. H., *Linear Transformations in Hilbert Space and Their Applications to Analysis*, Amer. Math. Soc. Colloquim Publications, Vol. *15*, New York, 1932, Chapter 10, Section 3.
9. Collin, R. E., *Field Theory of Guided Waves*, McGraw-Hill Book Company, Inc., New York, 1960, p. 473.
10. Winchell, A. N., *Elements of Optical Mineralogy*, John Wiley, New York, 1937, Part I, Principles and Methods, pp. 77-78.
11. Gelfand, I. M., and Shilov, G. E., *Generalized Functions*, Academic Press, New York, 1964, Vol. *I*, p. 360.

Demagnetizing Fields in Thin Magnetic Films

By D. B. DOVE

(Manuscript received January 31, 1967)

Demagnetizing fields play an important role in the operation of many thin magnetic film devices. A requirement of high packing density leads to strong localization of induced changes in magnetization; and, therefore, to correspondingly large demagnetizing fields and drive currents. A treatment of the demagnetizing field problem for thin film materials is given here for film properties and fields which are nonuniform along the hard anisotropy axis. Specifically considered are saturating fields, variations in film thickness and anisotropy constant, interaction between films, and the effect of easy direction bias fields.

I. INTRODUCTION

The behavior of the magnetization in thin magnetic films of large lateral extent subject to a uniform applied field may be calculated directly from a knowledge of film properties and field strength. The calculation of the behavior of magnetization in the presence of nonuniformity of film properties or of applied field, however, must take into account the demagnetizing field that arises from a local nonuniformity of magnetization. Such a situation occurs in many problems of practical interest. Internally generated fields give rise to a number of effects when nonuniform fields are applied to thin uniaxially anisotropic films.[1,2] For example, the hard axis field required for saturation may be several times the anisotropy field and the induced magnetization component may spread to regions where the applied field is very small. The occurrence of such effects in thin films has been considered by Rosenberg[3] using a calculus of variations approach and by Kump and Greene[4] and Kump[5] using an iterative numerical procedure. More recently Dove and Long[6] have shown that there is a simple solution to the nonuniform field problem in the case of nonsaturating spatially periodic applied fields, and have treated localized

fields by using a Fourier series technique. Good agreement was found with Kerr-effect probe measurements on flat and cylindrical permalloy films.

The purpose of the present work is to show how the Fourier series technique permits straightforward solution of a number of thin film magnetostatic problems. Flat and cylindrical film geometries are treated; however, the results are of special interest to the case of cylindrical films with axial hard direction, owing to the circumferential flux closure. Specifically, we consider the cases of;

(*i*) nonuniform hard axis field,
(*ii*) nonuniform saturating field,
(*iii*) variation in film thickness,
(*iv*) variation in anisotropy constant,
(*v*) external fields due to magnetization distribution in film, flux linkage with conductors, magnetic shielding,
(*vi*) interaction between parallel films, keepers, and
(*vii*) nonuniform hard axis field in presence of easy direction bias field.

It is assumed that the quantities of interest vary along the film hard axis only and that properties and fields are uniform along the easy axis. Film thickness is taken to be sufficiently small that the direction of magnetization always lies in the plane of the film, exchange forces are neglected, being insignificant for cases considered, and anisotropy dispersion effects are not included.

II. GENERAL CONSIDERATIONS

We consider demagnetizing field effects that arise in thin uniaxially anisotropic films when relevant parameters vary only along the hard anisotropy axis. Many applications fall within this category and will be treated in following sections. Many of the results may be applied to thin films of other types of magnetic materials in the range where they exhibit a constant permeability, if the effective anisotropy field is taken to be equal to the saturation magnetization divided by the permeability.

Although the demagnetizing field may be found if the magnetization distribution is known, and conversely a knowledge of the field enables the distribution to be found, there is considerably greater difficulty in determining both distribution and field directly. In the thin film case, the Fourier series technique provides a means of representing the field

distribution for which the demagnetizing field can be found quite generally. The rotation of magnetization within a film may then be found by balancing, for example, (for nonsaturating fields) anisotropy torque versus the torque due to applied field and demagnetizing field. This leads to equations relating the coefficients of the various series which in a practical application may be most conveniently evaluated by computer.

The number of terms included in the series determines the resolution with which a particular curve may be delineated. However, a series with, say, 100 terms may be made to fit ordinates at 100 locations exactly, with oscillations about the required curve elsewhere. The procedure followed here is to use the series to calculate ordinates at the 100 locations, and a smooth curve is then drawn through the calculated ordinates. Refs. 7 and 8 have been found of value for the evaluation of integrals occurring in the following sections.

Numerical examples, where given, refer to nonmagnetostrictive 80/20 NiFe films. The films are finely polycrystalline and are characterized by a uniaxial anistropy. The easy direction is taken to be circumferential in the cylindrical film case.

III. NONUNIFORM HARD AXIS FIELD

This case has been discussed previously[6] but is included here briefly for completeness. Let x represent distance along the film hard direction, M is the value of saturation magnetization, T the film thickness, K the anisotropy constant and $\theta(x)$ the angle which the direction of magnetization (at x) makes with the film easy anisotropy direction. We now assume that the applied field $H(x)$ may be adequately represented over a range $-\lambda/2$ to $+\lambda/2$ by the series

$$H(x) = \sum_{n=-\infty}^{\infty} h_n \exp(2\pi i n x/\lambda) \tag{1}$$

and that the resulting hard direction component of magnetization $M(x)$ may be similarly represented,

$$M(x) = M \sum_{n=-\infty}^{\infty} m_n \exp(2\pi i n x/\lambda). \tag{2}$$

The distribution $M(x)$ gives rise to a local (positive) pole density at location (X, Y) of amount $-\text{div } \mathbf{M}(X, Y)$. This gives rise to a field $\mathbf{dH}$ at (x, y) distance R from (X, Y) given by

$$\mathbf{dH}(x, y) = -\text{div } \mathbf{M}(X, Y) \cdot \left(\frac{d\text{vol}}{R^2}\right) \cdot (\mathbf{R}) \left(\frac{1}{R}\right)$$

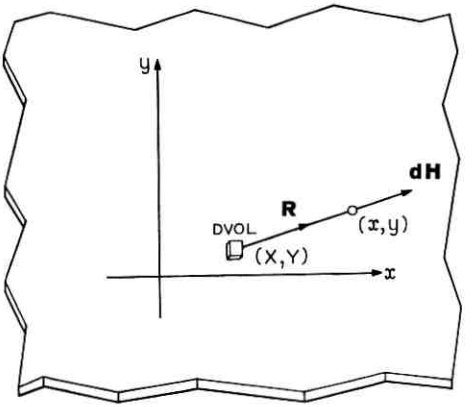


Fig. 1—A divergence of magnetization at (X,Y) gives rise to a field **dH** at (x, y). The x direction is taken to coincide with the film hard (anisotropy) direction. Under no applied field the direction of magnetization lies along the y, or easy, direction.

where **dH** is parallel to **R**, as in Fig. 1. Since the only variation of magnetization is along the x direction, variation with thickness being neglected, then div **M** reduces to $dM(X)/dx$ where $M(X)$ is the x direction component of **M**, at X.

The field **dH** has both easy and hard direction components, however, symmetry ensures that the resultant field $H_m(x)$, obtained by integrating over the film volume, lies along the hard direction. Then, we find, for a flat film

$$H_m(x) = -\int_{Y=-\infty}^{\infty}\int_{X=-\infty}^{\infty} \frac{dM(X)}{dx}\frac{(x-X)}{R^3}\,dX\,dY\,T, \tag{3}$$

where T is the film thickness. Substituting $R = [(x-X)^2 + (y-Y)^2]^{\frac{1}{2}}$ and integrating over Y we have

$$H_m(x) = -2T\int_{X=-\infty}^{\infty} \frac{dM(X)}{dx}\frac{1}{x-X}\,dX.$$

Now substituting for $M(X)$ in terms of the Fourier series, we have

$$H_m(x) = +2TM\int_{-\infty}^{\infty}\sum_{n=-\infty}^{\infty} m_n\left(\frac{2\pi in}{\lambda}\right)\frac{\exp(2\pi inX/\lambda)}{x-X}\,dX$$

and evaluating the integral,

$$H_m(x) = \sum_{-\infty}^{\infty} \alpha_n m_n \exp(2\pi inx/\lambda), \tag{4}$$

where $\alpha_n = 4\pi^2 TMn/\lambda$, $n > 0$, and $\alpha_{-n} = \alpha_n$. A similar result holds for cylindrical films having a circumferential easy direction, where x now refers to distance along the cylinder axis. In this case, we find

$$\alpha_n = 4\pi M(T/a)(2\pi na/\lambda)^2 I_0(2\pi na/\lambda)K_0(2\pi na/\lambda),$$

where a is the cylinder radius and I_0, K_0 are modified Bessel functions.

The local rotation $\theta(x)$ of magnetization away from the easy direction due to the applied field is determined by balancing the torque due to the applied field against the torques due to anisotropy and the demagnetizing field

$$2K \sin \theta(x) \cos \theta(x) + MH_m(x) \cos \theta(x) = MH(x) \cos \theta(x), \quad \text{all } x. \tag{5}$$

We note that $\sin \theta(x) = M(x)/M$, and providing $\cos \theta(x) \neq 0$, we may rewrite (5) as

$$\frac{2K}{M}\frac{M(x)}{M} + H_m(x) = H(x). \tag{6}$$

If the field is sufficiently large that $\theta(x)$ becomes equal to $\pi/2$ then the film is said to have saturated (at x) and the torque equation (5) is replaced by $M(x) = M$. In the nonsaturating case the series representations (1), (2), (4) are now substituted in (6) giving

$$H_K \sum m_n \exp(2\pi inx/\lambda) + \sum \alpha_n m_n \exp(2\pi inx/\lambda) = \sum h_n \exp(2\pi inx/\lambda),$$

where $H_K = 2K/M$. Equating coefficients of corresponding terms gives the result,

$$m_n = h_n/(H_K + \alpha_n).$$

Hence, the series for the $M(x)$ distribution may be obtained in terms of the coefficients of the applied field and geometrical parameters α_n which automatically take into account the demagnetizing field,

$$M(x) = M \sum_{-\infty}^{\infty} \frac{h_n}{H_K + \alpha_n} \exp(2\pi inx/\lambda). \tag{7}$$

As an example, we consider a wire at distance d from a flat film, lying parallel to the film easy direction. A current I along the wire produces a hard direction field component given by $H(x) = CdI/(d^2+x^2)$, where the origin for x is taken directly beneath the wire, and C is a calibration constant whose value depends on the units used, ($C = 78.8$ for d and x in mil inches, I in amperes, H in oersteds). It is

next assumed that the field is repeated at intervals λ along the hard direction in such a way that the field over one wavelength is given by

$$H(x) = \frac{C\,dI}{d^2 + x^2}, \qquad -\frac{\lambda}{2} \leqq x \leqq \frac{\lambda}{2}.$$

To determine the Fourier coefficients we proceed in the usual way, and find that for λ sufficiently large $H(x)$ is given to a good approximation by the cosine series,

$$H(x) = \frac{CI\pi}{\lambda} + \frac{2CI\pi}{\lambda} \sum_{n=1}^{\infty} e^{-2\pi nd/\lambda} \cos 2\pi nx/\lambda.$$

Substituting into (7) we have

$$M(x) = \frac{CIM\pi}{\lambda H_K} + \frac{2CIM\pi}{\lambda} \sum_{n=1}^{\infty} \frac{e^{-2\pi nd/\lambda}}{H_K + \alpha_n} \cos 2\pi nx/\lambda. \tag{8a}$$

If such a drive wire arrangement is used to apply a field to a cylindrical film, there is some variation in axial field strength across the cylinder. In many cases of interest, the cylinder diameter is small compared with axial dimensions and there is very tight magnetostatic coupling around the circumference. We therefore take the effective axial field as that applied along the wire axis, a reasonable approximation for many cases. The result (8a) then applies to the cylindrical film case provided α_n is given the appropriate value.

When a field is applied by a circular loop of radius d around the film (of radius a), it may be shown that the axial field at the surface is given by the series, for λ sufficiently large,

$$H(x, a) = \frac{CI\pi}{\lambda} + \frac{2CI\pi}{\lambda} \sum_{n=1}^{\infty} \frac{2\pi nd}{\lambda} K_1\left(\frac{2\pi nd}{\lambda}\right) I_0\left(\frac{2\pi na}{\lambda}\right) \cos \frac{2\pi nx}{\lambda},$$

where K_1, I_0 are modified Bessel functions. The field is defined over $-\lambda/2$ to $+\lambda/2$ and $d > a$. The axial component of magnetization in a cylinder excited by such a field is then,

$$M(x) = \frac{CIM\pi}{\lambda H_K} + \frac{2CIM\pi}{\lambda} \sum_{n=1}^{\infty} \frac{\frac{2\pi nd}{\lambda} K_1\left(\frac{2\pi nd}{\lambda}\right) I_0\left(\frac{2\pi na}{\lambda}\right) \cos \frac{2\pi nx}{\lambda}}{H_K + \alpha_n}. \tag{8b}$$

Similar results may be derived for fields applied by more complicated drive wire or drive strap arrangements. It can be noted that the effect of superimposing several applied fields results simply in superimposing the magnetization distributions obtained for the fields separately. Hence, one approach to designing a magnetization distribution of a required shape is to approximate the shape by superimposing a set of

known distributions. Many distributions of practical interest may be described by a cosine series and discussion in the following sections is, for clarity, limited to the cosine rather than the full series. Results for the full series may be readily derived, if required.

Fig. 2(a) to (f) shows the relative fall off in applied field $H(x)$ and in axial magnetization component $M(x)$ for a range of drive strap geometries. The plots are for a 1 μm thick cylindrical permalloy film of 5.0 mil diameter. Curves a, b, c, d correspond to drive strap half widths of 1.0, 5.0, 10.0, 20.0 mils, respectively. In Fig. 2(a), (b) the distance between drive strap (or return strap) and film axis is 3.5 mils. Fig. 2(c), (d) and (e), (f) correspond, respectively, to a distance of 5.0 and 10.0 mils. It can be noted that the magnetization distributions extend to a considerable distance and do not vary as strongly as the applied field. The fields of Fig. 2(a), (c), (e) are shown to normalized scale, however, the peak field or drive current required to just saturate the axial component at $x = 0$ varies significantly with geometry, and is shown in Fig. 3.

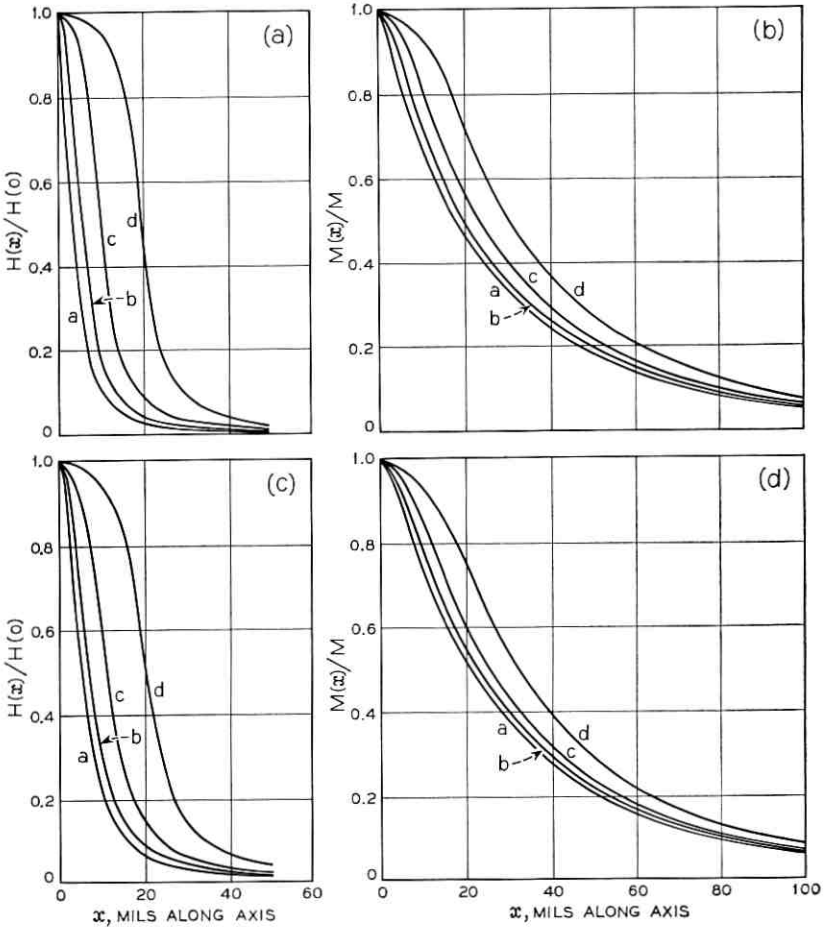
In a plated wire memory, the local state of a region of film may be assigned as positive or negative depending on the remanent circumferential component of magnetization. To read out the circumferential component in a nondestructive manner, a local axial field is applied by a drive strap surrounding the wire at the location of interest, and the signal appearing across the ends of the plated wire is measured. The signal is due to the circumferential flux change integrated along the wire (neglecting capacitive or other emfs). The circumferential component distribution is obtained simply from the axial component using the relation, $M(\text{circumferential}) = (M^2 - M(\text{axial})^2)^{\frac{1}{2}}$. The total area under this curve is proportional to the signal obtained when the circumferential component has been set completely into one direction. It is convenient to equate the integrated circumferential component to an equivalent length of film that has everywhere a 90° rotation of magnetization. Fig. 4 shows the equivalent lengths of film for the curves of Fig. 2.

If now a locally reversed region is established and the readout field applied again, the signal will have decreased, since the reversed region contributes to the signal with reversed sign. It has been found previously[6] that the presence of a domain wall has little effect on the macroscopic magnetization distribution; hence, the curves of Fig. 2 may be used to estimate the new signal. In this case, the area under the circumferential plot is taken negatively over the length of the reversed

region and positively for the remainder. Fig. 5 shows curves of net equivalent length versus width of reversed region. Curves a, and b correspond to strap half width of 1.0 mil but half separations of 3.5 and 5.0 mils, respectively. Curves c and d correspond to strap half width of 10.0 mils, and half separations of 5.0 and 10.0 mils, respectively.

IV. NONUNIFORM FIELDS LARGE ENOUGH TO PRODUCE LOCAL SATURATION

When the local effective field reaches the value H_K then the local magnetization rotation has the value $\pi/2$; hence, $M(x) = M$, the



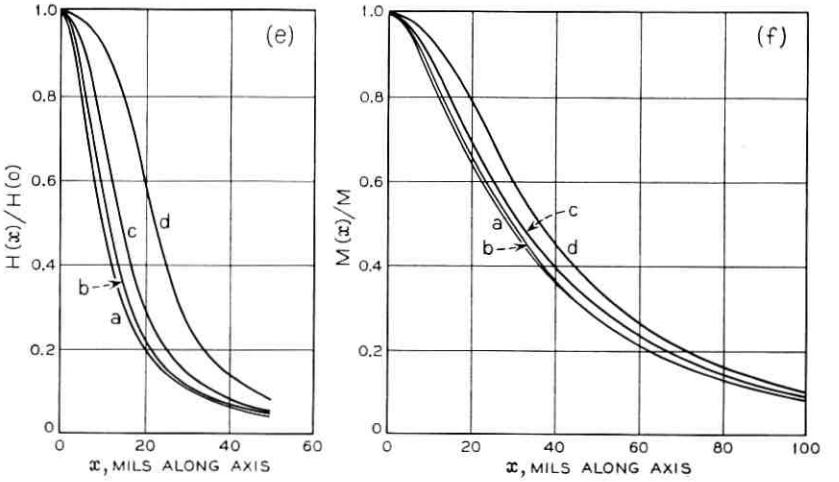


Fig. 2 — The curves denoted a, b, c, d refer, respectively, to a parallel drive strap arrangement of half widths 1.0, 5.0, 10.0, and 20.0 mils. (a) and (b) correspond to a strap-to-film axis distance of 3.5 mils, (c) and (d) correspond to 5.0 mils and (e) and (f) to 10.0 mils. (a), (c), and (e) give to normalized scale the field $H(x)/H(0)$ applied along the axis of a 5.0 mil diameter, 1μm thick cylindrical permalloy film with $H_K = 3.0$. (b), (d), and (f) show the resulting axial magnetization components $M(x)/M$ due to the actual (i.e., non-normalized) applied field.

saturation value. A further increase in the field cannot therefore, produce any further increase in $M(x)$ and it is necessary to modify the preceding discussion to take the effect of saturation into account.

We assume that the magnetization distribution is monotonic, and the width of the saturated region is specified at the outset. The current required to produce this degree of saturation may then be found for a given drive strap geometry, and the resulting magnetization distribution is calculated. This somewhat arbitrary procedure renders the problem tractable.

If the film has saturated over a region $-R \leqq x \leqq R$ then the material within this region has $M(x) = M$ a constant; hence, $dM(x)/dx$ vanishes within this region. It is convenient to introduce a modifying function $S(x)$, having period λ, that is zero over the range $-R \leqq x \leqq R$, but is otherwise unity. The product $S(x)dM(x)/dx$ then has the property of being zero over $-R \leqq x \leqq R$ but is otherwise equal to $dM(x)/dx$. By introducing this product into the integral for the demagnetizing field in place of $dM(x)/dx$, we have effectively modified the integral without changing the limits of integration. Let

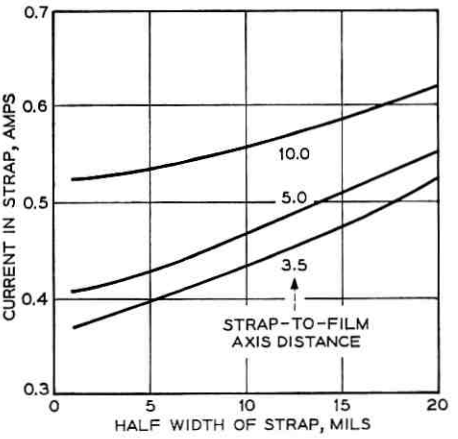


Fig. 3 — Current in drive strap required to just saturate the film of Fig. 2 at $x = 0$, for the several drive strap geometries of Fig. 2.

$H(x)$ and $M(x)$ be represented by the finite series

$$H(x) = \sum_{0}^{N} h_n \cos 2\pi nx/\lambda, \qquad M(x) = M \sum_{0}^{N} m_n \cos 2\pi nx/\lambda,$$

also let $S(x)$ be represented by a cosine series, then

$$S(x) = \sum_{n=0}^{\infty} s_n \cos 2\pi nx/\lambda,$$

where for the required step function

$$s_0 = 1 - (2R/\lambda), \qquad s_n = -\frac{4R}{\lambda}\left(\frac{\sin 2\pi nR/\lambda}{2\pi nR/\lambda}\right), \qquad n > 0.$$

Differentiating the series for $M(x)$, we have

$$\frac{dM(x)}{dx} = -\frac{2\pi M}{\lambda} \sum_{n=0}^{N} nm_n \sin 2\pi nx/\lambda.$$

Then the product may be written,

$$S(x)\frac{dM(x)}{dx} = -\frac{2\pi M}{\lambda} \sum_{j=0}^{\infty} \sum_{n=0}^{\infty} s_j nm_n \cos 2\pi jx/\lambda \sin 2\pi nx/\lambda$$

$$= -\frac{\pi M}{\lambda} \sum_{j=0}^{\infty} \sum_{n=0}^{N} s_j nm_n(\sin 2\pi(j+n)x/\lambda - \sin 2\pi(j-n)x/\lambda).$$

This represents a series of the form $A_0 + A_1 \sin 2\pi x/\lambda + \ldots$ and we may rearrange by grouping the coefficients to obtain

$$S(x)\frac{dM(x)}{dx} = -\frac{\pi M}{\lambda}\sum_{p=1}^{N} p s_p m_p - \frac{\pi M}{\lambda}\sum_{n=1}^{N}\left[\sum_{p=1}^{N}(s_{|p-n|} - s_{p+n} + s_0\delta_p^n)pm_p\right]\sin 2\pi nx/\lambda,$$

where $\delta_p^n = 1$ when $p = n$, but is otherwise zero, and the series for $S(x)$ is terminated for subscripts greater than $2N$. Using this final series in place of the series for $dM(x)/dx$ in the integral (3) for the demagnetizing field we obtain,

$$H_m(x) = \sum_{n=1}^{N}\left\{\frac{1}{2n}\sum_{p=1}^{N}(s_{|n-p|} - s_{n+p} + s_0\delta_p^n)pm_p\right\}\alpha_n \cos 2\pi nx/\lambda, \qquad (9)$$

where the α_n have the values calculated previously for the nonsaturating case. There are now several conditions that the magnetization distribution must satisfy: it has the value $M(x) = M$ over the range $-R \leqq x \leqq R$ and satisfies the torque equation (6) outside this range, and finally, the amplitude of the applied field is such that $M(x)$ determined from (6) has also the value M at $x = \pm R$. The required field value is given by the calculation for any particular drive strap

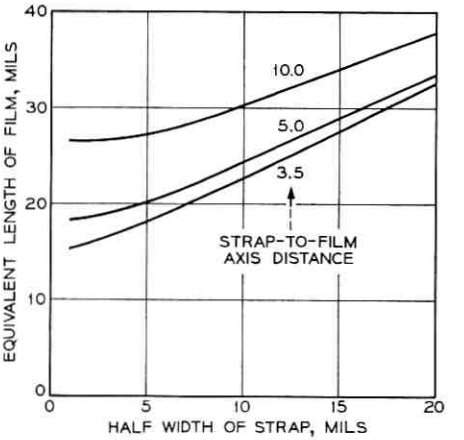


Fig. 4 — The change in circumferential component of magnetization averaged along the film is proportional to the signal obtained during readout. This is expressed in terms of equivalent length of film that would produce the same signal when uniformly excited to saturation. The plots are derived from the axial component distributions of Fig. 2.

configuration. We now substitute the series (1), (2), and (9) into the torque equation (6) and gathering coefficients, we obtain,

$$H_K m_0 = h_0 \quad \text{for} \quad n = 0$$

and the set of N equations,

$$H_K m_n + \sum_{p=1}^{N} \frac{p\alpha_n}{2n} (s_{|n-p|} - s_{n+p}) m_p + \frac{s_0}{2} \alpha_n m_n = h_n ,$$
$$n = 1, 2, \cdots , N. \qquad (10)$$

These N equations constitute a set of linear simultaneous equations in the N unknown coefficients m_n. These equations may be expressed,

$$\sum_{p=1} c_{np} m_p = h_n , \qquad n = 1, 2, \cdots , N,$$

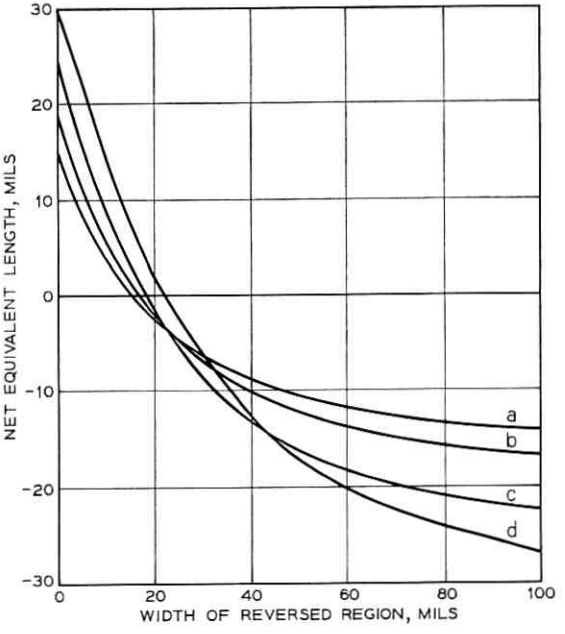


Fig. 5 — Change in net equivalent length of film (proportional to output signal during NDRO), versus width of reversed domain established beneath drive strap. Curves a, b refer to strap half width of 1.0 mils, and strap to film axis distances of 3.5 and 5.0 mils, respectively. Curves c and d refer to strap half width of 10.0 mils and strap to film axis distances of 5.0 and 10.0 mils, respectively. The curves are derived from the axial distributions of Fig. 2.

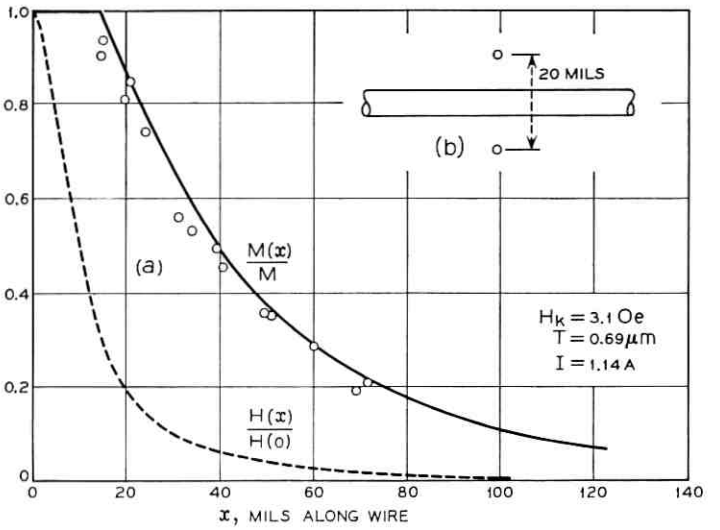


Fig. 6 — (a) Theoretical curve and experimental points taken with the Kerr effect probe[9] for a saturated cylindrical film. The broken curve shows the relative fall off of the axial applied field. (b) The field is applied by a parallel drive wire arrangement shown in cross section. The current I applied in the drive wires is 1.14 A.

where the c_{np} are given by

$$c_{np} = \left\{\frac{p\alpha_n}{2n}\left(s_{|n-p|} - s_{n+p}\right) + \left(\frac{s_0}{2}\alpha_n + H_K\right)\delta_n^p\right\}.$$

Such a set of equations may be conveniently inverted by computer for any particular case giving the m_n coefficients in terms of the h_n's. Since the m_n and h_n coefficients are linearly related, a scale factor, e.g., current in drive strap, is applied to $H(x)$ to ensure that the distribution has a value M at $x = \pm R$. The resulting series indicates a non-uniform distribution for $M(x)$ within the range $-R \leqq x \leqq R$, but, by the action of $S(x)$, this produces no demagnetizing field and therefore does not influence the distribution obtained outside the range. The value of $M(x)$ is therefore set equal to M inside the saturation range. The plot obtained within this range reflects instead the value of $(H - H_m)/H_K$.

Fig. 6(a) shows a plot of the axial magnetization distribution where the film has saturated over a length of 30 mils, for a cylindrical film of 5.2 mil diameter, 0.69μm thickness and $H_K = 3.1$ Oe. The broken

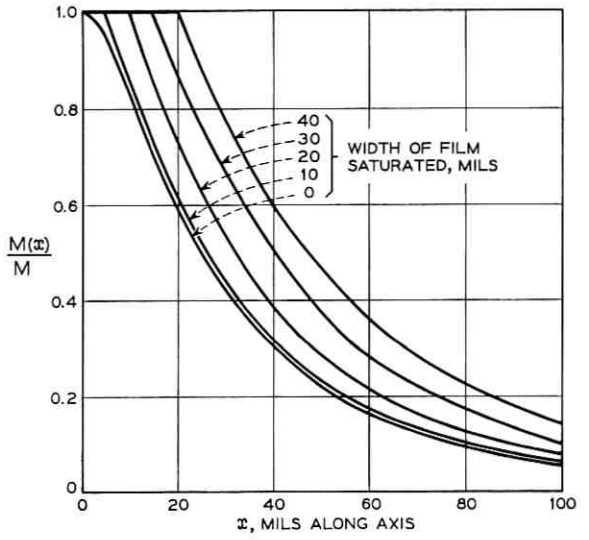


Fig. 7 — Axial component of magnetization for the cylindrical film of Fig. 6 when driven to different degrees of saturation.

curve of Fig. 6(a) shows a normalized plot of the applied field. The field is applied by a drive wire, and the separation between drive and return wire is 20 mils as shown in 6(b). The calculation indicates a current of 1.14 amps to produce this degree of saturation. The points represent measurements made previously[6] using the Kerr Effect probe.

Fig. 7 shows the axial magnetization component for the geometry of Fig. 6 where the film has saturated to widths of 0, 10, 20, 30, 40 mils. The applied field is shown in Fig. 8, curve a, versus width of saturated region produced by the field. Curve b is for a drive strap of half width 10 mils and strap to film axis distance of 10 mils. The shape of the curve does not appear to vary markedly with drive strap geometry. It can be noted that little increase in current is required to extend the saturated region from 1 to 10 mils, but that saturation to greater widths requires increasingly larger currents.

V. FILM THICKNESS VARIATION

Now let $T(x)$ be the variable film thickness and assume that $T(x)$ and $H(x)$ have the same periodic distance λ, then we may write

$$T(x) = \sum_{n=0}^{\infty} t_n \cos(2\pi n x/\lambda).$$

In the thin film approximation, magnetization variations within the thickness of the film are neglected and demagnetizing fields are calculated from the net pole density per unit area of film. To take into account a variation in thickness we take the product $T(x)M(x)$ as the total magnetization component in the hard direction and evidently the pole density is then given by $-(d/dx)\,[T(x)M(x)]$.

Taking the product of the series, we obtain

$$T(x)M(x) = \frac{M}{2}\left\{t_0 m_0 + \sum_{p=0}^{N} t_p m_p + \sum_{n=1}^{N}\sum_{p=0}^{N} m_p[(t_{n+p} + t_{|n-p|}) + t_0\delta_n^p]\cos 2\pi nx/\lambda\right\},$$

hence, replacing $M(x)$ by $T(x)M(x)$ in (3), the demagnetizing field is given by

$$H_m(x) = \sum_{n=0}^{N}\frac{\alpha_n}{2}\sum_{p=0}^{N} m_p(t_{n+p} + t_{|n-p|} + t_0\delta_n^p)\cos 2\pi nx/\lambda, \qquad (11)$$

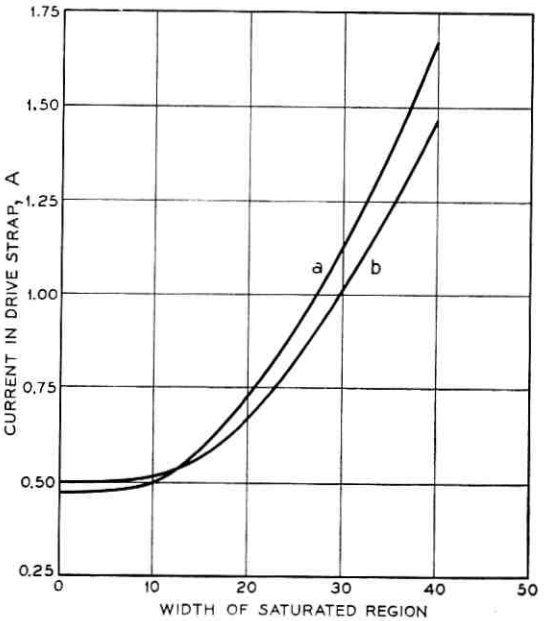


Fig. 8—Current required to produce a given width of saturated region along a cylindrical film of radius 2.6 mils, thickness 0.69μm, H_K = 3.1 Oe. Curve a is for the arrangement of Fig. 6. Curve b is for a parallel conductor drive strap of width 20 mils situated at $\pm$10 mils from the film axis.

where $\delta_n^p = 1$ when $n = p$ but is otherwise zero. Substituting into the torque equation (6), and equating coefficients we have finally

$$H_K m_0 = h_0$$

and

$$\sum_{p=0}^{N} \left[\frac{\alpha_n}{2} (t_{n+p} + t_{|n-p|} + t_0 \delta_n^p) + H_K \delta_n^p \right] m_p = h_n , \qquad n = 1, 2, \cdots, N$$

i.e.,

$$\sum_{p=1}^{N} \left[\frac{\alpha_n}{2} (t_{n+p} + t_{|n-p|} + t_0 \delta_n^p) + H_K \delta_n^p \right] m_p = h_n - \alpha_n t_n h_0 / H_K . \qquad (12)$$

This last expression represents a set of linear simultaneous equations which may be solved numerically to give the coefficients m_n in terms of t_n and h_n. The calculation, when applied to the case of a flat film strip having an ellipsoidal cross section along the hard direction, subject to a uniform field, predicts a uniform demagnetizing field of magnitude very close to that indicated by the tables of Osborne[9] based on the solution of Maxwell's equation for the general ellipsoid. Fig. 9 shows the magnetization distribution near an edge of a uniform thickness (0.22 μm) flat film with $H_K = 2.62$ Oe. The points represent data taken with the Kerr effect probe.

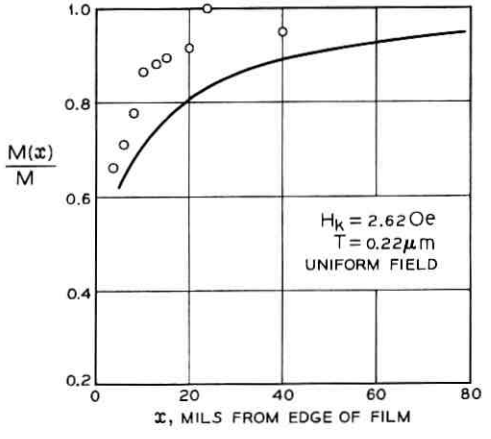


Fig. 9 — Magnetization component near the edge of a flat film of thickness 0.22μm, and $H_K = 2.62$ Oe. The applied field is uniform and equal to H_K. The edge runs parallel to the film easy direction. The points show measurements taken with the Kerr effect probe.

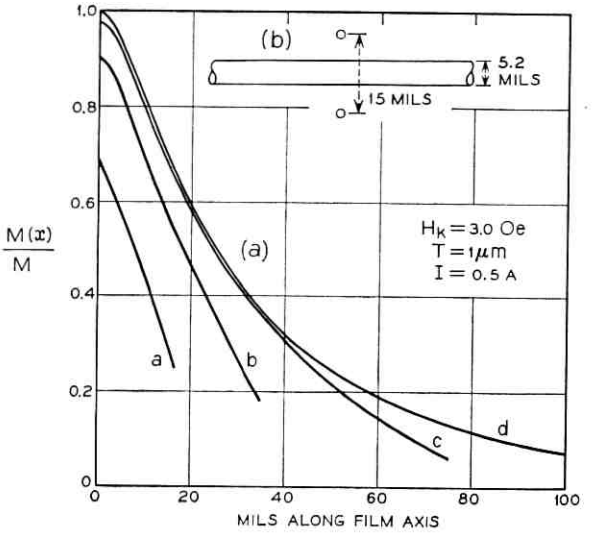


Fig. 10 — Axial magnetization component for cylindrical film segments of differing length due to the field from a parallel wire drive strap at distance ±7.5 mils from film axis. Curves a, b, and c refer to segments of length 40, 80, 160 mils, respectively. d refers to a continuous film. The current in the drive wire is 0.5 A. (b) shows a cross section of the drive wire arrangement.

