ARO Report 78-1

(12)

TRANSACTIONS OF THE TWENTY-THIRD CONFERENCE OF ARMY MATHEMATICIANS

Sponsored by

The Army Mathematics Steering Committee

on behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT

AND ACQUISITION

US ARMY RESEARCH OFFICE

Report No. 78-1

February 1978

TRANSACTIONS OF THE TWENTY-THIRD CONFERENCE OF ARMY MATHEMATICIANS

Sponsored by the Army Mathematics Steering Committee

Host

U.S. Army Mobility Research and Development Laboratory

Langley Research Center, Hampton, Virginia

11-13 May 1977

US Army Research Office

PO Box 12211

Research Triangle Park, North Carolina

FOREWORD

The regularity observed in natural processes can often be expressed and explained in mathematical language. Galileo put forth this idea in stating that the language of science is mathematics. In many scientific fields mathematical descriptions have provided the tools for fundamental advances and important discoveries. These conferences, which are sponsored by the Army Mathematics Steering Committee (AMSC), provide an opportunity for engineers, physicists and other scientists to get together with mathematicians to originate conceptual and analytic tools to treat problems in these various fields.

As in previous meetings, the Twenty-Third Conference of Army Mathematicians gave its attendees a chance to see the developments taking place in the various Army laboratories. The Army scientist's contributions covered a broad spectrum of scientific areas. Through these meetings, techniques developed at one installation are brought to the attention of scientists at other places, thus reducing duplication of effort. Another important phase of these meetings is presenting the members of the audience an opportunity to hear nationally known scientists discuss recent developments in their own fields. This year there were five invited speakers who gave addresses in the areas listed below, and who were more than willing to discuss various problems with scientists in the Army agencies.

Speaker and Institution	Area of Talk
Prof. M.D. Kruskal Princeton University	What's All This About Solitons
Prof. D.H. Sattinger University of Minnesota	Group Theoretic Methods in Bifurcation Theory
Prof. Mike Crandall Mathematics Research Center	Evolution Governed by Accretive Operators
Prof. H.O. Kreiss Uppsala University, Sweden (Visiting NYU)	Numerical Solution of Problems with Different Time Scales
Prof. Edward Kamen Georgia Institute of Technology	Use of Algebraic Methods in the Design of Controllers and Observers for Systems with Time Delays

The success of the the benefits derived from these conferences depend a great deal on the host installation. This year we were pleased to have the U.S. Army Air Mobility Research and Development Laboratory, Langley Directorate, Hampton, Virginia, serve in this capacity. Those in attendance were fortunate to have Mr. Robert L. Tomaine as Chairman on Local Arrangements. He, together with members of his staff, provided all those things, such as projection equipment, travel information, etc., needed for a successful symposium.

The AMSC is pleased to be able to provide the Proceedings of this Conference. It is hoped the scientific ideas contained herein will benefit, not only those who were able to attend the conference, but also many others that did not enjoy that privilege.

TABLE OF CONTENTS*

*This table of contents contains only the papers that are published in this technical manual. For a list of all papers presented at the Twenty-second Conference of Army Mathematicians, see the Program of the meeting.

PROGRAM

THE 23rd CONFERENCE OF ARMY MATHEMATICIANS

Langley Directorate, USAAMRDL, NASA-Langley Research Center
Hampton, Virginia

General Sessions and Technical Sessions I, III, V, and VII will be held in Room 200 on the second floor of Bldg. 1212. Technical Sessions II, IV, VI, and VIII will be held in Room 185 on the first floor of Bldg. 1192D.

Wednesday, 11 May 1977

0800 BUS FROM HOLIDAY INN TO NASA-LANGLEY RESEARCH CENTER

0815-0845 REGISTRATION - ROOM 200, 2nd FLOOR, BLDG 1212, LANGLEY RESEARCH CENTER

0845-0900 OPENING REMARKS - ROOM 200, BLDG 1212

0900-1000 GENERAL SESSION I - ROOM 200, BLDG 1212

CHAIRPERSON: Dr. Francis G. Dressel
US Army Research Office
Research Triangle Park, NC

SPEAKER: Professor M. D. Kruskal
Princeton University
Princeton, New Jersey

TITLE: What's All This About Solitons

1000-1020 BREAK

Wednesday AM

1020-1200 TECHNICAL SESSION I - ROOM 200 - BLDG 1212

CHAIRPERSON: Mr. Joseph M. Kirshner
Harry Diamond Laboratories
Adelphi, Maryland

SOLITON PERTURBATION OF FLUXON DYNAMICS
Alwyn C. Scott, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

A MODEL FOR NORMAL MODE EXCITATION OF MOLECULES SUBJECTED TO AN INTENSE INFRARED LASAR RADIATION FIELD
R. A. Shatas, L. M. Narducci, C. A. Coulter, and S. S. Mitra, US Army Missile Research and Development Command, Redstone Arsenal, Alabama

NONNEUTRAL PLASMA EXPANSION
K. E. Lonngren, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

THE MATHEMATICL DESCRIPTION OF A MOVING BOUNDARY PROBLEM IN AN ELLIPTIC-PARABOLIC SYSTEM OF PARTIAL DIFFERENTIAL EQUATIONS IN THE HYDRODYNAMICS OF POROUS MEDIA
Yoshisuke Nakaro, US Army Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire

CRITICAL REVIEW OF ONE-DIMENSIONAL TUBE FLOW EQUATIONS
Aivars Celmins, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

1020-1200 TECHNICAL SESSION II - ROOM 185 - BLDG 1192D

CHAIRPERSON: Dr. Dean J. Weidman
NASA-Langley Research Center
Hampton, Virginia

FINITE ELEMENT STRESS ANALYSIS OF AXISYMMETRIC BODIES UNDER TORSION
Tien-Yu Tsui, Army Materials and Mechanics Research Center, Watertown, Mass.

THREE-DIMENSIONAL ELASTIC STRESS AND DISPLACEMENT ANALYSIS OF TENSILE FRACTURE SPECIMENS CONTAINING CRACKS
Jon Kring, USAAMRDL, Lewis Directorate; John Gyekenyesi and Alexander Mendelson, NASA-Lewis Research Center, Cleveland, Ohio

FULLY PLASTIC DEFORMATION IN ANISOTROPIC ANNULAR PLATES UNDER INTERNAL PRESSURE
P. C. T. Chen, Benet Weapons Laboratory, Watervliet, New York

COMPUTER SIMULATION OF SHOCK PROPAGATION IN THE ONE-DIMENSIONAL LATTICE
John D. Powell and Jad H. Batteh, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

Wednesday AM

A PERTURBATION EXPANSION OF THE NAVIER-STOKES EQUATIONS FOR SHOCK WAVES
John D. Powell and Jad H. Batteh, US Army Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

Wednesday PM

1200-1330 LUNCH (NASA Cafeteria, Bldg 1213)

1330-1510 TECHNICAL SESSION III - ROOM 200, BLDG 1212

CHAIRPERSON: Dr. Edward W. Ross, Jr.
US Army Natick R&D Command
Natick, Maryland

A MONOTONE METHOD FOR NONLINEAR BOUNDARY VALUE PROBLEMS IN ARBITRARY BANACH SPACES
S. R. Bernfeld, V. Lakshmikantham and S. Leela, University of Texas, Arlington, Texas

COMPARISON THEOREMS FOR SECOND-ORDER LINEAR DIFFERENTIAL EQUATIONS
Leon Katon, US Army Electronics Command, Fort Monmouth, New Jersey

THE COLLAPSED CUBIC ISOPARAMETRIC ELEMENT AS A SINGULAR ELEMENT FOR CRACK PROBLEMS
S. L. Pu, M. A. Hussain and W. E. Lorensen, Benet Weapons Laboratory, Watervliet, New York

BOUNDARY CONDITION SOLUTIONS OF THE GENERALIZED FELLER EQUATION
Siegfried H. Lehnigk, US Army Missile Research and Development Command, Redstone Arsenal, Alabama

BIVARIATIONAL BOUNDS
Peter D. Robinson, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

1330-1510 TECHNICAL SESSION IV - ROOM 185 - BLDG 1192D

CHAIRPERSON: Mr. T. N. E. Greville
Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

SOME GENERIC PROPERTIES OF A LOGIC MODEL FOR ANALYZING HARDWARE MAINTENANCE AND DESIGN CONCEPTS
James T. Wong and William L. Andre, USAAMRDL, Ames Research Center, Moffett Field, California

Wednesday PM

COMPUTER GRAPHICS IN A PRODUCTION ENVIRONMENT
William D. Johnston, National Range Operations Directorate, US Army White Sands Missile Range, New Mexio