Fig. 10 shows, for comparison the magnetization distribution for a nonsaturating hard direction field applied to 5.2 mil diameter cylindrical film segments of differing lengths, but uniform thickness of 0.7μm, and $H_K = 3.0$. The field is applied by a parallel drive wire arrangement of separation 15 mils. Finally, Fig. 11 shows the axial magnetization distribution for a uniform field applied to a cylindrical film having a circumferential cut. Film radius is 2.6 mils, thickness is 1.0 μm and $H_K = 3.0$ Oe. It is to be noted that the present technique has a spatial resolution limited both by the number of terms of the series that can be retained for computation, and by the basic limitation that exchange forces are neglected. We cannot, therefore, expect to obtain detail of magnetic behavior very close to an edge, for example, or for an extremely narrow scratch.

VI. ANISOTROPY MAGNITUDE VARIATION

Let us assume that the anisotropy constant is represented by a cosine series, i.e.,

$$K(x) = \sum k_n \cos 2\pi nx/\lambda.$$

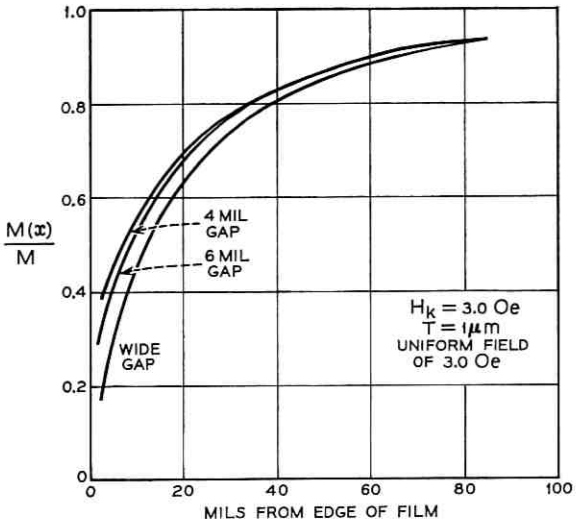


Fig. 11 — Plot of axial magnetization component for a 5.2 mil diameter cylindrical film with a circumferential gap. The curves show the result for a 4 mil, 6 mil and wide gap. The axial applied field is uniform and equal to 3.0 oe. Film thickness is 1.0μm and $H_K = 3.0$ Oe.

Then substituting into the torque equation (6), and gathering terms we find

$$\sum_{p=0}^{N} \frac{1}{M} (k_p + k_0 \delta_p^0) m_p = h_0 \tag{13}$$

and

$$\sum_{p=0}^{N} \left\{ \frac{1}{M} (k_{n+p} + k_{|n-p|} + k_0 \delta_p^n) + \alpha_p \delta_p^n \right\} m_p = h_n \,,$$

$$n = 1, 2, \cdots, N. \tag{14}$$

Together these equations represent $N + 1$ linear simultaneous equations in $N + 1$ unknown coefficients m_p, and may be solved by computer. This calculation may be used for example to find the local behavior of **M** at the junction between two regions with differing anisotropy constants, or to find the effective permeability of a film having some systematic variation in anisotropy constant. A simplified discussion of this latter problem has been given previously.[10] Fig. 12 shows the effect of using a high H_K buffer region surrounding a normal H_K section of film. Curve a shows the distribution for a uniform wire with $H_K = 3.0$, b shows the

modification when H_K is increased to a value $H_K = 15$ for all distances beyond $x = 10$ mils and c shows the result when H_K is further increased to 30 oe in the buffer region. The effectiveness of the high H_K buffer region in sharpening the distribution can be noted. This is achieved, however, at the expense of greater current required to just saturate at $x = 0$. For curves a, b, c the currents are 0.50, 0.79, and 0.93 A, respectively. Fig. 12(b) shows a cross section of the parallel conductor drive strap arrangement.

VII. FIELD EXTERNAL TO FILM

Combs and Wujek[11] have calculated the field external to a thin film rectangular slab assuming a pole distribution concentrated at the edges of the slab. We now calculate the field external to a continuous film subject to various applied field conditions where the details of the effective pole distribution form the essential part of the problem. The results of previous sections may be adapted to find the field external to films which have a hard axis variation in thickness or anisotropy

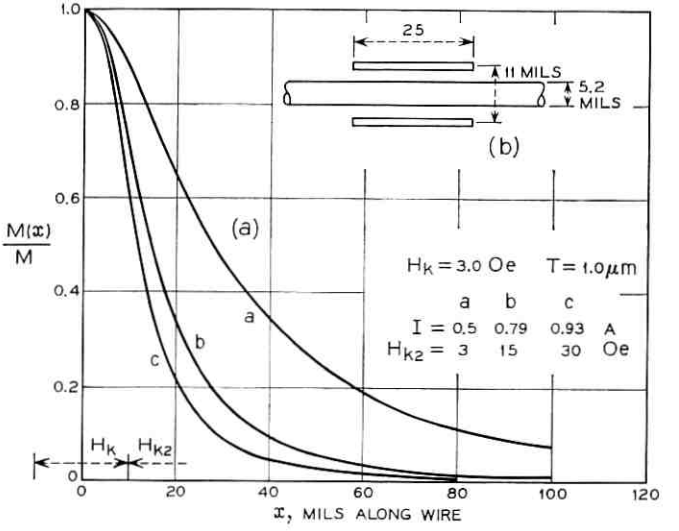


Fig. 12 — (a) Effect of high H_K buffer region surrounding a normal H_K section of cylindrical film. Curve a shows the magnetization component for a uniform film with $H_K = 3.0$. Curves b and c show the result when H_K is increased to 15 and 30 Oe, respectively for distances greater than 10 mils to either side of the drive strap centerline. (b) Details of drive strap arrangement. The currents required to just saturate the film at $x = 0$ are 0.5, 0.79, and 0.93 A for the cases a, b, and c, respectively.

but these cases are not considered in detail here. Consider the field at some distance d from the surface of a flat film and at distance x along the hard axis. The external field H_m parallel to the film due to the distribution of poles over the film surface may be found by evaluating the integral

$$H_{me}(x, d) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} -\frac{\frac{dM(x)}{dx}(x - X)\,dX\,dY\,T}{[d^2 + Y^2 + (X - x)^2]^{\frac{3}{2}}}. \tag{15}$$

Substituting for $M(x)$ and performing the integration we find

$$H_{me}(x, d) = -\frac{4\pi^2 MT}{\lambda}\sum_{n=1}^{\infty} n m_n e^{-2\pi dn/\lambda}\cos 2\pi nx/\lambda. \tag{16a}$$

This is the external field parallel to the plane of the film given as a function of distance d from the film. For a cylindrical film the result is

$$H_{me}(x, d) = -4\pi aTM\sum_{n=1}^{\infty}\left(\frac{2\pi n}{\lambda}\right)^2 K_0(2\pi nd/\lambda) \cdot I_0(2\pi na/\lambda)m_n\cos 2\pi nx/\lambda, \tag{16b}$$

where a is the cylinder radius, and d is the distance from cylinder axis to the location at which the axial component of field is measured, $(d > a)$. The field inside the cylinder may be similarly derived, the result is

$$H_{me}(x, d) = -4\pi aTM\sum_{n=1}^{\infty}\left(\frac{2\pi n}{\lambda}\right)^2 K_0(2\pi na/\lambda)I_0(2\pi nd/\lambda)m_n\cos 2\pi nx/\lambda,$$

where now $d < a$. Along the cylinder axis $I_0(0) = 1$. Fig. 13 shows a plot of the axial component of the demagnetizing field for several values of distance from film axis. The cylindrical film is assumed to have a diameter of 5.2 mils, $H_K = 3.0$ Oe, thickness is 1.0μm, and is excited by a one turn loop of radius 7.5 mils.

The flux coupling a parallel wire loop parallel to a flat film surface and to the film easy direction with the conductors at $\pm D$ from the surface may now be found. The flux F per unit length of the parallel conductor loop is then

$$F = 4\pi M(x)T - 2\int_0^D H_{me}(x, z)\,dz.$$

Substituting for H_{me} and rearranging, we find

$$F = 4\pi MT\sum_{n=0}^{N} m_n e^{-2\pi nD/\lambda}\cos 2\pi nx/\lambda. \tag{17}$$

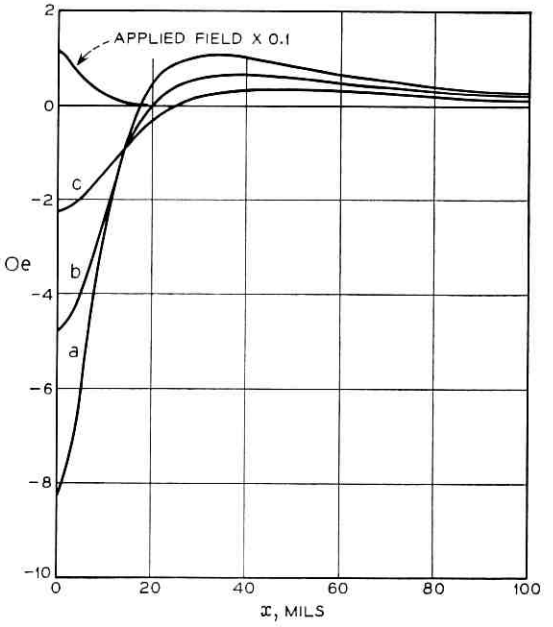


Fig. 13 — External axial component of field due to the distribution of magnetization along a cylindrical film. (The field due to the drive strap is not included.) The field is plotted along lines parallel to the film axis, at several distances from the axis. The film has a thickness of 1μm, $H_K = 3.0$ Oe, diameter 5.2 mils, and is subject to the field from a one turn circular loop of diameter 15 mils. Curves a, b, and c refer to distances of 2.6, 5.0, and 10 mils from the axis, respectively.

If the magnetization distribution is due to the field from a parallel wire loop with conductors at $\pm d$ from the film surface, then using expression (8a), we have

$$\frac{F(x)}{4\pi T} = \frac{\pi CIM}{\lambda H_K} + \frac{2\pi CIM}{\lambda} \sum \frac{e^{-2\pi n(D+d)} \cos 2\pi nx/\lambda}{H_K + \alpha_n}. \qquad (18a)$$

It can be noted that $F(x)/4\pi T$ is formally equivalent to the magnetization component in the film at the plane of the loop due to a current I in a loop with conductors at $\pm(D + d)$ from the film. The mutual inductance between two loops (not necessarily enclosing the film) may then be found directly from the above results.

The flux linkage between the film and drive loop is obtained by setting $x = 0$ and $D = d$. A current I in the loop gives rise to a magnetization component $M(0, I, d)$ at $x = 0$, and the flux linking the loop is

given by $M(0, I, 2d)$, using (18a). The fractional flux linkage is therefore $M(0, I, 2d)/M(0, I, d)$.

At $x = 0$, the expression (18a) may be evaluated in closed form; the result is,

$$\frac{F(0)}{4\pi T} = -\frac{2IC}{4\pi MT} \exp(2dH_K/4\pi MT)E_i(-2dH_K/4\pi MT).$$

Hence the fractional flux linkage (FFL) is

$$\text{FFL} = \exp(\mu d)E_i(-2\mu d)/E_i(-\mu d),$$

where $\mu = 2H_K/4\pi MT$ and E_i is the exponential integral. This is a useful parameter which shows the degree of coupling between loop and film, and is plotted in Fig. 14 as a function of d, for a flat film of thickness 0.1μm, $H_K = 4.0$ Oe.

The result for cylindrical films is more complicated. In this case it can be shown that

$$\frac{F(x)}{4\pi T} = \frac{2CI\pi M}{\lambda} \cdot \sum_1^\infty \frac{d\left(\frac{2\pi n}{\lambda}\right)^3 K_0\left(\frac{2\pi nD}{\lambda}\right) I_0\left(\frac{2\pi na}{\lambda}\right) K_1\left(\frac{2\pi nd}{\lambda}\right) I_0\left(\frac{2\pi na}{\lambda}\right) \cos\frac{2\pi nx}{\lambda}}{H_K + \frac{4\pi MT}{a}\left(\frac{2\pi na}{\lambda}\right)^2 K_0\left(\frac{2\pi na}{\lambda}\right) I_0\left(\frac{2\pi na}{\lambda}\right)}, \quad (18b)$$

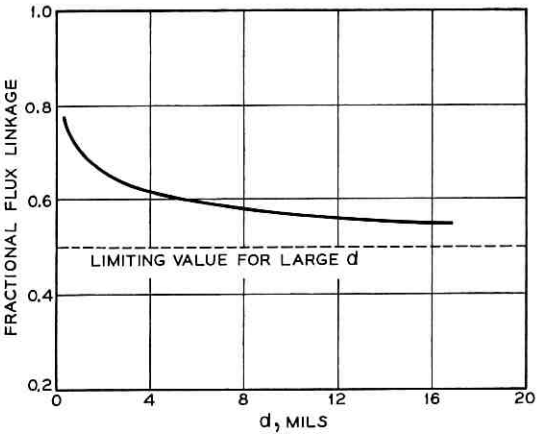


Fig. 14 — Fractional flux linkage between a flat film of thickness 0.1μm, H_K = 4.0 Oe, and a pair of parallel wire conductors as a function of distance from film to the conductors. The parallel wire conductors serve as both drive and sense windings.

where the cylinder has radius a, thickness T and is excited by the field from a circular loop of radius d. $F(x)$ gives the amount of flux picked up by a loop of radius D at an axial distance x from the drive loop.

VIII. INTERACTION BETWEEN PARALLEL FILMS

Consider two plane parallel films (denoted 1 and 2) of thickness T and T' and anisotropy fields H_K and H_K', respectively, separated by a distance w along a normal to the film's surface. A nonuniform field is applied along the (parallel) hard directions by a drive strap. Let the hard direction fields be $H(x)$ and $H'(x)$. The field acting on film 1 due to the distribution within film 2 we denote by $H(x)_{12}$, and similarly the field acting on 2 due to film 1 is $H(x)_{21}$. These fields are taken to act along the film's common hard direction, and the films are assumed to be sufficiently thin that fields normal to the surface have negligible effect.

The torque equation determining the local rotation of magnetization within the two films may be written

$$H_K \sin \theta(x) = H(x) + H_m(x) + H_{12}(x), \quad \text{film 1} \tag{19}$$

$$H_K' \sin \theta'(x) = H'(x) + H_m'(x) + H_{21}(x), \quad \text{film 2.} \tag{20}$$

Let $M(x)$, $M'(x)$ be the hard direction components of magnetization in the two films, then from previous sections we have (assuming cosine distributions)

$$H(x) = \sum h_n \cos 2\pi nx/\lambda$$

$$H'(x) = \sum h_n' \cos 2\pi nx/\lambda$$

$$H_m(x) = -\beta T \sum n m_n \cos 2\pi nx/\lambda, \quad H_m'(x) = -\beta T' \sum n m_n' \cos 2\pi nx/\lambda$$

$$H_{12}(x) = -\beta T' \sum n m_n' \exp(-2\pi nw/\lambda) \cos 2\pi nx/\lambda$$

$$H_{21}(x) = -\beta T \sum n m_n \exp(-2\pi nw/\lambda) \cos 2\pi nx/\lambda,$$

where $\beta = 4\pi^2 M/\lambda$. Noting that $\sin \theta(x) = M(x)/M$ and $\sin \theta'(x) = M'(x)/M$, we substitute the above series into the two torque equations and equating coefficients, we obtain,

$$\left.\begin{aligned} H_K m_n &= h_n - \beta n T m_n - \beta n T' m_n' \exp(-2\pi nw/\lambda) \\ H_K' m_n' &= h_n' - \beta n T' m_n' - \beta n T m_n \exp(-2\pi nw/\lambda) \end{aligned}\right\}.$$

Solving for m_n and m_n', we have finally

$$m_n = \left[h_n - \frac{\beta n T' h_n' \exp(-2\pi nw/\lambda)}{H_K' + \beta n T'}\right] \cdot \left[H_K + \beta n T - \frac{\beta^2 n^2 T T' \exp(-4\pi nw/\lambda)}{H_k' + \beta n T'}\right]^{-1} \tag{21}$$

$$m_n' = \left[h_n' - \frac{\beta n T h_n \exp(-2\pi n w/\lambda)}{H_K + \beta n T}\right] \cdot \left[H_K' + \beta n T' - \frac{\beta^2 n^2 T T' \exp(-4\pi n w/\lambda)}{H_K + \beta n T}\right]^{-1}. \quad (22)$$

These expressions can be compared with the results when the films are present singly, i.e., at large separations,

$$m_n = (h_n)(H_K + \beta n T)^{-1}$$
$$m_n' = (h_n')(H_K' + \beta n T')^{-1}.$$

Evidently the calculation can be extended to a greater number of layers and it is immaterial whether the drive fields are applied positively or negatively provided the fields are appropriately assigned, that is, the field may be generated by conductors located between or completely to one side of the films. The equations relating the coefficients m_n, m_n' may be concisely expressed in matrix form,

$$\left\{\begin{bmatrix} H_K & 0 \\ 0 & H_K' \end{bmatrix} - \beta n \begin{bmatrix} T & 0 \\ 0 & T' \end{bmatrix} - \beta n \exp(-2\pi n w/\lambda) \begin{bmatrix} 0 & T \\ T' & 0 \end{bmatrix}\right\} \begin{bmatrix} m_n \\ m_n' \end{bmatrix} = \begin{bmatrix} h_n \\ h_n' \end{bmatrix}. \quad (23)$$

The three matrix terms of the left-hand side represent in turn the effect of anisotropy, demagnetizing field, and interaction between films. The extension to three or more films is straightforward. Fig. 15 shows the effect of flux closure between two films only 2 mils apart subjected to the field from a drive wire sandwiched between them. The films have equal thickness of 0.1 μm and anisotropy field $H_K = 4.0$ Oe. Since the fields are applied in opposite directions in the two films the demagnetizing fields tend to cancel and the magnetization distribution widths are smaller than for similar films well spread apart. Curve a shows the coupled distribution, and b shows the distribution with one film removed. The current required to just saturate the films is 0.127 A, with one film removed the current required rises to 0.170 A. With films of thickness 1000 Å, separations of order a few mils are essential for this effect to be appreciable.

We may use the results (21) and (22) to examine the effect of a keeper layer. The action of the keeper is to modify the field applied to the film and to provide some degree of flux closure. Consider the case of a flat film situated between two drive wires, distance d from the film, with a keeper layer distance $w > d$ from the film. Let primed quantities refer to the keeper, and unprimed refer to the film. The keeper typically has a

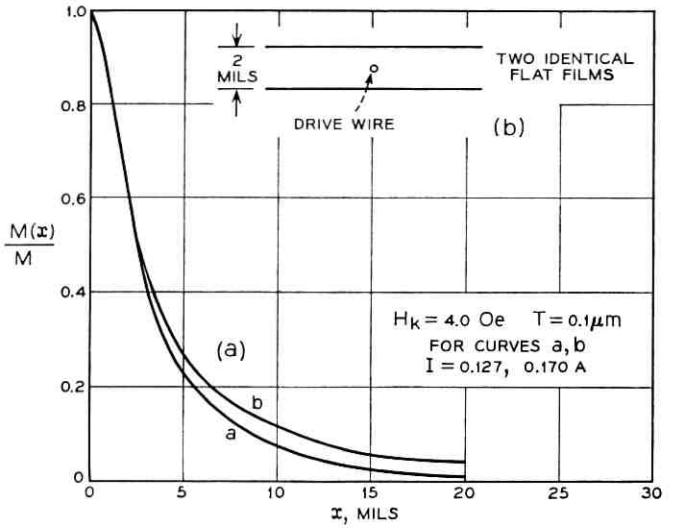


Fig. 15 — (a) Effect of flux closure between two identical flat films, separated by a distance of 2 mils. The field is applied by a single wire placed between the films as shown in (b). The films have a thickness of 0.1μm and $H_K = 4.0$ Oe. Curve b shows the result when one of the films is removed. The current required to just saturate the films at $x = 0$ now rises from the bifilm value 0.127 A to 0.170 A for a single film.

thickness of mils or tens of mils and hence $4\pi^2 MT'/\lambda \gg H_K'$ for reasonable values of M and λ. Equation (21) then reduces to,

$$m_n = [h_n - h_n' \exp(-2\pi nw/\lambda)]/[H_K + \beta nT(1 - \exp(-4\pi nw/\lambda)]. \quad (24)$$

The field applied to the film in the absence of the keeper is $H(x) = \sum h_n \cdot \cos 2\pi nx/\lambda$, where for the present case

$$h_0 = \frac{2CI\pi}{\lambda}, \qquad h_n = \frac{4CI\pi}{\lambda} \exp(-2\pi nd/\lambda).$$

I is the current in the drive wires. The field applied to the keeper is given by $\sum h_n' \cos 2\pi nx/\lambda$ where $h_0' = 0$,

$$h_n' = \frac{2CI\pi}{\lambda} \{\exp(-2\pi n(w + d)/\lambda) - \exp(-2\pi n(w - d)/\lambda)\}.$$

Then, $m_0 = 2CI\pi/\lambda H_K$, and

$$m_n = CI(2\pi/\lambda)[2 \exp(-2\pi nd/\lambda) - \exp(-2\pi n(2w + d)/\lambda) + \exp(-2\pi n(2w - d)/\lambda)]/[H_K + \beta nT(1 - \exp(-4\pi nw/\lambda))]. \quad (25)$$

It can be noted that the terms in the numerator are equivalent to the coefficients of the field due to the drive strap directly, and to images of the drive straps, with the keeper as mirror. The image property of the keeper layer is well known and has had considerable application to the discussion of keepers, see, for example, Refs. 12 and 13. The effect of the mutual interaction between keeper and film is to modify the α_n factors ($\alpha_n = \beta nT$ for a flat film) by a term $1 - \exp(-4\pi nw/\lambda)$. The influence of this term is two fold, (*i*) the spreading of the magnetization component is reduced and (*ii*) the drive field required is reduced.

Fig. 16 shows the effect of a keeper layer on the distribution in a flat film of thickness 0.2μm, $H_K = 4.0$ Oe. Field is supplied by a pair of drive straps of width 10 mils carrying a current of 0.22 A, at a distance of 5 mils from the film. The keeper layer is taken to be 6 mils from the film. Curve a shows the hard direction component in the absence of the keeper, b shows the effect only of the image fields due to the presence of the keeper, and c shows the final result when image fields and partial flux closure are taken into account.

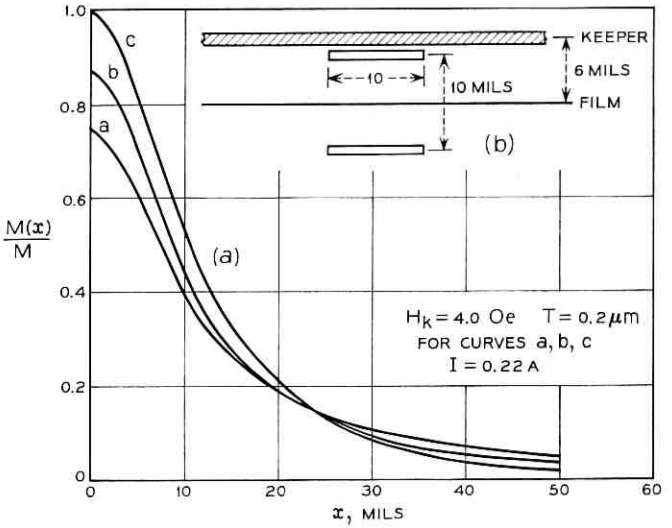


Fig. 16—(a) Effect of a keeper layer on the magnetization distribution in a flat film of thickness 0.2μm, $H_K = 4.0$ Oe. Field is applied by parallel drive straps of width 10 mils at ± 5 mils from the film. The keeper layer is taken to be 6 mils from the film as shown in (b). Curve a shows the hard direction component in the absence of the keeper, Curve b shows the effect of the image fields only when the keeper is present, and Curve c shows the final result when image fields and flux closure are taken into account.

The effects of a flat keeper layer on the response of a cylindrical film are not amenable to calculation by the present method owing to the mixed geometry.

The case of a cylindrical film with a concentric cylindrical keeper is next considered. The discussion closely parallels that for flat films and leads to a result analogous to (24),

$$m_n = \left[h_n - h_n' \frac{I_0\left(\frac{2\pi na}{\lambda}\right)}{I_0\left(\frac{2\pi nA}{\lambda}\right)} \right] \cdot \left[H_K + \alpha_n \left[1 - \frac{I_0\left(\frac{2\pi na}{\lambda}\right) K_0\left(\frac{2\pi nA}{\lambda}\right)}{I_0\left(\frac{2\pi nA}{\lambda}\right) K_0\left(\frac{2\pi na}{\lambda}\right)} \right] \right]^{-1} \tag{26}$$

where for cylindrical geometry $\alpha_n = 4\pi M(T/a)(2\pi na/\lambda)^2 I_0(2\pi na/\lambda) \cdot K_0(2\pi na/\lambda)$. The field is applied by a loop (of radius d) around the cylindrical film (of radius a), and h_n, h_n' are the Fourier coefficients of the field at the surface of the film and at the keeper (radius A), respectively. The axial field from a circular loop of radius d, at distance a from the axis and x from the plane of the loop, is given by[14,15]

$$H(x, a) = CI\left[K(k) + \frac{d^2 - a^2 - x^2}{(d - a)^2 + x^2} E(k) \right] / [(a + d)^2 + x^2]^{\frac{1}{2}},$$

where K and E are complete elliptic integrals of the first and second kinds, respectively, and $k^2 = 4da/[(a + d)^2 + x^2]$.

It can be noted that the effect of the keeper is to modify the applied field and to reduce the demagnetizing field. Fig. 17 shows a practical approximation to such a keeper geometry. Fig. 18 shows a plot of axial magnetization component in a 1μm thick permalloy film with H_K = 3.0 Oe plated on a 5.2 mil diameter wire, subject to the field from a one turn circular loop of diameter 7.5 mils carrying a current 0.3 amps.

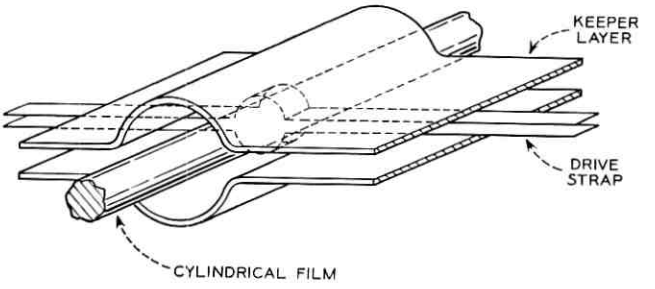


Fig. 17 — A possible practical approximation to a cylindrical keeper geometry.

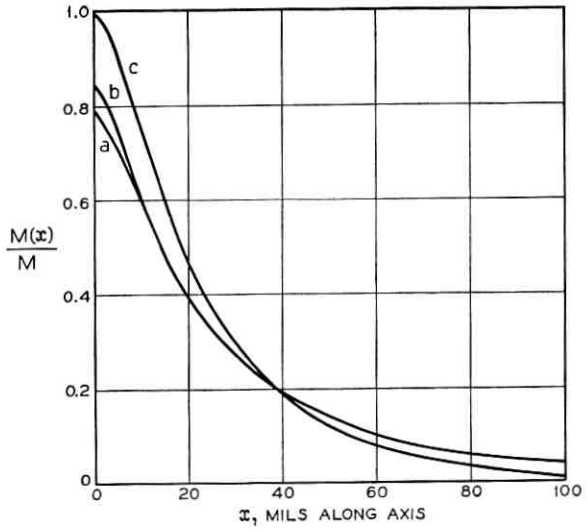


Fig. 18 — Effect of a cylindrical keeper layer on the axial magnetization distribution in a cylindrical film of thickness 1.0μm, $H_K = 3.0$ oe, diameter 5.2 mils. Field is applied by a one turn loop of radius 7.5 mils. Keeper radius is taken to be 10 mils. Curve a shows the distribution with no keeper present, curve b shows the effect of the keeper in modifying the applied field, and curve c shows the final result when field modification and flux closure are taken into account.

The keeper radius is taken to be 10 mils. Curve a shows the distribution with no keeper present, b shows the effect of field modification alone when a keeper cylinder of diameter 20 mils is in place, and c shows the final result when field modification and flux return are taken into account.

IX. NONUNIFORM HARD DIRECTION FIELD IN PRESENCE OF EASY DIRECTION BIAS FIELD

In this case the torque equation has to be modified to include the easy direction field $H_E(x)$, then

$$2K \sin \theta(x) \cos \theta(x) = M(H(x) - H_m(x)) \cos \theta(x) - MH_E(x) \sin \theta(x). \quad (27)$$

Providing $\cos \theta \neq 0$, we may write,

$$H_K \sin \theta(x) = H(x) - H_m(x) - H_E(x) \tan \theta(x), \quad (28)$$

where $H_K = 2K/M$ and it is assumed that H_E is parallel to the easy direction component of magnetization. It is convenient to represent $H_E(x) \tan \theta(x)$ by a series

$$H_E(x) \tan \theta(x) = \sum_{n=0}^{N} d_n \cos 2\pi n x/\pi.$$

Substituting into the torque equation (28), and gathering coefficients, we have

$$(H_K + \alpha_n)m_n = h_n - d_n, \qquad n = 0, 1, 2, \cdots, N.$$

The coefficients d_n are now complicated functions of the m_n's and this equation cannot be solved directly. Instead we use an iterative procedure as follows: $H(x)$ is given a peak value insufficient to produce saturation in the case $H_E = 0$ and then successive approximations are found for the m_n coefficients. In the first approximation we take

$$m_n = \frac{h_n}{H_K + \alpha_n}.$$

$\tan \theta(x)$ may now be found from $\sin \theta(x) = M(x)/M$, and the Fourier coefficients d_n of the product $H_E(x) \tan \theta(x)$, may be obtained. In the next approximation, we take

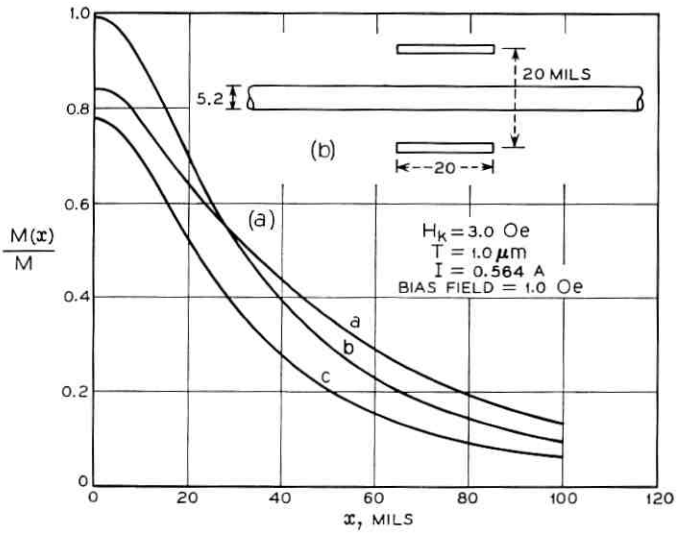


Fig. 19 — (a) Axial magnetization component for a cylindrical film with uniform easy direction bias field of 1.0 oe. The nonuniform hard direction field is applied by the drive strap arrangement shown in (b). In curve a, the bias field aids the rotation of magnetization for large x. A reverse domain is assumed to have been written into a width 20 mils, for $x < 10$ mils therefore the bias field opposes the rotation of magnetization. Curve b corresponds to zero bias field. Curve c corresponds to a reversal of bias field where it is assumed that the reversed region has been erased.

$$m_n = \frac{h_n - d_n}{H_K + \alpha_n}.$$

We now find, as before, new coefficients d_n ; hence, new coefficients m_n , until the m_n coefficients change by less than, say 5 percent per iteration. The curves for $H(x)$ and $M(x)$ are then plotted. The whole procedure may be repeated as necessary. The bias field may be a constant H_E or be a step function changing from H_E to $-H_E$ at some location $x = R$. The step function corresponds to the case of a domain wall being present at $x = R$. The use of the step function provides a formal way of treating the modification to the torque equation, due to H_E and the easy component of **M** being parallel for $x < R$, and antiparallel for $x > R$.

It is to be noted that the torque balance becomes unstable for certain combinations of applied fields. The critical fields are related by $[H(x) - H_m(x)]^{\frac{2}{3}} + H_E^{\frac{2}{3}} = H_K^{\frac{2}{3}}$, where it is assumed that H_E is antiparallel to the easy direction component of M. This limitation does not apply when H_E and the easy direction component of **M** are parallel.

Fig. 19 shows a typical axial magnetization distribution for a cylindrical film, and corresponds to the procedure of "writing" into a region of film. A current in the plated wire produces a uniform easy direction bias field of 1.0 oe and an external drive strap produces a nonuniform hard direction field. The greater spread of the curve *a* compared with the zero bias field distribution [shown by curve b] is due to the bias field lowering the effective anisotropy to $H_K - H_E$ for rotations less than about 40°. The attempt to "erase" by reversing the bias field, curve c, raises the apparent anisotropy to $H_K + H_E$ over much of the curve, and hence the film response is generally reduced. In curve c it is assumed that the reversed region has been erased. It will be appreciated that the present calculation assumes at the outset that a domain wall has some given location. The resulting distribution must then be inspected to decide whether the location chosen was appropriate or even stable under the applied field. In a practical case, wall location is affected by additional factors such as dispersion and creep, and is not discussed further here. Experiments on flat films show that the reversed region is not totally erased by simple reversal of bias field. Fig. 20(a) is a Kerr effect picture showing a reverse domain of width 20 mils, written in by a bias field of 1 Oe and a peak drive field of 5.0 Oe (11 mil strap, 10 mils from film). Fig. 20(b), shows the result of reapplying the fields with reversed bias. Fig. 20(c) shows the result of first demagnetizing the film into a fine domain structure, the width of the domain established is now much wider. In this case, the effect of the bias field changing the

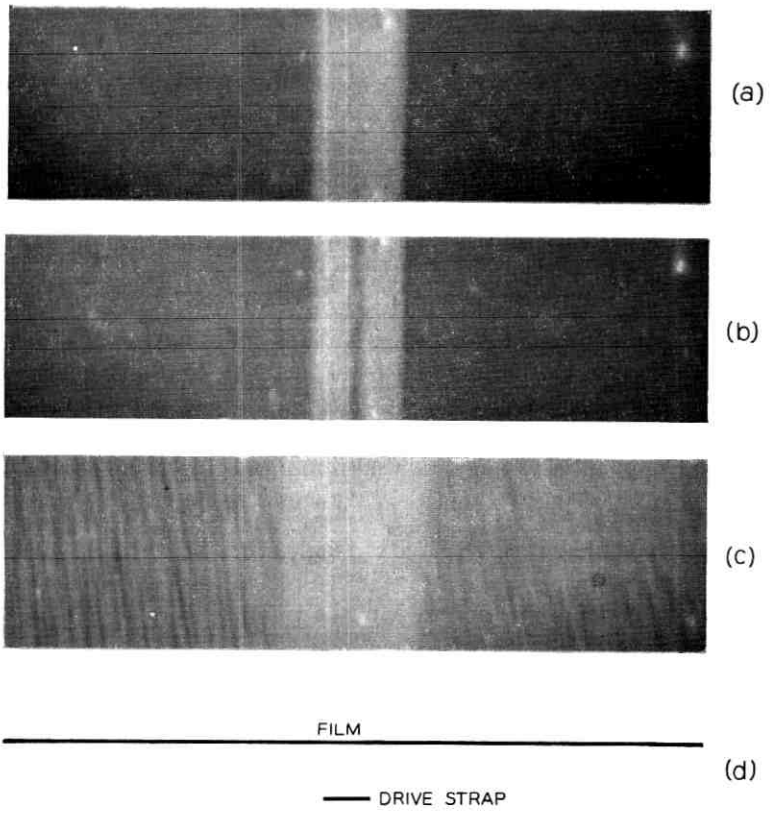


Fig. 20 — (a) Kerr effect picture showing reversed domain (light) in a flat film written in by an 11 mil drive strap situated 10 mils beneath the film. (b) When bias field is reversed, the domain is not completely erased. (c) Width of domain written after first demagnetizing film with a large uniform hard axis field. (d) Shows the drive strap arrangement to the same scale.

apparent anisotropy is much reduced, but the film now has an appreciable remanent state; hence, significant hard direction local demagnetizing fields exist in addition to the field introduced by the effect of the external fields. The relevance of such considerations to domain wall creep processes, under practical operating conditions, warrants further study but is not pursued here.

X. CONCLUSION

Demagnetizing fields play an important role in the operation of many thin magnetic film devices. The requirement of high packing density as in a memory, leads to strong localization of induced changes

in magnetization, and to correspondingly large demagnetizing fields and drive currents.

In an open flux structure attempts to confine magnetization changes by using segmented films or high anisotropy buffer regions are successful only at the expense of a considerable increase in drive field requirement. To some extent flux keeper layers may be used to modify applied fields and to permit partial flux closure, with in consequence, both a lowering of drive currents and a reduced spread in induced magnetization component.

The method of calculation given here permits a detailed examination to be made of the effectiveness of such procedures, and has been applied to a variety of thin film demagnetizing field problems. Kerr effect probe measurements[6] are in good agreement with calculation although relatively little data is at present available. The results have particular applicability to cylindrical film problems, where axial variation of field or properties is of primary concern.

XI. ACKNOWLEDGMENT

Acknowledgments are due to Mrs. L. Calamia for valuable assistance with the computer programs.

REFERENCES

1. Leaver, K. D. and Prutton, M., The Effect of Applied Field Inhomogeneity on the Reversal Behavior of Thin Magnetic Films, J. Elect. Control, *15*, 1963, pp. 173–181.
2. Edwards, J. G., Apparent Anisotropy Field of Continuous Magnetic Films Subjected to Inhomogeneous Drive Fields, Nature, *7*, 1963, pp. 130–134.
3. Rosenberg, R., unpublished work.
4. Kump, H. J. and Greene, T. G., Magnetization of Uniaxial Cylindrical Thin Films, IBM J., *7*, 1963, p. 130.
5. Kump, H. J., Demagnetization of Flat Unixial Thin Films Under Hard Direction Drive, IBM J., *9*, 1965, p. 118.
6. Dove, D. B. and Long, T. R., Magnetization Distribution in Flat and Cylindrical Films Subject to Nonuniform Hard Direction Fields, IEEE Trans. on Magnetics, *2*, 1966, pp. 194–197.
7. Gradshteyn, I. S. and Ryzhik, I. M., *Tables of Integrals, Series and Products*, Editor, A. Jeffrey, Academic Press, New York, 1965.
8. Erdélyi, A., Editor, *Bateman Manuscript Project*, McGraw-Hill Book Co., New York, 1954.
9. Osborne, J., Demagnetizing Factors of the General Ellipsoid, Phys. Rev., *67*, 1945, p. 351–357.
10. Dove, D. B., Anisotropy Magnitude Dispersion in Thin Films, Electronics Letters, *2*, 1966, pp. 15–16.
11. Combs, C. A. and Wujek, J. H., On the Static Magnetic Field Associated with a Thin Film Ferromagnetic Slab, J. Franklin Institute, *277*, 1964, pp. 305–312.
12. Pohm, A. V., Heller, L. G., and Smay, T. A., Adjacent Element Coupling in Continuous Film Memories, IEEE Trans. on Magnetics, *2*, 1966, pp. 512–515.

13. Feltl, H. and Harloff, H. J., Flux Keepers in Magnetic Thin Film Memories, IEEE Trans. on Magnetics, *2*, 1966, pp. 516–520.
14. Stratton, J. A., *Electromagnetic Theory,* McGraw-Hill Book Co., New York, p. 263.
15. Nagaoko, N., Magnetic Field of Circular Currents, Phil. Mag. Series 6, *41*, 1920, pp. 377–388.

Some Properties and Limitations of Electronically Steerable Phased Array Antennas*

By D. VARON and G. I. ZYSMAN

(Manuscript received April 4, 1967)

This paper is a treatment on linear and planar phased arrays of current sources, whose amplitudes are uniform and scan-invariant. By recognition that the radiation impedance of an array element is an analytic function of a complex scan variable, a powerful mathematical tool becomes available for the investigation of some important properties of the impedance as a function of scan. For example, it is proven that in a finite array the impedance seen by such a scan-invariant current source cannot be perfectly matched over a continuous scanning range using lossless, linear, passive and time-invariant elements. This result is extended to the infinite-array case by treating the latter as a pericdic structure, and assuming that the Green's function of the unit cell is analytic with respect to the scan variable. The theory includes both linear and planar arrays. Among other results it is shown that the element impedance in an infinite array must be of a specific mathematical form. It is hoped that by recognizing the limitations imposed thereby, useful guidelines will be established for achieving optimal match of an array into space.

I. INTRODUCTION

The class of antennas widely known as phased arrays includes essentially two types of radiators: stationary and steerable ones. The first operates at fixed amplitude and fixed relative phase between the array elements. Consequently, the antenna characteristics, such as radiation pattern, input impedance, and mutual coupling between elements, remain unchanged during the entire operational lifetime of the antenna. The steerable antenna is characterized by time varying ex-

* This work was supported by the U. S. Army under contract DA-30-069-AMC-333(Y).

citation. The relative phase between adjacent elements is varied either mechanically or electronically to bring about a variation in the orientation of the beam. In most instances scanned arrays are large in size and may contain several thousand elements. Their illumination has a linear phase taper. As a result the antenna characteristics become scan dependent. The relationship between scan angle and various parameters of interest such as gain, element impedance, and mutual coupling between elements have been the subject of intense investigation in recent years.[1, 2] One particular direction has been towards improvement of the impedance match over wide scanning ranges.[3] At present the merit of a matching technique can be determined only relatively to other techniques. To the best of the authors' knowledge an absolute mathematical criterion, based on physical realizability requirements, has not been formulated. Some investigators[4, 5] claim that a perfect match of an infinite array for all scan angles (at which the active impedance is not infinite, zero or purely reactive) can be achieved by an infinite set of interconnecting network elements. However, the proof is based on the assumption that the scan-dependent equivalent load impedance at the array-space interface remains unchanged after the sources have been interconnected by coupling elements. Although this assumption has been successfully applied[3, 5] to improve the matching capability of an infinite array, it is incorrect to use it in a perfect matching scheme.

In this paper a new mathematical approach to phased array analysis is presented. The model for the analysis is a phased array of ideal current sources (electric or magnetic) of scan-invariant uniform amplitude. This model is further discussed in Section II. The analysis itself is based on the general laws of antenna theory and on those properties which are common to all phased arrays represented by the model.

The first part of the theory is devoted to finite arrays and is treated in Section III. The starting point of the theory is a theorem which establishes that the radiation impedance of an element in a finite array is an analytic function of the scan angle. Further, it is shown that an element in a linear or planar phased array cannot be perfectly matched over a continuous scanning range by using lossless, linear, passive and time-invariant elements. Then it is demonstrated that the directions in space of the beams' maxima are eigenvalues of a Laplacian differential operator with periodic boundary conditions which are related to the phase taper of the array, and several useful properties of those eigenvalues are derived.

The second part of the theory appears in Section IV and is devoted to infinite arrays, which play an important role in the analysis of large phased arrays. The investigation is based on a transformation between the scan angle and a complex variable $s = \alpha + j\beta$, which can be interpreted on $0 < \alpha \leqq 1$, $\beta = 0$ as the trigonometric sine function of the angle between the plane of the array and the direction in which a chosen grating lobe propagates. It is subsequently shown that the element impedance, as a function of s, is restricted to a specific mathematical form. Recognition of the limitations imposed thereby may provide new insight into the behavior of such arrays.

II. PRELIMINARY REMARKS

The model chosen for the following treatment is a linear or planar phased array excited by a set of ideal current generators of uniform amplitude and linear phase taper. The description *ideal* implies that the sources have no internal impedance and are invariant under any loading. This means that except for the relative phasing between contiguous generators the currents are scan independent. Frequently in antenna analysis induced currents are replaced by equivalent sources by application of the equivalence principle.[6] Such currents are not part of the sources. The induced currents are accounted for automatically by fulfillment of the requirement that the tangential component of the electric field has to vanish on all conductors. In general, the source-current amplitude in each element of the array may be a function of scan. However, this dependence is generally unknown and is often neglected in theoretical work. The types of excitations commonly used are the "free excitation" and "forced excitation".* The first assumes a generator with a scan-invariant internal impedance which is capable of delivering scan-invariant incident power. In the latter a constant terminal voltage or current is maintained. As pointed out by Oliner and Malech free excitation is easier to realize in high-frequency technology than forced excitation. The latter, however, is more tractable here. The results of this study remain valid for scan-dependent excitation as well, provided the current density of the source is a smoothly varying function of scan angle and can be analytically continued into a complex scan-angle plane.

Under the assumption that the array is excited by a uniform amplitude and a linear phase taper, the current density excitation function

---

* A. A. Oliner and R. G. Malech, Ref., 1, pp. 209–211.

of an M-element linear array (Fig. 1) is given by

$$\mathbf{J}(x, y, z, \psi) = \begin{cases} \mathbf{J}_0(x - ma, y, z)e^{jm\psi}, & ma \leqq x \leqq (m+1)a, \\ & m = 0, 1, \cdots, M-1, \\ 0 \quad \text{otherwise} \end{cases} \tag{1}$$

and that of an $M \times N$ element planar array of rectangular symmetry (Fig. 2) is given by

$$\mathbf{J}(x, y, z, \psi_x, \psi_y) = \begin{cases} \mathbf{J}_0(x - ma, y - nb, z)e^{j(m\psi_x + n\psi_y)} \\ \qquad ma \leqq x \leqq (m+1)a, \\ \qquad nb \leqq y \leqq (n+1)b, \\ \qquad m = 0, 1, 2, \cdots, M-1, \\ \qquad n = 0, 1, 2, \cdots, N-1, \\ 0 \quad \text{otherwise.} \end{cases} \tag{2}$$

The above currents can be either electric or magnetic the latter being regarded as equivalent to ideal electric voltage sources.

Note that the spherical coordinate systems in Fig. 1 and 2 differ from those commonly used in phased array analysis. The poles are located at endfire instead of broadside and the ranges of colatitude and azimuth are such that the upper hemisphere is spanned by $0 \leqq \theta \leqq \pi$, $0 \leqq \varphi < \pi$. This convention is chosen for reasons of mathematical convenience. The results derived in Section III are valid for linear as well as planar arrays. The inclusion of both cases in a single treatment is facilitated by a generalized notation for the current density excitation function. The steering phases $m\psi$ and $m\psi_x + n\psi_y$ are replaced by an equivalent "steering coefficient" $\sigma_{mn}(\varphi_{pq})$ in the plane of scan oriented at azimuth angle φ_{pq}. The steering coefficient

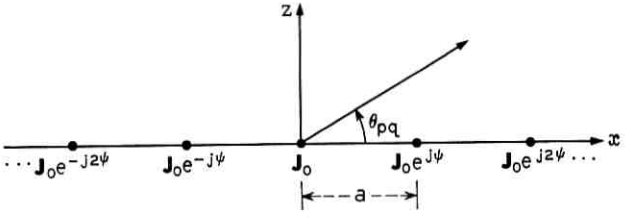


Fig. 1 — Linear phased array.

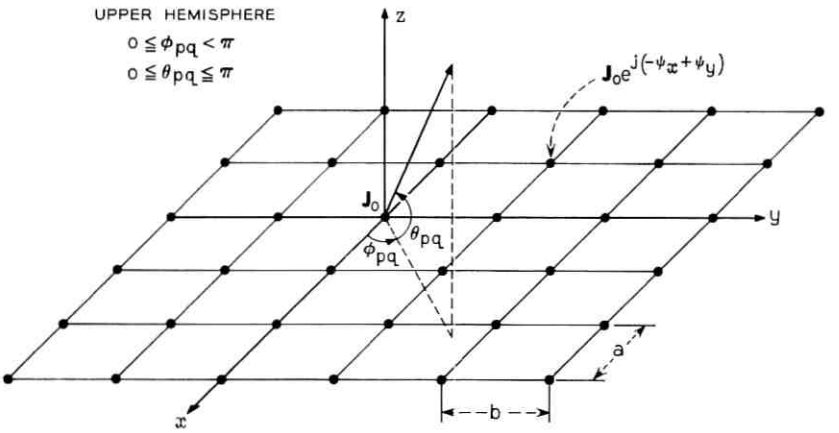


Fig. 2 — Planar phased array.

is derived by its relationship to the direction of a beam's maximum, which is determined for linear arrays by the equation

$$\psi + 2p\pi = ka \cos \theta_{p0} \qquad p = 0, \pm 1, \pm 2, \cdots \pm \infty \tag{3}$$

and for planar arrays by

$$\psi_x + 2p\pi = ka \cos \theta_{pq} \cos \varphi_{pq} \qquad p = 0, \pm 1, \pm 2, \cdots \pm \infty \tag{4a}$$

$$\psi_y + 2q\pi = kb \cos \theta_{pq} \sin \varphi_{pq} \qquad q = 0, \pm 1, \pm 2, \cdots \pm \infty, \tag{4b}$$

where k is the wave number in the medium, and θ_{pq} is as shown in Fig. 1 and 2. The steering coefficient is then defined by

$$\sigma_{mn}(\varphi_{pq}) = k(ma \cos \varphi_{pq} + nb \sin \varphi_{pq})$$

$$p, q = 0, \pm 1, \pm 2, \cdots \pm \infty. \tag{5}$$

Equations (1) and (2) can now be written as

$$\mathbf{J}(x, y, z, \theta_{pq}) = \begin{cases} J_0(x - ma, y - nb, z) \exp(j\sigma_{mn} \cos \theta_{pq}), \\ \qquad ma \leqq x \leqq (m+1)a, \\ \qquad nb \leqq y \leqq (n+1)b, \\ \qquad m = 0, 1, \cdots, M-1, \\ \qquad n = 0, 1, \cdots, N-1, \\ 0 \qquad \text{otherwise} \end{cases} \tag{6}$$

at $\varphi_{pq} = \text{const.}$

Under the above generalization the excitation function for the linear array becomes a special case, $q = 0$, $N = 1$, $\varphi_{pq} = 0$, and the period in the y-direction extends from $-\infty$ to $+\infty$; or alternatively $p = 0$, $M = 1$, $\varphi_{pq} = \pi/2$ and the period in the x-direction extending from $-\infty$ to $+\infty$. Since the phase constant $\exp\{j\sigma_{mn}(\varphi_{pq}) \cos \theta_{pq}\}$ is independent of (p, q), any θ_{pq} may be chosen as the independent variable of scan. The subscript pq will henceforth be omitted whenever the mathematical expressions are independent of (p, q).

The time dependence $e^{j\omega t}$ is assumed throughout the analysis. In a steerable array the phase taper is time dependent. However, it is understood that the rate of change of the phase taper is very small in comparison to the angular frequency, i.e., $d\psi/dt \ll \omega$, since only under that condition do the classical concepts of directivity and radiation impedance remain meaningful. If $\psi(t)$ is a step function it is assumed that the time interval is long enough to allow all transients to reach negligible values before a new step is initiated.

The formal solution of the array problem is obtained from Maxwell's Equations via a vector potential $\mathbf{A}(x, y, z, \theta)$ which is a solution of the inhomogeneous reduced wave equation

$$\nabla^2\mathbf{A} + k^2\mathbf{A} = -\mu\mathbf{J}(x, y, z, \theta), \tag{7}$$

where μ is the permeability of the medium. The magnetic field is given by

$$\mathbf{H} = \frac{1}{\mu} \nabla \times \mathbf{A}, \tag{8a}$$

and the electric field (under Lorentz gauge) by

$$\mathbf{E} = -j\omega\left(\mathbf{A} + \frac{1}{k^2} \nabla\nabla\cdot\mathbf{A}\right). \tag{8b}$$

The solution to (7) over infinite space V can be written in closed form in terms of a dyadic Green's function[7]

$$\mathbf{A}(x, y, z, \theta) = \mu \int_V \bar{\mathcal{G}}(x, y, z \mid \xi, \eta, \zeta)\cdot\mathbf{J}(\xi, \eta, \zeta, \theta)\, d\xi\, d\eta\, d\zeta, \tag{9}$$

where $\bar{\mathcal{G}}(x, y, z \mid \xi, \eta, \zeta)$ is a solution of

$$\frac{\partial^2\bar{\mathcal{G}}}{\partial x^2} + \frac{\partial^2\bar{\mathcal{G}}}{\partial y^2} + \frac{\partial^2\bar{\mathcal{G}}}{\partial z^2} + k^2\bar{\mathcal{G}} = -\bar{I}\,\delta(x - \xi)\,\delta(y - \eta)\,\delta(z - \zeta), \tag{10}$$

$\bar{I}$ being the unit dyadic $\mathbf{a}_x\mathbf{a}_x + \mathbf{a}_y\mathbf{a}_y + \mathbf{a}_z\mathbf{a}_z$. The boundary conditions which $\bar{\mathcal{G}}$ has to satisfy are derivable via the Vector Green's Theorem*

* P. M. Morse and H. Feshbach, Ref. 7, p. 1767.

by imposition of the requirement that the tangential component of the electric field has to vanish on all conductors. This guarantees that all induced currents are accurately determined.

It can be shown[8] that the average complex power delivered by the mnth element in the array is

$$P_{mn} = -\frac{1}{2}\int_{V_{mn}} \mathbf{E}\cdot\mathbf{J}^*_{mn}\, dv, \tag{11}$$

where

$$\mathbf{J}_{mn}(x, y, z, \theta) = \mathbf{J}(x, y, z, \theta),$$
$$ma \leqq x \leqq (m+1)a, \qquad nb \leqq y \leqq (n+1)b \tag{12}$$

the asterisk (*) denotes complex conjugate, and V_{mn} is a simply connected volume occupied by $\mathbf{J}_{mn}$. If S_{mn} is a surface obtained by taking a cross section through V_{mn}, the total current, I_{mn}, flowing through the cross section S_{mn} is

$$I_{mn} = \iint_{S_{mn}} \mathbf{J}\cdot\mathbf{ds}. \tag{13}$$

The element radiation impedance, Z_{mn}, is defined in terms of the complex power by

$$P_{mn} = \tfrac{1}{2}\,|\,I_{mn}\,|^2 Z_{mn}\,. \tag{14}$$

By (10) and (13) via (8b) and (9), the element radiation impedance can be defined directly in terms of the array geometry and the excitation:

$$Z_{mn}(\theta) = \frac{1}{|\,I_{mn}\,|^2}\int_{V_{mn}}\int_{V} \mathbf{J}^*_{mn}(x, y, z, \theta)\cdot\bar{\bar{G}}(x, y, z\,|\,\xi, \eta, \zeta)$$
$$\cdot\mathbf{J}(\xi, \eta, \zeta, \theta)\, d\tau\, dv, \tag{15a}$$

where $d\tau = d\xi d\eta d\zeta$, $dv = dxdydz$, and

$$\bar{\bar{G}}(x, y, z\,|\,\xi, \eta, \zeta) = j\omega\mu\left(\bar{\bar{I}} + \frac{1}{k^2}\nabla\nabla\right)\cdot\bar{\bar{G}}(x, y, z\,|\,\xi, \eta, \zeta). \tag{15b}$$

Operator ∇ operates on (x, y, z). The quantity $|I_{mn}|^2$ is introduced for the purpose of normalization, and may depend on the choice of the cross section S_{mn}.

The definition of the impedance includes both linear and planar array elements. It is consistent with the commonly known definition of impedance[9] if the latter is viewed as a relation between the average

complex power delivered by the generator and the rms current flowing into the load. The definition given by (15) is necessary in view of the fact that in a system excited by distributed currents, a terminal voltage in the time domain is not always uniquely defined. In a system excited by magnetic currents, (15) defines the element admittance if the permeability μ is replaced by the permittivity ϵ and the electric currents by their magnetic counterparts.

In the following theoretical discussion, it is assumed that the phased arrays are excited by a uniform amplitude and a linear phase taper.

III. FINITE ARRAYS

Theorem 1: The element radiation impedance in a finite, steerable, linear or planar phased array of scan-invariant current sources, radiating into a linear, lossless, passive and time-invariant system, is an entire function[10] *of the scan angle θ in any given plane of scan, with an essential singularity** *at $\theta \to \infty$.*

Proof: By (15a)

$$| I_{rs} |^2 Z_{rs} = \int_{V_{rs}} \int_{V} \mathbf{J}_{rs}^*(x, y, z, \theta) \cdot \bar{\bar{G}}(x, y, z \mid \xi, \eta, \zeta) \cdot \mathbf{J}(\xi, \eta, \zeta, \theta)\, d\tau\, dv. \tag{16}$$

On expanding (16) in a double sum of integrals over all cells $\{(m, n)\}$, $m = 0, 1, \cdots, M-1$; $n = 0, 1, \cdots, N-1$, and using the relationships of (6) followed by a change of variable in each term of the sum, one obtains

$$Z_{rs}(\theta) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \tilde{z}_{mnrs} \exp\,[j(\sigma_{mn} - \sigma_{rs}) \cos\theta], \tag{17a}$$

where

$$\tilde{z}_{mnrs} = \frac{1}{| I_{rs} |^2} \int_{V_{oo}} \int_{V_{oo}} \mathbf{J}_0^*(x, y, z) \cdot \bar{\bar{G}}(x + ra, y + sb, z \mid \xi + ma, \eta + nb, \zeta) \cdot \mathbf{J}_0(\xi, \eta, \zeta)\, d\tau\, dv. \tag{17b}$$

In any given plane of scan φ is constant, so that

$$\sigma_{mn} - \sigma_{rs} = k[(m - r)a \cos\varphi + (n - s)b \sin\varphi] = \sigma_{m-r,n-s} \tag{18}$$

is independent of θ. Both $\cos\theta$ and the exponential function are entire functions.† Consequently, the exponential function appearing in (17a)

* R. V. Churchill, Ref. 10, Sec. 68, p. 157; Sec. 112, p. 270.
† R. V. Churchill, Ref. 10, Sec. 21, p. 47; Sec. 23, p. 50.

is an entire function of an entire function, which is likewise entire[11] (entire functions are also called integral functions). $Z_{rs}(\theta)$ is a finite sum of entire functions and is also entire.

The nature of the essential singularity at $\theta \to \infty$ is obtained by first expanding *cos* θ in the complex θ-plane

$$\cos(\theta_r + j\theta_i) = \cos\theta_r \cosh\theta_i - j \sin\theta_r \sinh\theta_i . \tag{19}$$

Then, if $|\theta_i| \to \infty$ in such a way that $(\sigma_{mn} - \sigma_{rs})\theta_i > 0$, the *mn*th term behaves as $\exp\{|\sigma_{mn} - \sigma_{rs}| \sin\theta_r \exp[|\theta_i|]\}$ Q.E.D. Note that even when $\mathbf{J}_0(x, y, z, \theta)$ is scan dependent, $Z_{rs}(\theta)$ is analytic provided $\mathbf{J}_0(x, y, z, \theta)$ is analytic. However, other isolated singularities may exist.

Corollary 1a: Re $\{Z_{rs}\}$ *and* $\frac{\partial}{\partial\theta} Re\{Z_{rs}\}$ *are entire functions of* θ *each with an essential singularity at* $\theta \to \infty$. Proof appears in Appendix A.

Theorem 2: The power radiated by an element in a finite, steerable, linear or planar phased array of scan-invariant current sources, radiating into a lossless, linear, passive and time-invariant system cannot be kept constant over a continuous scanning range with lossless, linear, passive and time-invariant network elements and scatterers only.

Proof: Let $\bar{\bar{G}}(x, y, z \mid \xi, \eta, \zeta)$ be the dyadic Green's function of the entire system including all equalizing elements. The radiation impedance of the *mn*th element of the array is given by (15a) for a lossless, linear, passive, time-invariant system. If the array is radiating constant power over a continuous scanning range, the real part of the radiation impedance, $R_{rs}(\theta) = \mathrm{Re}\{Z_{rs}\}$, must remain constant in that range and

$$\frac{\partial}{\partial\theta}[R_{rs}(\theta)] = 0, \qquad \theta_1 \leqq \theta_r \leqq \theta_2 , \qquad \theta_i = 0 \tag{20}$$

where $\theta = \theta_r + j\theta_i$. By Corollary 1a, $\frac{\partial}{\partial\theta}[R_{rs}(\theta)]$ is analytic in the closed θ-plane and has an essential singularity at $\theta \to \infty$. However, if the derivative vanishes along the line $\theta_1 \leqq \theta_r \leqq \theta_2$ it must vanish everywhere in the θ-plane*. Hence, it cannot have an essential singularity at infinity. The contradiction implies that $R_{rs}(\theta)$ cannot be constant over a continuous scanning range. Q.E.D.

Equations (3) and (4) specify the directions of the beams' maxima, however, not all of them correspond to real directions in space. Whereas φ_{pq} is real for all (p, q), θ_{pq} can be either real or imaginary, as may be

* P. M. Morse and H. Feshbach, Ref. 7, Vol. I, p. 390.

seen from the solution of (4):

$$\varphi_{pq} = \tan^{-1} \frac{(\psi_y + 2q\pi)a}{(\psi_x + 2p\pi)b}, \qquad 0 \leqq \varphi_{pq} < \pi \quad (21a)$$

$$\theta_{pq} = \cos^{-1} \frac{\psi_x + 2p\pi}{ka \cos \varphi_{pq}} = \cos^{-1} \frac{\psi_y + 2q\pi}{kb \sin \varphi_{pq}}, \quad 0 \leqq \theta_{pq} \leqq \pi. \quad (21b)$$

If θ_{pq} is real it is said that the beam is in real space. By way of mathematical generalization it is said that all those beams having an imaginary θ_{pq} are in "imaginary space". If $\theta_{pq} = 0$, or $\theta_{pq} = \pi$, it is said that the beam is in a grazing position between real and imaginary space. It can easily be verified from (21) that for a given phasing (ψ_x, ψ_y) every pair (p, q) corresponds to a unique direction $(\varphi_{pq}, \theta_{pq})$ in the complex domain $0 \leqq \varphi < \pi$, $0 \leqq \mathrm{Re}\{\theta\} \leqq \pi$. These directions are the characteristic directions of the system. They are directly related, through (4), to the eigenvalues of

$$\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 F}{\partial y^2} + \Gamma^2 F(x, y) = 0 \quad (22)$$

with the following periodic boundary conditions

$$F(x, y) = F(x + a, y) \exp(-j\psi_x), \quad (23a)$$

$$\frac{\partial F}{\partial x}(x, y) = \frac{\partial F}{\partial x}(x + a, y) \exp(-j\psi_x), \quad (23b)$$

$$F(x, y) = F(x, y + b) \exp(-j\psi_y), \quad (23c)$$

$$\frac{\partial F}{\partial y}(x, y) = \frac{\partial F}{\partial y}(x, y + b) \exp(-j\psi_y). \quad (23d)$$

The eigenfunctions, which form a complete orthogonal set in the interval $0 \leqq x \leqq a$, $0 \leqq y \leqq b$ are

$$F_{pq}(x, y) = \exp\left[j(\psi_x + 2p\pi)\frac{x}{a}\right] \exp\left[j(\psi_y + 2q\pi)\frac{y}{b}\right], \quad p, q = 0, \pm 1, \pm 2, \cdots \pm \infty. \quad (24)$$

By (4) they can also be written as

$$F_{pq}(x, y) = \exp\{jk \cos \theta_{pq}(x \cos \varphi_{pq} + y \sin \varphi_{pq})\}. \quad (25)$$

The eigenvalues $\{\Gamma_{pq}\}$ are

$$\Gamma_{pq} = k \cos \theta_{pq} \qquad p, q = 0, \pm 1, \pm 2, \cdots \pm \infty. \quad (26)$$

The results thus derived lead to several interesting conclusions which are summarized in the following lemmas.

Lemma 1: Every steerable linear or planar phased array with a linear phase taper has only a finite number of beams in real space. Proof appears in Appendix *B*.

For every pair of phasing (ψ_x, ψ_y) there exists an infinite set of characteristic directions $\{\theta_{pq}, \varphi_{pq}\}$. As the array is scanned by varying the values of (ψ_x, ψ_y) in the intervals $-\pi \leqq \psi_x \leqq \pi$, $-\pi \leqq \psi_y \leqq \pi$ some characteristic directions will go through a grazing position going from imaginary to real space or vice versa. We shall call such characteristic directions "transitive characteristic directions".* Since the condition for a grazing position is $|\cos \theta_{pq}| = 1$, it follows from Lemma 1 that the number of transitive characteristic directions is finite.

Lemma 2: The radiation impedance of an element in a linear or planar phased array can be expanded by an infinite series over all characteristic directions of the system. Proof appears in Appendix C.

IV. INFINITE ARRAYS

In analyzing large arrays it has been found useful to approximate the behavior of the center elements by the behavior of identical elements in an infinite array of the same geometry.[12] This approximation is motivated by the fact that the performance of the center elements is strongly affected through mutual coupling by contiguous elements, but very weakly by elements far away.[13]

The formulation of the infinite array problem may be obtained from the results derived for finite-size arrays by letting the number of elements M and N approach infinity. The infinite array problem can also be treated as a periodic structure by application of Floquet's theorem. In the following, the latter approach is adopted, but first it is demonstrated that both methods are consistent.

The electric field of an infinite array as given by (8b) must satisfy the same periodicity conditions as the source function, i.e.,

$$\mathbf{E}(x + ma, y + nb, z) = \mathbf{E}(x, y, z) \exp [j\sigma_{mn}(\varphi) \cos \theta]. \quad (27)$$

* Note the distinction made between "grazing position" and "transitive characteristic direction". A beam associated with a transitive characteristic direction may attain a grazing position for a particular phasing, but may also point in other directions.

On the other hand, the electric field

$$\mathbf{E}(x, y, z, \theta) = -\int_V \bar{\bar{G}}(x, y, z \mid \xi, \eta, \zeta) \cdot \mathbf{J}(\xi, \eta, \zeta, \theta)\, d\tau \tag{28}$$

can be expanded in an infinite sum of integrals using the relationships of (6):

$$\mathbf{E}(x, y, z, \theta) = -\sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \exp(j\sigma_{mn} \cos\theta) \cdot \int_{V_{00}} \bar{\bar{G}}(x, y, z \mid \xi + ma, \eta + nb, \zeta) \cdot \mathbf{J}_0(\xi, \eta, \zeta)\, dv, \tag{29}$$

where V_{00} is the volume occupied by $\mathbf{J}_0$. Define a new Green's function

$$\bar{\bar{G}}_0(x, y, z \mid \xi, \eta, \zeta) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \exp(j\sigma_{mn} \cos\theta) \bar{\bar{G}}(x, y, z \mid \xi + ma, \eta + nb, \zeta) \tag{30}$$

and notice that

$$\bar{\bar{G}}_0(x, y, z \mid \xi + Ma, \eta + Nb, \zeta) = \exp(-j\sigma_{MN} \cos\theta) \bar{\bar{G}}_0(x, y, z \mid \xi, \eta, \zeta) \tag{31}$$

since by (5)

$$\sigma_{m+M, n+N} = \sigma_{mn} + \sigma_{MN}\,. \tag{32}$$

From (27) and (31) it follows that $\bar{\bar{G}}_0(x, y, z \mid \xi, \eta, \zeta)$ can be expanded by the eigenfunctions (25) as

$$\bar{\bar{G}}_0 = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \bar{g}_{pq}(z \mid \zeta) F_{pq}(x, y) F_{pq}^*(\xi, \eta), \tag{33a}$$

where

$$\bar{g}_{pq} = \frac{1}{ab} \int_0^a \int_0^b \bar{\bar{G}}_0 F_{pq}(\xi, \eta) F_{pq}^*(x, y)\, dx\, dy\, d\xi\, d\eta. \tag{33b}$$

Substituting (30) via (33a) into (15a) for the center element, $m = n = 0$, one obtains

$$Z_{00} = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} z_{pq}\,, \tag{34a}$$

where

$$z_{pq} = \frac{1}{|I_{00}|^2} \int_{V_{00}} \int_{V_{00}} \mathbf{J}_0^*(x, y, z) \cdot \bar{g}_{pq}(z \mid \zeta) F_{pq}^*(\xi, \eta) F_{pq}(x, y) \cdot \mathbf{J}_0(\xi, \eta, \zeta)\, d\tau\, dv. \tag{34b}$$

Equation (34) is an alternate representation to (86) for the radiation impedance of the infinite array element and it demonstrates that Lemma 2 is valid for infinite arrays as well.

By substituting the new representation for $\bar{\bar{G}}_0$, (30), (33), into (29) and noting that the electric field satisfies the homogeneous reduced wave equation in the source-free region, one obtains for the unbounded space

$$\mathbf{E}(x, y, z) = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \mathcal{E}_{pq} F_{pq}(x, y) \exp(-\gamma_{pq} | z |) \qquad | z | > d_{\max}, \quad (35a)$$

where

$$\gamma_{pq} = \sqrt{\Gamma_{pq}^2 - k^2} = jk \sin \theta_{pq} \quad (35b)$$

$$\mathcal{E}_{pq} e^{-\gamma_{pq}|z|} = -\int_{V_{00}} \bar{g}_{pq}(z \mid \zeta) \cdot \mathbf{J}_0(\xi, \eta, \zeta) F_{pq}^*(\xi, \eta)\, d\tau, \quad (35c)$$

and $d_{\max}$ is the projection on the z-axis of the largest distance between two points on the surface enclosing V_{00}. It can be seen that the electric field in the source-free region, above the central area of a large array may be approximated by a finite number of homogeneous plane waves propagating in the real characteristic directions, and an infinite number of nonhomogeneous plane waves, exhibiting exponential decay in the direction perpendicular to the plane of the array. The latter are interpreted as waves propagating in the imaginary characteristic directions.

In an infinite array all elements are embedded in an identical environment, and therefore the power radiated by each element is the same. There is no net power flow into a unit cell through the "side walls". Consequently, the quantity $\mathrm{Re}\{|I_{00}|^2 z_{pq}\}$ of (34b) is equal to the power propagated by the plane wave (p, q) within a unit cell in the direction perpendicular to the plane of the array. By Lemma 1 there is only a finite number of plane waves with transitive characteristic directions (see footnote p. 1571). Let them be distinguished from all other plane waves by assignment of the subscript $(p, q) = (\tau, \nu)$.

$$\mathbf{E}_{\tau\nu} = \mathcal{E}_{\tau\nu} F_{\tau\nu}(x, y) \exp(-jk | z | \sin \theta_{\tau\nu}), \qquad | z | > d_{\max} \quad (36)$$

$$\mathbf{H}_{\tau\nu} = \mathcal{H}_{\tau\nu} F_{\tau\nu}(x, y) \exp(-jk | z | \sin \theta_{\tau\nu}), \qquad | z | > d_{\max}, \quad (37)$$

where $F_{\tau\nu}(x, y)$ is given by (26), and $\mathcal{E}_{\tau\nu}$ by (35c). If

$$\mathfrak{D}_{\tau\nu} = jk[\cos \theta_{\tau\nu} \cos \varphi_{\tau\nu} \mathbf{a}_x + \cos \theta_{\tau\nu} \sin \varphi_{\tau\nu} \mathbf{a}_y - \sin \theta_{\tau\nu} \mathbf{a}_z] \quad (38)$$

then

$$\mathfrak{IC}_{\tau\nu} = \frac{j}{\omega\mu} \mathfrak{D}_{\tau\nu} \times \mathfrak{E}_{\tau\nu} . \tag{39}$$

The power radiated by a (τ, ν) plane wave per unit cell into the upper hemisphere is

$$P_{\tau\nu} = \frac{1}{2} \mathrm{Re} \int_0^a \int_0^b (\mathbf{E}_{\tau\nu} \times \mathbf{H}_{\tau\nu}^*) \cdot \mathbf{a}_z \, dx \, dy. \tag{40}$$

Substitution of (36) through (39) into (40) gives

$$P_{\tau\nu} = \frac{ab}{\eta_0} \sin \theta_{\tau\nu}^* \left[| \mathfrak{E}_{\tau\nu} \cdot \mathbf{a}_x |^2 + | \mathfrak{E}_{\tau\nu} \cdot \mathbf{a}_y |^2 + \frac{\sin \theta_{\tau\nu}}{\sin \theta_{\tau\nu}^*} | \mathfrak{E}_{\tau\nu} \cdot \mathbf{a}_z |^2 \right] \tag{41}$$

where $\eta_0 = (\mu/\epsilon)^{\frac{1}{2}}$. From (41) a radiation resistance per wave is defined as

$$R_{\tau\nu} = \frac{P_{\tau\nu}}{| I_{00} |^2}. \tag{42}$$

Since the entire system is passive and lossless, then by conservation of energy, the power $P_{\tau\nu}$ must originate from the element itself. Hence,

$$R_{\tau\nu} = \mathrm{Re} \{z_{\tau\nu}\}, \qquad \theta_{\tau\nu} \quad \text{real}, \tag{43}$$

where $z_{\tau\nu}$ is given by (34b).

From (41) it follows that when a wave (τ, ν) is in real space $R_{\tau\nu}$ is real, and when it is in imaginary space $R_{\tau\nu}$ is imaginary (in which case $\mathrm{Re}\{z_{\tau\nu}\} = 0$). Hence, of all the elements comprising the source's load, $R_{\tau\nu}$ appears either resistive or reactive, depending upon the scan angle. Such properties of a load, which are unknown in lumped network theory, are a consequence of the losslessness postulate. When propagation is possible power is carried away from the source. When propagation is inhibited there is no net loss of power and the load must be reactive. By Lemma 1 only $\mathrm{Re}\{z_{\tau\nu}\}$ has those properties. All other z_{pq}, $(p, q) \neq (\tau, \nu)$ and $\mathrm{Im}\{z_{\tau\nu}\}$ always retain their dissipative or reactive characteristics. Further, there is only a finite number of terms having $\mathrm{Re}\{z_{pq}\} > 0$. In practical phased arrays the spacing between the elements and the scanning range are such that only one such term exists at a time.

The following two definitions summarize the properties described above:

Definition 1: An O-type network function is a scan-dependent immittance (impedance or admittance) which is seen by the source as resis-

tive when the beam is in real space and as reactive when the beam is in imaginary space, and it behaves like an open circuit for impedance and like a short circuit for admittance in the grazing position.

Definition 2: An E-type network function is a scan-dependent immittance (impedance or admittance) which remains either resistive or reactive when the beam passes through the grazing position.

The motivation behind the nomenclature introduced by the two definitions will become clear later, in Theorems 4 and 5. The O-type and E-type immittances are of distinct mathematical form. To arrive at it consider first the following transformation:*

$$s = \sin \theta_{mn} \tag{44}$$

$$\chi = \cos \varphi_{mn} , \tag{45}$$

where (m, n) is one particular transitive characteristic direction out of all (τ, ν). Given s and χ all other characteristic directions are uniquely determined. By (4)

$$\psi_x = ka\chi\sqrt{1 - s^2} - 2m\pi \tag{46}$$

$$\psi_y = kb\sqrt{1 - \chi^2}\ \sqrt{1 - s^2} - 2n\pi, \tag{47}$$

where $(1 - \chi^2)^{\frac{1}{2}} \geqq 0$ for all possible χ and $(1 - s^2)^{\frac{1}{2}} > 0$ if $0 \leqq \theta_{mn} < \pi/2$, and $(1 - s^2)^{\frac{1}{2}} < 0$ if $\pi/2 < \theta_{mn} \leqq \pi$. Then by substitution of (47) into (22) all other characteristic directions are found:

$$\tan \varphi_{pq} = \frac{kab(1 - \chi^2)^{\frac{1}{2}}(1 - s^2)^{\frac{1}{2}} + 2(q - n)\pi a}{kab\chi(1 - s^2)^{\frac{1}{2}} + 2(p - m)\pi b} \tag{48a}$$

$$\cos \theta_{pq} = f_{pq}(s), \tag{48b}$$

where

$$f_{pq}(s) = \frac{ka\chi(1 - s^2)^{\frac{1}{2}} + 2(p - m)\pi}{ka \cos \varphi_{pq}}. \tag{48c}$$

This suggests that when characteristic direction (m, n) is scanned in a plane $\chi = \text{const}$, each of the components z_{pq} of the total input impedance as given by (34) can be expressed as a function of the same variable s. The conformal mapping between the θ_{mn}-plane and the s-plane is shown in Fig. 3. In view of the branch cut $-1 \leqq \alpha \leqq 1$ it will be understood that $s = \alpha$ denotes $s = \alpha - j0$ if $0 \leqq \theta_{mn} \leqq \pi/2$ and $s = \alpha + j0$ if $\pi/2 \leqq \theta_{mn} \leqq \pi$. Let $s = s_{\tau\nu}$ be the value at which characteristic direc-

* Recall that θ_{mn} and φ_{mn} are not in the conventional spherical coordinate system (see p. 1564).