MATHEMATICAL TRADE-OFFS FOR MANAGERIAL CONTROL
John L. Lazaruk, US Army Communications Command, Fort Huachuca, Arizona

RADAR CROSS-SECTION DATA REDUCTION
Ernest J. Sanchez, National Range Operations Directorate, US Army White Sands Missle Range, New Mexio

1510-1530 BREAK

1530-1630 GENERAL SESSION II - ROOM 200 - BLDG 1212

CHAIRPERSON: Dr. Ben Noble
Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

SPEAKER: Professor Heinz Otto Kreiss
Courant Institute of Mathematical Sciences
New York University
New York, New York

TITLE: NUMERICAL SOLUTION OF PROBLEMS WITH DIFFERENT TIME SCALES

****

Thursday, 12 May 1977

0900-1040 TECHNICAL SESSION V - ROOM 200 - BLDG 1212

CHAIRPERSON: Dr. William F. White
Langley Directorate, USAAMRDL
NASA-Langley Research Center
Hampton, Virginia

A NONLINEAR SINGULARLY PERTURBED VOLTERRA INTEGRODIFFERENTIAL EQUATION OCCURING IN POLYMER RHEOLOGY
A. L. Lodge, J. B. McLeod and J. A. Nohel, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

A COMPARISON OF NUMERICAL TECHNIQUES USED IN TRANSIENT ROTOR DYNAMICS
Albert F. Kascak, Lewis Directorate, USAAMRDL, Lewis Research Center, Cleveland, Ohio

VIBRATIONS OF A HELICOPTER ROTOR BLADE USING FINITE ELEMENT-UNCONSTRAINED VARIATIONAL FORMULATIONS
J. J. Wu and C. N. Shen, Benet Weapons Laboratory, Watervliet, New York

EFFECTS OF DAMPING AT THE SUPPORT OF A ROTATING BEAM ON VIBRATIONS
J. D. Vasilakis and J. J. Wu, Benet Weapons Laboratory, Watervliet, New York

AN ELEMENTARY METHOD FOR BOUNDING THE ERROR IN AN APPROXIMATE EIGENSYSTEM OF A MATRIX, WITH AN APPLICATION TO A DYNAMICAL SYSTEM
Ben Noble, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

0900-1040 TECHNICAL SESSION VI - ROOM 185 - BLDG 1192D

CHAIRPERSON: Dr. Shunsuke Takagi
US Army Cold Regions
R&E Laboratory
Hanover, New Hampshire

AN EVALUATION PROCEDURE FOR INCOMPLETE GAMMA FUNCTIONS
W. Gautschi, Mathematics Research Center, University of Wisconsin, Madision, Wisconsin

A METHOD OF EVALUATING LAPLACE TRANSFORMS WITH SERIES OF COMPLETE OR INCOMPLETE GAMMA FUNCTIONS
Alexander S. Elder and Emma M. Wineholt, US Army Ballistics Research Laboratory, Aberdeen Proving Ground, Maryland

APPROXIMATION OF IRREGULAR SURFACES
H. M. Sassenfeld, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, New Mexio

EVALUATION OF PADE' AND OTHER DENSITY FUNCTIONS IN DETERMINING RELIABILITY OF TURBINE STATOR VANES
R. Beeuwkees, E. Lenoe and D. M. Neal, The Army Materials and Mechanics Research Center, Watertown, MA

1040-1100 BREAK

1100-1200 GENERAL SESSION III - ROOM 200, BLDG 1212

CHAIRPERSON: Mr. Romas Shatas
US Army Missile Command
Redstone Arsenal, Alabama

Thursday AM

SPEAKER: Professor David H. Sattinger
University of Minnesota
Minneapolis, Minnesota

TITLE: GROUP THEORETIC METHODS IN BIFURCATION THEORY

Thursday PM

1200-1330 LUNCH (NASA Cafeteria - Bldg 1213)

1330-1430 GENERAL SESSION IV - ROOM 200 - BLDG 1212

CHAIRPERSON: Dr. Siegfried Lehnigk
US Army Missle Command
Redstone Arsenal, Alabama

SPEAKER: Professor Michael Crandall
Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

TITLE: EVOLUTION GOVERNED BY ACCRETIVE OPERATORS

1430-1500 BREAK

1500-1600 TOUR OF NASA VISITORS CENTER

****

Friday, 13 May 1977

0900-1040 TECHNICAL SESSION VII - ROOM 200 - BLDG 1212

CHAIRPERSON: Dr. Aivars K. Celmins
Ballistics Research Laboratories
Aberdeen Proving Ground, MD

HARMONIC FUNCTIONS ON REGIONS WITH REENTRANT CORNERS
J. Barkley Rosser, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

ADAPTIVE ACCELERATION OF SSOR FOR SOLVING LARGE LINEAR SYSTEMS
Vitalius Benokraitis, Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

APPLICATION OF MACSYMA IN THE SOLUTION OF BOUNDARY VALUE PROBLEMS
E. Cuthill and Kent Meals, David W. Taylor Naval Ship Research and Development Center, Bethesda, Maryland

THE INFLUENCE OF BOUNDARY MODELING ON THE NUMERICAL STABILITY OF A NONLINEAR FINITE DIFFERENCE PROGRAM
J. M. Santiago and J. D. Wortman, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

0900-1040 TECHNICAL SESSION VIII - ROOM 185 - BLDG 1192D

CHAIRPERSON: Mr. Alexander S. Elder
Ballistics Research Laboratory
Aberdeen Proving Ground, Maryland

MOVING-WEIGHTED-AVERAGE SMOOTHING CONTINUED TO THE EXTREMITIES OF THE DATA
T. N. E. Greville, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

OPTIMIZATION OF THE MEMORY CAPACITY OF A STORE AND FORWARD RELAY
W. Pressman and J. Benson, Communications/Automatic Data Processing Laboratory, Fort Monmouth, New Jersey

NUMERICAL CALCULATION OF THE SOLUTION OF THE VISCOELASTIC DEFORMATION OF AN INFINITELY-WIDE FLOATING ICE PLACED UNDER A CIRCULAR LOAD
Shunsuke Takagi, US Army Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire

STATISTICAL MODELS OF TIRE WEAR
Edward Saibel, US Army Research Office, Durham, North Carolina

1040-1100 BREAK

1100-1200 GENERAL SESSION V - ROOM 200 - BLDG 1212

CHAIRPERSON: Dr. Walter Pressman
US Army Electronics Command
Fort Monmouth, New Jersey

SPEAKER: Professor Edward W. Kamen
Georgia Institute of Technology
Atlanta, Georgia

TITLE: USE OF ALGEBRAIC METHODS IN THE DESIGN OF CONTROLLERS AND OBSERVERS FOR SYSTEMS WITH TIME DELAYS

1215 ADJOURN

NONNEUTRAL PLASMA EXPANSION

K.E. Lonngren*
Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

ABSTRACT

The self similar expansion of nonneutral plasmas is examined using various models. Analytical solutions are obtained for: (1) The expansion of a Maxwellian electron cloud governed by Ohm's Law; (2) The expansion of an electron cloud using a mobility model; (3) The expansion of an electron cloud using a cold plasma model; and (4) The expansion of a charge particle cloud with a temporally decaying nonlinear diffusion coefficient.

AMS(MOS) Subject Classification: 78.35

Key Words: Self similar solution of nonneutral plasma expansion

Work Unit Number 3 - Applications of Mathematics

* On leave from the University of Iowa.

Sponsored by
the United States Army under Contract No. DAAG29-75-C-0024 and
the National Science Foundation Grant #ENG 76-15645.

NONNEUTRAL PLASMA EXPANSION

K. E. Lonngren

In a previous report,[1] we presented a methodical procedure for obtaining the self similar variables using some of the techniques of Lie Group Theory. General physical phenomena which occur in plasmas and could be modeled with equations amenable to a self-similar treatment were presented along with their solutions. An extensive list of references was also given. The purpose of this report is to summarize a pot-pourri of further examples which fall under the umbrage of "Nonneutral Plasma Expansion".

With the increased interest in relativistic electron beam fusion devices, considerable attention has been given to the study of nonneutral plasmas. The recent monograph by Davidson[2] summarizes the past work and motivates our interest in examining various aspects of this problem. In this report, we shall examine three sets of fluid equation models and using the technique of "Self Similar Solution of Partial Differential Equations, obtain analytical solutions.

In addition, a recent experiment on the Wisconsin Multipole suggests that the cross-field diffusion coefficient can be modeled with a one-dimensional diffusion equation where the diffusion coefficient is nonlinear and exponentially decaying in time.[3] A self-similar solution of this problem shall also be given.

In Section II, we describe the various models and present the self-similar solutions. The models are: (1) The expansion of a Maxwellian electron cloud governed by Ohm's Law; (2) The expansion of an electron cloud using a mobility model; (3) The expansion of an electron cloud using a cold plasma model; and (4) The expansion of a charge particle cloud with a temporally decaying nonlinear diffusion coefficient. Section III is the conclusion.

II. Various Problems

To analyze the problems, we follow the procedure given in reference 1 and only describe the physical phenomena, list the PDE, the self similar variables, the ODE and the solution to the ODE without repeating the details of the procedure in each case.

1) Expansion of a Maxwellian electron cloud governed by Ohm's Law[4]

In this example, we examine the expansion into a vacuum of a thermalized electron cloud, described by an isothermal Maxwellian distribution. It is assumed that collisions are sufficiently frequent such that we can speak of a conductivity for the medium.

The governing equations are:

I) Equation of continuity

$$j_x + \rho_t = 0 \tag{1}$$

II) Ohm's Law

$$j = -\sigma\phi_x \quad \text{and} \tag{2}$$

III) Maxwellian electrons

$$\rho = n_0 q \exp[q\phi/k_B T_e] \tag{3}$$

where all symbols are standard. By differentiating (3) with respect to x and substituting (2), we obtain

$$j = -\sigma \frac{k_B T_e}{q} \frac{1}{\rho} \rho_x \tag{4}$$

In normalized units, (4) and (1) lead to

$$\left(\frac{1}{\psi} \psi_y\right)_y - \psi_\tau = 0 \tag{5}$$

where

$$\psi = \rho/n_0 q, \quad y = x/\lambda_D, \quad \tau = t/(\varepsilon_0/\sigma)$$

and λ_D is the Debye length and ε_0/σ is a relaxation time.

We find that (5) admits a self-similar solution of the form

$$\eta(\xi) = \tau\psi(y,\tau), \ \xi = y/\tau \ . \tag{6}$$

The solution satisfies the conservation law that electron charge is conserved in space.

The boundary conditions which are germane to this problem are:

a) $\psi(0,\tau) = \psi_0/\tau \rightarrow \eta(\xi=0) = \psi_0$

b) $\psi(\infty,\tau) = 0 = \psi(y,0) \rightarrow \eta(\xi=\infty) = 0$ i.e. "consolidation". (7)

Substituting (6) into (5), we obtain

$$[\frac{1}{\eta}\,\eta_\xi]_\xi = -[\xi\eta]_\xi \quad . \tag{8}$$

The first integral of (8) is

$$-\xi\eta + C_1 = \frac{1}{\eta}\,\eta_\xi \quad . \tag{9}$$

We shall further impose the condition that the current $j \rightarrow 0$ as $y \rightarrow \infty$. From (4), this transforms to $\frac{1}{\eta}\,\eta_\xi \rightarrow 0$ as $\xi \rightarrow \infty$. Using this and (7b), $C_1 = 0$.

The integral of (9) is

$$\eta = \frac{1}{\xi^2/2 + C_2} \tag{10}$$

The constant C_2 is determined from (7a) to be $C_2 = 1/\psi_0$. In terms of y and τ , the final result is

$$\psi = \frac{\rho}{n_0 q} = \frac{1}{\tau[y^2/2\tau^2 + 1/\psi_0]} \quad . \tag{11}$$

Charge is conserved as is shown below.

$$\int_0^\infty \frac{\rho}{n_0 q}\,dy = \frac{1}{n_0 q}\int_0^\infty \frac{d\xi}{[\xi^2/2 + 1/\psi_0]} = \frac{\pi}{n_0 q}\sqrt{\frac{\psi_0}{2}} = \text{const.} \tag{12}$$

In conclusion, we have examined the expansion of an electron cloud in a vacuum. Under conditions where Coulomb forces can be neglected, this calculation could model an electron cloud expansion in a plasma in time scales short with respect to ion motion.

2) Expansion of an electron cloud using a mobility model[5]

Recently, considerable attention has been given to the problem of the transient behavior of the bulk electric field and space charge distribution in semi-conductors and in the conduction in dielectric and insulating fluids. It has been found prudent to use a

mobility model where the velocities of charge carriers injected from an emitting electrode are proportional to the electric field through their mobilities and the electric field is related to the charge densities of the carriers through Gauss's law. Many and Rakavy [6] and Helfrich and Mark[7] were probably the first to suggest that the problem could be modeled with the set of dimensionless equations

$$\begin{aligned} E_x &= \rho \\ i_x + \rho_t &= 0 \\ i &= \rho E \end{aligned} \tag{13}$$

which are Poisson's equation, the equation of continuity and a mobility definition for current respectively. The subscript x and t denote a partial differentiation with respect to space and time.

In their original paper, Many and Rakavy[6] obtained a solution to (13) by looking for the characteristics of the problem. Subsequently, this approach was extended by Batra, Schechtman and Seki,[8] Zahn, Tsang and Pao,[9] de Oliveira and Ferreira[10] and others. An extensive list of relevant experimental observations is given in reference 9.

As the problem is extremely important, we suggest an alternative technique for solution which will describe the spatial and temporal evolution of: I) a fixed electric field and II) a constant source of current which are both governed by (13). The technique that we shall apply is to find the "self-similar solution" of this set of partial differential equations.

Equation (13) can be written as

$$EE_{xx} + (E_x)^2 + E_{xt} = 0 \ . \tag{14}$$

The self similar variables are

$$\phi = \frac{E}{t^{\alpha/\gamma}} \quad \text{and} \quad \xi = \frac{x}{t^{\beta/\gamma}} \tag{15}$$

where α/γ and β/γ are constants which will be specified by invariance and conservation requirements. Invariance specifies that $\alpha - \beta = -\gamma$

Substituting (15) in (14), we write

$$(\phi - \frac{\beta}{\gamma}\xi)\phi_{\xi\xi} + (\frac{\alpha}{\gamma} - \frac{\beta}{\gamma})\phi_{\xi} + (\phi_{\xi})^2 = 0 \quad . \tag{16}$$

We shall obtain solutions for (16) subject to: I) A fixed electric field and II) a constant current at $x = 0$ requirement.

I) Electric field is constant at $x = 0$. We choose:

$$\alpha = 0$$

$$\beta = \gamma \, .$$

With these values, $\xi = x/t$ and $\phi = E$. The integral of (16) with these constants is

$$\phi\phi_{\xi} - \xi\phi_{\xi} = k_1 \tag{17}$$

where k_1 is a constant of integration. The constant is set equal to zero since $E(x = 0,t) = 0$ in order to satisfy space charge limited conditions. This specifies $\phi(\xi = 0) = 0$. The solution of (17) is $\phi = \xi$ from which we compute that

$$E = \frac{x+x_0}{t+t_0}$$

$$\rho = \frac{1}{t+t_0} \tag{18}$$

$$i = \frac{(x+x_0)}{(t+t_0)^2}$$

where the constants x_0 and t_0 have been explicitely included since (13) is invariant to translation.

II) Current is constant at $x = 0$. We choose:

$$\frac{2\alpha}{\gamma} = \frac{\beta}{\gamma} \, .$$

With this choice, we have $\xi = \frac{x}{t^2}$ and $\phi = \frac{E}{t}$ where ϕ satisfies (16) which becomes

$$(\phi - 2\xi)\phi_{\xi\xi} - \phi_{\xi} + (\phi_{\xi})^2 = 0 \, . \tag{20}$$

This can be integrated once to

$$(\phi - 2\xi)\phi_\xi + \phi = k_2 \quad . \tag{21}$$

The constant of integration k_2 is set equal to zero since we require that $E(x = 0,t)$ be zero for space charge limited conditions which specifies $\phi(\xi = 0) = 0$. The integral of (21) is

$$\phi = \xi \tag{22}$$

from which we compute that

$$E = \frac{x+x_0}{t+t_0}$$

$$\rho = \frac{1}{t+t_0} \tag{23}$$

$$i = \frac{x+x_0}{(t+t_0)^2}$$

where again the constants x_0 and t_0 have been introduced since (13) is invariant to translation. Note that this is identical to (18).

In (21), we can also obtain the solution for a non-space charge limited condition $(E(x = 0,t) \neq 0)$ by setting the constant of integration k_2 equal to, say, $2i_0$. The integral of (21) can then be written as

$$\phi = \xi + i_0 \tag{24}$$

from which we compute that

$$E = (t + t_0)\left(i_0 + \frac{x+x_0}{(t+t_0)^2}\right)$$

$$\rho = \frac{1}{t+t_0} \tag{25}$$

$$i = \frac{x+x_0}{(t+t_0)^2} + i_0$$

where again the constants x_0 and t_0 have been reintroduced. Note that (25) reduces to the space charge limited case for $i_0 = 0$.