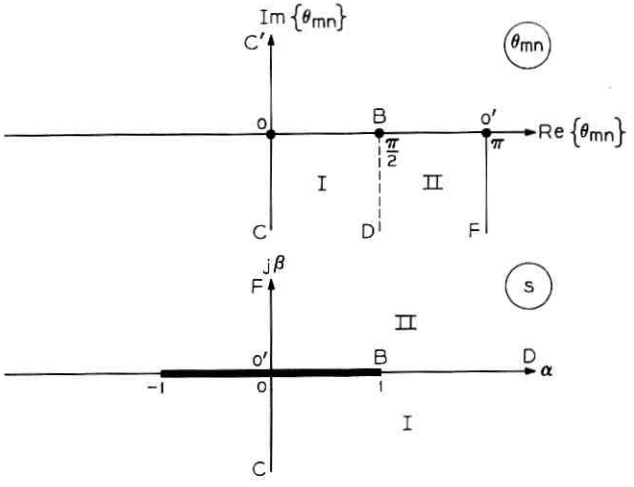


Fig. 3 — Conformal mapping $s = \sin \theta_{mn}$.

tion (τ, ν) is in grazing position. At this value

$$f^2_{\tau\nu}(s_{\tau\nu}) = 1. \tag{49}$$

Of all values $\{s_{\tau\nu}\}$ there is at least one which satisfies (49) for $s_{\tau\nu} = 0$. From (48) it is obvious that $f^2_{mn}(0) = 1$, and there may be other transitive characteristic directions $(\tau, \nu) \neq (m, n)$ which attain their grazing positions at $s_{\tau\nu} = 0$.

Theorem 3: In an obstacle-free space, the impedance function $z_{pq}(s)$, associated with the characteristic direction (p, q), is an analytic function of the complex variable $s = \alpha + j\beta$, with branch points at $s = s_{\tau\nu}$ and an essential singularity at $|s| \to \infty$. If $(p, q) = (m, n)$ then $z_{mn}(s)$ may have a simple pole at $s = s_{mn} = 0$.

Proof: The general definition of z_{pq} is given by (34b) in which the θ_{pq} dependence is contained in the Green's function component $\bar{g}_{pq}(z \mid \zeta)F^*_{pq}(\xi, \eta)F_{pq}(x, y)$. The Green's function is derived from (10) via (15b). Green's function $\bar{\mathcal{G}}(x, y, z \mid \xi, \eta, \zeta)$ satisfies the same periodic boundary conditions as $\bar{G}_0(x, y, z \mid \xi, \eta, \zeta)$ and can be expanded in a series similar to (33a):

$$\bar{\mathcal{G}} = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \bar{C}_{pq}(z \mid \zeta)F_{pq}(x, y)F^*_{pq}(\xi, \eta). \tag{50}$$

By substitution of (50) into (10) and use of the orthogonality property of $F_{pq}(x, y)$ one obtains a differential equation for $\bar{\bar{C}}_{pq}(z \mid \zeta)$

$$\frac{d^2\bar{\bar{C}}_{pq}}{dz^2} - \gamma_{pq}^2\bar{\bar{C}}_{pq}(z \mid \zeta) = -\bar{\bar{I}}\frac{\delta(z-\zeta)}{ab}$$
$$\gamma_{pq} = jk \sin \theta_{pq} \tag{51}$$

with the additional requirement that as $\mid z \mid \rightarrow \infty$, $\bar{\bar{C}}_{pq}$ behaves as an outgoing or evanescent wave. The solution of (51) for free space is

$$\bar{\bar{C}}_{pq}(z \mid \zeta) = \bar{\bar{I}}\frac{1}{2jabk \sin \theta_{pq}} \exp\{-jk \mid z - \zeta \mid \sin \theta_{pq}\} \tag{52}$$

$\bar{\bar{g}}_{pq}(z \mid \zeta)$ is obtained from $\bar{\bar{C}}_{pq}(z \mid \zeta)$ through an operator $\bar{\bar{\mathfrak{R}}}_{pq}$:

$$\bar{\bar{g}}_{pq}(z \mid \zeta) = j\omega\mu\bar{\bar{\mathfrak{R}}}_{pq}\cdot\bar{\bar{C}}_{pq}(z \mid \zeta), \tag{53a}$$

where

$$\bar{\bar{\mathfrak{R}}}_{pq} = \bar{\bar{I}} + \frac{1}{k^2}\mathfrak{D}_{pq}\mathfrak{D}_{pq} \tag{53b}$$

$\mathfrak{D}_{pq}$ being given by (38). Substitution of (52) into (53a) followed by substitution into (34b) gives

$$z_{pq} = \frac{\omega\mu}{2abk \mid I_{00} \mid^2 \sin \theta_{pq}} \int_{V_{00}}\int_{V_{00}} \mathbf{J}_0^*(x, y, z)$$
$$\cdot\bar{\bar{\mathfrak{R}}}_{pq}\cdot\mathbf{J}_0(\xi, \eta, \zeta) \exp\{jk \cos \theta_{pq}[(x - \xi) \cos \varphi_{pq}$$
$$+ (y - \eta) \sin \varphi_{pq}] - jk \mid z - \zeta \mid \sin \theta_{pq}\}\, d\tau\, dv. \tag{54}$$

The integrand is an entire function of θ_{pq} with an essential singularity at $\mid \text{Im}\{\theta_{pq}\} \mid \rightarrow \infty$. Hence,* if $\mathbf{J}_0(x, y, z)$ is piecewise continuous, the integral is also an entire function with the same essential singularity. By (48)

$$\sin \theta_{pq} = [1 - f_{pq}^2(s)]^{\frac{1}{2}}. \tag{55}$$

By Lemma 1,

$$f_{pq}^2(s) \not\equiv 1 \quad \text{if} \quad (p, q) \not\equiv (\tau, \nu). \tag{56}$$

From Fig. 3 it is readily seen that $\mid s \mid < \infty$ when $\mid \theta_{pq} \mid < \infty$ which implies, via (48), (55) that $\mid \cos \theta_{pq} \mid < \infty$ and $\mid \sin \theta_{pq} \mid < \infty$ as long as $\mid s \mid < \infty$. Thus, the singularities introduced by the transformation (44), (45) are the branch points at $s = s_{\tau\nu}$. Also if $(p, q) = (m, n)$†

* E. J. Copson, Ref. 11, Sec. 5.5, pp. 107–109.

† Recall that (m, n) is the characteristic direction which defines the transformation from (ψ_x, ψ_y) into (s, χ), (44)–(47).

z_{pq} may have a simple pole at $s = 0$. (Note, for example, that for horizontal polarization, $\mathbf{J}_0 = \mathbf{a}_x J_0$, there is a simple zero in the plane of scan corresponding to $\varphi_{mn} = 0$, at $s = 0$.) Q.E.D.

The above proof can be applied separately to the real and imaginary parts of the right-hand side of (54). If $z_{pq} = R_{pq}(s) + jX_{pq}(s)$, then $R_{pq}(s)$ and $X_{pq}(s)$ are analytic functions of s, real on the real axis of s, with an essential singularity at $| s | \to \infty$, branch points at $s = s_{\tau\nu}$, and possibly simple poles at $s = 0$.

In systems other than obstacle-free space, the normalized complex power $z_{pq}(s)$ has different forms. Except for isolated values of s, the radiated power and the stored energy per unit cell are bounded and continuous functions of s over those portions of the real and imaginary axes of the s-plane which have physical meaning. Hence, it is reasonable to postulate that an analytic continuation of z_{pq} as a function of scan can be made into a region of the complex s-plane which includes portions of both the real and imaginary axes. It may be of interest to note that the impedance function $z_{pq}(s)$ derived by L. Stark[14] for the planar dipole array over a ground plane is analytic. The regularity of $z_{pq}(s)$ depends directly on the regularity of $\bar{g}_{pq}(z \mid \zeta; s)$. The singularities of z_{pq} in the s-plane are determined by the boundary conditions which $\bar{g}_{pq}(z \mid \zeta; s)$ satisfies.

Theorem 4: An E-type immittance function $V(s)$ is an even function of s.

Proof: Let the complex variable s be defined with respect to the transitive characteristic direction (m, n). Once (m, n) is chosen, the proper branch of $(1 - s^2)^{\frac{1}{2}}$ in (48) is uniquely determined. Let (k, l) denote all other transitive characteristic directions which reach their transitive position simultaneously with (m, n). Formally, this implies

$$f^2_{\tau\nu}(0) = 1 \qquad (\tau, \nu) = (m, n), (k, l). \tag{57}$$

As a consequence of Definition 2 and Lemma 1, $V(s)$ is recognizable as

$$V(s) = \begin{cases} R_{pq}(s) & (p, q) \neq (m, n), (k, l), \\ X_{pq}(s) & \text{all } (p, q), \end{cases} \tag{58}$$

where $R_{pq}(s) + jX_{pq}(s) = z_{pq}(s)$, z_{pq} given by (54). Thus, (58) establishes the connection between the defined E-type function and physical quantities corresponding to $R_{pq}(s)$ and $X_{pq}(s)$. Consider Definition 2 which states

$$V(s) - V^*(s) = 0 \qquad s = \alpha \qquad 0 < \alpha < 1, \tag{59a}$$

$$V(s) - V^*(s) = 0 \qquad s = j\beta. \tag{59b}$$

Since $V(s)$ is analytic and also real on the real axis of s, (59) may be rewritten as*

$$V(s) - V(s^*) = 0 \qquad s = \alpha \qquad 0 < \alpha < 1, \tag{60a}$$

$$V(s) - V(s^*) = 0 \qquad s = j\beta. \tag{60b}$$

On the real axis

$$V(\alpha) - V(\alpha) = 0. \tag{61a}$$

On the imaginary axis

$$V(j\beta) - V(-j\beta) = 0. \tag{61b}$$

By analytic continuation† of (61b) from the imaginary axis to a point s in the complex plane one obtains

$$V(s) - V(-s) = 0. \tag{62}$$

Hence, $V(s)$ is an even function of s. Q.E.D.

Theorem 5: An O-type immittance $W(s)$ is an odd function of s. The proof is similar to that of Theorem 4 and it appears in Appendix D.

It has been shown in Theorem 2 that a finite phased array cannot be perfectly matched over a continuous scanning range. The proof is limited to finite arrays and cannot be directly extended to infinite arrays since the representation of the element impedance by (17a) does not guarantee convergence in the complex θ-plane if the limits of the summations are extended to infinity. In treating the infinite array, the element impedance is derived by symmetry considerations from which it is concluded that the net complex power radiated from each element is conserved entirely within the unit cell of that element. It has been shown that the two definitions are consistent. Although the problem of whether an infinite array can be perfectly matched is of academic interest only, it is worthwhile noting that as for finite arrays, the answer in this case is *also* negative. To show this the reader may recall that the impedance has been defined as normalized power and postulated to be an analytic function of the scan variable $s = \alpha + j\beta$. The normalization constant is $|I_{00}|^2$ given by (13). If the complex power as a function of scan is represented by

$$\hat{P}(s) = |I_{00}|^2 [R(s) + jX(s)], \tag{63}$$

* P. M. Morse and H. Feshbach, Ref. 7, Vol. I, p. 393.
† Morse and Feshbach, *Op. Cit.*, p. 389.

then by Lemma 1, the term $R(s)$ is a finite sum of analytic functions of the complex variable s. Consequently, $R(s)$ is an analytic function of s. In general, it may be represented as

$$R(s) = E(s) + \mathcal{O}(s), \tag{64}$$

where $E(s)$ is an even function of s and $\mathcal{O}(s)$ is an odd function of s. Under conditions of perfect match over a continuous range, constant power, P_r, is radiated over that range. Since $R(s)$ is analytic it implies $R(s) = P_r \mid I_{00} \mid^{-2}$ everywhere in the s-plane. Since a constant is even, $\mathcal{O}(s) = 0$. Further, $E(s)$ must have a branch cut on the real axis of the s-plane in the interval $[-1,1]$. But the branch cut does not exist if $E(s) = P_r \mid I_{00} \mid^{-2}$. The contradiction implies that $\hat{P}(s)$ in (63) cannot equal a constant over a continuous range of s.

Theorem 6: The resistance and reactance functions of an element, or their derivatives, in an infinite linear or planar phased array of current sources are discontinuous when a grating lobe is in a grazing position.

Proof: In an infinite array the grating lobes are plane waves propagating in the characteristic directions. By Theorems 4 and 5 the element impedance $Z(s)$ in an obstacle-free space can be written as

$$Z(s) = P(s) + \frac{Q(s)}{s}. \tag{65}$$

For real values of s, $P(s)$ is an even complex function of s bounded at $s = 0$, and $Q(s)$ is an even real function of s nonzero at $s = 0$. On the real axis of s

$$Z(\alpha) = P(\alpha) + \frac{Q(\alpha)}{\alpha}. \tag{66a}$$

On the imaginary axis of s

$$Z(j\beta) = P(j\beta) - j\frac{Q(j\beta)}{\beta}. \tag{66b}$$

A grating lobe is in its transitive position at $s = 0$. The pole discontinuities are established by showing that

$$\text{Re}\{\lim_{\alpha\to 0} Z(\alpha) - \lim_{\beta\to 0} Z(j\beta)\} = \lim_{\alpha\to 0}\frac{Q(\alpha)}{\alpha} = \infty \tag{67a}$$

$$\text{Im}\{\lim_{\alpha\to 0} Z(\alpha) - \lim_{\beta\to 0} Z(j\beta)\} = \lim_{\beta\to 0}\frac{Q(j\beta)}{\beta} = \infty. \tag{67b}$$

The pole discontinuity has to be interpreted as an invalid mathematical solution at the transitive position. It is a result of the idealization introduced by the concept of an "infinite array." If $R_{mn}(s)$ has a simple zero at $s = 0$, as is the case when a horizontally polarized array is placed above a ground plane, then the active impedance in the neighborhood of $s = 0$ can be written as

$$Z(s) = R(s) + jX(s), \tag{68a}$$

where $R(s)$ and $X(s)$ are real functions of s (real for s real).

$$R(s) = \sum_{i=0}^{\infty} a_i s^i \tag{68b}$$

$$X(s) = \sum_{i=0}^{\infty} b_{2i} s^{2i}. \tag{68c}$$

When the beam whose transitive characteristic direction is in real space, $s = \alpha$

$$R_\alpha \triangleq R(\alpha) = \sum_{i=0}^{\infty} a_i \alpha^i \tag{69a}$$

and when it is in imaginary space, $s = j\beta$

$$R_\beta \triangleq \text{Re}\,\{R(j\beta)\} = \sum_{i=0}^{\infty} (-1)^i a_{2i} \beta^{2i}. \tag{69b}$$

The discontinuity in the derivative of the resistance is

$$\lim_{\alpha \to 0} \frac{dR_\alpha}{d\alpha} - \lim_{\beta \to 0} \frac{dR_\beta}{d\beta} = a_1. \tag{70}$$

Similarly, the reactance

$$X_\alpha \triangleq X(\alpha) = \sum_{i=0}^{\infty} b_{2i} \alpha^{2i} \tag{71a}$$

$$X_\beta \triangleq \text{Im}\,\{Z(j\beta)\} = \sum_{i=0}^{\infty} (-1)^i [b_{2i} \beta^{2i} + a_{2i+1} \beta^{2i+1}] \tag{71b}$$

$$\lim_{\alpha \to 0} \frac{dX_\alpha}{d\alpha} - \lim_{\beta \to 0} \frac{dX_\beta}{d\beta} = -a_1 . \tag{72}$$

The proof can be generalized for any order algebraic singularity or zero at $s = 0$. For example, if there is a zero of multiplicity N the discontinuity will be in the Nth derivatives of the resistance and reactance. A noninteger order zero yields a discontinuity after a sufficient number of differentiations. Q.E.D.

V. SUMMARY AND CONCLUSIONS

A new mathematical approach to phased arrays has been adopted to investigate and discover various properties of the radiation impedance of an array element as a function of scan angle. The underlying idea of the method is the treatment of the impedance as an analytic function of a complex scan variable, which enables one to prove that an array element subject to the model chosen cannot be perfectly matched over a continuous scanning range by using lossless, linear, passive and time-invariant elements.

The first half of the theory is devoted to finite arrays. It is shown that the directions (in space) of the beams' maxima are eigenvalues of a Laplacian differential operator with periodic boundary conditions, which are related to the phase taper of the array. It is proven that there exists only a finite number of real eigenvalues. The known concept of imaginary space is then adopted to accommodate the imaginary eigenvalues. Furthermore, it is demonstrated that all beams except a finite number are completely confined either to real space or to imaginary space, and that only a finite number of beams may attain a grazing position. The unique properties of the latter beams have been found to play an important role in the investigation of infinite arrays, to which the second half of the theory is devoted.

The interest in infinite arrays, apart from its academic aspect, stems from the good approximation it provides for the behavior of the center portion of a large finite array. It has been found that the infinite array element impedance as a function of scan is restricted to a specific mathematical form. It is the authors' hope that recognition of the limitations imposed by that form may provide useful guidelines in achieving optimal match of an array to space.

VI. ACKNOWLEDGMENT

The authors wish to thank N. Amitay, E. R. Nagelberg, and R. G. Pecina for their critical reading of the manuscript and for valuable comments and suggestions regarding this work.

APPENDIX A

Proof of Corollary 1a

Corollary 1a: $Re\{Z_{rs}\}$ *and* $\frac{\partial}{\partial\theta} Re\{Z_{rs}\}$ *are entire functions of* θ *each with an essential singularity at* $\theta \to \infty$.

Proof: Denoting

$$\tilde{z}_{mnrs} = \rho_{mnrs} + j\chi_{mnrs} \tag{73}$$

one obtains from (17a)

$$\begin{aligned}\mathrm{Re}\{Z_{rs}(\theta)\} &\triangleq R_{rs}(\theta) \\ &= \sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\{\rho_{mnrs}\cos[B_{mnrs}(\theta)] - \chi_{mnrs}\sin[B_{mnrs}(\theta)]\},\end{aligned} \tag{74}$$

where

$$B_{mnrs}(\theta) = (\sigma_{mn} - \sigma_{rs})\cos\theta \tag{75}$$

and

$$\begin{aligned}\frac{\partial}{\partial\theta}R_{rs}(\theta) = -\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}B'_{mnrs}(\theta)\{&\rho_{mnrs}\sin[B_{mnrs}(\theta)] \\ &+ \chi_{mnrs}\cos[B_{mnrs}(\theta)]\}.\end{aligned} \tag{76}$$

Since $\cos\theta$ is an entire function of θ, $\cos[B_{mnrs}(\theta)]$ and $\sin[B_{mnrs}(\theta)]$ are entire functions of an entire function, and are therefore entire. The existence of the essential singularity can be demonstrated in a similar fashion to that in Theorem 1. Q.E.D.

APPENDIX B

Proof of Lemma 1

Lemma 1: Every steerable linear or planar phased array with a linear phase taper has only a finite number of beams in real space.

Proof: A beam (p, q) is in real space if $|\cos\theta_{pq}| \leqq 1$. Dividing (4a) by ka and (4b) by kb, squaring and adding, one obtains

$$\left(\frac{\psi_x + 2p\pi}{ka}\right)^2 + \left(\frac{\psi_y + 2q\pi}{kb}\right)^2 \leqq 1 \tag{77}$$

or

$$\left(\frac{\psi_x}{2\pi} + p\right)^2\left(\frac{\lambda}{a}\right)^2 + \left(\frac{\psi_y}{2\pi} + q\right)^2\left(\frac{\lambda}{b}\right)^2 \leqq 1. \tag{78}$$

Necessary conditions for the above inequality to be satisfied are

$$\left|\frac{\psi_x}{2\pi} + p\right|\frac{\lambda}{a} \leqq 1 \tag{79}$$

$$\left|\frac{\psi_y}{2\pi} + q\right|\frac{\lambda}{b} \leqq 1. \tag{80}$$

Since

$$-\frac{1}{2} \leqq \frac{\psi_x}{2\pi} \leqq \frac{1}{2} \quad \text{and} \quad -\frac{1}{2} \leqq \frac{\psi_y}{2\pi} \leqq \frac{1}{2},$$

$$| p | \leqq \frac{a}{\lambda} + \frac{1}{2} \tag{81}$$

$$| q | \leqq \frac{b}{\lambda} + \frac{1}{2}. \tag{82}$$

Hence, both p and q are bounded. Q.E.D.

APPENDIX C

Proof of Lemma 2

Lemma 2: The radiation impedance of an element in a linear or planar phased array can be expanded by an infinite series over all characteristic directions of the system.

Proof: The current density excitation function of a finite-size array given by (1), (2) satisfies the periodic boundary conditions (23) in the finite domain occupied by the array. Let this domain be denoted by D. The current density can, therefore, be uniquely expanded in D in terms of the eigenfunctions (25):

$$\mathbf{J}(x, y, z) = U(x, y, D) \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \mathbf{j}_{pq}(z) F_{pq}(x, y), \tag{83}$$

where

$$\mathbf{j}_{pq}(z) = \frac{1}{ab} \int_0^a \int_0^b \mathbf{J}_0(x, y, z) F_{pq}^*(x, y)\, dx\, dy \tag{84}$$

and $U(x, y, D)$ is a two-dimensional unit step function

$$U(x, y, D) = \begin{cases} 1 & (x, y) \quad \text{in} \quad D, \\ 0 & \text{otherwise.} \end{cases} \tag{85}$$

Substitution of (31a) into (15a) yields

$$Z_{mn} = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} \hat{z}_{mnpq}, \tag{86a}$$

where

$$\hat{z}_{mnpq} = \frac{1}{| I_{mn} |^2} \int_{V_{mn}} \int_{V} \mathbf{J}_{mn}^*(x, y, z) \cdot \bar{\bar{G}}(x, y, z \mid \xi, \eta, \zeta) \cdot \mathbf{j}_{pq}(\zeta) F_{pq}(\xi, \eta) U(\xi, \eta, D)\, d\tau\, dv. \tag{86b}$$

Q.E.D.

APPENDIX D

Proof of Theorem 5

Theorem 5: An O-type immittance $W(s)$ is an odd function of s.

Proof: Let the complex variable s be defined with respect to the transitive characteristic direction (m, n). Let (k, l) be all other transitive characteristic directions which reach their transitive position simultaneously with (m, n). Then as a consequence of Definition 1 and Lemma 1

$$W(s) = R_{pq}(s) \qquad (p, q) = (m, n), (k, l), \tag{87}$$

where $R_{pq}(s) = \mathrm{Re}\{z_{pq}\}$, z_{pq} given by (54). Thus, (87) establishes the connection between the defined O-type function and a physical quantity corresponding to $R_{pq}(s)$. From Definition 1

$$W(s) - W^*(s) = 0 \qquad s = \alpha, \qquad 0 < \alpha < 1 \tag{88}$$

$$W(s) + W^*(s) = 0 \qquad s = j\beta. \tag{89}$$

Since $W(s)$ is real on the real axis of s, (88), (89) may be rewritten as

$$W(s) - W(s^*) = 0 \qquad s = \alpha \tag{90}$$

$$W(s) + W(s^*) = 0 \qquad s = j\beta. \tag{91}$$

On the real axis

$$W(\alpha) - W(\alpha) = 0. \tag{92}$$

On the imaginary axis

$$W(j\beta) + W(-j\beta) = 0. \tag{93}$$

By analytic continuation of (93) from the imaginary axis to a point s in the complex plane one obtains

$$W(s) + W(-s) = 0. \tag{94}$$

Hence, $W(s)$ is an odd function of s. Q.E.D.

REFERENCES

1. Hansen, R. C., Editor, *Microwave Scanning Antennas,* Vol. II, Academic Press, 1964.
2. Wu, C. P., private communication.
3. Amitay, N., Improvement of Planar Array Match by Compensation through Contiguous Element Coupling, IEEE Trans. Antennas Prop., *AP-14,* No. 5, September, 1966, pp. 580–586.
4. Hannan, P. W., Proof that a Phased-Array Antenna Can Be Impedance

Matched for all Scan Angles, Radio Science, *2* (new series), No. 3, March, 1967, pp. 361–369.
5. Hannan, P. W., Lerner, D. S., and Knittel, G. H., Impedance Matching a Phased-Array Antenna Over Wide Scan Angles by Connecting Circuits, IEEE Trans. Antennas Prop., *AP-13,* No. 1, January, 1965, pp. 28–34.
6. Schelkunoff, S. A., Some Equivalence Theorems of Electromagnetics and their Application to Radiation Problems, B.S.T.J., *15,* January, 1936, pp. 92–112.
7. Morse, P. M., and Feshbach, H., *Methods of Theoretical Physics,* McGraw-Hill Book Co., Inc., 1953, Vol. II, p. 1769.
8. Stratton, J. A., *Electromagnetic Theory,* McGraw-Hill Book Co., Inc., 1941, pp. 132–133.
9. Guillemin, E. A., *Communication Networks,* John Wiley and Sons, Inc., 1935, Vol. II, pp. 474–476.
10. Churchill, R. V., *Complex Variables and Applications,* McGraw-Hill Book Co., Inc., 1960, Sec. 19, p. 40.
11. Copson, E. J., *Theory of Functions of a Complex Variable,* Oxford University Press, 1935, Sec. 7.1, p. 159.
12. Wheeler, H. A., The Radiation Resistance of an Antenna in an Infinite Array or Waveguide, Proc. IRE, *36,* April, 1948, pp. 478–488.
13. Galindo, V. and Wu, C. P., Asymptotic Behavior of the Coupling Coefficients for an Array of Thin-Walled Rectangular Waveguides, IEEE Trans. Antennas Prop., *AP-14,* No. 2, March, 1966, pp. 248–249.
14. Stark, L., Radiation Impedance of a Dipole in an Infinite Planar Phased Array, Radio Science, *1* (new series), No. 3, March, 1966, pp. 361–377.

An Energy-Density Antenna for Independent Measurement of the Electric and Magnetic Field

By W. C.-Y. LEE

An energy-density antenna which can measure both the E field and H field of a plane wave simultaneously has been developed, consisting of two small orthogonal semiloops over a ground plane. Hybrids were used to take the sum and difference of the loop outputs, giving voltages uniquely proportional to the E and H fields. The loop dimensions and optimum configurations were experimentally determined by measurements at a frequency of 836 MHz in a man-made free-space environment. Energy-density computation from the measured E and H fields of a standing wave in free space showed that the maximum-to-minimum range of the energy density is much less than that of either the E or H fields alone.

I. INTRODUCTION

A new way of reducing the signal fading encountered on a mobile radio transmission path is being investigated.[1] One source of fading is due to the fact that plane waves propagating in opposite directions at the same frequency produce a standing wave with nulls in the electric field every half free-space wavelength. The magnetic field also has nulls like the electric field but displaced a quarter wavelength from the electric field nulls. The electromagnetic energy density of such a pure standing wave is constant. If we sample E and H in free space and amplify the signals by the appropriate relative gains, square and add them, we obtain a signal proportional to electromagnetic energy density

$$w = \tfrac{1}{2}(\epsilon E^2 + \mu H^2). \tag{1}$$

The resulting output would be constant as we move through this idealized standing wave pattern. This method of energy-density utilization may be helpful in overcoming the rapid fading due to motion through the more complicated standing wave patterns in the mobile radio electromagnetic field. To utilize the energy concept, we need an

antenna that has three outputs independently proportional to the field components E_z, H_x, and H_y at any point in the field (assuming vertical polarization). Since neither the ordinary loop antenna nor the shielded loop antenna can be used in this particular case, an investigation was undertaken to develop a suitable antenna.

This paper describes a particular antenna* which satisfactorily meets these requirements. The antenna consists of two small orthogonal loops and will be described later. Measurements on such an antenna and several other comparable ones were made in a simulated free-space environment.

II. METHOD OF TESTING THE PROBES

First of all, we need a method of test which tells us how well the antenna is responding to the H field alone. As mentioned before, the nulls of the E and H field in an ideal standing wave pattern are $\lambda/4$ apart. Therefore, if we can establish such an ideal pattern, the E nulls can be located accurately by a whip antenna; then the positions of the H nulls are known. Then we can test the magnetic probe in this environment, looking for nulls at these H-null positions.

A conducting ground plane 16 feet × 3 feet was surrounded with commercially available absorbers (minimum absorption is 17 dB one-way) to provide a man-made free space. Two waves traveling in opposite directions were produced by exciting two identical transmitting antennas from a common source. These two transmitting antennas "S" and "N," approximately 12 λ apart, were $\lambda/4$ whip antennas operating at 836 MHz over the ground plane as shown in Fig. 1(a). The receiving antenna under test could slide in a slot about 2 λ long which is in between the two transmitting antennas. E fields were first tested separately from the two transmitting antennas in order to make sure that the reflections in the man-made free space were small, and that the individual fields were sensibly constant along the length of the slot. The two curves shown in Fig. 2 are the amplitudes of the signal from each of the transmitting antennas. The field from the "N" antenna had a maximum-to-minimum variation range of about 2.5 dB, and that from the "S" antenna a variation range of about 3.5 dB.

*A brief description of this antenna appears in two papers: (1) Theoretical and Experimental Study of the Properties of the Signal from an Energy Density Mobile Radio Antenna, presented at the IEEE Vehicular Communications Conference on December 2, 1966, in Montreal, Canada. (2) Statistical Analysis of the Level Crossings and Duration of Fades of the Signal from an Energy Density Mobile Radio Antenna, B.S.T.J., *46*, February, 1967, pp. 417–448.

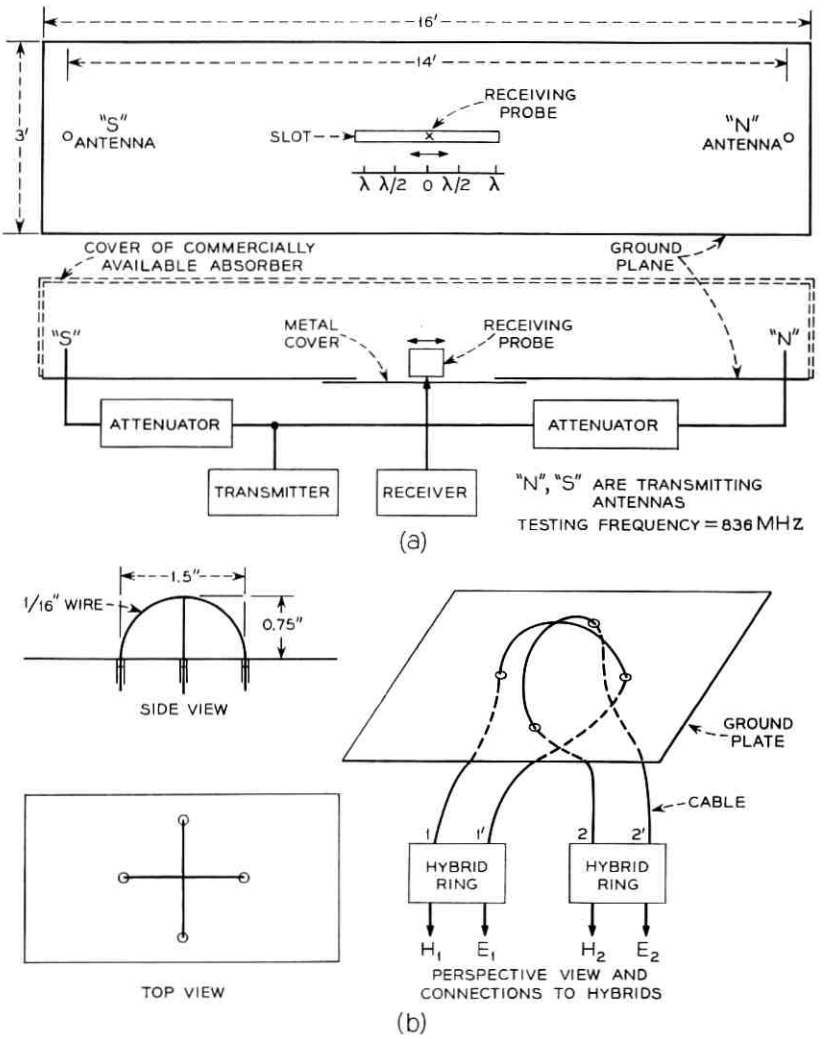


Fig. 1 — (a) Experimental set-up. (b) Energy-density antenna—double orthogonal loop antenna.

These variations, due to residual reflections, were felt to be acceptable. Since the average amplitudes of signal strength of two transmitting antennas were not quite the same, 11-dB attenuation was put on "*S*" antenna, and 10 dB on "*N*" antenna in order to get a good standing wave. The peak-to-null value of the standing wave produced when

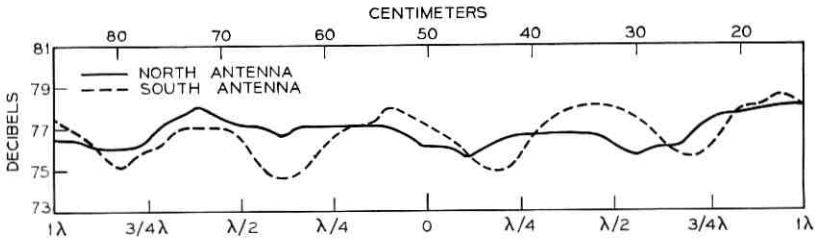


Fig. 2 — Amplitude of signal strength along the slot receiving from one transmitting antenna only.

both transmitting antennas were excited was then 23 dB, as shown in Fig. 3. We should remember that the measured standing wave was obtained from two E fields. Then we know a standing wave of the H field exists which will have the same peak-to-null value but a $\lambda/4$ shift from the standing wave of the E field.

III. TYPE OF ANTENNAS TESTED

3.1 *Single-Ended Loop*

A semiloop with one end grounded and the other end as output can be used as a magnetic field probe. However, the size of the loop is critical. Large errors are obtained in measuring the magnetic fields unless its diameter is less than 0.01 λ (about 0.14 inch diameter at 836 MHz).[2]

3.2 *Double-Ended Loop*

A semiloop with two output ends can be used as a combined electric and magnetic probe.[3] If the double-ended loop is in the field of a plane wave, the sum of the two outputs of the semiloop is proportional to the E field, and their difference to the H field. If the plane of the loop is in line with the direction of propagation the output is proportional to the total H field, otherwise only to a component of H field. This would be a limitation in using this type of probe for general purposes.

3.3 *Two Orthogonal Loops*

This antenna has been proposed for receiving a linearly polarized wave coming from a remote source which may not necessarily be in line with the plane of the loop. It consists of two double-ended loops with their planes perpendicular to each other. The "orthogonal loop

antenna" has two pairs of outputs. Adding two pairs of outputs separately gives two values which should be identical and expressed theoretically as proportional to the total E field. Subtracting two pairs of outputs separately gives the two components of the H field. These two components are the components along the rectangular coordinates which have been defined by the planes of the two loops. The orthogonal loop antenna is an electric and magnetic field probe which appears to be promising for probing the energy density of the total field. Hence, it is called the energy-density antenna.

3.3.1 *Connected Loops*

The two loops are electrically connected at the top point. Since this configuration can allow the two loops to be identical, the two values of E field obtained from the two loops are expected to be equal, the currents in the two loops are correspondent to the two components of the H field which are normal to the planes of two loops. However, the connection at the top points is not exactly at the middle, which may introduce some errors.

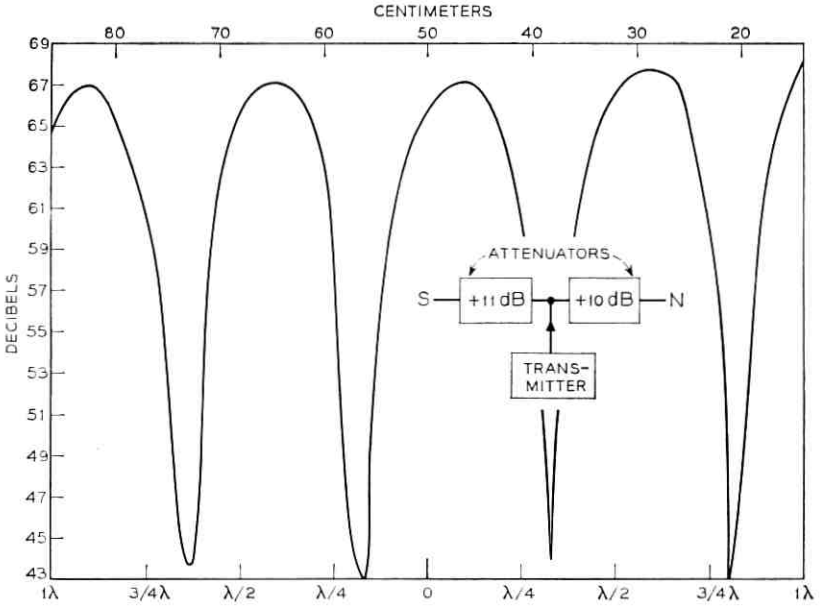


Fig. 3 — Standing wave along the slot by using a whip antenna as a receiving probe.

3.3.2 *Unconnected Loops*

The two loops are not connected electrically at the top points. In this configuration the two loops cannot be identical. One loop must be bent at the top in order to disengage the top point from that of the other loop. Therefore, the E field obtained from two loops may be different; also the two H component fields. However, the current in one loop may not be affected by the other due to the fact that the two loops are unconnected.