In conclusion, we have shown that the set of equations which describe the Transient Space Charge Limited Current Problem admit self-similar solutions for two physically interesting boundary conditions. These solutions are valid in the initial stages before

the particles reach a second electrode which may be placed at $x = L$.

3. Expansion of an electron cloud using a cold plasma model

A model which can describe the behavior of an electron cloud expansion into a fixed ion background is to assume that Poisson's equation is an initial condition. The electrostatic approximation for the $\nabla \times \bar{B}$ Maxwell equation and the continuity equation assure that Poisson's equation is satisfied for all time.[11]

The basic equations are:

$$\begin{aligned} (\rho v)_x + \rho_t &= 0 \\ mv_t + mvv_x &= -eE \\ \nabla \times \bar{B} \sim 0 &= \varepsilon_0 E_t - \rho v \end{aligned} \tag{26}$$

which are the equations of continuity and motion and Maxwell's equation respectively.

The self similar variables which satisfy the conservation law that $\int_0^\infty ndx = \text{constant}$ are

$$\varepsilon = E;\ N = nt^2;\ u = v/t;\ \xi = x/t^2 . \tag{27}$$

These are the same self similar variables that were obtained in an earlier study of the set (26) where the Ansatz that Maxwell's equation could replace Poisson's equation had not been made.[12] In the earlier study, it was not a pedestrian task to integrate the ODE.

Substituting (27) into (26), we now obtain the ODE:

$$\begin{aligned} -2N - 2\xi N_\xi + (NU)_\xi &= 0 \\ U - 2\xi U_\xi + UU_\xi &= -\varepsilon \\ 2\xi \varepsilon_\xi + NU &= 0 . \end{aligned} \tag{28}$$

A solution of this set is

$$U = 2\xi ,\quad N = 2 \quad \text{and} \quad \varepsilon = -2\xi \tag{29}$$

from which we write the solution of (26) using (27) and (29) as

$$E = -\frac{2x}{t^2} \qquad n = \frac{2}{t^2} \qquad v = \frac{2x}{t} . \tag{30}$$

In conclusion, we find that in the final self similar solution, the density is

independent of position at the end rather than making it an a priori assumption in the calculation as did Gintsburg who treated a similar problem.[13]

4. Expansion of charged particles with a temporally decaying nonlinear diffusion coefficient

In recent experiments on the Wisconsin Multipole, it was confirmed that the cross-field diffusion coefficient depended on time and amplitude as[3]

$$D \sim \frac{\epsilon^{-\alpha t}}{\sqrt{n}} \quad . \tag{31}$$

Incorporating this in the one dimensional diffusion equation, we obtain

$$n_t = \left[\epsilon^{-\alpha t} \frac{1}{\sqrt{n}} n_x\right]_x \tag{32}$$

where all constants except α have been suitably normalized away. A change of variables

$$\tau = \frac{1}{\alpha}\left[1 - \epsilon^{-\alpha t}\right] \tag{33}$$

transforms (32) to

$$n_\tau = \left[\frac{1}{\sqrt{n}} n_x\right]_x \tag{34}$$

The self similar treatment of (34) is straightforward, at least for the case where the conservation law $\int_0^\infty n dx = \text{constant}$ is valid. The self similar variables are

$$N = n\tau^{2/3} \quad \text{and} \quad \xi = x/\tau^{2/3} \tag{35}$$

and the resulting ODE is

$$\frac{2}{3}(\xi N)_\xi = \left[\frac{1}{\sqrt{N}} N_\xi\right]_\xi \quad . \tag{36}$$

If the burst of particles is symmetric at $x = 0$ such that $n_x\big|_{x=0} = 0$, and $n(x=0,\tau) = \tau^{-2/3}$ (36) can be integrated twice to yield

$$\sqrt{N} = \frac{2}{2 + \xi^2/3} \quad . \tag{37}$$

Using (33) and (35) in (37), we finally obtain

$$n(x,t) = \frac{4[\frac{1}{\alpha}(1-\epsilon^{-\alpha t})]^2}{[2(\frac{1}{\alpha}(1-\epsilon^{-\alpha t}))^{4/3} + \frac{x^2}{3}]^2} \quad . \tag{38}$$

We note that (38) gives a reasonably accurate qualitative description to the experimental results. This seems true even though the time scale in the experiment is sufficiently long such that normal modes have been exited.

III. Conclusion

The self similar behavior of four plasma phenomena have been described.

Acknowledgment

The author acknowledges Drs. J. Berryman, A. Hirose, G. A. Navratil, R. Post, H. Shen and M. Schonbeck for discussions of various aspects of this work.

REFERENCES

(1) K. E. Lonngren, Proc. of NSF sponsored workshop on Plasma Physics, Ahmedabad, India, 1976, to be published in Pramāna. Mathematics Research Center Technical Summary Report #1698.

(2) R. C. Davidson, Theory of Nonneutral Plasmas, W. A. Benjamin, Inc. (1974).

(3) J. R. Drake, J. R. Greenwood, G. A. Navratil and R. S. Post, Phys. Fluids 20, 148(1977).

(4) K. E. Lonngren and A. Hirose, Phys. Letters, 59A, 285(1976).

(5) K. E. Lonngren, J. Appl. Phys., to be published.

(6) A. Many and G. Rakavy, Phys. Rev. 126, 1980(1962).

(7) W. Helfrich and P. Mark, Z. Physik 166, 370(1962).

(8) I. P. Batra, B. H. Schechtman and H. Seki, Phys. Rev. B 2, 1592(1970).

(9) M. Zahn, C. F. Tsang and S. C. Pao, J. Appl. Phys. 45, 2432 (1974).

(10) L. Nunes de Oliveira and G. F. Leal Ferreira, J. Electrostatics 1, 371(1975).

(11) R. C. Davidson, Methods in Nonlinear Plasma Theory, Academic Press (1972), 33-34.

(12) H. Shen and K. E. Lonngren, IEEE Trans. PS-4, 144(1976).

(13) M. A. Gintsburg, Sov. Phys. Dokl., 19, 216(1974).

THE MATHEMATICAL DESCRIPTION OF A MOVING BOUNDARY PROBLEM IN AN ELLIPTIC - PARABOLIC SYSTEM OF PARTIAL DIFFERENTIAL EQUATION IN THE HYDRODYNAMICS OF POROUS MEDIA

Yoshisuke Nakano
U.S. Army Cold Regions
Research and Engineering Laboratory
Hanover, New Hampshire

ABSTRACT. The simultaneous solution of two types of partial differential equations, a parabolic equation for unsaturated flow and an elliptic equation for saturated flow is required for analysis of water movement in partly unsaturated and partly saturated porous media. A new and complete mathematical description of the boundary is obtained. It is proven that the boundary is generally a singular surface and the existing theory, which neglects such singularity is incorrect.

I. INTRODUCTION. The analysis of water movement in a partly unsaturated and partly saturated porous medium requires the simultaneous solution of two types of partial differential equations: a parabolic equation for the unsaturated part and an elliptic equation for the saturated part. Raats (1972, cf. Raats and Gardner, 1974) studied the boundary condition between these two parts and found that no condition necessarily had to be imposed on the boundary. This finding has been accepted as the complete mathematical description of the boundary. However, in the present work it is proven that the boundary condition derived by Raats (1972), although correct, is not complete, because it does not account for the discontinuity of a certain physical variable. A new and complete mathematical description accounting for this discontinuity is derived. It is also proven that a similar boundary condition holds true for a wetting front.

II. THEORY. Consider a material volume V intersected by a moving boundary σ with a velocity vector $\underset{\sim}{u}$. The material volume V consists of a saturated part V_s and an unsaturated part V_u as shown in Figure 1. V_s is bounded by surfaces S_s and σ_s while V_u is bounded by surfaces S_u and σ_u.

The signs of unit normal vectors $\underset{\sim}{n}$ on these surfaces are indicated by superscripts in Figure 1. We follow the standard method of derivation used by Raats and Gardner (1974). For the sake of simplicity we assume that the medium neither exhibits the capillary fringe phenomenon nor contains either a source or a sink. We also assume that all variables are smooth, in other words, all variables including their derivatives of all orders are continuous in V-σ.

Applying the Reynolds' transport theorem to the two volumes V_s and V_u by the use of a Cartesian coordinate system (x, y, z) attached to the solid phase of the medium, we get

$$\frac{D}{Dt}\int_{V_s}\theta dV = \int_{S_s}\theta v_n dV + \int_{\sigma_s}\theta u_n dS \qquad (1a)$$

and

$$\frac{D}{Dt}\int_{V_u}\theta dV = \int_{V_u}\frac{\partial\theta}{\partial t}dV + \int_{S_u}\theta v_n dS + \int_{\sigma_u}\theta u_n dS \qquad (1b)$$

where

θ = volumetric water content. In V_s $\theta=\theta_s$, where θ_s is the saturated volumetric water content.