IV. EXPERIMENTAL RESULTS

4.1 *Single-Ended Loop*

The standing wave along the slot was measured by using different sizes of the single-ended loop. Investigation of three loops, 1, 1.5, and 2 inches in Fig. 4 shows that a 1.5-inch loop is better than the other two. The nulls of H field of the 1.5-inch loop are located more like the true H field though the amplitude of H field is 2 dB less than the 2-inch

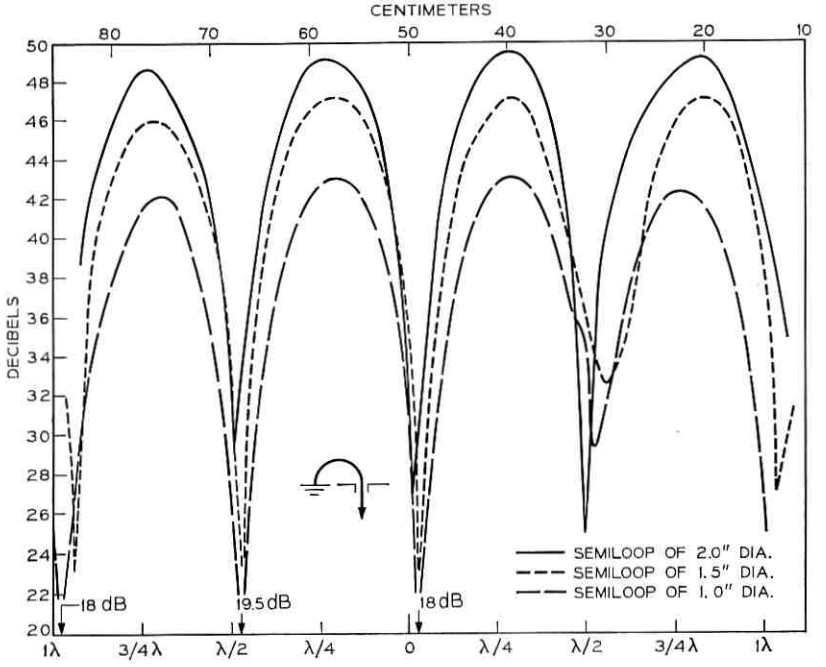


Fig. 4 — Standing wave of H fields along the slot by using a semiloop as a receiving probe (one end output).

loop. Comparing the 1.5-inch loop with the 1-inch loop, the amplitude of the 1.5-inch loop is 3 dB higher and the nulls are still located slightly better than in the 1-inch loop. Hence, the 1.5-inch loop is chosen even though it is 1/36 λ off the true H field (the standing wave of the true H field should be exactly a $\lambda/4$ shift from the E field). This was due to the effect of the electric field. The 1.5-inch loop is approximately 0.1 λ in diameter. This size of the loop was selected and used in the other types of loop configurations.

4.2 *Double-Ended Loop*

The standing wave along the slot was measured by using a semiloop as a receiving probe. The two outputs from the 1.5-inch semiloop were connected to a hybrid ring where the sum port gave the E field, and the difference port gave the H field. Since the plane of the loop was in line with the two transmitting antennas, the H field was a total H field. The E field and the H field outputs are shown in Fig. 5. The first null of the H field on the right had a slight disturbance which was probably due to the imperfect free space.

4.3 *Two Orthogonal Loops (unconnected)*

This probe consisted of two semiloops 1.5 inches in diameter. The size of the loop was chosen from Fig. 4. The circuit arrangement is shown in Fig. 1(b), except the top points of two loops were not connected.

4.3.1 *45° Orientation*

A double orthogonal semiloop was tested at an orientation of 45° to the line between the two transmitting antennas. Fig. 6 shows the two components of H field: H_1 and H_2. The two H components should be equal since the two loops were oriented 45° to the axis. However, the two loops, due to the fact they were roughly hand-made, were not precisely 45° to the axis. They were also not connected at the top points. So the fact that H_2 was higher from loop 2 than H_1 was from loop 1 was not a surprise. Fig. 6 also shows the E fields from the two loops, and we note that the nulls of the E field from loop 2 were lower than loop 1. The difference between the two loops was that loop 2 had more cross section area than loop 1.

4.3.2 *90° Orientation*

A double orthogonal semiloop was oriented at 90° to the line between the two transmitting antennas. In this case, H_2 should equal

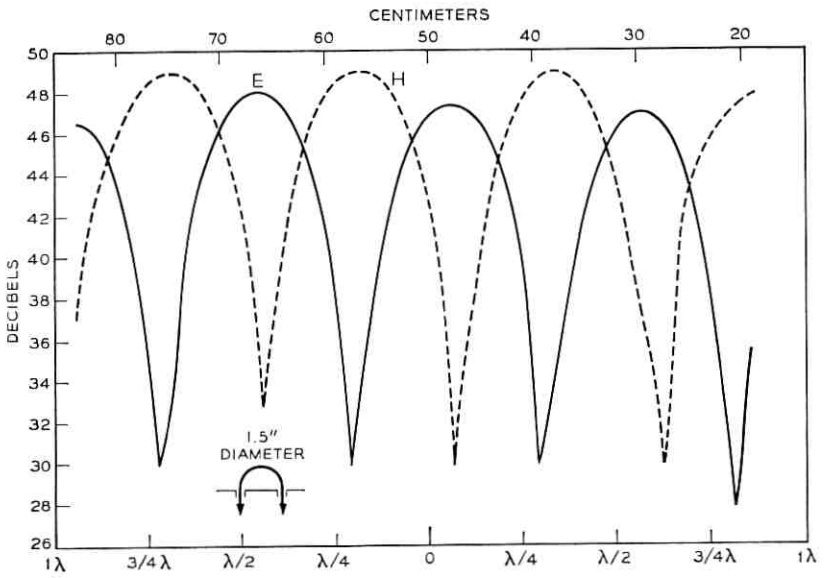


Fig. 5 — Standing waves of E and H fields along the slot by using a semiloop as a receiving probe.

the H field and H_1 should read zero output. From Fig. 7 we see that H_1 is 18 dB down compared with H_2, but has apparently picked up some E field since the peaks of H_1 are almost located at the nulls of H_2 and vice versa. H_2 in Fig. 7 is almost equal to the vector sum of the two components, H_1 and H_2, in Fig. 6 (45° case) as one would expect. E_1 and E_2 in Fig. 7 should be identical. They both represent the E field. In an ideal situation, E_1 and E_2 in Fig. 7 and in Fig. 6 should all be the same. Since the two loops were not connected at the top point, the maximum output from loop 1 was slightly lower than loop 2. Hence, the nulls of the four E's were not the same.

4.4 *Two Orthogonal Loops (connected) — Energy-Density Antenna*

The two orthogonal loops (1.5 inches in dia.) were connected at the top point of two loops, shown in Fig. 1(b).

4.4.1 *45° Orientation*

A double orthogonal semiloop was oriented at 45° to the two transmitting antennas. Fig. 8 shows the two components H_1 and H_2. Since the loops, due to the fact they were roughly hand-made, were not

oriented precisely 45° to the axis and were not actually quite symmetrical to the center, the two components H_1 and H_2 were not equal. There was no remarkable difference between Fig. 8 and Fig. 6. Fig. 8 shows the two E fields: E_1 and E_2. Their peaks are almost the same, which might be due to the fact that the two loops were connected at the top points, but the nulls did not coincide with each other due to the two unsymmetrical loops. Comparing Fig. 8 and Fig. 6, we found that we had better results when there was a connection at the top points of the two loops in that the nulls of E_1 were somewhat deeper.

4.4.2 *90° Orientation*

A double orthogonal semiloop was oriented at 90° to the two transmitting antennas. Fig. 9 shows H_2 which is the amplitude of the total H. Loop 1 picked up some E field, as H_1 shows, of about the same value as in the unconnected case. H_1 was almost 20 dB down compared with H_2. There was no remarkable difference between Fig. 9 and Fig. 7 except that H_1 in Fig. 9 picked up more like a pure E field although

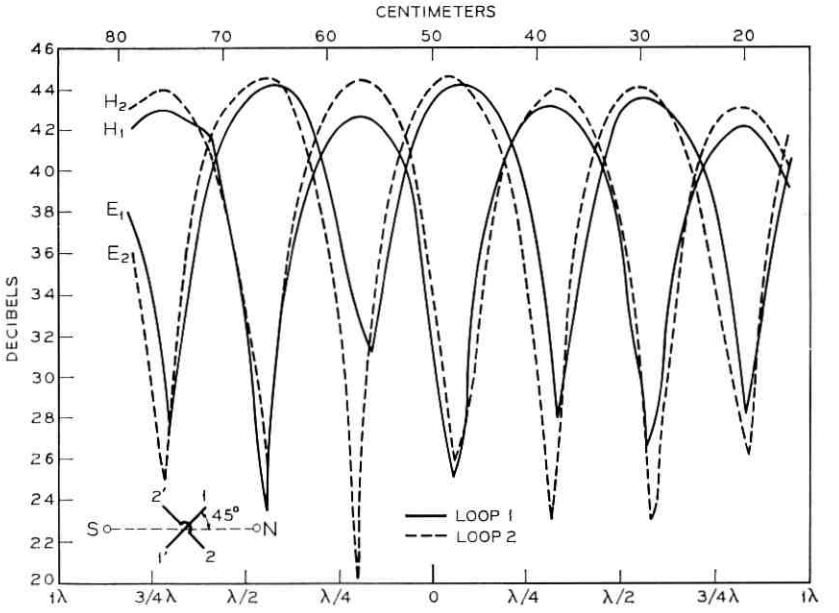


Fig. 6 — Standing waves of H fields and E fields along the slot by using a double orthogonal semiloop antenna unconnected at the top point (oriented at 45°).

it has small field strength. Fig. 9 shows that E_1 and E_2 almost coincided, but in Fig. 7 they did not. Hence, it is better when the two loops connect at the top points than when they do not.

V. ENERGY-DENSITY COMPUTATION

We used the H and E components of two connected orthogonal loops oriented at 45° (Fig. 8) and 90° (Fig. 9) to compute two sets of energy density from the measurements made in the free-space environment. Since both E and H were measured in volts, the energy density we computed from (1) is

$$w = \frac{\epsilon}{2}\left(E^2 + \left(\frac{\mu}{\epsilon}\right)H^2\right)$$

$$= \frac{\epsilon}{2}\left[E^2(\text{volts}^2/\text{m}^2) + (377\ \text{ohm})^2 \times H^2(\text{amp}^2/\text{m}^2)\right]$$

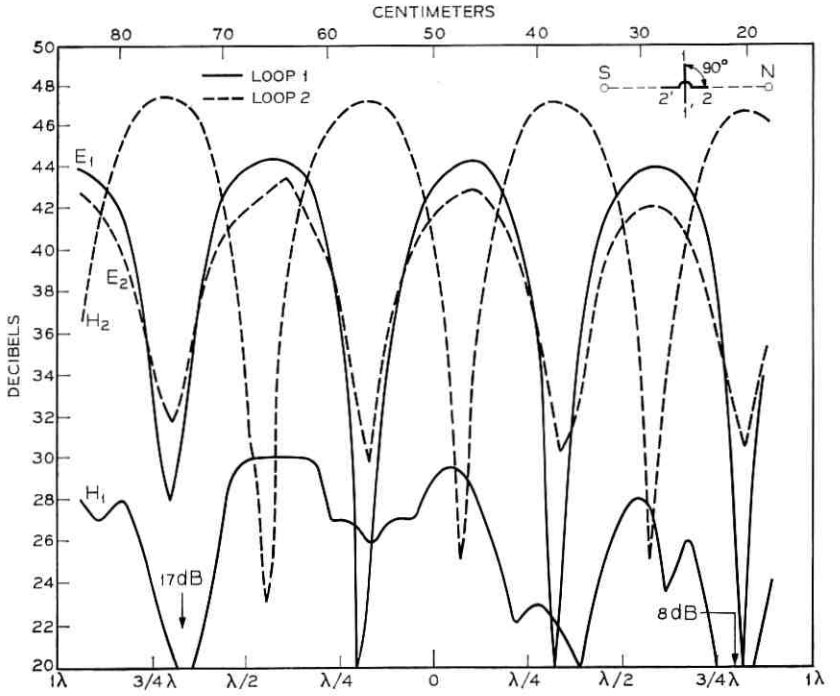


Fig. 7 — Standing waves of H fields and E fields along the slot by using a double orthogonal semiloop antenna unconnected at the top point (oriented at 90°).

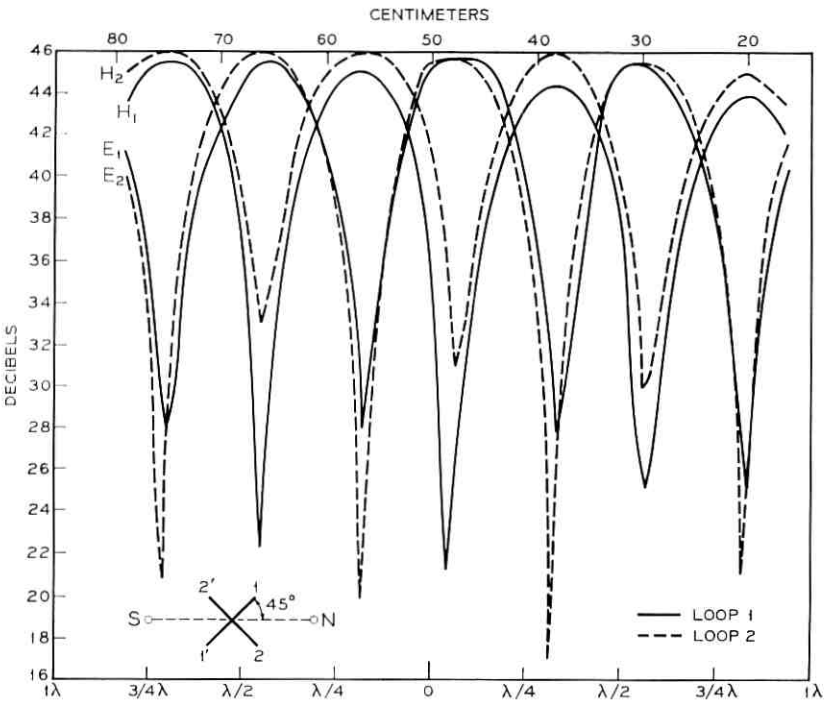


Fig. 8 — Standing waves of H fields and E fields along the slot by using a double orthogonal semiloop antenna connected at the top point—an energy-density antenna (oriented at 45°).

$$= \frac{\epsilon}{2}\,[E^2(\text{volts/m})^2 + H^2(\text{volts/m})^2]$$

$$= \frac{\epsilon}{2}\,(w'), \tag{2}$$

where

$H^2 = \alpha_1 H_1^2 + \alpha_2 H$,

$E^2 = E_1^2$ or E_2^2 ,

α = a weighting factor (a factor relating the level of average peak values of H_1 and H_2 components to the E field), and

w' = the energy density in our calculation.

From Fig. 8 we found that the maximum value of H_1 was about 1 dB less than H_2. Also from Fig. 9 we found that the maximum value

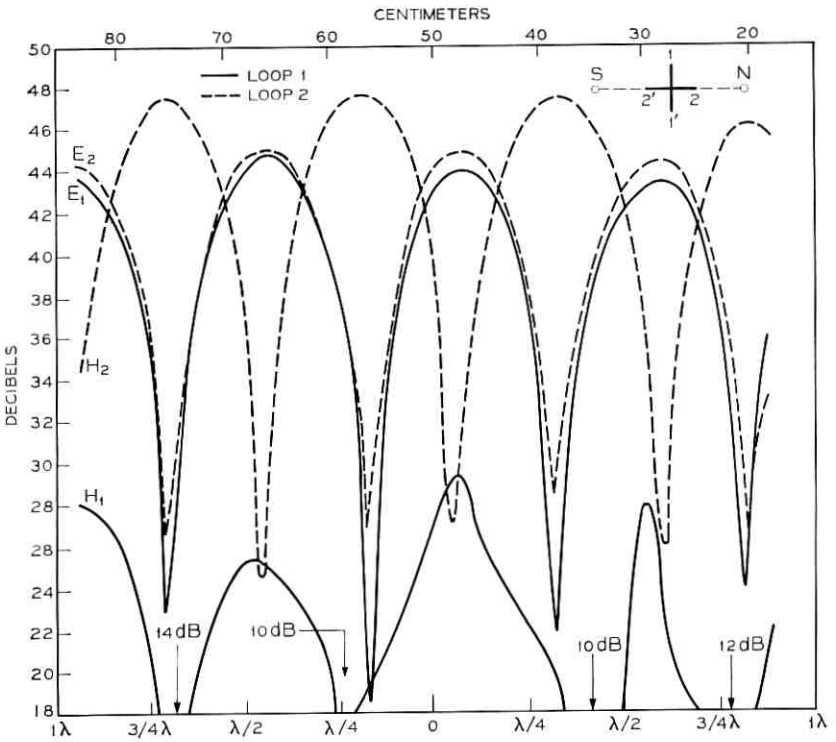


Fig. 9—Standing waves of H fields and E fields along the slot by using a double orthogonal semiloop antenna connected at the top point—an energy-density antenna (oriented at 90°).

of either of two E fields was about 2 dB less than H_2. Hence, we might suggest the following equation representing the energy density obtained from this particular antenna:

$$w' = (1.122H_1)^2 + H_2^2 + (1.26E_2)^2$$
$$= (1.26)^2[(0.89H_1)^2 + (0.795H_2)^2 + E_2^2], \quad (3)$$

where $\alpha_1 = 0.89$ and $\alpha_2 = 0.795$. From (3) we can calculate two energy-density curves, one shown in Fig. 10 for the orientation of antenna at 45° and another also shown in Fig. 10 for the orientation of antenna at 90°. From both curves, the maximum-to-minimum range was only about 2.4 dB, compared to 18–20 dB in Fig. 8 and 9 for the E and H fields alone.

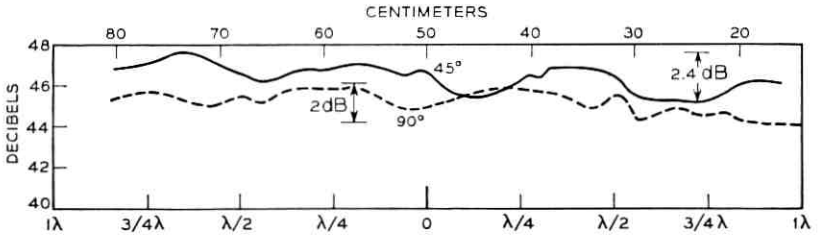


Fig. 10 — The energy-density calculation of an energy-density antenna (oriented at 45° and 90°).

VI. CONCLUSION AND COMMENTS

An energy-density antenna with loops of 1.5 inches in diameter was selected from the measurements as the one to test in the mobile radio field. The connected orthogonal loops were somewhat better than unconnected ones. For two orientations of the loop in the standing wave field in the test environment, the computed energy density varied much less than any of the field components. The configuration of the energy-density antenna could be used at other frequency ranges by scaling the diameter of the loops. After an energy-density antenna was made, a calibration to obtain the weighting factors α_1 and α_2 was needed to set up a proper energy-density equation for this particular antenna.

I wish to take this opportunity to thank W. C. Jakes, Jr., for his advice and suggestions.

REFERENCES

1. Pierce, J. R. and Jakes, W. C., Jr., private communication.
2. Whiteside, H. and King, R. W. P., The Loop Antenna as a Probe, IEEE Trans. Ant. Prop., *AP-12*, May, 1964.
3. Harrison, C. W., Jr., Qualitative Analysis of Loop Antenna Behavior in Linearly and Elliptically Polarized Electric Fields, J. Amer. Soc. Naval Engr., May, 1957, pp. 369–374. Whiteside, H. and King, R. W. P., Loc. Cit.

Error Probability for Binary Signaling Through a Multipath Channel

By R. T. AIKEN

(Manuscript received February 27, 1967)

Error probability is considered for binary signaling through a multipath channel in which (i) the receiver observes a waveform comprising white Gaussian noise and the sum of (perhaps several) time-delayed, frequency-shifted, Rayleigh-faded versions of the transmitted waveform, (ii) the receiver decides with minimum error probability which of the two possible transmissions was sent. Results given herein for the exact minimum error probability necessarily depend upon a number of parameters and are cumbersome to use. By introducing bounds on the error probability, depending upon bounds on spectra of certain matrices, the number of parameters is reduced and the less cumbersome results become applicable to any one of a set of channels rather than to just one channel. The error-probability bounds are presented in terms of values of the distribution function, derived herein, of the difference of two chi-square random variables. The bounds are sharp when the spectra are narrow. For the case of widely orthogonal signals, any version of one possible transmission being orthogonal to any version of the other transmission, the bounds are given as a set of universal curves plotted versus signal-to-noise ratio for various values of the number of paths and of the spectral width of certain matrices. Spectral bounds can easily be computed when the versions for each transmission are nearly orthogonal. Returning to the general case, another bound is derived, by a technique due to Chernoff, which does not explicitly require spectral bounds which may neither be readily available nor be accurate approximations of eigenvalues. This bound is not as sharp as the previous bound for the case of small spectral width, but has promise for the large-width case.

I. INTRODUCTION

This paper considers error probability for the optimum reception of binary signals transmitted through a multipath channel having

P paths.* One of two possible signals is transmitted; the received waveform is the sum of P Rayleigh-faded, time-delayed, frequency-shifted versions of the transmitted signal, plus white Gaussian noise. That is to say, if the complex signal $\sqrt{2E_m}\, x_m(t)$ is transmitted, $m = 1, 2$, the contribution to the received waveform from the pth path is

$$y_m(t; p) = \sqrt{2E_m}\, a_p x_m(t - \tau_p) \exp [i(2\pi f_p t + \varphi_p)],$$

where a_p, φ_p, τ_p, and f_p are the Rayleigh-distributed amplitude, the uniformly-distributed phase, the fixed time delay, and the fixed frequency shift associated with the pth path. The received waveform is

$$z_m(t) = \sum_{p=1}^{P} y_m(t; p) + n(t),$$

where $n(t)$ is white Gaussian noise.

The above multipath situation is a special case of a more general communications situation in which a receiver observes a sample $z(t)$ of a zero-mean complex Gaussian process on the time interval [O, T], the covariance function $\langle z(s)z^*(t)\rangle_m$ having been selected from a set of two distinct functions by chance according to the prior probabilities $\{\alpha_m\}$, $m = 1, 2$, and the other second-moment function $\langle z(s)z(t)\rangle_m$ being zero. The receiver is to be designed so that its decision upon one of the two possible hypotheses is made with minimum average error probability P_e, where $P_e = \Sigma \alpha_m P_e(m)$ and $P_e(m)$ is the probability, when covariance indexed m is true, of deciding otherwise.

The receiver-design problem has been treated in Ref. 1, rigorously demonstrating that optimum processing involves quadratic filtering. However, the filter kernels, being the solutions of integral equations, are difficult to determine in general; moreover, the error probability is not evaluated. For the multipath channel, the first difficulty is overcome in Ref. 2 and the evaluation of binary error probability is considered in the present paper.

Section II presents the theory of a method that can be used to calculate error probability exactly. However, it is quickly appreciated that error probability depends in a cumbersome fashion upon a large number of parameters including the path strengths and the scalar products of the versions. To simplify this situation, this paper introduces bounds on the error probability which depend upon bounds on the spectra of certain matrices, the eigenvalues of which determine

* Each path could comprise a multitude of randomly phased subpaths having essentially the same delay and frequency-shift parameters.

error probability exactly. Thus, the bounds are applicable to any one of a set of channels rather than to just one channel.

Section III presents these error-probability bounds in terms of values of the distribution function of the difference of two chi-square random variables and then derives this distribution function. More specific results are obtained in Section IV for the case of widely orthogonal signals, any path's version of one of the two possible transmitted waveforms being orthogonal to any path's version of the other waveform. Here, easily computed spectral bounds can be given for the case in which the versions under each hypothesis are nearly orthogonal. Section V considers the case of well-resolved paths, making contact with diversity theory (Ref. 3, Chap. 7), and the case of on-off keying.

The error-probability bounds considered above require spectral bounds which may not always be easily computed and which may not be accurate approximations of the eigenvalues. A bound that circumvents these difficulties is obtained in Section VI with a technique due to Chernoff. Comparison of this bound with previous bounds is carried out analytically only for the case of well-resolved paths, but qualitative comparison is made for more general cases.

II. PROCEDURE TO OBTAIN ERROR PROBABILITY IN THE GENERAL CASE

2.1 *Notation*

The binary situation is a specialization of the case of M-ary signaling through the multipath channel in which the received process $z(t)$ can have one of M possible covariance functions, $\langle z(s)z^*(t)\rangle_m$, $m = 1, 2, \ldots M$, of the form

$$2E_m \sum_{p=1}^{P} \sigma_p b_p(s, m) b_p^*(t, m) + N_0 \delta(s - t),$$

a degenerate kernel plus a white-noise kernel (Ref. 2). Here $b_p(t, m) = \exp(i2\pi f_p t) x_m(t - \tau_p)$ is a time-doppler-shifted normalized version of the transmitted signal $\sqrt{2E_m}\, x_m(t)$; the path with index p has an average cross section of σ_p units, a delay of τ_p seconds, and a doppler-shift of f_p Hz. We put

$$\int dt \mid x_m(t) \mid^2 = \int dt \mid b_p(t, m) \mid^2 = 1,$$

so that the average energy received from the medium is

$$\frac{1}{2}\int dt\, 2E_m \sum \sigma_p \mid b_p(t, m) \mid^2 = E_m \sum \sigma_p = E_m ,$$

since we put $\sum \sigma_p = 1$.

The above covariance function can be written

$$2E_m \tilde{b}(s, m)\sigma b^*(t, m) + N_0\delta(s - t),$$

where $b(t, m)$ is a vector with pth component $b_p(t, m)$ and σ is the diagonal matrix with pth entry σ_p, with $tr\ \sigma = 1$.

The optimum receiver decides according to the value of m that corresponds to the largest of M test statistics computed as follows. For each value of m, the receiver first generates the column vector $Z(m) = N_0^{-\frac{1}{2}} \int dt\, z(t)b^*(t, m)$ and then evaluates a test statistic comprising a Hermitian form in $Z(m)$ plus a bias constant. This test statistic is

$$[(N_0/2E_1)^{\frac{1}{2}}Z(m)]^\dagger(2E_m/N_0)H(m)[(N_0/2E_1)^{\frac{1}{2}}Z(m)] + (N_0/2E_1)\theta(m),$$

where the Hermitian combining matrix is

$$(2E_m/N_0)H(m) = (2E_m/N_0)[(2E_m/N_0)B(m) + \sigma^{-1}]^{-1},$$

the bias is given by,

$$\theta(m) = \log \frac{\alpha_m \det [(2E_1/N_0)B(1) + \sigma^{-1}]}{\alpha_1 \det [(2E_m/N_0)B(m) + \sigma^{-1}]},$$

$B(m)$ is the correlation-function matrix $\int dt\, b^*(t, m)\tilde{b}(t, m)$, and the hypotheses are ordered so that $E_1 = \max E_m$. The above test statistic is obtained from that given in Ref. 2 by subtracting $\log [\alpha_1 \det^{-1} \sigma \det^{-1} H^{-1}(1)]$ and multiplying all resulting terms by $N_0/2E_1$.

The above test statistic has a certain intuitive appeal. The components of the vector $Z(m)$ are the correlations of the received signal against the noise-free versions of the transmitted signal that would occur when message m is sent. That is to say, $Z(m)$ provides a measure of the projection of $z(t)$ on the P-dimensional subspace spanned by these versions. Moreover, the test statistic is a measure of the likelihood that this P-dimensional subspace is in fact the correct subspace. Then the optimum receiver strategy is decision according to the most likely of the M possible subspaces. Also, since P dimensions are involved, it might be anticipated that the results are related to the case of P-fold diversity, cf. Section 5.1.

Henceforth, only the binary case, $M = 2$, is considered. In this case, decision according to the larger of two test statistics is equivalent

to decision according to the sign of their difference. The decision events can then be written in terms of one Hermitian form in a composite Gaussian vector

$$Z = (N_0/2E_1)^{\frac{1}{2}}\begin{pmatrix} Z(1) \\ Z(2) \end{pmatrix}$$

as follows. Let

$$Q = \begin{bmatrix} \frac{2E_1}{N_0} H(1) & O_{P\times P} \\ O_{P\times P} & -\frac{2E_2}{N_0} H(2) \end{bmatrix},$$

where $0_{P\times P}$ is the $p \times p$ zero matrix. Then the receiver decides upon $m = 2$ when $Z^\dagger QZ$ is less than $(N_0/2E_1)\theta(2)$, and decides upon $m = 1$ otherwise.

The conditional error probabilities are thus

$$P_e(1) = \Pr\{Z^\dagger QZ < (N_0/2E)\theta \mid 1\} = F_1\left[\left(\frac{N_0}{2E}\right)\theta\right],$$

$$P_e(2) = \Pr\{Z^\dagger QZ > (N_0/2E)\theta \mid 2\} = 1 - F_2\left[\left(\frac{N_0}{2E}\right)\theta\right],$$

where $E = E_1$, $\theta = \theta(2)$, and $F_m(x)$ is the distribution function of $Z^\dagger QZ$ conditioned upon the mth hypothesis.

2.2 *The Fundamental Matrices*

Since $Z^\dagger QZ$ is a function of a Gaussian vector, the distribution function $F_m(x)$ is determined by the conditional mean, $\langle Z\rangle_m$, which is the zero vector, and by the conditional covariance $L(m) = \langle ZZ^\dagger\rangle_m$, the other second-moment matrix $\langle Z\tilde{Z}\rangle_m$ being the $2P \times 2P$ zero matrix.

The conditional covariance matrix $L(m)$ is evaluated as follows. Let

$$L(m) = \begin{bmatrix} L^{11}(m) & L^{12}(m) \\ L^{21}(m) & L^{22}(m) \end{bmatrix},$$

where $L^{jk}(m) = (N_0/2E_1)\langle Z(j)Z^\dagger(k)\rangle_m$. Then, by the definition of $Z(j)$ and interchange of operations, we obtain

$$\langle Z(j)Z^\dagger(k)\rangle_m = \frac{1}{N_0}\iint ds\, dt\; b^*(s, j)\langle z(s)z^*(t)\rangle_m \tilde{b}(t, k)$$

$$= \frac{1}{N_0} \iint ds\, dt\, b^*(s, j) 2E_m \tilde{b}(s, m) \sigma b^*(t, m) \tilde{b}(t, k)$$

$$+ \iint ds\, dt\, b^*(s, j)\, \delta(s - t) \tilde{b}(t, k)$$

$$= \frac{2E_m}{N_0} B(j, m) \sigma B(m, k) + B(j, k),$$

where $B(j, m) = \int ds\, b^*(s, j) \tilde{b}(s, m)$ is a cross-correlation matrix. Hence,

$$L^{ik}(m) = (E_m/E_1) B(j, m) \sigma B(m, k) + (N_0/2E_1) B(j, k).$$

Similarly, it is found that $\langle Z\tilde{Z} \rangle_m$ is the $2P \times 2P$ zero matrix.

For future computations, it is convenient to write

$$Q = \begin{pmatrix} Q^{11} & 0 \\ 0 & Q^{22} \end{pmatrix},$$

where

$$Q^{11} = B^{-1}(1) \{ I + (N_0/2E_1) [B(1)\sigma]^{-1} \}^{-1},$$

$$Q^{22} = -(E_2/E_1) B^{-1}(2) \{ (E_2/E_1) I + (N_0/2E_1) [B(2)\sigma]^{-1} \}^{-1}.$$

2.3 *The Characteristic-Function Method*

To obtain the distribution, consider the conditional characteristic function

$$\varphi_m(t) = \langle \exp (itZ^\dagger QZ) \rangle_m .$$

It is well known, e.g., Ref. 4, that

$$\varphi_m(t) = \det{}^{-1} [I - itL(m)Q] = \prod_k [1 - it\lambda_k(m)]^{-1},$$

where $\{\lambda_k(m)\}$ is the set of eigenvalues of the matrix $L(m)Q$. The eigenvalues are real, since $L(m)Q$ is similar to the Hermitian matrix $L^{\frac{1}{2}}(m) Q L^{\frac{1}{2}}(m)$.

The distribution function can now be obtained from the characteristic function. As a preliminary, it is noted that the characteristic function $(1 - it\lambda)^{-n}$ corresponds to one of two distribution functions, according to the sign of λ. When λ is positive, the distribution function is

$$\int_{-\infty}^{y} dx\, U(x) \frac{x^{n-1} \exp(-x/\lambda)}{\lambda^n (n-1)!} = \begin{cases} I(y/\lambda, n-1) & (y > 0), \\ 0 & (y < 0), \end{cases}$$

$$= U(y) I(y/\lambda, n - 1),$$

where $U(x)$ is the unit step function (unity for $x > 0$, zero for $x < 0$, one-half for $x = 0$) and where

$$I(y, n) = \frac{1}{n!}\int_0^y dx\, x^n e^{-x} = 1 - e^{-y}\sum_{k=0}^{n}\frac{y^k}{k!}$$

is the incomplete gamma function. Similarly, when λ is negative, the distribution function is

$$\int_{-\infty}^{y} dx\, U(-x)\frac{(-x)^{n-1}\exp(x\mid\lambda\mid^{-1})}{\mid\lambda\mid^n (n-1)!}$$

$$= \begin{cases} 1 & (y > 0), \\ 1 - I(-y\mid\lambda\mid^{-1}, n-1) & (y < 0), \end{cases}$$

$$= 1 - U(-y)I(y/\lambda, n-1).$$

To obtain the distribution function of $Z^\dagger QZ$, the characteristic function is expanded into its partial fractions. Each term will be proportional to $(1 - it\lambda)^{-n}$ for some n, and corresponds to a term in the expansion of the distribution function. For example, when all eigenvalues are distinct, the expansion of the characteristic function is

$$\varphi_m(t) = \sum_k \frac{d_k(m)}{1 - it\lambda_k(m)},$$

where

$$d_k(m) = \prod_{j\neq k}\left(1 - \frac{\lambda_j(m)}{\lambda_k(m)}\right)^{-1}.$$

The expansion of the distribution function $F_m(x)$ is then

$$\sum_{[k:\lambda_k(m)>0]} d_k(m)U(x)I(y/\lambda_k(m), 0)$$

$$+ \sum_{[k:\lambda_k(m)<0]} d_k(m)[1 - U(-x)I(y/\lambda_k(m), 0]$$

In the case of a degenerate spectrum, an eigenvalue λ with multiplicity r contributes the sum $\sum_{n=1}^{r} A_n(1 - it\lambda)^{-n}$ to the expansion of the characteristic function, and the corresponding part of the distribution function involves $I(\cdot, n)$ for $n = 0, 1, 2, \cdots, r-1$.

It should be observed that the general approach of summing distribution functions corresponding to partial fractions is fully equivalent to inverting the characteristic function by contour integration, the approach used by Turin[5] for a similar problem. (When all poles are simple, the expansion coefficients $\{d_k(m)\}$ are residues of the poles.)

III. UPPER AND LOWER BOUNDS ON THE ERROR PROBABILITY

3.1 *Error-probability Bounds from Degenerate-spectrum Variables*

Exact computation of error probability involves considerable numerical work in computing eigenvalues followed by evaluation of cumbersome formulas. Moreover, an often inordinately large number of independent parameters must be specified. To simplify this situation, we consider bounds on the spectrum of $L(m)Q$ rather than the spectrum itself. With a technique suggested in Ref. 6, we can obtain error-probability bounds. Although we do not obtain the error probability itself, the error-probability bounds apply to not just one channel but rather to any channel for which the spectral bounds are met.

Observe that the characteristic function is precisely specified by the spectrum of $L(m)Q$. This spectrum is the same as the spectrum of I diag $[\lambda_1(m), \cdots, \lambda_{2P}(m)]$, where I plays the role of a covariance matrix and the diagonal matrix plays the role of a matrix of a Hermitian form. Hence, the distribution of $Z^\dagger QZ$ is the same as the distribution of

$$q(m) = \sum_{k=1}^{2P} \lambda_k(m) \mid z_k \mid^2,$$

where $\{z_k\}$ are complex zero-mean Gaussian variates with covariance matrix $\langle z_i z_k^* \rangle = \delta_{ik}$, $\langle z_i z_k \rangle$ being zero.

Suppose bounds on the eigenvalues are available. That is to say, suppose it is known that the positive eigenvalues satisfy

$$\underline{\mu} \leqq \lambda_k(m) \leqq \bar{\mu}, \tag{1a}$$

and that the negative eigenvalues satisfy

$$-\bar{\nu} \leqq \lambda_k(m) \leqq -\underline{\nu}, \tag{1b}$$

where the μ's and ν's are positive numbers that depend on m. Then, a lower bound on $q(m)$ is the degenerate-spectrum random variable $\underline{q}(m)$, defined by

$$\underline{q}(m) = \underline{\mu} \sum_{k=1}^{P} \mid z_k \mid^2 - \bar{\nu} \sum_{k=P+1}^{2P} \mid z_k \mid^2.$$

Note that we have used the fact that the number of positive eigenvalues and the number of negative eigenvalues are the same, see Appendix A. Similarly, an upper bound on $q(m)$ is provided by the random variable

$$\bar{q}(m) = \bar{\mu} \sum_{k=1}^{P} \mid z_k \mid^2 - \underline{\nu} \sum_{k=P+1}^{2P} \mid z_k \mid^2.$$

Since $\underline{q}(m) \leqq q(m) \leqq \bar{q}(m)$, it follows that

$$\Pr\{\bar{q}(m) \leqq y\} \leqq F_m(y) = \Pr\{q(m) \leqq y\} \leqq \Pr\{\underline{q}(m) \leqq y\}.$$

Evaluation of these bounds requires the distribution function $G(y; P, \alpha)$ of the degenerate-spectrum random variable

$$\sum_{k=1}^{P} |z_k|^2 - \alpha \sum_{k=P+1}^{2P} |z_k|^2,$$

which is the difference of two chi-square variables each with an even number of degrees of freedom. The bounds become

$$G[(\bar{\mu})^{-1}y; P, \underline{\nu}(\bar{\mu})^{-1}] \leqq F_m(y) \leqq G[(\underline{\mu})^{-1}y; P, \bar{\nu}(\underline{\mu})^{-1}],$$

where we use $y = (N_0/2E)\theta$ and reiterate that the μ's and ν's depend on m.

It is anticipated that these bounds are sharp when the spectrum is narrow, the spread of the positive spectrum being much less than any positive eigenvalue and similarly for the negative spectrum. Also, when θ itself is not precisely known, but bounds $\underline{\theta} \leqq \theta \leqq \bar{\theta}$ are available, the distribution function is bounded by

$$G[(\bar{\mu})^{-1}\underline{y}; P, \underline{\nu}(\bar{\mu})^{-1}] \leqq F_m(y) \leqq G[(\underline{\mu})^{-1}\bar{y}; P, \bar{\nu}(\underline{\mu})^{-1}], \tag{2}$$

where $\underline{y} = (N_0/2E)\underline{\theta}$ and $\bar{y} = (N_0/2E)\bar{\theta}$.

3.2 *Distribution of a Degenerate-Spectrum Variable*

It will be demonstrated that $G(y; P, \alpha)$ equals

$$\left(\frac{\alpha}{1+\alpha}\right)^P \sum_{k=0}^{P-1} \binom{P-1+k}{k} \left(\frac{1}{1+\alpha}\right)^k \left[1 - I\left(\frac{|y|}{\alpha}, P-1-k\right)\right] \tag{3a}$$

when $y < 0$, and equals

$$\sum_{k=0}^{P-1} \binom{P-1+k}{k} \left(\frac{1}{1+\alpha}\right)^k \left[\left(\frac{\alpha}{1+\alpha}\right)^P + \frac{\alpha^k}{(1+\alpha)^P} I(y, P-1-k)\right] \tag{3b}$$

when $y > 0$.

Before doing so, note that when $y < 0$, the parameter α serves as a scale size for y in the argument of $I(x, n)$, but that this is not true when $y > 0$. Nevertheless, α does act as a scale size in the following way. A power-series expansion of $I(x, n)$ yields

$$\frac{\alpha^k}{(1+\alpha)^P} I(y, P-1-k) = \left(\frac{\alpha}{1+\alpha}\right)^P \left(\frac{y}{\alpha}\right)^{P-k} \frac{1}{(P-k)!} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \frac{P-k}{P-k+n} y^n,$$

and when $\alpha \ll 1$, the factor $(y/\alpha)^{P-k}$ determines the small-y behavior. Also, this result exhibits $[\alpha/(1+\alpha)]^P$ as a factor for the case $y > 0$, in agreement with the expression for $y < 0$.

To find $G(y; P, \alpha)$, we consider its characteristic function

$$(1 - it)^{-P}(1 + it\alpha)^{-P}.$$

Let the partial-fraction expansion of this characteristic function be

$$\sum_{m=0}^{P-1} A_{P-m}(1 - it)^{-(P-m)} + \sum_{n=0}^{P-1} B_{P-n}(1 + it\alpha)^{-(P-n)}.$$

To evaluate A_{P-m}, multiply by $(1 - it)^P$ and let $1 - it = \tau$ to obtain

$$(1 + \alpha - \alpha\tau)^{-P} = \sum_{m=0}^{P-1} A_{P-m}\tau^m + \tau^P \sum_{n=0}^{P-1} B_{P-n}(1 + \alpha - \alpha\tau)^{-(P-n)}.$$

Since the second sum is analytic at $\tau = 0$, we have exhibited the Taylor expansion with remainder. But

$$\begin{aligned}(1 + \alpha - \alpha\tau)^{-P} &= (1 + \alpha)^{-P}\left(1 - \frac{\alpha}{1+\alpha}\tau\right)^{-P} \\ &= \left(\frac{1}{1+\alpha}\right)^P \sum_{k=0}^{\infty} \binom{P + k - 1}{k}\left(\frac{\alpha}{1+\alpha}\right)^k \tau^k,\end{aligned}$$

where we have used (7) on page 2 of Ref. 7. Hence,

$$A_{P-m} = \left(\frac{1}{1+\alpha}\right)^P \binom{P + m - 1}{m}\left(\frac{\alpha}{1+\alpha}\right)^m.$$

Similarly, to obtain B_{P-n}, multiply by $(1 + it\alpha)^P$ and let $1 + it\alpha = \tau$ to obtain

$$\left(1 + \frac{1}{\alpha} - \frac{\tau}{\alpha}\right)^{-P} = \sum_{n=0}^{P-1} B_{P-n}\tau^n + \tau^P \sum_{m=0}^{P-1} A_{P-m}\left(1 + \frac{1}{\alpha} - \frac{\tau}{\alpha}\right)^{-(P-m)}.$$

Reasoning as before, it is seen that

$$B_{P-n} = \left(\frac{\alpha}{1+\alpha}\right)^P \binom{P + n - 1}{n}\left(\frac{1}{1+\alpha}\right)^n.$$

Collecting these results, it is seen that the characteristic function is

$$\sum_{k=0}^{P-1} \binom{P + k - 1}{k}\left(\frac{1}{1+\alpha}\right)^k \left[\left(\frac{\alpha}{1+\alpha}\right)^P (1 + it\alpha)^{-(P-k)} + \frac{\alpha^k}{(1+\alpha)^P}(1 - it)^{-(P-k)}\right].$$

This immediately establishes the distribution function $G(y; P, \alpha)$.

IV. WIDELY ORTHOGONAL SIGNALS

4.1 *Matrices for the Two Hypotheses*

We consider the special case in which the signals are widely orthogonal, $B(1, 2) = B(2, 1) = O_{P\times P}$. That is to say, all time-doppler shifted versions of one signal are orthogonal to all such versions of the other signal, a situation that would prevail in frequency-shift keying with widely separated frequencies. In this case,

$$L^{jk}(m) = \delta_{jk}\left[\frac{E_m}{E_1} B(j, m)\sigma B(m, k) + \frac{N_0}{2E_1} B(j, k)\right].$$

The "diagonal" form of the covariance matrix $L(m)$ and of the matrix Q implies that the spectrum of $L(m)Q$ comprises the specturm of $L^{11}(m)Q^{11}$ together with the spectrum of $L^{22}(m)Q^{22}$. This can be seen by employing the formulas of Schur (Ref. 8, pp. 45–46) to reduce the determinantal equation $\det [L(m)Q - \lambda I] = 0$ from order $2P$ to order P. For $m = 1$,

$$L^{11}(1)Q^{11} = B(1)\sigma,$$

$$L^{22}(1)Q^{22} = -(N_0/2E_1)(E_2/E_1)\{(E_2/E_1)I + (N_0/2E_1)[B(2)\sigma]^{-1}\}^{-1}.$$

For $m = 2$,

$$L^{11}(2)Q^{11} = (N_0/2E_1)\{I + (N_0/2E_1)[B(1)\sigma]^{-1}\}^{-1},$$

$$L^{22}(2)Q^{22} = -(E_2/E_1)B(2)\sigma.$$

It should be observed that the spectra of the above matrices are simply related to the spectra of $B(1)\sigma$ and of $B(2)\sigma$. When $E_2 = E_1 = E$, the spectrum of $L^{22}(1)Q^{22}$ is $\{-(N_0/2E)(1 + (N_0/2E)\delta_k^{-1})^{-1}\}$, where $\{\delta_k\}$ is the spectrum of $B(2)\sigma$. Similarly, the spectrum of $L^{11}(2)Q^{11}$ is $\{(N_0/2E)(1 + (N_0/2E)\omega_k^{-1})^{-1}\}$, where $\{\omega_k\}$ is the spectrum of $B(1)\sigma$.

Second, it should be observed that when $E_2 = E_1 = E$, the forms of the matrices for the cases $m = 1$ and $m = 2$ are the same, with the roles of positive and negative matrices interchanged. To compute error probability for $m = 1$, we use the distribution function $F_1(x)$; for $m = 2$, we use the conjugate distribution $1 - F_2(x)$ which can be expressed as $P\{-Z^\dagger QZ < -x \mid 2\}$, the distribution function of the negative of the original variable evaluated at $-x$. Introduction of this random variable for the case $m = 2$ reverses the roles of positive and negative matrices, the net effect being that for both $m = 1$ and $m = 2$ the positive and negative matrices have the same forms.

4.2 *Bounds on Spectra, θ, and Error Probability*

It is clear that spectral bounds on $L(1)Q$ can be obtained from spectral bounds on $B(m)\sigma$, $m = 1, 2$, and similarly for $L(2)Q$. Consider the bounds on $L(1)Q$ when $E_1 = E_2 = E$. The positive spectrum is bounded as follows:

$$\underline{\mu} = \underline{\omega} \leqq \min \omega_k \leqq \lambda(1) \leqq \max \omega_k \leqq \bar{\omega} = \bar{\mu},$$

and the negative spectrum is bounded as follows:

$$-\bar{\nu} = -(N_0/2E)[1 + (N_0/2E)(\bar{\delta})^{-1}]^{-1} \leqq \lambda(1)$$

$$\lambda(1) \leqq -(N_0/2E)[1 + (N_0/2E)(\underline{\delta})^{-1}]^{-1} = -\underline{\nu},$$

where $\underline{\delta} \leqq \min \delta_k \leqq \max \delta_k \leqq \bar{\delta}$.

Moreover, bounds on θ can also be obtained. When $E_1 = E_2 = E$ and $\alpha_1 = \alpha_2 = \frac{1}{2}$ (equilikely signals),

$$y = (N_0/2E)\theta = (N_0/2E) \log \frac{\det [B(1)\sigma + (N_0/2E)I]}{\det [B(2)\sigma + (N_0/2E)I]}.$$

Since a determinant is the product of the eigenvalues of the matrix, we have

$$(N_0/2E)\theta = (N_0/2E) \log \prod_{k=1}^{P} \frac{\omega_k + (N_0/2E)}{\delta_k + (N_0/2E)}.$$

Thus, an upper bound is

$$(N_0/2E)\bar{\theta} = (N_0/2E)P \log \frac{\bar{\omega} + (N_0/2E)}{\underline{\delta} + (N_0/2E)},$$

and a lower bound is

$$(N_0/2E)\underline{\theta} = (N_0/2E)P \log \frac{\underline{\omega} + (N_0/2E)}{\bar{\delta} + (N_0/2E)}.$$

Recall that the distribution function $F_1[(N_0/2E)\theta]$ is bounded from above by $G[(\underline{\mu})^{-1}(N_0/2E)\bar{\theta};\ P,\ \bar{\nu}(\underline{\mu})^{-1}]$. Further, suppose that the spectra of $B(1)\sigma$ and $B(2)\sigma$ are narrow about the nominal value $(1/P)$ tr $B(m)\sigma = (1/P)$ tr $\sigma = (1/P)$. We can put

$$\bar{\omega} = \bar{\delta} = \frac{1 + \beta}{P}, \qquad \underline{\omega} = \underline{\delta} = \frac{1 - \beta}{P}, \tag{4}$$

where β is the fractional spectral half width. Then, the parameters required to compute the upper bound on the distribution function are

$$(\underline{\mu})^{-1}(N_0/2E)\bar{\theta} = \frac{1}{1-\beta}(N_0P/2E)P \log \frac{1+\beta+(N_0P/2E)}{1-\beta+(N_0P/2E)}, \quad (5a)$$

$$\bar{\nu}(\underline{\mu})^{-1} = \frac{1}{1-\beta}(N_0P/2E)\left[1+\frac{1}{1+\beta}(N_0P/2E)\right]^{-1}. \quad (5b)$$

Similarly, the distribution function $F_1[(N_0/2E)\theta]$ is bounded from below by $G[(\bar{\mu})^{-1}(N_0/2E)\underline{\theta}; P, \underline{\nu}(\bar{\mu})^{-1}]$. The parameters required for this bound are

$$(\bar{\mu})^{-1}(N_0/2E)\underline{\theta} = \frac{1}{1+\beta}(N_0P/2E)P \log \frac{1-\beta+(N_0P/2E)}{1+\beta+(N_0P/2E)}, \quad (5c)$$

$$\underline{\nu}(\bar{\mu})^{-1} = \frac{1}{1+\beta}(N_0P/2E)\left[1+\frac{1}{1-\beta}(N_0P/2E)\right]^{-1}. \quad (5d)$$

Having considered the case $m = 1$, the bounds for the case $m = 2$ are apparent. Considering the random variable $-Z^{\dagger}QZ$ with θ assumed known, the positive and negative spectral bounds are precisely the same as for the case $m = 1$, and the upper bound is

$$G[-(\underline{\mu})^{-1}(N_0/2E)\theta; P, \bar{\nu}(\underline{\mu})^{-1}]$$

whereas the lower bound is $G[-(\bar{\mu})^{-1}(N_0/2E)\theta; P, \underline{\nu}(\bar{\mu})^{-1}]$. But θ is unknown, and the upper bound is given by replacing $-\theta$ by $\bar{\theta}$, and the same result is obtained as previously; similarly, the lower bound is given by replacing $-\theta$ by $\underline{\theta}$. In short, the bounds apply to both cases, $m = 1$ and 2.

The numerical values of these bounds are given in Figs. 1 to 3 as functions of $2E/N_0P$ (the signal-to-noise ratio per path) for various fixed values of β (the fractional spectral half-width) and P (the number of paths). The curves are nested with respect to values of the fractional spectral half-width β; an increase of β always yields an increase of the upper bound and a decrease of the lower bound. A measure of the sharpness of the bounds (given a nominal value of error probability P_e) is provided by the difference of the upper-bound and lower-bound values of $2E/N_0P$ (in dB) for given values of β and P. For $P_e = 10^{-4}$ and $P = 4$, the sharpness is $1\frac{1}{4}$ dB for $\beta = 0.05$ and $2\frac{1}{4}$ dB for $\beta = 0.1$. This measure of sharpness appears to be relatively insensitive to the value of P. An alternate measure would be the difference in error probability for a given value of $2E/N_0P$, and this measure is indeed markedly sensitive to P.

In the region of the curves corresponding to high signal-to-noise ratio, there is an improvement in error probability associated with

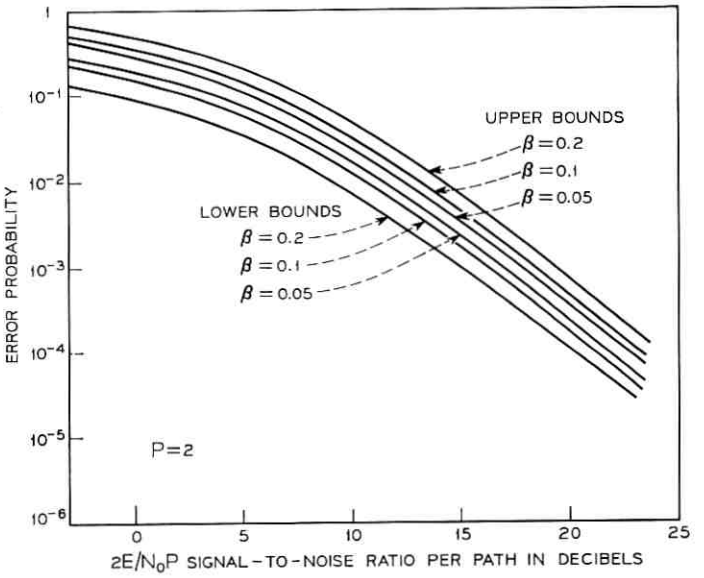


Fig. 1 — Error-probability bounds for widely-orthogonal signaling, $P = 2$.

larger P; the curves become straight lines since P_e becomes proportional to $(2E/N_0P)^{-P}$. However, this improvement is in part attributable to choosing $2E/N_0P$, the average per-path signal-to-noise ratio, as the abscissa rather than $2E/N_0$, the total signal-to-noise ratio. To obtain plots vs $2E/N_0$, one moves the $P = 2^n$ curves to the right by $3n$ dB; then, the improvement with increased P is less dramatic in this region of high signal-to-noise ratio.

4.3 *Computing Spectral Bounds*

It has been observed that bounds on the error probability for the case of widely orthogonal signals can be obtained from bounds on the spectrum of $B(m)\sigma$, $m = 1, 2$. We now give several easily computed formulas for these bounds.

Recall that $B(m)$ is defined to be $\int dt\ b^*(t, m)\tilde{b}(t, m)$, a matrix of scalar products or a Gram matrix. In general, this is uninformative, since a matrix is a Gram matrix if and only if the matrix is positive semidefinite. However, we will shortly use the fact that in our case the diagonal entries of $B(m)$ are unity because of the normalization. Next, note that $B(m)\sigma$ is similar to $\sigma^{\frac{1}{2}}B(m)\sigma^{\frac{1}{2}}$, a hermitian matrix which has real roots (since σ is a real diagonal matrix with positive

entries, the matrices $\sigma^{\frac{1}{2}}$ and $\sigma^{-\frac{1}{2}}$ exist; then $\sigma^{\frac{1}{2}}[B(m)\sigma]\sigma^{-\frac{1}{2}} = \sigma^{\frac{1}{2}}B(m)\sigma^{\frac{1}{2}})$. When $B(m)$ is diagonal or nearly so, the roots of $B(m)\sigma$ should be close to the entries of σ; this is justified by the following theorem.[9] The characteristic roots of any matrix A lie in the closed region of the z-plane consisting of all the disks $\{z: | z - A_{ii} | \leqq \sum_{i \neq i} | A_{ij} |, i = 1, 2, \cdots, P\}$. In our case, the region must be on the real line, and we obtain a set of not necessarily nonoverlapping intervals centered about $\{\sigma_i\}$, the half-widths being $\{\sum_{j \neq i} | B_{ij}(m) | \sigma_i\}$ when we take $A = B(m)\sigma$. The spectral bounds are then the rightmost right-end point

$$\max_i [A_{ii} + \sum_{j \neq i} | A_{ij} |],$$

and the leftmost left-end point

$$\min_i [A_{ii} - \sum_{j \neq i} | A_{ij} |]$$

(when it is positive).

A family of spectral bounds is obtainable from this theorem by applying it to $B(m)\sigma$ and to matrices similar to $B(m)\sigma$, e.g., $\sigma^{\frac{1}{2}}B(m)\sigma^{\frac{1}{2}}$, $\sigma B(m)$, and more generally $\sigma^{\alpha}B(m)\sigma^{1-\alpha}$, $0 \leqq \alpha \leqq 1$. Thus, we have the family

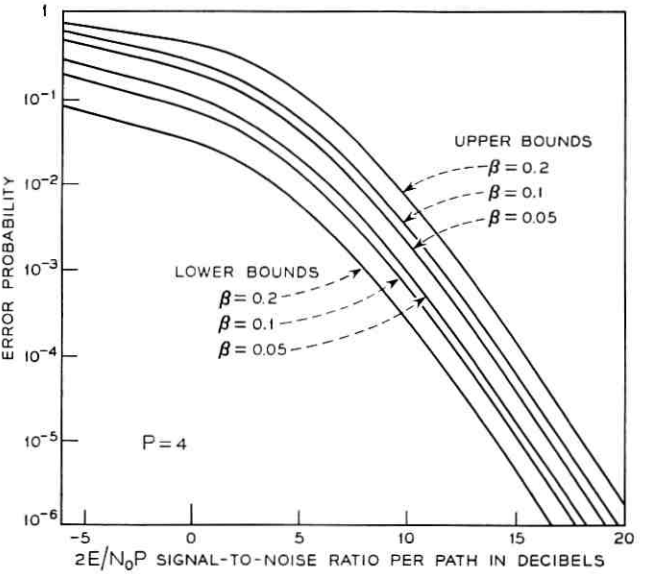


Fig. 2 — Error-probability bounds for widely-orthogonal signaling, $P = 4$.

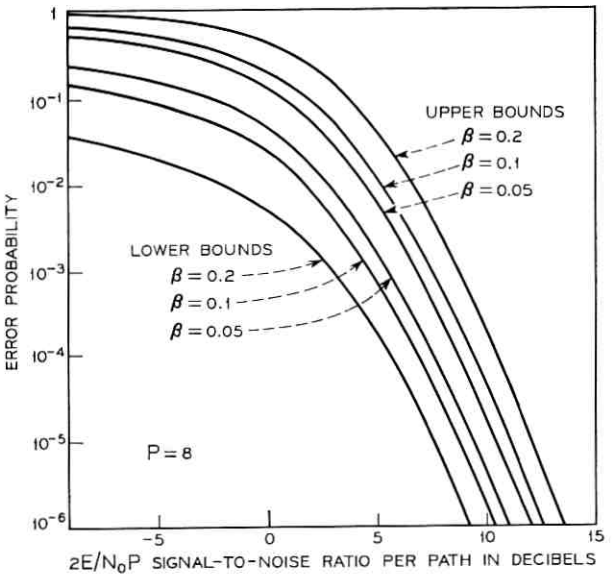


Fig. 3 — Error-probability bounds for widely-orthogonal signaling, $P = 8$.

of upper spectral bounds

$$\max_i \{\sigma_i + \sum_{j \neq i} \sigma_i^{\alpha} \mid B_{ij}(m) \mid \sigma_j^{1-\alpha}\}, \qquad 0 \leqq \alpha \leqq 1. \tag{6}$$

The question arises: which is the smallest upper bound? It is not true in general that a bound is attained for the value of i that maximizes σ_i, but suppose this is the case when $\alpha = 0$. That is to say, suppose $\sigma_i = \max_k \sigma_k$ and that

$$\sigma_i\left[1 + \sum_{j \neq i} \mid B_{ij}(m) \mid \frac{\sigma_j}{\sigma_i}\right] = \max_k \sigma_k\left[1 + \sum_{j \neq k} \mid B_{kj}(m) \mid \frac{\sigma_j}{\sigma_k}\right].$$

Then it follows that this is the smallest bound in the family, for $\sigma_j/\sigma_i \leqq 1$ implies that

$$\sum_{j \neq i} \mid B_{ij}(m) \mid \frac{\sigma_j}{\sigma_i} \leqq \sum_{j \neq i} \mid B_{ij}(m) \mid \left(\frac{\sigma_j}{\sigma_i}\right)^{1-\alpha},$$

and hence

$$\sigma_i\left[1 + \sum_{j \neq i} \mid B_{ij}(m) \mid \frac{\sigma_j}{\sigma_i}\right] \leqq \sigma_i\left[1 + \sum_{j \neq i} \mid B_{ij}(m) \mid \left(\frac{\sigma_j}{\sigma_i}\right)^{1-\alpha}\right]$$
$$\leqq \max_k \left\{\sigma_k\left[1 + \sum_{j \neq k} \mid B_{kj}(m) \mid \left(\frac{\sigma_j}{\sigma_k}\right)^{1-\alpha}\right]\right\}.$$

Similarly, we have the family of lower spectral bounds

$$\min_i \{\sigma_i - \sum_{j \neq i} \sigma_i^{\alpha} \mid B_{ij}(m) \mid \sigma_i^{1-\alpha}\}, \qquad 0 \leqq \alpha \leqq 1. \tag{7}$$

The largest lower bound is obtained when $\alpha = 1$ provided that $\sigma_i = \min_k \sigma_k$ and that

$$\sigma_i[1 - \sum_{j \neq i} \mid B_{ij}(m) \mid] = \min_k \{\sigma_k[1 - \sum_{j \neq k} \mid B_{kj}(m) \mid]\}.$$

To see this, observe that $\sigma_j/\sigma_i \geqq 1$ implies

$$\sum_{j \neq i} \mid B_{ij}(m) \mid \left(\frac{\sigma_j}{\sigma_i}\right)^{1-\alpha} \geqq \sum_{j \neq i} \mid B_{ij}(m) \mid,$$

and hence

$$\sigma_i[1 - \sum_{j \neq i} \mid B_{ij}(m) \mid] \geqq \sigma_i\left[1 - \sum_{j \neq i} \mid B_{ij}(m) \mid \left(\frac{\sigma_j}{\sigma_i}\right)^{1-\alpha}\right]$$

$$\geqq \min_k \left\{\sigma_k\left[1 - \sum_{j \neq k} \mid B_{kj}(m) \mid \left(\frac{\sigma_j}{\sigma_k}\right)^{1-\alpha}\right]\right\}.$$

It should be noted that less sharp bounds are easily obtained. For example, the matrix $\sigma B(m)$ yields the upper bound

$$\max_i \{\sigma_i[1 + \sum_{j \neq i} \mid B_{ij}(m) \mid]\} \leqq \max_i \sigma_i \max_i [1 + \sum_{j \neq i} \mid B_{ij}(m) \mid],$$

and the right-hand side is easily computed. The corresponding lower bound is

$$\min_i \{\sigma_i[1 - \sum_{j \neq i} \mid B_{ij}(m) \mid]\} \geqq [\min_i \sigma_i][1 - \max_i \sum_{j \neq i} \mid B_{ij}(m) \mid].$$

These less sharp bounds are easier to compute than those obtained in a similar fashion from $B(m)\sigma$ or from $\sigma^{\alpha}B(m)\sigma^{1-\alpha}$.

Also, it should be noted that sharper bounds can be obtained by employing a sharper theorem of matrix theory:[9] The characteristic roots of any matrix A lie in the closed region of the z-plane consisting of all the ovals $\mid z - A_{ii} \mid \mid z - A_{jj} \mid \leqq (\sum_{k \neq i} A_{ik})(\sum_{k \neq j} A_{jk})$, $i \neq j$. We do not pursue these bounds, but note that simple formulas are obtained only when all paths have equal strength, $\sigma_i = 1/P$.

It is now clear that when $B(m)$ is essentially diagonal, with $\sum_{j \neq i} \mid B_{ij}(m) \mid \ll 1$ for all i, the path gains σ_i are good nominal values for the characteristic roots of $B(m)\sigma$. If, moreover, these path gains are equal, or approximately equal, then the upper and lower spectral

bounds are close to one another. When this narrow-spectrum condition prevails, the positive and negative portions of the spectrum of $L(m)Q$ are also narrow, and the bounds on error probability are sharp.

V. OTHER SPECIAL CASES

5.1 *Well-resolved Paths and the Theory of Diversity*

We consider the case in which the signals are resolvable, $B(1) = B(2) = I$, i.e., the paths are well separated in time and frequency so that any time-Doppler shifted version of a signal is orthogonal to any other version of itself. Moreover, we also assume that $B(1, 2) = \int dt\ b^*(t, 1)\tilde{b}(t, 2)$ becomes a diagonal matrix, $B(1, 2) = \rho I$ where $\rho = \int dt\ x_1^*(t)x_2(t)$, i.e., the paths are sufficiently separated so that any version of one signal is orthogonal to all but the same-path version of the other signal.

It is then easily seen that the covariance matrix is comprised of diagonal submatrices. For $m = 1$,

$$L^{11}(1) = \sigma + (N_0/2E_1)I \qquad L^{12}(1) = \rho[\sigma + (N_0/2E_1)I]$$

$$L^{21}(1) = \rho^*[\sigma + (N_0/2E_1)I] \qquad L^{22}(1) = |\rho|^2\sigma + (N_0/2E_1)I.$$

For $m = 2$, assuming $E_2 > 0$,

$$L^{11}(2) = (E_2/E_1)[|\rho|^2\sigma + (N_0/2E_2)I]$$

$$L^{12}(2) = \rho(E_2/E_1)[\sigma + (N_0/2E_2)I]$$

$$L^{21}(2) = \rho^*(E_2/E_1)[\sigma + (N_0/2E_2)I]$$

$$L^{22}(2) = (E_2/E_1)[\sigma + (N_0/2E_2)I].$$

Moreover, the matrix Q is diagonal, being related to

$$(2E_m/N_0)H(m) = (2E_m/N_0)[(2E_m/N_0)I + \sigma^{-1}]^{-1} = \sigma[\sigma + (N_0/2E_m)I]^{-1}.$$

It then follows that $L(m)Q$ is comprised of diagonal submatrices. To find the spectrum, the order of the determinantal equation can be reduced from $2P$ to P. Then the argument of the determinant is quadratic in λ. For the case $E_1 = E_2$, a method of Turin [(22)–(23) in Ref. 5] can be used relating the λ_k to the eigenvalues (elements) of σ.

The above example brings the present analysis in contact with the theory of diversity combining, see e.g., Ref. 3, Sec. 7.4. Turin,[5] for example, considered the case in which separate waveforms are available and the fading is nonindependent in general. In our analysis, only one waveform is in general available. But in the case of well-separated paths,

we may assume P separate signal waveforms have been observed. However, these separate waveforms must fade independently in keeping with our general discrete-path model, and the on-diagonal component matrices of $L(m)$, *viz.*, $L^{11}(m) = (N_0/2E_1)\langle Z_1 Z_1^\dagger\rangle_m$ and $L^{22}(m) = (N_0/2E_1)\langle Z_2 Z_2^\dagger\rangle_m$, are themselves diagonal matrices. It is still entirely possible that $L^{12}(m)$, the off-diagonal component matrix of $L(m)$, is not a diagonal matrix; e.g., when $x_2(t)$ is a delayed version of $x_1(t)$, then time overlap may preclude $B(1, 2)$ being diagonal even though $B(1)$ and $B(2)$ are diagonal. But when we assume that $B(1, 2)$ is also diagonal, then we obtain the form for $L(m)$ exhibited above. It can be observed that this is precisely the result Turin obtained for the case of optimum diversity combining, where his not necessarily diagonal A becomes our diagaonal σ. When $B(1, 2)$ is not diagonal, then our results do not specialize to the form given by Turin, a reflection of the fact that the multipath channel is not in general fully equivalent to a diversity channel.

5.2 *On-Off Keying*

Another example is the case of on-off keying in which $E_2 = 0$. The test statistic $Z^\dagger QZ$ becomes

$$[(N_0/2E_1)^{\frac{1}{2}}Z(1)]^\dagger(2E_1/N_0)H(1)[(N_0/2E_1)^{\frac{1}{2}}Z(1)], \quad \text{since} \quad Q^{22} = 0.$$

Thus, the distribution is determined by the spectrum of the matrix $L^{11}(m)Q^{11}$, where

$$L^{11}(m) = \delta_{m1}B(1, m)\sigma B(m, 1) + (N_0/2E_1)B(1),$$
$$Q^{11} = B^{-1}(1)\{I + (N_0/2E_1)[B(1)\sigma]^{-1}\}^{-1}.$$

Observe that we no longer have the difference of positive-definite forms, the test statistic now being a positive random variable. The threshold $(N_0/2E_1)\theta(2)$ is

$$(N_0/2E_1)\log\det\left[\left(\frac{2E_1}{N_0}\right)B(1)\sigma + I\right]$$

which is positive since the eigenvalues of $(2E_1/N_0)B(1)\sigma + I$ are greater than unity.

Assuming that the spectrum of $L^{11}(m)(2E_1/N_0)H(1)$ lies in the interval $(\underline{\mu}, \bar{\mu})$, where $\underline{\mu}$ and $\bar{\mu}$ are functions of m, the bounds on the distribution function are

$$G[(\bar{\mu})^{-1}(N_0/2E)\theta; P, 0] \leqq F_m\left[\left(\frac{N_0}{2E}\right)\theta\right] \leqq G[(\underline{\mu})^{-1}(N_0/2E)\theta; P, 0].$$

Recall that $G(x; P, 0)$ is related to the incomplete gamma function,

$$G(x; P, 0) = I(x, P - 1).$$

The spectral bounds must exhibit two forms of $(2E_1/N_0)$-dependence. When $m = 1$, $L^{11}(1)Q^{11} = B(1)\sigma$, and bounds on $B(1)\sigma$ become $\underline{\mu}$ and $\bar{\mu}$. When $m = 2$, $L^{11}(2)Q^{11} = (N_0/2E_1)\{I + (N_0/2E_1)[B(1)\sigma]^{-1}\}^{-1}$, so that

$$\underline{\mu} = (N_0/2E_1)[1 + (N_0/2E_1)(\underline{\omega})^{-1}]^{-1},$$

$$\bar{\mu} = (N_0/2E_1)[1 + (N_0/2E_1)(\bar{\omega})^{-1}]^{-1},$$

where the spectrum of $B(1)\sigma$ is confined to $(\underline{\omega}, \bar{\omega})$.

Collecting our results, when $m = 1$,

$$F_1[(N_0/2E)\theta] \leqq I\{(\underline{\omega})^{-1}(N_0/2E)P \log [(2E/N_0)\bar{\omega} + 1]; P - 1\}$$

$$F_1[(N_0/2E)\theta] \geqq I\{(\bar{\omega})^{-1}(N_0/2E)P \log [(2E/N_0)\underline{\omega} + 1]; P - 1\}.$$

Similarly, when $m = 2$

$$F_2[(N_0/2E)\theta] \leqq I\{[1 + (N_0/2E)(\underline{\omega})^{-1}]P \log [(2E/N_0)\bar{\omega} + 1]; P - 1\}$$

$$F_2[(N_0/2E)\theta] \geqq I\{[1 + (N_0/2E)(\bar{\omega})^{-1}]P \log [(2E/N_0)\underline{\omega} + 1]; P - 1\}.$$

These results permit the computation of error-probability-bound curves that would be universal in the same sense as the curves for widely-orthogonal signaling, i.e., the curves would apply to any element of the set of channels for which the spectral bounds are met.

VI. CHERNOFF BOUNDS

6.1 *General Case*

Up to this point, consideration of spectral bounds has lead to error-probability bounds which are sharp when the spectrum comprises narrow positive and negative portions. These bounds are easy to employ when $B(1, 2) = 0$ and $B(1)$, $B(2)$ are nearly diagonal matrices. But in more general cases, the estimation of spectral bounds may be difficult and bounds may be poor approximations of eigenvalues. We turn to another technique of bounding error probability which does not explicitly require spectral bounds.

Consider the error probability when hypothesis $m = 2$ is true, $P_e(2) = \Pr\{Z^\dagger QZ > (N_0/2E)\theta \mid 2\}$. Recall that the unit step function $U(x)$ is unity for $x > 0$, zero for $x < 0$, and one-half for $x = 0$. Then

$$P_e(2) = \Pr\{U[Z^\dagger QZ - (N_0/2E)\theta] = 1 \mid 2\}$$
$$= \mathcal{E}_2\{U[Z^\dagger QZ - (N_0/2E)\theta]\},$$

where $\mathcal{E}_2$ denotes expectation under hypothesis $m = 2$. But since $U(x) \leqq \exp(\mu_2 x)$ for any $\mu_2 > 0$, we have

$$P_e(2) \leqq \mathcal{E}_2\{\exp \mu_2[Z^\dagger QZ - (N_0/2E)\theta]\}.$$

This average can readily be computed, since $Z^\dagger QZ$ has the same distribution as $\sum \lambda_k(2) \mid z_k \mid^2$, where $\{\lambda_k(2)\}$ is the spectrum of $L(2)Q$. Since $\mathcal{E}z_k = 0$, $\mathcal{E}z_i z_k^* = \delta_{ik}$, and $\mathcal{E}z_i z_k = 0$, the Gaussian variables $\{\mathrm{Re}\ z_i\}$, $\{\mathrm{Im}\ z_i\}$ are independent with zero mean and variance equal to $\frac{1}{2}$. Thus, $P_e(2)$ is bounded from above by

$$\exp[-\mu_2(N_0/2E)\theta]\left[\prod_{k=1}^{2P} \mathcal{E} \exp(\mu_2\lambda_k(2) \mid \mathrm{Re}\ z_k \mid^2)\right]^2,$$

where the outer square appears because the product involving $\{\mathrm{Im}\ z_k\}$ has been suppressed. But a standard calculation shows,

$$\mathcal{E} \exp(\mu_2\lambda_k(2) \mid \mathrm{Re}\ z_k \mid^2) = [1 - \mu_2\lambda_k(2)]^{-\frac{1}{2}}, \quad \text{when} \quad \mu_2\lambda_k(2) < 1,$$

and our bound is

$$\exp[-\mu_2(N_0/2E)\theta] \prod_{k=1}^{2P} [1 - \mu_2\lambda_k(2)]^{-1}.$$

Thus,

$$P_e(2) \leqq \exp[-\mu_2(N_0/2E)\theta] \det^{-1}[I - \mu_2 L(2)Q], \tag{8}$$

which holds for all μ_2 such that $0 < \mu_2 < [\max \lambda_k(2)]^{-1}$.

The above procedure is adopted from the technique due to Chernoff (see Ref. 3, Sec. 2.5 and 7.4). Here, we do not have identically distributed variables; indeed, half are positive and half are negative random variables.

To find the best value of μ_2, we write the bound as

$$\exp\left\{-\mu_2 \frac{N_0}{2E}\theta - \ln \prod_{k=1}^{2P} [1 - \mu_2\lambda_k(2)]\right\}$$

and differentiate the argument of the exponential. A necessary condition for an extremum is that the derivative be zero, and this yields

$$(N_0/2E)\theta = \sum_{k=1}^{2P} \frac{\lambda_k(2)}{1 - \mu_2\lambda_k(2)}$$
$$= \sum_{k=1}^{2P} \frac{1}{\lambda_k^{-1}(2) - \mu_2} = \mathrm{tr}\{[(L(2)Q)^{-1} - \mu_2 I]^{-1}\}.$$

If the value of μ_2 that satisfies this equation lies within the allowable interval $[0, \max^{-1}\lambda_k(2)]$, then this value of μ_2 minimizes the upper bound. A minimum occurs because the second derivative of the argument of the exponential is positive, being

$$\sum_{k=1}^{2P}\left[\frac{\lambda_k(2)}{1-\mu_2\lambda_k(2)}\right]^2.$$

In a similar fashion, the error probability for $m = 1$ can be overbounded.

$$\begin{aligned} P_e(1) &= \Pr\{-Z^\dagger QZ > -(N_0/2E)\theta \mid 1\} \\ &\leqq \mathcal{E}_1\{\exp \mu_1[-Z^\dagger QZ + (N_0/2E)\theta]\} \\ P_e(1) &\leqq \exp[\mu_1(N_0/2E)\theta]\det^{-1}[I + \mu_1 L(1)Q]. \end{aligned}$$

The best value of μ_1 satisfies

$$(N_0/2E)\theta = \operatorname{tr}\{L(1)Q[I + \mu_1 L(1)Q]^{-1}\},$$

provided this value lies in the allowable interval $[0, \max^{-1}(-\lambda_k(1))]$.

6.2 *Widely-orthogonal Signals*

Consider the case in which the signals are widely orthogonal, $B(1,2) = 0$, but have equal energy, $E_1 = E_2 = E$, and are equilikely, $\alpha_1 = \alpha_2 = \frac{1}{2}$. The overbound on $P_e(1)$ is obtained from the spectrum of $L(1)Q$ which comprises the spectrum of $L^{11}(1)Q^{11}$ together with the spectrum of $L^{22}(1)Q^{22}$. Thus,

$$P_e(1) \leqq \exp[\mu_1(N_0/2E)\theta]\det^{-1}[I + \mu_1 L^{11}(1)Q^{11}]\det^{-1}[I + \mu_1 L^{22}(1)Q^{22}].$$

But the matrices used here were related in Paragraph 4.1 to $B(1)\sigma$ and $B(2)\sigma$, and our bound becomes

$$\exp[\mu_1(N_0/2E)\theta]\det^{-1}[I + \mu_1 B(1)\sigma] \cdot \det^{-1}\{I - \mu_1(N_0/2E)[I + (N_0/2E)(B(2)\sigma)^{-1}]^{-1}\}.$$

After some manipulation, this bound becomes

$$\left(\frac{E}{2N_0}\right)^{-P}\left\{\exp\left[\mu_1\left(\frac{N_0}{2E}\right)\theta\right]\right\} \cdot \frac{\det\left[B(2)\sigma + \left(\frac{N_0}{2E}\right)I\right]}{\det\left[2\mu_1\left(\frac{N_0}{2E}\right)B(1)\sigma + \left(\frac{N_0}{E}\right)I\right]\det\left\{2\left[1 - \mu_1\left(\frac{N_0}{2E}\right)\right]B(2)\sigma + \left(\frac{N_0}{E}\right)I\right\}},$$

where

$$\exp\left[\mu_1(N_0/2E)\theta\right] = \left\{\frac{\det\left[B(1)\sigma + (N_0/2E)I\right]}{\det\left[B(2)\sigma + (N_0/2E)I\right]}\right\}^{\mu_1(N_0/2E)}.$$

The maximum allowable value of μ_1 is determined by the largest eigenvalue of $L^{22}(1)Q^{22}$ which in turn is determined by the largest eigenvalue of $B(2)\sigma$:

$$0 < \mu_1 < \frac{2E}{N_0} + \max^{-1}(\delta_k),$$

where $\{\delta_k\}$ is the spectrum of $B(2)$.