$\theta\underset{\sim}{v}$ = flux of water, where $\underset{\sim}{v}$ is the velocity vector of water relative to the solid phase.

$\theta v_n = \theta(\underset{\sim}{v}\cdot\underset{\sim}{n})$ = component of the flux in the direction of $\underset{\sim}{n}$.

$u_n = \underset{\sim}{u}\cdot\underset{\sim}{n}$ = component of the velocity vector $\underset{\sim}{u}$ in the direction of $\underset{\sim}{n}$. The positive direction of $\underset{\sim}{n}$ on σ is defined as outward from the saturated part.

$\frac{D}{Dt}$ = material derivative.

The Reynolds' transport theorem is essentially a mathematical relation between time derivatives in terms of two different space coordinate systems (Eringen, 1967).

Upon adding these two equations and letting σ_u and σ_s approach σ while considering that the sign of $\underset{\sim}{n}$ on σ_u is opposite to the one on σ, we obtain

$$\frac{D}{Dt}\int_{V-\sigma}\theta dV = \int_{V_u}\frac{\partial\theta}{\partial t}dV + \int_{S_u}\theta v_n dS - \int_{\sigma}[\![\theta u_n]\!]dS + \int_{S_s}\theta v_n dS \tag{2}$$

where

$$[\![\theta A]\!] = \lim_{\sigma_u \to \sigma}(\theta A) - \lim_{\sigma_s \to \sigma}(\theta A)$$

A = any vector component

Using the Green-Gauss theorem

$$\int_V \operatorname{div}\underset{\sim}{\chi}\, dV = \int_S v_n\, dS \tag{3}$$

to replace the second and fourth terms on the right-hand side of Eq. (2), we get

$$\begin{aligned}\frac{D}{Dt}\int_{V-\sigma}\theta dV &= \int_{V_u}\left[\frac{\partial\theta}{\partial t} + \operatorname{div}(\theta\underset{\sim}{\chi})\right]dV \\ &+ \int_{V_s}\operatorname{div}(\theta\underset{\sim}{\chi})\,dV + \int_{\sigma}[\![\theta(v_n - u_n)]\!]\,dS\end{aligned} \tag{4}$$

From Eq. (4) we get the following boundary condition:

$$[\![\theta(v_n - u_n)]\!] = 0 \tag{5}$$

Reats (1972) obtained Eq. (5). In the next step we apply the same procedure to $\partial\theta/\partial t$ instead of θ and derive another more stringent condition than Eq. (5). By the Reynolds' transport theorem, we get

$$\frac{D}{Dt}\int_{V_s}\frac{\partial\theta}{\partial t}\,dV = \int_{S_s}\frac{\partial}{\partial t}(\theta v_n)\,dS + \int_{\sigma_s}\theta\frac{\partial u_n}{\partial t}\,dS \tag{6a}$$

$$\frac{D}{Dt}\int_{V_u}\frac{\partial\theta}{\partial t}\,dV = \int_{V_u}\frac{\partial^2\theta}{\partial t^2}\,dV + \int_{S_u}\frac{\partial}{\partial t}(\theta v_n)\,dS + \int_{\sigma_u}\frac{\partial}{\partial t}(\theta u_n)\,dS \tag{6b}$$

Upon adding these two equations, applying the Green-Gauss theorem, and letting σ_s and σ_u approach σ, we get

$$\frac{D}{Dt}\int_{V-\sigma}\frac{\partial\theta}{\partial t}\,dV = \int_{V_u}\frac{\partial}{\partial t}\left[\frac{\partial\theta}{\partial t} + \operatorname{div}(\theta\underset{\sim}{v})\right]dV$$

$$+\int_{V_s}\frac{\partial}{\partial t}\operatorname{div}(\theta\underset{\sim}{v})\,dV - \int_{\sigma}\left[\!\left[\theta\frac{\partial u_n}{\partial t}\right]\!\right]dS \tag{7}$$

$$+\int_{\sigma}\frac{\partial}{\partial t}\left[\!\left[\theta v_n\right]\!\right]dS - \int_{\sigma_u\to\sigma}\frac{\partial\theta}{\partial t}\,u_n\,dS$$

From Eq. (7), and if $\left[\!\left[\theta\frac{\partial u_n}{\partial t}\right]\!\right] = 0$, we get a new boundary condition:

$$\frac{\partial}{\partial t}\left[\!\left[\theta v_n\right]\!\right] = \lim_{\sigma_u\to\sigma} u_n\frac{\partial\theta}{\partial t} \tag{8}$$

If $\left[\!\left[\theta\frac{\partial u_n}{\partial t}\right]\!\right] \neq 0$, then

$$\lim_{\sigma_u\to\sigma}\theta = \theta^* \neq \theta_s \tag{9a}$$

and

$$\frac{\partial u_n}{\partial t} \neq 0 \tag{9b}$$

Eq. (9a) implies that θ is discontinuous on σ and (5) holds true. Suppose Eqs. (9a) and (9b) hold true. We write the flux of water as

$$F(\theta) = \theta v_n \tag{9c}$$

From Eq. (5) we get

$$F(\theta^*) - F(\theta_s) = (\theta^* - \theta_s)\, u_n \tag{9d}$$

There are two possible cases, either Case 1, $F(\theta^*) \neq F(\theta_s)$, or Case 2, $F(\theta^*) = F(\theta_s)$. For Case 1 if we let u_n approach zero, the right-hand side of Eq. (9d) approaches zero, while the left-hand side remains constant. This is contradictory. For Case 2 the left-hand side of Eq. (9d) vanishes, while the right-hand side does not vanish. This is also contradictory. Therefore the assumption is wrong and θ must be continuous on σ. In other words $\left[\theta \frac{\partial u_n}{\partial t}\right]$ must vanish.

Eq. (8) can be derived by simple partial differentiation of Eq. (5) in terms of t. Upon differentiation, we get

$$\frac{\partial}{\partial t}\left[\theta v_n\right] = \lim_{\sigma_u \to \sigma} \left(\theta \frac{\partial u_n}{\partial t} + u_n \frac{\partial \theta}{\partial t}\right) - \lim_{\sigma_s \to \sigma} \left(\theta \frac{\partial u_n}{\partial t}\right) \tag{9e}$$

where the assumption that all variables are smooth in V-σ is used to exchange the order of operations, differentiation and limit. It is easy to see that Eq. (9e) reduces to Eq. (8). Now we examine Eq. (8).

First, for the case in which the boundary moves with a constant speed:

$$\frac{\partial u_n}{\partial t} = 0 \tag{10a}$$

Then Eq. (8) reduces to

$$\frac{\partial}{\partial t}\left[\!\left[\theta v_n\right]\!\right] = \lim_{\sigma_u \to \sigma} \frac{\partial}{\partial t}(u_n \theta) \tag{10b}$$

Upon integrating Eq. (10b), we get

$$\left[\!\left[\theta v_n\right]\!\right] = \lim_{\sigma_u \to \sigma}(u_n \theta) + f(x, y, z) \tag{10c}$$

where f is an arbitrary function of the space coordinates.

Since $\left[\!\left[\theta v_n\right]\!\right] = 0$ when $u_n = 0$,

$$\left[\!\left[\theta v_n\right]\!\right] = u_n \theta_s \qquad \text{if } \frac{\partial u_n}{\partial t} = 0 \tag{11}$$

Next we prove that $\partial\theta/\partial t$ does not approach zero as $\sigma_u \to \sigma$. If we assume the contrary, then from Eq. (8) we get

$$\frac{\partial}{\partial t}\left[\theta v_n\right] = 0 \tag{12a}$$

Upon integrating,

$$\left[\theta v_n\right] = f(x, y, z) \tag{12b}$$

Eq. (12b) should apply also to the case where $\partial u_n/\partial t = 0$. Particularly for the special case where $u_n = 0$, we get

$$\left[\theta v_n\right] = 0 \tag{12c}$$

Eq. (12c) does not agree with Eq. (11) for the case where $\partial u_n/\partial t = 0$. This is obviously contradictory. Therefore, the initial assumption is wrong and $\partial\theta/\partial t$ does not approach zero as $\sigma_u \to \sigma$.