The best value of μ_1 is found from the relation

$$\begin{aligned}(N_0/2E)\theta &= \sum \frac{\lambda_k(1)}{1 + \mu_1\lambda_k(1)} \\ &= \operatorname{tr}\left\{L^{11}(1)Q^{11}[I + \mu_1 L^{11}(1)Q^{11}]^{-1}\right\} \\ &\qquad + \operatorname{tr}\left\{L^{22}(1)Q^{11}[I + \mu_1 L^{22}(1)Q^{11}]^{-1}\right\},\end{aligned}$$

where we again have exploited the decomposition of the spectrum of $L(1)Q$. After some manipulation, we find

$$\begin{aligned}(N_0/2E)\theta &= \operatorname{tr}\left\{B(1)\sigma[I + \mu_1 B(1)\sigma]^{-1}\right\} \\ &\qquad - \operatorname{tr}\left\{B(2)\sigma\left[I + \left(\frac{2E}{N_0} - \mu_1\right)B(2)\sigma\right]^{-1}\right\}.\end{aligned}$$

An approximate solution can be obtained for the case of high signal-to-noise ratio. Let $\mu_1 = \bar{\mu}_1(2E/N_0)$; the relation becomes

$$(N_0/2E)\theta = \sum \frac{\omega_k}{1 + (2E/N_0)\bar{\mu}_1\omega_k} - \sum \frac{\delta_k}{1 + (2E/N_0)(1 - \bar{\mu}_1)\delta_k}.$$

Suppose $\bar{\mu}_1(2E/N_0)\omega_k \gg 1$ and $(1 - \bar{\mu}_1)(2E/N_0)\delta_k \gg 1$. Then the right side becomes approximately

$$\frac{P}{(2E/N_0)\bar{\mu}_1} - \frac{P}{(2E/N_0)(1 - \bar{\mu}_1)}.$$

Equating this to $(N_0/2E)\theta$ and solving the resulting quadratic for the root applicable for the case $\theta = 0$ yields

$$\bar{\mu}_1 = \left[\left(1 + \frac{\theta}{2P}\right) + \left(1 + \frac{\theta^2}{4P^2}\right)^{\frac{1}{2}}\right]^{-1}.$$

When $\theta/2P$ is small, this value of $\bar{\mu}_1$ is approximately

$$\frac{1}{2}\left[1 - \frac{\theta}{4P}\right],$$

and the corresponding value of μ_1 is $(E/N_0)[1 - (\theta/P)]$ which is approximately at the midpoint of the allowable interval.

In a similar fashion, the overbound on $P_e(2)$ is

$$\exp\left[-\mu_2(N_0/2E)\theta\right] \det^{-1}\left[I - \mu_2 L^{11}(2)Q^{11}\right] \det^{-1}\left[I - \mu_2 L^{22}(2)Q^{22}\right],$$

which becomes

$$\exp\left[-\mu_2(N_0/2E)\theta\right] \det^{-1}\left\{I - \mu_2(N_0/2E)[I + (N_0/2E)(B(1)\sigma)^{-1}]^{-1}\right\} \cdot \det^{-1}\left[I + \mu_2 B(2)\sigma\right],$$

or

$$\left(\frac{E}{2N_0}\right)^{-P} \left\{\exp\left(-\mu_2 \frac{N_0}{2E}\theta\right)\right\} \cdot \frac{\det\left[B(1)\sigma + \left(\frac{N_0}{2E}\right) I\right]}{\det\left\{2\left[1 - \mu_2\left(\frac{N_0}{2E}\right)B(1)\sigma + \left(\frac{N_0}{E}\right) I\right\} \det\left[2\mu_2\left(\frac{N_0}{2E}\right) B(2)\sigma + \left(\frac{N_0}{E}\right) I\right]\right.},$$

where

$$\exp\left[-\mu_2(N_0/2E)\theta\right] = \left\{\frac{\det\left[B(2)\sigma + (N_0/2E)I\right]}{\det\left[B(1)\sigma + (N_0/2E)I\right]}\right\}^{\mu_2(N_0/2E)}.$$

The maximum allowable value of μ_2 is determined by the largest eigenvalue of $L^{11}(2)Q^{11}$ in turn determined by the largest eigenvalue of $B(1)\sigma$:

$$0 < \mu_2 < \frac{2E}{N_0} + \max^{-1}(\omega_k),$$

where $\{\omega_k\}$ is the spectrum of $B(1)\sigma$. The best value of μ_2 satisfies

$$(N_0/2E)\theta = \operatorname{tr}\left\{L^{11}(2)Q^{11}[I - \mu_2 L^{11}(2)Q^{11}]^{-1}\right\} + \operatorname{tr}\left\{L^{22}(2)Q^{22}[I - \mu_2 L^{22}(2)Q^{22}]^{-1}\right\}$$

$$(N_0/2E)\theta = \operatorname{tr}\left\{B(1)\sigma\left[\left(\frac{2E}{N_0} - \mu_2\right)B(1)\sigma + I\right]^{-1}\right\} - \operatorname{tr}\left\{B(2)\sigma[I + \mu_2 B(2)\sigma]^{-1}\right\}.$$

Let $\mu_2 = \bar{\mu}_2(2E/N_0)$ and suppose $\bar{\mu}_2(2E/N_0)\delta_k \gg 1$, $(1-\bar{\mu}_2)(2E/N_0)\omega_k \gg 1$.

Then the right side becomes

$$\frac{P}{(2E/N_0)(1-\bar{\mu}_2)} - \frac{P}{(2E/N_0)\bar{\mu}_2},$$

and the approximation of the best value of $\bar{\mu}_2$ is

$$\bar{\mu}_2 = \left[\left(1 - \frac{\theta}{2P}\right) + \left(1 + \frac{\theta^2}{4P^2}\right)^{\frac{1}{2}}\right]^{-1}$$

which is approximately $\frac{1}{2}[1 - (\theta/4P)]$ when $(\theta/2P) \ll 1$.

The foregoing results can be specialized to the case in which the paths are resolvable, $B(1) = B(2) = I$. Then $\theta = 0$, and it is easily seen that the best value of $\bar{\mu}_m$ is $\frac{1}{2}$. Both overbounds become

$$(E/2N_0)^{-P} \frac{\det[\sigma + (N_0/2E)I]}{\det^2[\sigma + (N_0/E)I]},$$

and this agrees with equation 7.134 in Ref. 3.

It should be noted that $\bar{\mu}_m = \frac{1}{2}$ is always an allowed value of $\bar{\mu}_m$. For the case of resolvable paths, it is the best value, and whenever $\theta/P \ll 1$ and $2E/N_0$ is sufficiently large, it is close to the best value. Using $\bar{\mu}_m = \frac{1}{2}$, we can obtain an overbound for both error probabilities, i.e., for $P_e(m)$, $m = 1, 2$. This overbound is

$$(E/2N_0)^{-P} \exp(\tfrac{1}{2} \mid \theta \mid) \frac{\det[B(3-m)\sigma + (N_0/2E)I]}{\det[B(1)\sigma + (N_0/E)I] \det[B(2)\sigma + (N_0/E)I]}. \tag{9a}$$

The factor $\exp(\frac{1}{2} \mid \theta \mid)$ can also be written in terms of determinants. When $\det[B(1)\sigma + (N_0/2E)I]$ is larger than $\det[B(2)\sigma + (N_0/2E)I]$, we have

$$\exp(\tfrac{1}{2} \mid \theta \mid) = \left\{\frac{\det[B(1)\sigma + (N_0/2E)I]}{\det[B(2)\sigma + (N_0/2E)I]}\right\}^{\frac{1}{2}}, \tag{9b}$$

and when the reverse inequality holds, $\exp(\frac{1}{2} \mid \theta \mid)$ is the reciprocal of the above.

For the case in which the spectrum of $B(m)\sigma$ lies in the interval $(1-\beta/P, 1+\beta/P)$ where $\beta < 1$, the overbound can be further overbounded. The factor involving determinants is less than

$$\left\{\frac{P\left[1 + \beta + \left(\frac{N_0P}{2E}\right)\right]}{\left[1 - \beta + 2\left(\frac{N_0P}{2E}\right)\right]^2}\right\}^P,$$

and $|\theta|/2$ is less than

$$\frac{P}{2}\log\frac{1+\beta+(N_0P/2E)}{1-\beta+(N_0P/2E)}.$$

It follows that the Chernoff bound is less than

$$(N_0P/2E)^P\left\{\frac{4[1+\beta+(N_0P/2E)]^{\frac{1}{2}}}{[1-\beta+(N_0P/2E)]^{\frac{1}{2}}[1-\beta+2(N_0P/2E)]^{\frac{1}{2}}}\right\}^P. \quad (10)$$

Numerical values of this bound are given in Fig. 4, and it has the same general character as the spectral-related bounds. Rather than sharpness given a nominal value of error probability P_e, we consider the sensitivity measured by the change in $2E/N_0P$ (in dB) vs β; for $P_e = 10^{-4}$ and $P = 4$, the sensitivity is 2 dB for $\beta = 0.1$. The sensitivity does not markedly increase with an increase in P, in agreement with the behavior of the sharpness of the previous bounds.

Comparison of the Chernoff bound with the previous bounds is conveniently done for the case $\beta = 0$ (cf. Sec. 7.4 of Ref. 3). The Chernoff bound does not specify a signal-to-noise ratio (required to achieve a nominal P_e) excessively greater than the previous value; for $P = 4$, less than 2.2 dB difference is observed. This excess does decrease with increasing P. Moreover, it is entirely conceivable that in a broad-spectrum case with a large number of paths, an exact value of the Chernoff bound would be better than the spectral-bound result. Of course, our inexact (overbounded) Chernoff bound is poor in the

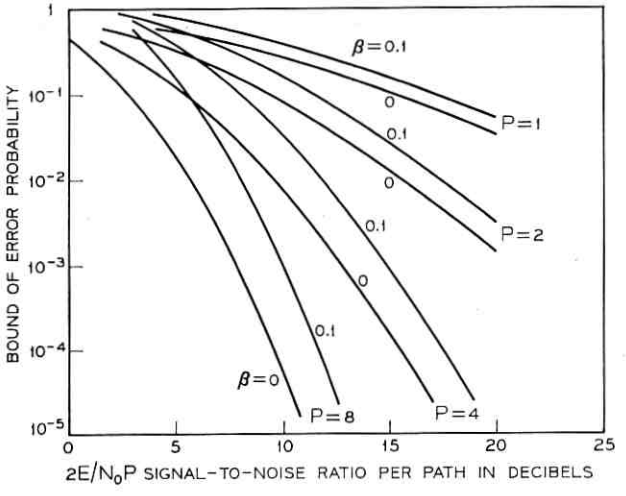


Fig. 4 — Overbounded Chernoff bounds for widely-orthogonal signaling.

broad-spectrum case, but a Chernoff bound using the proper values of the determinants should be good for two reasons. (*i*) Such a bound reflects the precise values of the eigenvalues of the matrix $L(m)Q$. (*ii*) When P is large, the probability density function is bell shaped with the probability "mass" being concentrated near the mean and most of the tail mass being at the leading portion of the tail; then the tail mass can be weighted by the exponential function with little error. On the other hand, the spectral-bound approach suffers in the broad-band case since the spectral bounds are not meaningful approximations of all the eigenvalues.

VII. DISCUSSION

Having observed that exact computation of error probability is cumbersome and depends upon an often inordinately large number of parameters, we considered error-probability bounds (2) that are universal in the sense that they apply to any one of a set of channels satisfying spectral bounds (1). Our bounds employ (3), the distribution function of the difference of chi-square variables. For the special case of widely orthogonal signals, we obtained bounds employing parameters (5) in terms of the spectral width β, see (4), of the matrices $B(m)\sigma$. Plots of these bounds showed that sharpness measured in dB change of $2E/N_0P$ with respect to β for a fixed value of error probability is not sensitive to the value of P. We presented a technique for obtaining spectral bounds for $B(m)\sigma$ when it is nearly diagonal, representative results being (6) and (7). This technique can also be applied to $L(m)Q$ for the more general case in which the signals are not widely orthogonal.

The case of resolvable signals ($B(m) = I$) made contact with the theory of diversity; we found that for the multipath channel to be a diversity channel, $B(1, 2)$ must also be a diagonal matrix. Of course, the previous results also were in contact with diversity theory. With $B(1, 2) = 0$ (a diagonal matrix) but $B(m)$ not necessarily diagonal, our results generalize those of diversity theory in the following sense. The special case $\beta = 0$ corresponds to a diversity channel with equal link gains, but the general case $\beta \neq 0$ can arise in the nondiversity situation when the matrix $B(m)$ is not diagonal. (If $B(m)$ were diagonal, $B(m) = I$ and the diversity case prevails.)

We then turned to the Chernoff bound (8) which does not explicitly employ spectral bounds. The overbounded form (10) for the case of widely orthogonal signals was poorer than the previous bound

when $\beta = 0$. Nevertheless, there is promise that in a broad-spectrum case, the original form (9) would be better than the spectral-related bounds. A further advantage is that once the determinants are evaluated, perhaps on an electronic computer, the error-probability bound is immediately obtained. In contrast, the spectral-related bounds require a certain amount of computation involving incomplete gamma functions even after spectral bounds are obtained.

VIII. ACKNOWLEDGMENT

The author is indebted to Ira Jacobs, M. I. Schwartz, and B. H. Bharucha for stimulating discussion and constructive comment. Miss J. Hoffspiegel wrote the computer programs to obtain the numerical results.

APPENDIX A

Here we show that the number of positive eigenvalues of LQ equals the number of negative eigenvalues.* Recall that L is positive definite and that Q can be written in the partitioned form

$$Q = \begin{bmatrix} Q^{11} & 0 \\ 0 & Q^{22} \end{bmatrix},$$

where Q^{11} and $-Q^{22}$ are positive definite. Clearly, the number of positive eigenvalues of Q equals the number of negative eigenvalues. We can construct a family of positive definite matrices L_t, $0 \leqq t \leqq 1$, such that $L_0 = I$, $L_1 = L$, and L_t is continuous in t. For example, let $L_t = (1 - t)I + tL$; L_t has positive eigenvalues $\{(1 - t) + t\gamma_k\}$, where $\{\gamma_k\}$ are the eigenvalues of L. Now the eigenvalues of L_tQ are real, for L_tQ is similar to the Hermitian matrix $L_t^{\frac{1}{2}}QL_t^{\frac{1}{2}} = L_t^{-\frac{1}{2}}(L_tQ)L_t^{\frac{1}{2}}$, where $L_t^{\frac{1}{2}}$ and $L_t^{-\frac{1}{2}}$ exist since L_t is positive definite, Moreover, the eigenvalues of L_tQ are continuous in t, since L_t is continuous in t. But L_tQ never has a zero eigenvalue, for L_t is positive definite and $(L_tQ)^{-1} = Q^{-1}L_t^{-1}$ always exists. Since the eigenvalues are real, continuous in t, and never zero, it follows that no positive eigenvalue of L_0Q can become negative as t varies on [0, 1], and no negative eigenvalue of L_0Q can become positive. The conclusion is established.

APPENDIX B

This appendix presents another derivation of the distribution function of $\sum_1^P |z_k|^2 - \alpha \sum_{P+1}^{2P} |z_k|^2$. This derivation makes contact with

* We are indebted to B. H. Bharucha for the conception of this proof.

the special functions that have appeared in analyses of diversity channels; also, this derivation appears to admit generalization to the case $z_k = \text{Re } z_k$ with $\langle z_j z_k \rangle = \delta_{jk}$. (An odd number of variables in the real case corresponds to half-integer P in the complex case.)

The density function of $\sum_1^P |z_k|^2$ is

$$f(x) = \begin{cases} \dfrac{x^{P-1}e^{-x}}{(P-1)!} & (x > 0), \\ 0 & (x < 0), \end{cases}$$

and the density function of $-\alpha \sum_{P+1}^{2P} |z_k|^2$ is

$$g(x) = \begin{cases} 0 & (x > 0), \\ \dfrac{(-x)^{P-1}e^{x/\alpha}}{\alpha^P(P-1)!} & (x < 0). \end{cases}$$

The density of the sum is the convolution of the densities,

$$h(x) = \int_{\max(0,x)}^{\infty} dy\, f(y)g(x-y),$$

where the first argument of max $(\cdot\,,\,\cdot)$ arises from the truncated form of f and the second argument arises from the truncated form of g. It follows that

$$h(x) = \frac{\exp(x/\alpha)}{\alpha^P(P-1)!(P-1)!} \int_{\max(0,x)}^{\infty} dy(y-x)^{P-1} \exp\left[-\left(\frac{1}{\alpha}+1\right)y\right].$$

For the case $x > 0$, the lower limit is x. For the case $x < 0$, the integral can be cast into the form of the integral for the case $x > 0$ by a change of variable. The result differs only in the exponential factor, i.e.,

$$h(x) = \frac{\exp(-x)}{\alpha^P(P-1)!(P-1)!} \cdot \int_{|x|}^{\infty} dy(y-|x|)^{P-1}y^{P-1} \exp\left[-\left(\frac{1}{\alpha}+1\right)y\right], \qquad x < 0.$$

The integral can be evaluated with the aid of relation (12) on page 202, Vol. II of Ref. 10, and the common result for the cases $x < 0$ and $x > 0$ is

$$h(x) = \frac{|x|^{P-\frac{1}{2}} \exp\left[\left(\frac{1}{\alpha}-1\right)\frac{x}{2}\right]}{\sqrt{\pi\alpha}\,(1+\alpha)^{P-\frac{1}{2}}(P-1)!} K_{P-\frac{1}{2}}\left(\frac{1+\alpha}{\alpha}\frac{|x|}{2}\right),$$

where $K_{P-\frac{1}{2}}(z)$ is the modified Bessel function of the third kind.

The above expression for the density is valid for all P, noninteger as well as integer. But in our application, P is an integer; a relation on page 80 of Ref. 11 yields

$$K_{P-\frac{1}{2}}(z) = \sqrt{\frac{\pi}{2z}}\, e^{-z} \sum_{k=0}^{P-1} \frac{(P-1+k)!}{k!(P-1-k)!(2z)^k}.$$

The density is then

$$h(x) = \left(\frac{\alpha}{1+\alpha}\right)^P \exp\left[\left(\frac{1-\alpha}{\alpha}\right)\frac{x}{2} - \frac{1+\alpha}{\alpha}\frac{|x|}{2}\right] \cdot \sum_{k=0}^{P-1} \frac{(P-1+k)!}{(P-1)!k!(P-1-k)!}\left(\frac{1}{1+\alpha}\right)^k \frac{1}{\alpha}\left(\frac{|x|}{\alpha}\right)^{P-1-k}.$$

When $x < 0$, the exponential becomes $\exp(x/\alpha)$, and when $x > 0$, it becomes $\exp(-x)$.

Observe that when $\alpha = 1$, the density is symmetric. When $\alpha < 1$, the factor $\exp[(1-\alpha/\alpha)x/2]$ shifts the mass to the right. When $\alpha \to 0$, it can be shown that $h(x) \to f(x)$.

To obtain the distribution function $G(y; P, \alpha)$, consider first the case $y < 0$. Since $\int_{-\infty}^{y} dx\, h(x)$ equals $\int_{|y|}^{\infty} dx\, h(-x)$, the following integral arises in each term of the sum,

$$\int_{|y|}^{\infty} \frac{dx}{\alpha}\, e^{-x/\alpha}\left(\frac{x}{\alpha}\right)^{P-1-k} = (P-1-k)![1 - I(|y|/\alpha, P-1-k)].$$

The case $y > 0$ is treated by considering $\int_{-\infty}^{0} dx\, h(x) + \int_{0}^{y} dx\, h(x)$. The integral that arises is just $(P-1-k)!I(y, P-1-k)$. These steps establish our final result, quoted above.

Our result could also have been obtained from the Fourier transform of the characteristic function $(1-it)^{-P}(1+it\alpha)^{-P}$. The Fourier transform of $(\alpha+it)^{-2\mu}(\beta-it)^{-2\nu}$ is given by relation (12) on page 119, Vol. *I* of Ref. 10 in terms of Whittaker functions that reduce to Bessel functions for the case $\mu = \nu = P/2$ in view of relation (14) on page 265, Ref. 12. The density function can thus be obtained.

REFERENCES

1. Kadota, T. T., Optimum Reception of *M*-ary Gaussian Signals in Gaussian Noise, B.S.T.J., *44*, November, 1965, pp. 2187–2197.
2. Aiken, R. T., Communication over the Discrete-Path Fading Channel, IEEE Trans. Inform. Theory, *IT-13*, April 1967, pp. 346–347.
3. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering*, John Wiley and Sons, Inc., New York, 1965.
4. Turin, G. L., The Characteristic Function of Hermitian Quadratic Forms in Complex Normal Variables, Biometrika, *47*, June, 1960, pp. 199–201.

5. Turin, G. L., On Optimal Diversity Reception, II, IRE Trans. Commun. Sys., *CS-10,* March, 1962, pp. 22–31.
6. Grenander, U., Pollak, H. O., and Slepian, D., Distribution of Quadratic Forms in Normal Variables, J. SIAM, *7,* December, 1959, pp. 374–401.
7. Dwight, H. B., Tables of Integrals and Other Mathematical Data, third edition, The Macmillan Company, New York, 1957.
8. Gantmacher, F. R., *The Theory of Matrices,* Vol. I, Chelsea Publishing Company, New York, 1959.
9. Marcus, M. and Minc, H., *Survey of Matrix Theory and Matrix Inequalities,* Allyn and Bacon, Inc., 1964.
10. Erdelyi, A., et al., *Tables of Integral Transforms,* McGraw-Hill Book Co., Inc., New York, 1954.
11. Watson, G. N., *A Treatise on the Theory of Bessel Functions,* Second edition, Cambridge University Press, 1958.
12. Erdelyi, A., et al., *Higher Transcendental Functions,* Vol. I, McGraw-Hill Book Co., Inc., New York, 1953.

Precoding for Multiple-Speed Data Transmission

By ROBERT W. CHANG

(Manuscript received April 24, 1967)

In certain applications, because of noise, compatibility, or other considerations, it is desirable that a data transmission system have the flexibility to operate at multiple speeds. In this paper, a precoding scheme for multiple-speed digital or analog data transmission is presented. The scheme has a flexibility which allows the data rate and overall channel characteristics to be changed simultaneously by simply changing the data format and some resistive elements. There is no change in the filters, the equalization, the transmitter signaling interval, or the receiver sampling time. By using partial response channels, a number of commonly used data rates are easily obtained, using a physically realizable precoder and correlator. With correct timing and the use of orthonormal siginals, the signal-to-noise ratio is maximized at each data rate for bandlimited white noise under the constraints of fixed line signal power and no intersymbol interference. Timing error is considered in a two-speed transmission scheme, and the selection of a precoding matrix using eye opening as the criterion is studied. This study clearly demonstrates the advantage of changing the overall channel characteristics when changing the data rate. Eye openings obtained are equal to or larger than those of two conventional schemes transmitting at the same data rates.

I. INTRODUCTION

In conventional pulse amplitude modulation (PAM) data transmission systems (digital or analog), the signal at the receiver input takes the form

$$s(t) = \sum_{k=1}^{N} a_k f(t - kT_0), \tag{1}$$

where $\{a_k\}$ are the information symbols, T_0 is the signaling interval, and the signals $f(t - kT_0)$, $k = 1, \cdots, N$, are time translates of each

other. It is well known[1] that in order for these systems to meet the criterion "Maximize the signal-to-noise ratio in the presence of band-limited white noise under the constraints of fixed line signal power and no intersymbol interference," the signals should be designed so that the overall channel characteristics are in the Nyquist I class and the overall amplitude characteristics are divided equally between the transmitting and the receiving side. Such a signal design scheme (hereafter referred to as Scheme I) is popular and is used even if the system designer is aware that the channel noise may not be white over the frequency band of interest. This is because the practical determination of the noise statistics and the realization of the corresponding optimum filters for a general communication complex are nearly impossible. A block diagram of Scheme I is shown in Fig. 1.

In this paper, a precoding signaling scheme (Scheme II) is presented for multiple-speed analog or digital data transmission. Scheme II also meets the signal-to-noise ratio criterion above. The very distinctive difference between Schemes I and II is that in I the signals $f(t - kT_0)$ are time translates, but in II the signals are not necessarily so. This property allows the data rate and overall channel characteristics (such as represented by the eye opening) of Scheme II to be changed simultaneously without changing the filters, the equalization, the signaling interval at the transmitter, or the sampling time at the receiver.

In Scheme II, a sequence of information symbols is divided into blocks and the blocks are transmitted sequentially. For clarity, we first consider in Section II the transmission of a single block at a fixed data rate and the precoder and the receiver structure. Multiple block multispeed transmission and the use of partial response channels are considered in Section III. A two-speed transmission scheme, sampling time error, and eye patterns are considered in Section IV.

II. TRANSMISSION OF A SINGLE BLOCK AT A FIXED DATA RATE

A block diagram of Scheme II is shown in Fig. 2. The quantities $H(j\omega)$ and $h(t)$ are, respectively, the transfer function and the impulse

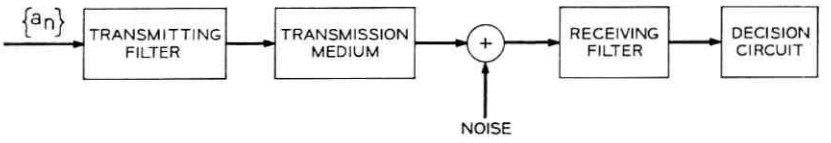


Fig. 1 — Block diagram of Scheme I.

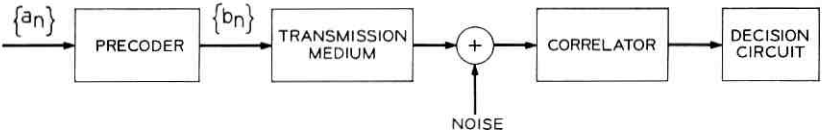


Fig. 2 — Block diagram of Scheme II.

response of the transmission medium. We shall consider $H(j\omega)$ to be bandlimited, and

$$\begin{aligned} H(j\omega) &\neq 0, \qquad |\omega| \leqq 2\pi f_c \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{2}$$

The time interval

$$T = \frac{1}{2f_c} \quad \text{seconds} \tag{3}$$

is the Nyquist interval.

Consider the transmission of a block of symbols $a_1, \cdots, a_N$. Each symbol can be an m-ary digit ($m \geqq 2$) or a real number. The precoder converts $a_1, \cdots, a_N$ into a sequence of numbers $b_1, \cdots, b_N$, and the number b_k, $k = 1, \cdots, N$, is transmitted at $t = kT$. This produces a signal at the input to the receiver given by

$$s(t) = \sum_{k=1}^{N} b_k h(t - kT). \tag{4}$$

From (2), the impulse responses $h(t - kT)$ are infinitely linearly independent, i.e.,

$$\sum_{k=1}^{N} b_k h(t - kT) = 0 \quad \text{for all } t \Rightarrow b_k = 0 \quad \text{for all } k, \tag{5}$$

where N can approach infinity. Equation (5) can be proven by noting that the equality

$$\sum_{k=1}^{N} b_k h(t - kT) = 0 \quad \text{for all } t$$

and (2) together imply that

$$\sum_{k=1}^{N} b_k e^{-j\omega kT} = 0 \quad \text{for} \quad |\omega| \leqq 2\pi f_c \,. \tag{6}$$

Equation (6) then implies that $b_k = 0$ for all k.

As is well known, a bandlimited signal, say $g(t)$, can be represented by its time samples. The vector whose elements are the time samples of $g(t)$ will be referred to as the time sample vector of $g(t)$. For convenience, we shall use time sample vectors in discussing the precoder and receiver structures, and use the signals themselves in analyzing the overall channel characteristics.

Let $\mathbf{h}_k$, a $M \times 1$ vector, be the time sample vector of $h(t - kT)$, where the value of M will be considered later. Then (4) is equivalent to the vector equation

$$\mathbf{S} = \sum_{k=1}^{N} b_k \mathbf{h}_k \tag{7}$$

The N vectors $\mathbf{h}_k$, $k = 1, \cdots, N$, are linearly independent since the impulse responses $h(t - kT)$ are. Hence, the N vectors $\mathbf{h}_k$, $k = 1, \cdots, N$, generate a real Euclidean vector space $\mathcal{E}_N$ of N dimensions. If the precoder were not used, we would have $b_k = a_k$ and $\mathbf{S} = \sum_{k=1}^{N} a_k \mathbf{h}_k$, and the information symbols a_k would be transmitted as coordinates of the basis vectors $\mathbf{h}_k$ of $\mathcal{E}_N$. It is well known that the basis can be changed by a linear transformation. A precoder can be used for this purpose so that a suitable set of basis vectors can be chosen for each transmission rate of a multi-speed system based on considerations such as signal-to-noise ratio and the effect of timing error.

Define

$$\mathbf{A} = \begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}, \quad \mathbf{H}' = \begin{bmatrix} \mathbf{h}_1' \\ \vdots \\ \mathbf{h}_N' \end{bmatrix}, \quad \mathbf{V}' = \begin{bmatrix} \mathbf{V}_1' \\ \vdots \\ \mathbf{V}_N' \end{bmatrix}, \tag{8}$$

where $\mathbf{V}$ represents a set of basis vectors for $\mathcal{E}_N$ and the prime notation represents transpose. Since $\mathbf{h}_k$, $k = 1, \cdots, N$, generate $\mathcal{E}_N$, $\mathbf{V}$ is related to $\mathbf{H}$ by

$$\mathbf{V} = \mathbf{H}\Lambda, \tag{9}$$

where $\Lambda = [\lambda_{ij}]$ is an $N \times N$ nonsingular matrix. If a_k is transmitted as a coordinate of $\mathbf{V}_k$, then

$$\mathbf{S} = \sum_{k=1}^{N} a_k \mathbf{V}_k = \mathbf{V}\mathbf{A} = \mathbf{H}\Lambda\mathbf{A}. \tag{10}$$

But, from (7)

$$\mathbf{S} = \mathbf{H}\mathbf{B}. \tag{11}$$

From (10) and (11), the precoder structure is

$$\mathbf{B} = \mathbf{\Lambda A}. \tag{12}$$

Since the noise statistics and the statistics of the customer's data are usually unavailable, we choose here not to carry out a usual optimization study on the choice of $\mathbf{V}$ using such statistics. In the sequel, $\mathbf{V}$ is chosen to be a set of orthonormal basis vectors. This enables the precoding signaling scheme (Scheme II) to meet the following requirements:

(*i*) The performance is optimum in the same sense as the popular Scheme I described in Section I.

(*ii*) The overall channel characteristics are controlled by the precoding matrix $\mathbf{\Lambda}$ and hence by resistive elements. (In Scheme I the overall channel characteristics are controlled by the transmitting and receiving filters.)

These requirements are met with a simple receiver structure. The noisy signal at the input of the receiver is

$$\mathbf{X} = \mathbf{S} + \mathbf{N} = \sum_{k=1}^{N} a_k \mathbf{V}_k + \mathbf{N}, \tag{13}$$

where $\mathbf{N}$ is the noise vector. A correlator at the receiver computes the decision statistics $\mathbf{X}'\mathbf{V}_1, \mathbf{X}'\mathbf{V}_2, \cdots, \mathbf{X}'\mathbf{V}_N$. Since $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_N$ are orthonormal, we have

$$\mathbf{X}'\mathbf{V}_k = a_k + \mathbf{N}'\mathbf{V}_k . \tag{14}$$

Because of orthonormality the decision statistic $\mathbf{X}'\mathbf{V}_k$ depends only on a_k and there is no intersymbol interference. A decision on the symbol a_k can be made from the decision statistic $\mathbf{X}'\mathbf{V}_k$ by a simple, standard decision rule.

A basic difference between Schemes I and II is that in I the signals $f(t - kT_0)$ are time translates of each other, but in II the orthogonal signals $\mathbf{V}_k$ are not necessarily time translates. A difference in operation between the two schemes is seen in the second requirement. In Scheme I the overall channel characteristics are controlled by the transmitting and receiving filters. But, in Scheme II, they are controlled by the precoding matrix. To illustrate this and also for use in Section IV, we derive the impulse responses of Scheme II. As shown in Fig. 3, the correlator can be implemented with a tapped delay line and N sets of attenuators. Only the jth set of attenuators is shown. The attenuation ratios $V_{j1}, \cdots, V_{jM}$ shown are the values of the elements of $\mathbf{V}_j$, and the decision statistic $\mathbf{X}'\mathbf{V}_j$ is obtained by sampling the output of the jth summing circuit. For analytical purposes, the tapped delay

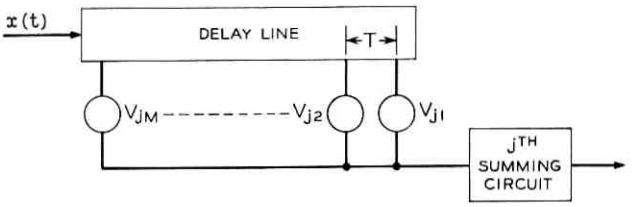


Fig. 3 — Diagram defining $h_{ij}(t)$.

line, the jth set of attenuators, and the jth summing circuit together are equivalent to a matched filter having impulse response $V_j(t_0 - t)$, where t_0 is the sampling instant and $V_j(t)$ is a signal whose time sample vector is $\mathbf{V}_j$. Now define

$$h_{ij}(t) = \text{output of the } j\text{th summing circuit when } a_i = 1 \text{ is applied to the precoder.} \quad (15)$$

Since a_i is transmitted by the signal $\mathbf{V}_i$ or $V_i(t)$, we have

$$h_{ij}(t) = \int_{-\infty}^{\infty} V_j(t_0 + \tau - t) V_i(\tau)\, d\tau. \quad (16)$$

From (8) and (9)

$$\mathbf{V}_j = \sum_{k=1}^{N} \lambda_{jk} \mathbf{h}_k . \quad (17)$$

From (16) and (17)

$$h_{ij}(t) = \sum_{k=1}^{N} \sum_{l=1}^{N} \lambda_{jk} \lambda_{il} \int_{-\infty}^{\infty} h(t_0 + \tau - t - kT) h(\tau - lT)\, d\tau. \quad (18)$$

It is seen from (18) that, for a given transmission medium, $h_{ij}(t)$ is controlled by the elements λ_{ij} of the precoding matrix. Since changing the precoding matrix requires only changing attenuation ratios in the precoder and the correlator, the overall channel characteristics are controlled by resistive elements.

III. PRECODING FOR MULTIPLE-SPEED TRANSMISSION

The transmission of a single block has been considered in Section II. Now consider the transmission of an infinite sequence of symbols. In Scheme II, a symbol sequence is divided into blocks with N symbols in each block. If the vectors $\mathbf{h}_1, \cdots, \mathbf{h}_N$ are $M \times 1$ as assumed in Section II, the blocks can be transmitted sequentially at MT seconds

intervals without interference between each other and the data rate is

$$R = \frac{N}{M} R_{\max} \quad \text{bauds}, \tag{19}$$

where $R_{\max}$ is the Nyquist rate.

Theoretically there is no limit on the block length N; however, we shall restrict N to be small number such as 3 so that the precoder and the correlator can be easily implemented. The parameter M must be restricted accordingly so that R [see (19)] can be a commonly used data rate such as 3/4 of the Nyquist rate. These requirements are satisfied by using the popular partial response channels.[2,3]

Table I of Ref. 3 illustrated five classes of partial response channels. From the table, it is clear that if $h(t)$ is in Class 1, then a set of sampling instants can be chosen (sampling time error will be considered later) such that $h(t - T), \cdots, h(t - NT)$ are simultaneously zero at all except $N + 1$ adjacent sampling points. This means that the vectors $\mathbf{h}_1, \cdots, \mathbf{h}_N$ are each $(N + 1) \times 1$ so that

$$\begin{aligned} M &= N + 1 \\ R &= \frac{N}{N + 1} R_{\max} \quad \text{bauds}. \end{aligned} \tag{20}$$

If $h(t)$ is in Class 2, or 3, or 4, sampling instants can be chosen such that $M = N + 2$. The rule can be easily extended to other classes.

Consider now multiple-speed operation. As will be shown it is desirable to change the overall channel characteristics when changing the data rate. To make these changes, it is necessary to change the data format; however, it is desired that the system not be altered significantly otherwise (such as changing the filters, the equalization, the signaling interval, the receiver sampling time, etc.).

The scheme developed allows the data rate and overall channel characteristics to be changed simultaneously by changing only the data format and some resistive elements. When the system operates as above, the data rate is $(N/M)R_{\max}$ bauds and the sequence of symbols

$$a_1 a_2 \cdots a_N a_{N+1} \cdots$$

is transmitted. If the particular channel is noisy, one may wish to reduce the baud rate so that signal energy per baud can be increased to combat noise (an adaptive technique). The data rate can be changed to

$$R = \frac{r}{M} R_{\max} \quad \text{bauds}, \tag{21}$$

where r can be any integer from 1 to N, by inserting $N - r$ zero digits into each block as follows

$$a_1 \cdots a_r \quad 0 \cdots 0 \quad a_{r+1} \cdots a_{2r} \quad 0 \cdots 0 \quad a_{2r+1} \cdots$$

and transmitting this sequence instead of the original symbol sequence. The r information symbols in each block are transmitted to the first r summing circuits of the correlator at the receiver, while the $N - r$ zero digits in each block are transmitted to the other summing circuits. For convenience, let us refer to the transmission path from the precoder to the jth summing circuit as the jth subchannel. Since there is no information transmission through the last $N - r$ subchannels, it is no longer necessary to consider their performances. The precoding matrix $\mathbf{A}$ can be changed to improve the performance of the first r subchannels (such as reducing the effect of timing error). This can be done by changing the resistive elements in the precoder and correlator.

To summarize, the multiple-speed transmission scheme has the following properties:

(*i*) Changing data rate and overall channel characteristics requires only changing the data format and some resistive elements. There is no change to the filters, the equalization, the signaling interval T, or the receiver sampling time.