It is noted that the surface σ is one of many equal water content surfaces in the unsaturated part. The boundary condition for any such surfaces $\sum$ except σ is given as

$$\frac{\partial}{\partial t}\left[\theta v_n\right] = \left[\frac{\partial}{\partial t}(\theta u_n)\right] \qquad \text{on } \sum \tag{13}$$

Now we prove that $\partial\theta/\partial t$ is continuous on $\sum$. By the use of an argument similar to that used before (Eqs. (9a) $\sim$ (9b), f.f.) it is easy to prove that θ is continuous on $\sum$. Assume that there exits a surface $\sum^*$ where $\partial\sigma/\partial t$ is discontinuous. Upon differentiating Eq. (5) in terms of t, we get

$$\frac{\partial}{\partial t}\left[\theta v_n\right] = \frac{\partial}{\partial t}\left[\theta u_n\right] \qquad \text{on } \sum^* \tag{14}$$

From Eqs. (13) and (14), we get

$$\frac{\partial}{\partial t}\left[\!\left[\theta u_n\right]\!\right] = \left[\!\left[\frac{\partial}{\partial t}\left(\theta u_n\right)\right]\!\right] \qquad \text{on } \Sigma^* \tag{15}$$

Eq. (15) implies that the limit operation and differentiation can be exchanged. From an elementary theorem of analysis $\partial(\theta u_n)/\partial t$ must be continuous on Σ^*. Since θ, u_n, and $\partial u_n/\partial t$ are all continuous, therefore $\partial\theta/\partial t$ must be continuous on Σ^*. This is contradictory. The assumption is wrong and $\partial\theta/\partial t$ is continuous in the unsaturated part.

Now it is clear that the discontinuity of $\partial\theta/\partial t$ on σ uniquely differentiates σ from any other equal water content surfaces. Eq. (8) has to be imposed on the boundary between the unsaturated part and the saturated part.

It is easy to see that Eq. (8) can be applied to the problem of unsaturated flow above a dry porous medium. Since $v_n = 0$ in the dry medium, Eq. (8) reduces to

$$\lim_{\sigma_u \to \sigma} \frac{\partial}{\partial t}\left(\theta v_n\right) = \lim_{\sigma_u \to \sigma} u_n \frac{\partial \theta}{\partial t} \tag{16}$$

If $\partial u_n/\partial t = 0$, Eq. (16) reduces to

$$\lim_{\sigma_u \to \sigma} \left(v_n - u_n\right) = 0 \tag{17}$$

Since $\partial\theta/\partial t$ is not continuous on the boundary between an unsaturated part and a dry part, Eq. (16) has to be imposed for the exact solution of unsaturated flow above a dry porous medium.

In summary, the complete mathematical description of the moving boundary between saturated and unsaturated flow in a porous medium is given by

$$\frac{\partial}{\partial t}\left[\!\left[\theta v_n\right]\!\right] = \lim_{\sigma_u \to \sigma} u_n \frac{\partial \theta}{\partial t} \tag{8}$$

Since $\partial\theta/\partial t$ is not continuous on σ, Eq. (8) has to be imposed for the exact solution of saturated and unsaturated flow. For the boundary between an unsaturated porous medium and a dry porous medium Eq. (8) reduces to

$$\lim_{\sigma_u \to \sigma} \frac{\partial}{\partial t} (\theta v_n) = \lim_{\sigma_u \to \sigma} u_n \frac{\partial \theta}{\partial t} \tag{16}$$

Eq. (16) has to be imposed on the boundary.

In the derivation of Eqs. (8) and (16) the only physical law used is the well established law of material balance, therefore Eqs. (8) and (16) require no experimental proof.

REFERENCES

Eringen, A.C., *Mechanics of Continua*, John Wiley and Sons, New York, p. 75, 1967.

Raats, P.A.C., Jump conditions in the hydrodynamics of porous media, *Fundamentals of Transport Phenomena in Porous Media*, Vol. 1, University of Guelph, Ontario, Canada, 155-173, 1972.

Raats, P.A.C. and W.R. Gardner, Movement of water in the unsaturated zone near a water table, In Drainage for Agriculture, edited by J.V. Schilfgaarde. Am. Soc. Agronomy, Inc., Madison, Wisconsin, 311-405, 1974.

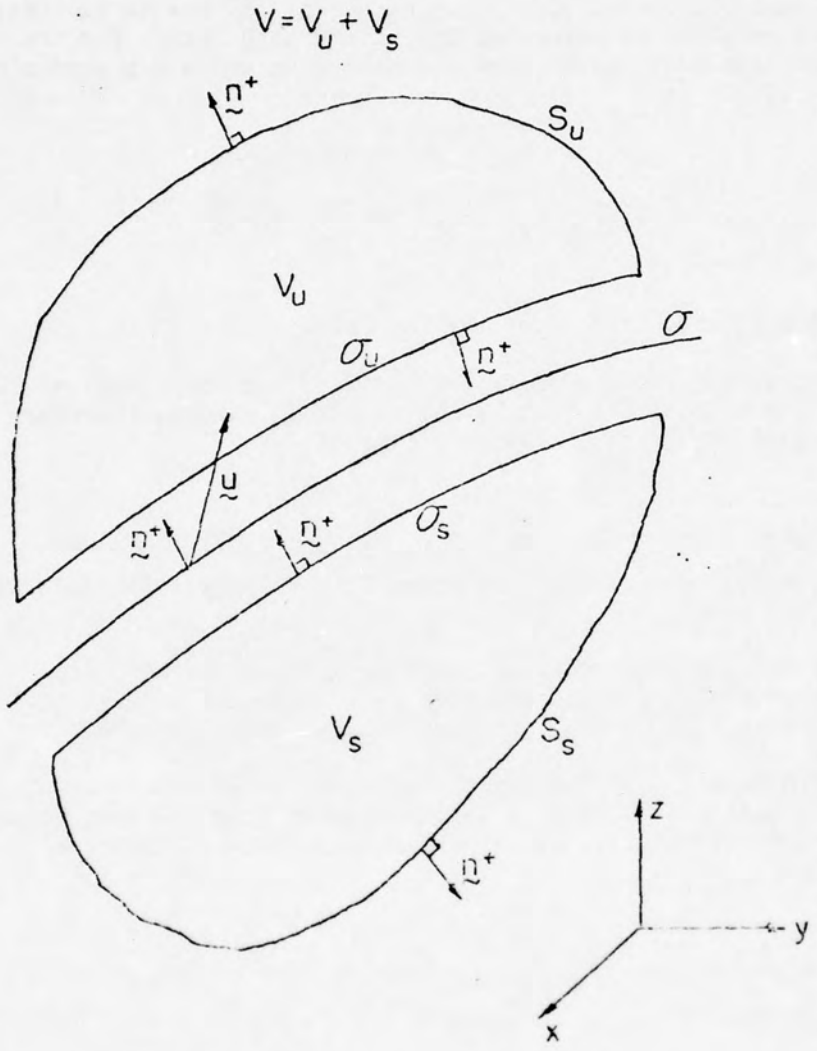


Figure 1. A material volume V intersected by a moving surface σ.

CRITICAL REVIEW OF ONE-DIMENSIONAL TUBE FLOW EQUATIONS

Aivars K.R. Celmiņš
Ballistic Modeling Division
Ballistic Research Laboratory, USARRADCOM
Aberdeen Proving Ground, Maryland 21005

ABSTRACT

Flows through ducts or pipes are often analyzed theoretically and numerically using one-dimensional flow equations. Generally it is assumed that the equations describe relations between average flow properties and that they are adequate if the axial component of the flow dominates. This paper reviews the derivation of the governing equations. It is shown that equations which are traditionally used for tube flows have a very limited scope of applicability. Their theoretical validity is in essence restricted to steady incompressible flows. In cases of more complicated flows some terms in the traditional momentum and energy equations can be in error by up to 50%. It is also shown that the popular approximation of the energy dissipation function by the product of the average velocity and average shear stress is appropriate for the simplest flows only. The paper reveals shortcomings of traditional methods of derivations of tube flow equations and provides explicit formulas for correction terms which should be used in the governing equations. An interesting property of the new equations for average flow properties is that the momentum equation and energy equation cannot be combined with the continuity equation to yield simple equations for velocity components and specific energy, respectively. Consequently a divergence form of the equations can be obtained only if momentum components and energy per unit volume are used as unknowns.

LIST OF SYMBOLS

C_{e1}	Correction term in energy equation ($J \cdot s^{-1} m^{-3}$). Definition by eq. (4.23)
C_{e2}	Correction term in energy equation ($Js^{-1} m^{-3}$) Definition by eq. (4.24)
C_m	Correction term in momentum equation (N/m^3). Definition by eq. (4.14)
C_Φ	Correction term in energy equation ($Js^{-1} m^{-3}$) Definition by eq. (4.28)
e	Specific internal energy (J/kg)
F	Force per volume (N/m^3)
$\overline{H}$	Heat source and heat flux terms in the energy equation ($J \cdot s^{-1} m^{-3}$)
k	Specific kinetic energy (J/kg)
n	Unit normal vector
p	Pressure (Pa)
q	Heat flux per volume ($J \cdot s^{-1} m^{-2}$)
Q	Heat source per volume ($J \cdot s^{-1} m^{-3}$)
r	Radial coordinate (m)
R	Radius of tube (m)
t	Time (s)
T	Viscous force per volume (N/m^3)
u	Velocity (m/s). (Axial velocity of a tube flow)
v	Velocity (m/s). (Radial velocity of a tube flow)
$\overline{W}_o$	Correction term in energy equation ($Js^{-1} m^{-3}$) Definition by eq. (4.27)
x	Cartesian coordinate (m)
X	Specific body force (N/kg)

$\dot{\varepsilon}$	Strain rate tensor (s^{-1}). Definition by eq. (2.9) and (A.6)
μ	Ordinary dynamic viscosity (Pa·s)
μ'	Dilatational dynamic viscosity (Pa·s)
ρ	Density (kg/m^3)
τ	Viscous stress tensor (Pa)
Φ	Heat dissipation function ($J \cdot s^{-1} m^{-3}$) Definition by eq. (2.10)

TABLE OF CONTENTS

1. INTRODUCTION

We consider in this paper the derivation of governing equations for fluid flows through ducts. Such flows are important elements in many mechanical systems. Most fluid mechanics textbooks present, therefore, a simple derivation of the governing equations, which reduces the general three-dimensional equations to a set of equations for one-dimensional flow. Experience has shown that these equations are adequate for many applications. Probably because of this success researchers sometimes tend to disregard the limits of applicability of the one-dimensional flow equations. In order to derive governing equations for more complicated flows they duplicate the steps used for simple duct flows. The resulting equations are not always adequate, e.g., in case of certain non-steady flows. Some textbooks discuss limitations of the usual tube flow equations. Often, however, the discussion is rather general, or limited to examples and exercise problems, and easily overlooked by casual readers. In this paper we will concentrate on the limitations. We will keep the discussions simple by considering in detail only a one-phase flow in a straight duct with a constant cross-section. The discussion of the example will provide a methodical approach to the derivation of flow equations for more general cases.

The starting point of our discussion is the set of general three-dimensional flow equations. In order to make this paper self-contained, we list the equations in Section 2. In Section 3 we specialize the equations for the case of a duct flow using a standard procedure, which is found in textbooks. In order to establish limits for the validity of the specialized equations, we carry out in Section 4 a more careful derivation of the duct flow equations. This derivation provides quantitative information about the errors which are introduced by the specialization of the equations. A comparison of the derivations and results of Sections 3 and 4 reveals that in standard derivations of the equations some non-zero terms are neglected. In Section 5 two examples are presented: a steady flow and an approximation to an interior ballistics flow. Quantitative estimates are given for some usually neglected terms in the governing equations. Section 6 contains some conclusions which can be drawn from the discussions of the equations.

2. BASIC GOVERNING EQUATIONS

We consider flows which satisfy conservation laws for mass, momentum and energy. Governing equations for such flows are derived and

discussed, e.g., by Tsien in Reference 1 and Batchelor in Reference 2. In this section we summarize the equations in order to make this paper self-contained. We use, in general, the same notation as Tsien, including the convention about the summation over equal indexes.

First we will consider the equations in integral form. In these equations the volume integrals are for an arbitrary control volume V, which need not be simply connected. We assume, however, for simplicity that its surface S has everywhere an outward pointing normal n_j. The conservation of mass can then be expressed by the equation

$$\frac{\partial}{\partial t}\int \rho \, dV + \oint \rho u_j n_j dS = 0 \tag{2.1}$$

The momentum equations are

$$\frac{\partial}{\partial t}\int \rho u_k dV + \oint \rho u_k u_j n_j dS + \int \frac{\partial p}{\partial x_k} dV = \int F_k dV. \tag{2.2}$$

The specific kinetic energy of the fluid is

$$k = \frac{1}{2} u_j u_j. \tag{2.3}$$

Combining eqs. (2.1) and (2.2) we obtain for the kinetic energy the equation

$$\frac{\partial}{\partial t}\int \rho \, k \, dV + \oint \rho \, k \, u_j n_j dS + \int u_j \frac{\partial p}{\partial x_j} dV = \int u_j F_j dV \, . \tag{2.4}$$

~

[1] *H.S. Tsien, "The Equations of Gas Dynamics," in Fundamentals of Gas Dynamics, edited by H.W. Emmons, Princeton University Press, 1958.*

[2] *G.K. Batchelor, An Introduction to Fluid Dynamics, Cambridge University Press, 1967.*

The first law of thermodynamics is

$$\frac{\partial}{\partial t}\int \rho\, e\, dV + \oint \rho\, e\, u_j n_j dS + \int p\, \frac{\partial u_j}{\partial x_j}\, dV = \int \left(Q - \frac{\partial q_j}{\partial x_j}\right) dV + \int \Phi\, dV. \quad (2.5)$$

By adding eqs. (2.4) and (2.5) we obtain an equation for the specific total internal energy e + k:

$$\frac{\partial}{\partial t}\int \rho(e + k)dV + \oint \rho(e + k)u_j n_j dS + \oint p\, u_j n_j dS = \int \left(Q - \frac{\partial q_j}{\partial x_j}\right) dV + \int (\Phi + u_j F_j)dV. \quad (2.6)$$

The last integral in eq. (2.6) is the contribution of external and viscous forces to the changes of the total internal energy. Its first part, $\int \Phi dV$, is the contribution of viscous forces to the internal energy e. The integrand Φ, i.e., the heat dissipation function, can be expressed in terms of the viscous stress tensor τ_{jk}:

$$\Phi = \tau_{kj}\, \frac{\partial u_k}{\partial x_j} \quad (2.7)$$

We assume that the viscous stress tensor is related by Stokes formula to the strain rate tensor ε_{kj} (Reference 1, page 13, Reference 3, page 132)

$$\tau_{kj} = 2\mu\, \varepsilon_{kj} + \left(\mu' - \frac{2}{3}\mu\right) \delta_{kj} \varepsilon_{ii} \quad (2.8)$$

This definition is not restricted to constant viscosities μ and μ', i.e., to homogenous fluids. However, it restricts the considerations to isotropic fluids. The viscosities μ and μ' must be positive or zero.

---

[3] G. Hamel, Mechanik der Kontinua, R.G. Teubner, Stuttgart, 1956.

The strain rate tensor ε_{kj} is defined by (Reference 2, page 80)

$$\varepsilon_{kj} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_j} + \frac{\partial u_j}{\partial x_k}\right) \tag{2.9}$$

Substituting (2.8) and (2.9) into eq. (2.7) we obtain the following expressions for Φ:

$$\begin{aligned}\Phi &= 2\mu\ \varepsilon_{kj}\varepsilon_{kj} + \left(\mu' - \frac{2}{3}\mu\right)\delta_{kj}\varepsilon_{ii}\varepsilon_{kj} = \\ &= 2\mu\ \{\varepsilon^2\}_{\text{trace}} + \left(\mu' - \frac{2}{3}\mu\right)\{\varepsilon\}^2_{\text{trace}} = \\ &= \frac{1}{2}\mu\left(\frac{\partial u_k}{\partial x_j} + \frac{\partial u_j}{\partial x_k}\right)^2 + \left(\mu' - \frac{2}{3}\mu\right)\left(\frac{\partial u_k}{\partial x_k}\right)^2.\end{aligned} \tag{2.10}$$

Eq. (2.10) shows that the heat dissipation function Φ is always positive for a stress tensor of the form (2.8).

The term $\int u_j F_j dV$ in eq. (2.6) is the contribution of viscous and body forces to the changes of the kinetic energy k. The force (per unit volume) F_j is a sum of body forces ρX_j and viscous forces T_j. The latter can be expressed in terms of the viscous stress tensor τ_{kj}. We thus have the equation

$$F_j = \rho X_j + T_j = \rho X_j + \frac{\partial \tau_{jk}}{\partial x_k} \tag{2.11}$$

Combining eqs. (2.7) and (2.11) we obtain

$$\begin{aligned}\Phi + u_j F_j &= \Phi + u_j T_j + \rho u_j X_j = \\ &= \frac{\partial}{\partial x_k}(u_j \tau_{jk}) + \rho u_j X_j\end{aligned} \tag{2.12}$$

In this form we have subdivided the contributions of forces to the changes of the total internal energy into contributions by viscous and by body forces. The corresponding volume integral in eq. (2.6) is

$$\int(\Phi + u_j F_j)dV = \int(\Phi + u_j T_j + \rho u_j X_j)\, dV =$$

$$= \oint \tau_{jk} u_j n_k dS + \int \rho u_j X_j dV \qquad (2.13)$$

We note that the surface integral in eq. (2.13) contributes to the internal as well as to the kinetic energy of the flow. It represents the viscous forces acting on the surface of the control volume. The volume integral over the body forces contributes to the kinetic energy only.