(*ii*) With correct timing and the use of orthonormal signals, signal-to-noise ratio is maximized at each data rate for band-limited white noise under the constraints of fixed line signal power and no inter-symbol interference.

(*iii*) By using partial response channels, commonly used data rates are easily obtained, using a physically realizable precoder and correlator.

The discussions so far are general. To show how the method can be applied, and, more important, to demonstrate the advantage of changing the overall channel characteristics when changing the data rate, we consider in detail a two-speed transmission scheme in Section IV.

IV. TWO-SPEED TRANSMISSION AND EYE PATTERNS

Consider the following problem: The transmission medium is equalized for transmission at half the Nyquist rate and

$$\begin{aligned} H(j\omega) &= \text{square root of full-cosine rolloff characteristic} \\ &= k \cos \frac{\pi f}{2f_c}, \qquad |\omega| \leqq 2\pi f_c \qquad (22) \\ &= 0, \qquad \text{otherwise,} \end{aligned}$$

where k is a gain factor and f_c is the bandwidth. It is recognized that, if Scheme I is used, the system is simply the popular full-cosine rolloff system transmitting at half the Nyquist rate.

The channel can be utilized more efficiently if the transmission rate can be changed according to the noise level. To compromise between efficiency and equipment complexity we choose to consider here two-speed transmission and two common data rates, $\frac{1}{2}$ and $\frac{3}{4}$ of the Nyquist rate.

We consider in detail how Scheme II can be used for this purpose. Note that $H(j\omega)$ in (22) is the Class 1 partial response system function. Therefore, from (20) and (21)

$$R = \frac{r}{N+1} R_{\max} \quad \text{bauds}, \tag{23}$$

where r can be any integer from 1 to N. To obtain $\frac{1}{2} R_{\max}$ and $\frac{3}{4} R_{\max}$ from (23), N can be 3, 7, etc. We choose $N = 3$ so that the precoder and correlator can be easily implemented.

To obtain the higher data rate, the sequence of information symbols is divided into blocks with three digits in each block, where the nth block contains the symbols a_{3n+1}, a_{3n+2}, and a_{3n+3}. The blocks are applied to the precoder sequentially at $4T$ intervals. The precoder converts the symbols a_{3n+1}, a_{3n+2}, and a_{3n+3} in the nth block into numbers b_{3n+1}, b_{3n+2}, and b_{3n+3} and transmits b_{3n+i} at $t = (4n + i)T$.

Consider the block containing a_1, a_2, and a_3. The precoder converts a_1, a_2, and a_3 into b_1, b_2, and b_3, and transmits b_1, b_2, and b_3 sequentially at $t = T, 2T, 3T$. This produces, as discussed in the previous section, a signal at the receiver input as

$$\mathbf{X} = b_1\mathbf{h}_1 + b_2\mathbf{h}_2 + b_3\mathbf{h}_3 + \mathbf{N} \tag{24}$$

where the time sample vectors $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ can be written as (omitting a gain factor and the common zero samples)

$$\mathbf{h}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{h}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{h}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}. \tag{25}$$

Equation (25) shows that if sampling time is correct (timing error will be considered later), $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ are limited to a $4T$ time interval.

Since each block is transmitted in a $4T$ time interval, there is no interference between adjacent blocks.

The vectors $\mathbf{h}_1$, $\mathbf{h}_2$, and $\mathbf{h}_3$ generate a three-dimensional real Euclidean vector space $\mathcal{E}_3$. Let

$$\mathbf{V}_1 = \begin{bmatrix} V_{11} \\ V_{12} \\ V_{13} \\ V_{14} \end{bmatrix}, \quad \mathbf{V}_2 = \begin{bmatrix} V_{21} \\ V_{22} \\ V_{23} \\ V_{24} \end{bmatrix}, \quad \mathbf{V}_3 = \begin{bmatrix} V_{31} \\ V_{32} \\ V_{33} \\ V_{34} \end{bmatrix} \tag{26}$$

be a set of orthonormal basis vectors for $\mathcal{E}_3$ and let a_1, a_2, and a_3 be transmitted as coordinates of $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$, respectively. Then the signal $\mathbf{X}$ at the input of the receiver must also be

$$\mathbf{X} = a_1\mathbf{V}_1 + a_2\mathbf{V}_2 + a_3\mathbf{V}_3 + \mathbf{N}. \tag{27}$$

The precoder structure then is

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{21} & \lambda_{31} \\ \lambda_{12} & \lambda_{22} & \lambda_{32} \\ \lambda_{13} & \lambda_{23} & \lambda_{33} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}, \tag{28}$$

where the λ_{ij}'s can be easily determined from (24), (25), (26), and (27). This precoder structure can be easily realized (Fig. 4).

The correlator at the receiver which computes the decision statistics $\mathbf{X}'\mathbf{V}_1$, $\mathbf{X}'\mathbf{V}_2$, and $\mathbf{X}'\mathbf{V}_3$ can also be easily realized (Fig. 3, $j = 1, 2, 3$; $M = 4$).

It is clear from Figs. 3 and 4 that the precoding matrix can be changed by simply changing the resistive elements (the attenuators) in the precoder and the correlator.

The transmission rate is $3/4\ R_{\max}$ when the system operates as above. To change the transmission rate to $1/2\ R_{\max}$, zero digits are inserted into the original data sequence as follows

$$\cdots\ a_1\ a_2\ 0\ a_3\ a_4\ 0\ a_5\ a_6\ 0\ \cdots,$$

and this new sequence is transmitted instead of the original sequence. Making use of the reduced baud rate to improve system performance, the overall channel characteristic is adjusted simultaneously by changing the precoding matrix. This is the subject of the following section.

4.1 *Timing Error and Eye Opening*

So far we have not specified which set of orthonormal basis vectors should be used. This is because with perfect timing the system meets

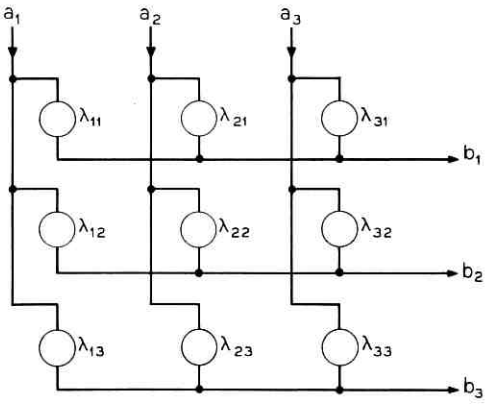


Fig. 4 — Precoder for two-speed transmission where λ_{ij}, $i, j = 1, 2, 3$, are attenuators.

the signal-to-noise ratio criterion in Section III regardless of which set of orthonormal basis vectors is chosen.

However, in practice, it is impossible to achieve zero sampling time error. In general, the receiver will sample the summing circuit outputs at $t = t_0 + \delta$ instead of the correct time t_0, where δ is a random timing error. Then the system's performance depends on the choice of $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$, i.e., depends on the choice of $\mathbf{\Lambda}$. To determine which $\mathbf{\Lambda}$ should be used, it is necessary to specify the type of transmission and choose a performance criterion accordingly.

In the sequel, we consider digital data transmission. Eye opening is adopted as the criterion since it is a widely accepted, practical one[4] (although considering eye openings in the presence of timing error leads to a difficult nonlinear mathematical problem).

Let $r_i(t)$ be the output of the ith summing circuit when an infinite sequence of digits is transmitted at the higher data rate $3/4\ R_{max}$. Then

$$r_i(t) = \sum_{n=-\infty}^{\infty} a_{3n+i} h_{ii}(t - 4nT) + \sum_{j \neq i} \sum_{n=-\infty}^{\infty} a_{3n+j} h_{ji}(t - 4nT), \quad (29)$$

where $h_{ij}(t)$, as defined in (15), is the output of the jth summing circuit when $a_i = 1$ is transmitted alone. From (18) and (22) it can be shown that

$$\begin{aligned} h_{ii}(t) = {} & [\lambda_{i1}\lambda_{j1} + \lambda_{i2}\lambda_{j2} + \lambda_{i3}\lambda_{j3}]I(t) + [\lambda_{i2}\lambda_{j1} + \lambda_{i3}\lambda_{j2}]I(t - T) \\ & + [\lambda_{i1}\lambda_{j2} + \lambda_{i2}\lambda_{j3}]I(t + T) \\ & + \lambda_{i3}\lambda_{j1}I(t - 2T) + \lambda_{i1}\lambda_{j3}I(t + 2T), \end{aligned} \quad (30)$$

where

$$I(t) = \frac{1}{f_c^2} \frac{\sin 2\pi f_c(t - t_0)}{2\pi(t - t_0)[1 - 4f_c^2(t - t_0)^2]}. \tag{31}$$

To evaluate eye opening of $r_i(t)$ at $t_0 + \delta$, we assume that the information digits $\{a_i\}$ are binary and that each can be $\frac{1}{2}$ or $-\frac{1}{2}$ (so that full eye opening = 1). Then

$$\begin{aligned} E_i(\delta) &= \text{Eye opening of } r_i(t) \text{ at } t_0 + \delta \\ &= | h_{ii}(t_0 + \delta) | - \sum_{n=\pm 1}^{\pm\infty} | h_{ii}(t_0 + \delta - 4nT) | \\ &\quad - \sum_{j \neq i} \sum_{n=-\infty}^{\infty} | h_{ji}(t_0 + \delta - 4nT) | \\ &\qquad\qquad i, j = 1, 2, 3. \end{aligned} \tag{32}$$

Similarly, let $r_i'(t)$ be the output of the ith summing circuit when an infinite sequence of digits is transmitted at the lower data rate $1/2\ R_{\max}$. Since zeros are inserted and no information digit is received at the third summing circuit, we need to consider only the eye openings $E_1'(\delta)$ and $E_2'(\delta)$ of $r_1'(t)$ and $r_2'(t)$, respectively.

4.2 *Selection of Precoding Matrix*

It is seen that at the higher data rate, we must consider simultaneously $E_1(\delta)$, $E_2(\delta)$, and $E_3(\delta)$, while at the lower data rate we need only to consider $E_1'(\delta)$ and $E_2'(\delta)$. This suggests that a different precoding matrix should be selected for each data rate.

The steps in selection of the precoding matrix are lengthy and are outlined in the Appendix. The results are summarized here.

The precoding matrix selected for the higher data rate is

$$\begin{aligned} \lambda_{11} &= 0.21, & \lambda_{12} &= 0.62, & \lambda_{13} &= -0.5 \\ \lambda_{21} &= -0.68, & \lambda_{22} &= 0.48, & \lambda_{23} &= -0.68 \\ \lambda_{31} &= -0.5, & \lambda_{32} &= 0.62, & \lambda_{33} &= 0.21. \end{aligned} \tag{33}$$

Eye openings obtained with this precoding matrix are given in Table I for $\delta = 0, \pm 0.1T, \pm 0.2T, \pm 0.3T$ (it is a reasonable expectation that the timing error δ will amount to no more than $\pm 0.2T$). Also given in Table I is the eye opening $E(\delta)$ of the popular "raised cosine" rolloff

TABLE I—COMPARISON OF EYE OPENINGS

Timing Error δ	$E_1(\delta)$	$E_2(\delta)$	$E_3(\delta)$	$E(\delta)$
$-0.3T$	0.312	0.418	0.351	0.312
$-0.2T$	0.559	0.625	0.575	0.551
$-0.1T$	0.790	0.821	0.793	0.783
0	1.000	1.000	1.000	1.000
$0.1T$	0.793	0.821	0.790	0.783
$0.2T$	0.575	0.625	0.559	0.551
$0.3T$	0.351	0.418	0.312	0.312

system[5] which transmits at the same data rate 3/4 R_{max} (i.e., which utilizes a 33.3 percent rolloff band). A glance shows that the eye openings $E_1(\delta)$, $E_2(\delta)$, and $E_3(\delta)$ are equal to or larger than the eye opening $E(\delta)$ of the conventional system.

The precoding matrix selected for the lower data rate is

$$\lambda_{11} = \frac{1}{\sqrt{2}}, \quad \lambda_{12} = 0, \quad \lambda_{13} = 0$$

$$\lambda_{21} = 0, \quad \lambda_{22} = 0, \quad \lambda_{23} = \frac{1}{\sqrt{2}},$$

where λ_{31}, λ_{32}, and λ_{33} can be arbitrary since no information digit is transmitted through the third subchannel. With this precoding matrix, the system is identical with the popular "full cosine" rolloff system at the lower data rate, and the eye openings $E_1'(\delta)$ and $E_2'(\delta)$ are both 1.00, 0.955, 0.896, and 0.823, respectively, for δ equal to 0, $\pm 0.1T$, $\pm 0.2T$, and $\pm 0.3T$. These eye openings are much larger than $E_1(\delta)$ and $E_2(\delta)$ in Table I. This clearly demonstrates the advantage of changing the precoding matrix when changing the transmission rate.

V. CONCLUSIONS

A precoding scheme is presented for multiple-speed digital or analog data transmission. The scheme has the following properties

(*i*) Changing data rate and overall channel characteristics requires only changing the data format and some resistive elements. There is no change to the filters, the equalization, the transmitter signaling interval, or the receiver sampling time.

(*ii*) With correct timing and the use of orthonormal signals, the signal-to-noise ratio is maximized at each data rate for band-limited white noise under the constraints of fixed line signal power and no intersymbol interference.

(*iii*) By using partial response channels, a number of commonly used data rates are easily obtained using a physically realizable precoder and correlator.

Timing error is considered in a two-speed transmission scheme. Eye openings are used as the criterion in selecting the precoding matrix. Eye openings obtained are equal to or larger than those of two conventional schemes transmitting at the same data rates. The study clearly demonstrates the advantage of changing the overall channel characteristics when changing the data rate.

APPENDIX

Selection of Precoding Matrix

As can be seen from (32), (30), and (31), the problem of finding a precoding matrix for maximizing the eye openings in some joint sense over a certain range of the random variable δ is nonlinear and mathematically intractable. In the following, we reduce the dimension and range of the precoding-matrix space $\mathcal{S} = \{\Lambda\}$ to a minimum by using constraints and properties of $\mathcal{S}$, then derive a guide for searching the reduced space. Eye openings are obtained equal to or larger than those of two conventional schemes transmitting at the same data rates.

Consider the higher data rate. The eye openings $E_1(\delta)$, $E_2(\delta)$, and $E_3(\delta)$ are determined by the nine parameters λ_{ij}, $i, j = 1, 2, 3$. We have from orthogonality of $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$

$$h_{ij}(t_0) = 0, \qquad i, j = 1, 2, 3; \qquad i \neq j. \tag{34}$$

Define for $i = 1, 2, 3$

$$W_i = \frac{\lambda_{i1}}{\lambda_{i2}} + \frac{1}{2}; \qquad C_i = \frac{\lambda_{i3}}{\lambda_{i2}} + \frac{1}{2}. \tag{35}$$

It can be shown from (30), (31), and (35) that (34) is equivalent to the constraints

$$C_iC_j = -W_iW_j - \tfrac{1}{2}, \qquad i, j = 1, 2, 3; \qquad i \neq j. \tag{36}$$

Equation (36) is satisfied if and only if one of the following conditions holds

$$(i) \qquad W_1W_2 \leqq -\tfrac{1}{2}, \qquad W_2W_3 \leqq -\tfrac{1}{2}, \qquad W_3W_1 \leqq -\tfrac{1}{2} \tag{37}$$

$$(ii) \qquad W_1W_2 \leqq -\tfrac{1}{2}, \qquad W_2W_3 \geqq -\tfrac{1}{2}, \qquad W_3W_1 \geqq -\tfrac{1}{2} \tag{38}$$

$$(iii) \qquad W_1W_2 \geqq -\tfrac{1}{2}, \qquad W_2W_3 \geqq -\tfrac{1}{2}, \qquad W_3W_1 \leqq -\tfrac{1}{2} \tag{39}$$

$$(iv) \qquad W_1W_2 \geqq -\tfrac{1}{2}, \qquad W_2W_3 \leqq -\tfrac{1}{2}, \qquad W_3W_1 \geqq -\tfrac{1}{2}. \tag{40}$$

Each of the four conditions specifies a subspace of S. Equation (37) corresponds to a null space because its requirements are conflicting. Equations (39) and (40) can be obtained from (38) by rotating the indexes of W; hence, for every point P in the subspace of (39) or (40), there is a point Q in the subspace of (38) such that P and Q produce eye openings differing only in indexes (for instance, P produces $E_1(\delta) = \alpha(\delta)$, $E_2(\delta) = \beta(\delta)$, and $E_3(\delta) = \gamma(\delta)$; Q produces $E_1(\delta) = \gamma(\delta)$, $E_2(\delta) = \alpha(\delta)$, and $E_3(\delta) = \beta(\delta)$). Since they are the same set of eye openings, we need only to cover the subspace of (38) in searching for $\mathbf{A}$.

The subspace of (38) can be further narrowed. It can be shown that (38) holds if and only if

$$W_1W_2 \leqq -\tfrac{1}{2}, \qquad W_2W_3 \geqq 0, \qquad -\tfrac{1}{2} \leqq W_3W_1 \leqq 0 \tag{41}$$

or

$$W_1W_2 \leqq -\tfrac{1}{2}, \qquad -\tfrac{1}{2} \leqq W_2W_3 \leqq 0, \qquad W_3W_1 \geqq 0. \tag{42}$$

Equation (42) can be obtained from (41) by exchanging W_1 and W_2. Thus, for the reason just cited, we need to search only the subspace of (41) instead of that of (38).

To further reduce S, we divide the subspace of (41) into two subspaces

$$(i) \qquad W_1W_2 \leqq -\tfrac{1}{2}, \qquad W_2W_3 \geqq 0, \qquad -\tfrac{1}{2} \leqq W_3W_1 \leqq -\tfrac{1}{4} \tag{43}$$

$$(ii) \qquad W_1W_2 \leqq -\tfrac{1}{2}, \qquad W_2W_3 \geqq 0, \qquad -\tfrac{1}{4} \leqq W_3W_1 \leqq 0. \tag{44}$$

From (36), (44) can be written as

$$C_2C_3 \leqq -\tfrac{1}{2}, \qquad C_1C_2 \geqq 0, \qquad -\tfrac{1}{2} \leqq C_1C_3 \leqq -\tfrac{1}{4}. \tag{44a}$$

It can be shown from (32), (30), and (35) that simultaneously exchanging W_1 and C_1, W_2 and C_2, and W_3 and C_3 does not change the eye openings. From this it can be shown that for every point P in the subspace of (43), there is a point Q in the subspace of (44a) such that P and Q have eye openings differing only in indexes. Since (44a) is equivalent to (44), this implies that only the subspace of (44) needs to be searched instead of that of (41).

The space S to be searched has been reduced to only that of (44). The $W_2 - W_3$ plane is reduced to a narrow strip for all $W_1 \neq 0$. For instance, for $W_1 = -1$, W_3 is bounded between 0 and $\frac{1}{4}$, and W_2 and W_3 are bounded in the very narrow strip shown in Fig. 5.

Each point (W_1, W_2, W_3) in the subspace of (44) determines a precoding matrix through (36), (35), and the orthonormality condition $h_{ii}(t_0) = 1$.

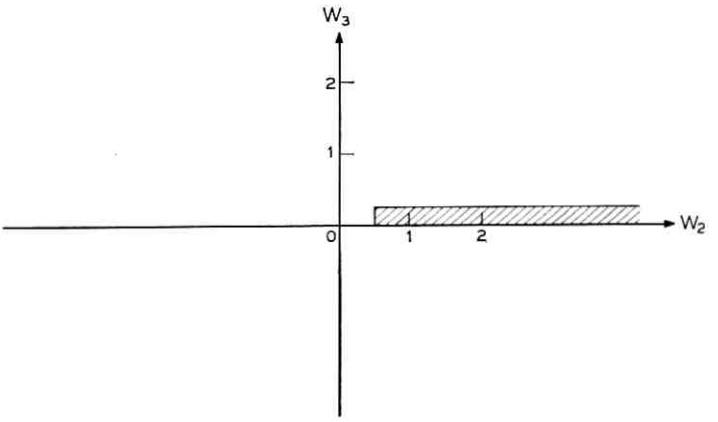


Fig. 5 — Region of W_2 and W_3 (shaded) when $W_1 = -1$.

Usually it is desirable that the three eye openings $E_1(\delta)$, $E_2(\delta)$, and $E_3(\delta)$ be approximately equal. It can be shown from (32), (30), and (31) that $E_1(\delta)$ and $E_3(\delta)$ are approximately equal if

$$W_1 = C_3 , \qquad W_2 = C_2 , \quad \text{and} \quad W_3 = C_1 . \tag{45}$$

Equation (45) defines the following region in the subspace of (44)

$$| W_1 | > \frac{1}{2} , \qquad W_2 = -\frac{2W_1}{4W_1^2 - 1} , \qquad W_3 = -\frac{1}{4W_1} . \tag{46}$$

$E_2(\delta)$ is larger than $E_1(\delta)$ and $E_3(\delta)$ at one extreme of the range $| W_1 | > \frac{1}{2}$, and is smaller at the other extreme. Therefore, in the region of (46), there are points at which $E_1(\delta)$, $E_2(\delta)$, and $E_3(\delta)$ are approximately equal. A simple search of this one-dimensional region gives one of such points as

$$W_1 = 0.84, \qquad W_2 = -0.92, \qquad W_3 = -0.3.$$

This point gives the precoding matrix in (33). Table I in Section IV shows that by using this precoding matrix for the higher data rate, the system has eye openings equal to or larger than the eye openings of a "raised cosine" rolloff system transmitting at the same data rate.

After the precoding matrix in (33) was obtained from the region of (46), the rest of the subspace of (44) was searched. About 5000 points were covered. It was found that no point had eye openings $E_1(\delta)$, $E_2(\delta)$, and $E_3(\delta)$ simultaneously larger than those in Table I.

A similar study for the lower data rate produced the result in Section IV.

REFERENCES

1. Bennett, W. R. and Davey, J. R., *Data Transmission,* McGraw-Hill Book Co., New York, 1965, pp. 99–107.
2. Lender, A., Correlative Level Coding for Binary Data Transmission, IEEE Spectrum, *3,* February, 1966, pp. 104–115.
3. Kretzmer, E. R., Generalization of a Technique for Binary Data Transmission, IEEE Trans. Commun. Tech., *COM-14,* February, 1966.
4. Lucky, R. W., Automatic Equalization for Digital Communication, B.S.T.J., *44,* April, 1965.
5. Bennett, W. R. and Davey, J. R., *Data Transmission,* McGraw-Hill Book Co., New York, 1965, p. 56.

Contributors to This Issue

R. T. Aiken, B.S., 1957, M.S., 1959, Ph.D., 1962, Carnegie Institute of Technology; U. S. Army 1961-1963; Bell Telephone Laboratories, 1963-. Mr. Aiken has engaged in a variety of radar- and communication-theory studies involving the effects of random media. Member, IEEE, Sigma Xi, Tau Beta Pi, Eta Kappa Nu.

W. L. Brown, B.S., 1945, Duke University; M.A., 1948, Harvard University; Ph.D., 1951, Harvard University; Bell Telephone Laboratories, 1950-. Mr. Brown has been concerned with studies of: semiconductor surface states, radiation produced defects in solids, geomagnetically trapped energetic particles in space and channeling of high-energy ions in crystal lattices. He is currently engaged in studies of temporal variations in the high-energy trapped electrons at synchronous satellite altitudes, and the channeling and implantation of heavy ions in solids. Member, Sigma Xi; Fellow, American Physical Society.

Robert W. Chang, B.S.E.E., 1955, National Taiwan University; M.S.E.E., 1960, North Carolina State College; Ph.D., 1965, Purdue University; Bendix Corporation, 1960-1963; Bell Telephone Laboratories, 1965-. Mr. Chang has been concerned with problems in data transmission and communication theory. Member, Eta Kappa Nu, Sigma Xi, Phi Kappa Phi, IEEE.

D. B. Dove, B.Sc., 1953, Ph.D., 1956, Imperial College of Science and Technology, London; Scientific Officer, Atomic Energy Research Establishment, England, 1955-1959; Senior Scientific Officer, 1959; Fellow, National Research Council of Canada 1959-1961; Bell Telephone Laboratories, 1961-. Mr. Dove was a member of the Device Research Department and has specialized in structural and magnetic properties of thin metal films. More recently he was a member of the Electronic Materials Laboratory with specialization in thin film problems. Mr. Dove is presently on leave of absence at the Department of Metallurgy and Materials, College of Engineering, University of Florida, Gainesville, Florida.

Louis H. Enloe, B.S.E.E., 1955, M.S.E.E., 1956, Ph.D., (E.E.), 1959, University of Arizona, Tucson; Bell Telephone Laboratories, 1959-. Mr. Enloe served as an Instructor in Electrical Engineering and as a member of the technical staff of the Applied Research Laboratory of the University of Arizona from 1956 to 1959. His work was primarily in transistor circuitry. Since 1959 he has been in the research division of Bell Telephone Laboratories. His early work was in modulation and noise theory in connection with space communications. Later work has been with lasers, coherent light, and holography with emphasis upon communication and display. He is presently Head, Opto-Electronics Research Department. Member, IEEE, Phi Kappa Phi, Sigma Xi, Tau Beta Pi, Pi Mu Epsilon, Sigma Pi Sigma.

John D. Gabbe, B.A., 1950, New York University; M.S., 1951, University of Illinois; Ph.D., 1957, New York University; Bell Telephone Laboratories, 1956-. Mr. Gabbe was first associated with the *Picturephone*® project, then with studies of the earth's magnetosphere. At present, he is engaged in research concerning the methodology of data analysis. Member, American Physical Society.

Karl R. Gardner, B.S. (Eng. Physics), 1960, M.S., (Physics), 1962, University of Illinois; Bell Telephone Laboratories 1962-. Mr. Gardner has been engaged in the development of several miniature silicon diodes. He is presently working on silicon integrated circuits. Member, Tau Beta Pi, American Physical Society.

William C.-Y. Lee, B.Sc. in Engineering, 1954, Chinese Naval Academy; M.Sc., E.E., 1960, and Ph.D. in E.E., in 1963, The Ohio State University; Bell Telephone Laboratories, 1964-. Mr. Lee has been concerned with the study of wave propagation in anisotropic medium and antenna theory. His present work has included studies of mobile radio antennas and signal fading problems. Member, Sigma Xi, IEEE.

James McKenna, B.Sc. (Math), 1951, Massachusetts Institute of Technology; Ph.D., (Math), 1961, Princeton University; Bell Telephone Laboratories, 1960-. Mr. McKenna has done research in quantum mechanics and classical electromagnetic theory. At present he is involved in a study of optical waveguides.

T. R. ROBILLARD, B. Physics, 1949, University of Minnesota; M.S., 1952, University of Illinois; Bell Telephone Laboratories, 1954-. Mr. Robillard has been engaged in the development of a variety of transistors, semiconductor diodes and integrated devices at both the Reading and the Allentown Laboratories. At present, he is supervisor at the Reading Laboratory responsible for the development of silicon integrated circuits. Member, Sigma Xi, Phi Beta Kappa, American Physical Society.

DAN VARON, B.S., 1957, Dipl. Ing. 1961, The Technion—Israel Institute of Technology; M.S. Electrophysics, 1963, Polytechnic Institute of Brooklyn; Eng. Sc.D., 1965, New York University; Israeli Air Force, 1957-1961; Bell Telephone Laboratories, 1965-. Mr. Varon has been engaged in applications of electromagnetic theory to studies of phased array antennas and microwave transmission devices. Member, IEEE, Eta Kappa Nu, Sigma Xi.

M. B. WILK, B. Eng. (Chem.), 1945, McGill University; M.S. (Statistics), 1953, Iowa State University; Ph.D. (Statistics), 1955, Iowa State University; National Research Council of Canada (Atomic Energy Project), 1945-1950; Iowa State University, 1951-1955; Princeton University, 1955-1957; Rutgers University, 1959-1963; Bell Telephone Laboratories, 1956-. Mr. Wilk has been involved in research into statistical methods and theory and applications in a variety of scientific areas. Presently, he is Head, Statistics and Data Analysis Research Department. Member, American Statistical Association (Fellow), Institute of Mathematical Statistics, Royal Statistical Society, Biometric Society, International Association for Statistics in the Physical Sciences.

G. I. ZYSMAN, B.E.E., 1959, Cooper Union; M.S.E.E., 1962, Ph.D. (Electrophysics), 1966, Polytechnic Institute of Brooklyn; Hazeltine Corp., N. Y., 1959-1960; Polytechnic Institute of Brooklyn, 1965-1966; Bell Telephone Laboratories, 1966-. Mr. Zysman has been concerned with the study of microwave circuits and phased array antennas. Member, IEEE.

B. S. T. J. BRIEFS

Axis-Crossing Intervals of Sine Wave Plus Noise

By A. J. RAINAL

I. INTRODUCTION

Let $I(t, a)$ denote the stationary random process consisting of a sinusoidal signal of amplitude $\sqrt{2a}$ and angular frequency q plus Gaussian noise, $I_N(t)$, of zero mean and unit variance. Thus,

$$I(t, a) = \sqrt{2a} \cos (qt + \theta_0) + I_N(t). \quad (1)$$

θ_0 denotes a random phase angle which is distributed uniformly in the interval $(-\pi, \pi)$. "a" denotes the signal-to-noise power ratio. When $a = 0$ Rice[1] presented some theoretical results which are very useful for studying statistical properties of the axis-crossing intervals and the axis-crossing points of $I(t, 0)$ at an arbitrary level I. The axis-crossing intervals and the axis-crossing points of $I(t, a)$ are defined in Fig. 1. In recent work Cobb[2] presented some theoretical results concerning the zero-crossing intervals, the axis-crossing intervals defined by the level $I = 0$, of $I(t, a)$. Some experimental and theoretical results concerning the zero-crossing intervals of $I(t, a)$ were reported by Rainal.[3] For the case when the power spectral density of $I_N(t)$ is narrow-band and symmetrical about the sine wave frequency, Blachman[4] presented some

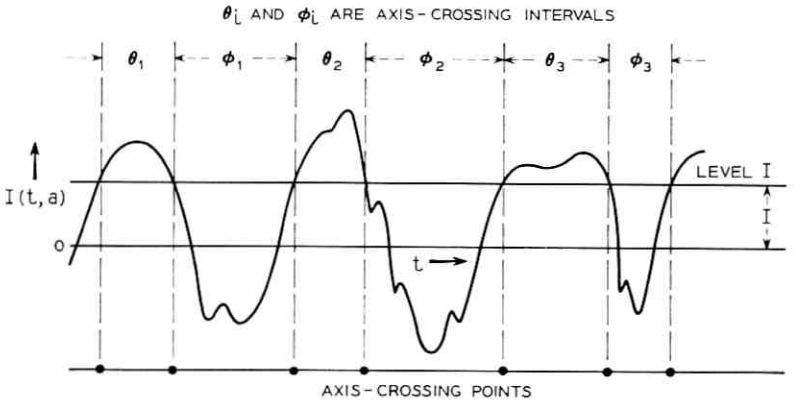


Fig. 1 — The level I defines the axis-crossing points and the axis-crossing intervals of $I(t, a) = \sqrt{2a} \cos (qt + \theta_0) + I_N(t)$.

theoretical results concerning the zero-crossing points, the axis-crossing points defined by the level $I = 0$, of $I(t, a)$.

The purpose of this brief is to present some theoretical results which are useful for studying statistical properties of the axis-crossing intervals and the axis-crossing points of $I(t, a)$ at an arbitrary level I. These results stem from a straightforward extension of Rice's[1] analysis.

II. THEORETICAL RESULTS

Using a notation consistent with Refs. 5 and 6 we define the following probability functions at an arbitrary level I and arbitrary signal-to-noise power ratio "a":

(*i*) $Q_2^+(\tau, I, a)d\tau$, the conditional probability that a downward axis-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given an upward axis-crossing at t.

(*ii*) $Q_2^-(\tau, I, a)d\tau$, the conditional probability that an upward axis-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given a downward axis-crossing at t.

(*iii*) $[U_2(\tau, I, a) - Q_2(\tau, I, a)]d\tau$, the conditional probability that an upward axis-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given an upward axis-crossing at t.

This latter conditional probability is also equal to the conditional probability that a downward axis-crossing occurs between $t + \tau$ and $t + \tau + d\tau$ given a downward axis-crossing at t.

The reader should refer to Rice[1] for the definition of all notation which is not defined in this brief. When $a \geqq 0$, Rice's[1] (38) becomes

$$Q_2^+(\tau, I, a) = -[2\pi N_I]^{-1} \int_{-\pi}^{\pi} d\theta \int_0^{\infty} dI_1' \int_{-\infty}^{0} dI_2' I_1' I_2' p(I, I_1', I_2', I), \quad (2)$$

where N_I = Rice's[7] equation (2.6) or (2.7)

$$p(I, I_1', I_2', I) = (2\pi)^{-2} M^{-\frac{1}{2}}$$

$$\cdot \exp\left\{-\frac{1}{2M}\left[M_{22}(I_1'^2 + I_2'^2) + 2M_{22}r_1 I_1' I_2' + 2D_1 I_1' + 2E_1 I_2' + F_1\right]\right\}$$

$$r_1 = \frac{M_{23}}{M_{22}} \qquad Q = \sqrt{2a}$$

$$D_1 = M_{12}[I - Q\cos\theta] + M_{13}[Q\cos(q\tau + \theta) - I]$$
$$+ M_{22}Qq\sin\theta + M_{23}Qq\sin(q\tau + \theta)$$

$$E_1 = M_{12}[Q\cos(q\tau+\theta) - I] + M_{13}[I - Q\cos\theta] + M_{22}Qq\sin(q\tau+\theta) + M_{23}Qq\sin\theta$$

$$F_1 = M_{11}\{2I^2 - 2QI[\cos\theta + \cos(q\tau+\theta)] + Q^2[\cos^2\theta + \cos^2(q\tau+\theta)]\}$$
$$+ 2M_{12}Qq\{[I - Q\cos\theta]\sin\theta + [Q\cos(q\tau+\theta) - I]\sin(q\tau+\theta)\}$$
$$+ 2M_{13}Qq\{[I - Q\cos\theta]\sin(q\tau+\theta) + [Q\cos(q\tau+\theta) - I]\sin\theta\}$$
$$+ 2M_{14}\{I[I - Q\cos\theta] + Q[Q\cos\theta - I]\cos(q\tau+\theta)\}$$
$$+ M_{22}(Qq)^2[\sin^2\theta + \sin^2(q\tau+\theta)] + 2M_{23}(Qq)^2\sin\theta\sin(q\tau+\theta).$$

The M's are given in Rice's[1] Appendix I with

$$m(\tau) = \int_0^\infty W(f)\cos 2\pi f\tau\, df, \tag{3}$$

where $W(f)$ = one-sided power spectral density of $I_N(t)$. When $I = 0$, $N_I Q_2^+(\tau, I, a)$ is equivalent to (9) of Cobb's[2] recent work.

Equation (2) can be put in a form analogous to Rice's[1] equation (47):

$$Q_2^+(\tau, I, a) = [4\pi^2 N_I]^{-1} M_{22}(1 - m^2)^{-\frac{1}{2}} \cdot \int_{-\pi}^{\pi} \exp(-G_1/2M) J(r_1, h_2, k_2)\, d\theta, \tag{4}$$

where

$$J(r_1, h_2, k_2) \equiv \frac{1}{2\pi\sqrt{1 - r_1^2}} \int_{h_2}^\infty dx \int_{k_2}^\infty dy (x - h_2)(y - k_2) e^z$$

$$z = -\frac{x^2 + y^2 - 2r_1 xy}{2(1 - r_1^2)}$$

$$h_2 = M_{22}^{-1}[1 - r_1^2]^{-1}[D_1 - r_1 E_1]\left[\frac{1 - m^2}{M_{22}}\right]^{\frac{1}{2}}$$

$$k_2 = -M_{22}^{-1}[1 - r_1^2]^{-1}[E_1 - r_1 D_1]\left[\frac{1 - m^2}{M_{22}}\right]^{\frac{1}{2}}$$

$$G_1 = M_{22}^{-1}[1 - r_1^2]^{-1}[2r_1 D_1 E_1 - D_1^2 - E_1^2] + F_1 .$$

$Q_2^-(\tau, I, a)$ is obtained from (2) by changing the signs of the ∞'s in the limits of integration. We find that $Q_2^-(\tau, I, a)$ is equal to the right-hand side of (4) with h_2, k_2 replaced by $-h_2$, $-k_2$.

$[U_2(\tau, I, a) - Q_2(\tau, I, a)]$ is obtained from (2) by changing the lower limit of integration of I_2' to $+\infty$. We find that $[U_2(\tau, I, a) - Q_2(\tau, I, a)]$

is equal to the right-hand side of (4) with the function $J(r_1, h_2, k_2)$ replaced by the function $J_1(r_1, h_2, k_2)$, where

$$J_1(r_1, h_2, k_2) \equiv \frac{1}{2\pi\sqrt{1 - r_1^2}} \int_{h_2}^{-\infty} dx \int_{k_2}^{\infty} dy(x - h_2)(y - k_2)e^z. \quad (5)$$

The functions $J(r_1, h_2, k_2)$ and $J_1(r_1, h_2, k_2)$ are expressed in terms of Karl Pearson's well-known tabulated function (d/N) in Ref. 5.

REFERENCES

1. Rice, S. O., Distribution of the Duration of Fades in Radio Transmission, B.S.T.J., *37*, May, 1958, pp. 581–635.
2. Cobb, S. M., The Distribution of Intervals Between Zero Crossings of Sine Wave Plus Random Noise and Allied Topics, IEEE Trans. Inform. Theor., *IT-11*, April, 1965, pp. 220–233.
3. Rainal, A. J., Zero-Crossing Intervals of Random Processes, Technical Report. No. AF-102, DDC No. AD-401-148, The Johns Hopkins University, Carlyle Barton Laboratory, Baltimore, Maryland, April, 1963. Abstracted in IEEE Trans. Inform. Theor., *IT-9*, Oct. 1963, p. 295.
4. Blachman, N. M., FM Reception and the Zeros of Narrowband Gaussian Noise, IEEE Trans. Inform. Theor., *IT-10*, July, 1964, pp. 235–241.
5. Rainal, A. J., Axis-Crossing Intervals of Rayleigh Processes, B.S.T.J., *44*, July-August, 1965, pp. 1219–1224.
6. Rainal, A. J., Axis-Crossings of the Phase of Sine Wave Plus Noise, B.S.T.J., *46*, April 1967, pp. 737–754.
7. Rice, S. O., Statistical Properties of a Sine Wave Plus Random Noise, B.S.T.J., *27*, January, 1948, pp. 109–157.