The governing equations (2.1), (2.2) and (2.4) through (2.6) can also be expressed in differential form as follows

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}(\rho\, u_j) = 0, \qquad (2.14)$$

$$\frac{\partial}{\partial t}(\rho\, u_k) + \frac{\partial}{\partial x_j}(\rho\, u_k u_j) + \frac{\partial p}{\partial x_k} = F_k\ , \qquad (2.15)$$

$$\frac{\partial}{\partial t}(\rho\, k) + \frac{\partial}{\partial x_j}(\rho\, k\, u_j) + u_j \frac{\partial p}{\partial x_j} = u_j F_j\ , \qquad (2.16)$$

$$\frac{\partial}{\partial t}(\rho\, e) + \frac{\partial}{\partial x_j}(\rho\, e\, u_j) + p\frac{\partial u_j}{\partial x_j} = Q - \frac{\partial q_j}{\partial x_j} + \Phi\ , \qquad (2.17)$$

$$\frac{\partial}{\partial t}[\rho(e+k)] + \frac{\partial}{\partial x_j}[\rho(e+k)u_j] + \frac{\partial}{\partial x_j}(p\, u_j) = Q - \frac{\partial q_j}{\partial x_j} + \Phi + u_j F_j\ . \qquad (2.18)$$

Eq. (2.16) is a consequence of eqs. (2.14) and (2.15), because k is defined by eq. (2.3). Also, eq. (2.18) is the sum of eqs. (2.16) and (2.17). We have, therefore, only five independent differential equations for the six quantities ρ, u_k, p, and e. To complete the system of equations we need another equation, which is provided by the equations of state for the fluid under consideration. We assume that such an equation is available, e.g., in the form

$$\Theta(e,p,\rho) = 0\ , \qquad (2.19)$$

and that eq. (2.19) can be solved explicitly for either of the arguments. For example, in case of an ideal gas with constant specific heats eq. (2.18) is

$$\frac{p}{\rho} - (\gamma-1)e = 0 \ . \qquad (2.20)$$

For the discussions in the rest of this paper we will not make use of eq. (2.19) or (2.20). The assumption of the existence of such an equation is made here only to close the set of governing equations.

3. APPROXIMATE GOVERNING EQUATIONS FOR DUCT FLOWS

In this section we derive approximate governing equations for duct flows. The dominant component of such flows is usually in the axial direction. Also, in many cases only the dependence of flow properties on the axial coordinate is of practical interest. Duct flows are therefore usually treated by one-dimensional equations which are derived from the general flow equations of Section 2.

A standard procedure for the derivation of these equations is to consider a control volume which consists of a length Δz of the duct. The integral forms of the governing equations are applied to this control volume and corresponding differential equations obtained by letting Δz approach zero. This method is used, e.g., in References 4 and 5, and we will follow these references closely.

Another possible approach is to start with governing equations for one-dimensional flow, i.e., a flow which depends on only one coordinate and which has a velocity component in the direction of that coordinate only. Three dimensional effects, e.g., from the wall friction, are then added to the equations by ad hoc procedures. We will not pursue this approach here because the former approach can be generalized more easily.

---

[4] *A.H. Shapiro, The Dynamics and Thermodynamics of Compressible Fluid Flow, Vol. I and II, Roland Press Company, New York, 1954.*

[5] *L. Crocco, "One-Dimensional Treatment of Steady Gas Dynamics" in Fundamentals of Gas Dynamics II, edited by H.W. Emmons, Princeton University Press, 1958.*

Let z be the axial coordinate and let for simplicity the cross-sectional area A of the duct be constant. The continuity equation (2.1) is then for the control volume

$$\frac{\partial}{\partial t}\left\{\int_{z}^{z+\Delta z} \bar{\rho}\, A\, dz\right\} + \left[\bar{\rho}\,\bar{u}\,A\right]_{z}^{z+\Delta z} = 0 \qquad (3.1)$$

The bars on ρ and u in eq. (3.1) indicate that we are dealing with average density and velocity, respectively. We apply now the mean value theorem to the first term in eq. (3.1) and use a Taylor series expansion for the second term. The result is

$$\frac{\partial}{\partial t}\{\rho(\hat{z})\ A\}\cdot\Delta z + \Delta z\cdot A\,\frac{\partial(\bar{\rho}\,\bar{u})}{\partial z} = O(\Delta z^2)\ , \qquad (3.2)$$

where $z \le \hat{z} \le z + \Delta z$. Letting Δz in eq. (3.2) approach zero we obtain the continuity equation

$$\frac{\partial\bar{\rho}}{\partial t} + \frac{\partial(\bar{\rho}\,\bar{u})}{\partial z} = 0\ . \qquad (3.3)$$

The momentum balance equation is considered in the z-direction only. First we obtain as above from eq. (2.2)

$$\frac{\partial}{\partial t}\{\bar{\rho}(\hat{z})\ \bar{u}(\hat{z})\ A\}\cdot\Delta z + \Delta z\cdot A\,\frac{\partial(\bar{\rho}\,\bar{u}^2)}{\partial z} + \Delta z\cdot A\,\frac{\partial\bar{p}}{\partial z} = \bar{F}\cdot A\cdot\Delta z + O(\Delta z^2)\ . \qquad (3.4)$$

The momentum equation for the average flow properties is obtained from eq. (3.4) by letting Δz approach zero. The result is

$$\frac{\partial}{\partial t}(\bar{\rho}\,\bar{u}) + \frac{\partial}{\partial z}(\bar{\rho}\,\bar{u}^2) + \frac{\partial\bar{p}}{\partial z} = \bar{F}\ . \qquad (3.5)$$

The force per unit volume, $\bar{F}$, can be expressed as a sum of two components in analogy to eq. (2.11). The momentum equation is then

$$\frac{\partial}{\partial t}(\bar{\rho}\,\bar{u}) + \frac{\partial}{\partial z}(\bar{\rho}\,\bar{u}^2) + \frac{\partial \bar{p}}{\partial z} = \bar{\rho}\,\bar{X} + \bar{T} \,. \tag{3.6}$$

The quantity $\bar{T}$ in eq. (3.6) is obtained from the resultant of the viscous boundary forces on the surface of the control volume. For simple tubes $\bar{T}$ can be expressed in terms of the pipe friction coefficient or, by experimental correlations, in terms of the surface roughness and perimeter of the tube. The term $\bar{\rho}\,\bar{X}$ usually represents the gravity force component in the axial direction of the tube.

A combination of eqs. (3.3) and (3.6) yields

$$\frac{\partial}{\partial t}\left(\frac{1}{2}\bar{\rho}\,\bar{u}^2\right) + \frac{\partial}{\partial z}\left(\frac{1}{2}\bar{\rho}\,\bar{u}^3\right) + \bar{u}\,\frac{\partial \bar{p}}{\partial z} = \bar{u}\,\bar{F} = \bar{u}\,\bar{\rho}\,\bar{X} + \bar{u}\,\bar{T} \,. \tag{3.7}$$

Eq. (3.7) can be considered as an equation for the kinetic energy, if the latter is approximated by

$$\bar{k} = \frac{1}{2}\bar{u}^2 \,. \tag{3.8}$$

However, eq. (3.7) is a mathematical consequence of the continuity and momentum equations, i.e., eqs. (3.3) and (3.6), and is independent of any assumptions about the kinetic energy.

Next we consider the energy balance. Following general practice (see, e.g., Reference 4) we start with the eq. (2.6) instead of using the first law of thermodynamics, i.e., eq. (2.5). For the control volume we obtain first

$$\frac{\partial}{\partial t}\{\bar{\rho}\cdot(\bar{e}+\bar{k})\cdot A\}\cdot\Delta z + \Delta z\cdot A\,\frac{\partial}{\partial z}\{\bar{\rho}\cdot\bar{u}\cdot(\bar{e}+\bar{k}) + \bar{u}\cdot\bar{p}\} =$$

$$= \left(\bar{Q} - \frac{\partial \bar{q}}{\partial z}\right)\cdot A\cdot\Delta z + (\bar{\Phi} + \bar{u}\cdot\bar{F})\cdot A\cdot\Delta z + O(\Delta z^2) \tag{3.9}$$

At the limit $\Delta z \to 0$ eq. (3.9) yields the energy equation

$$\frac{\partial}{\partial t}\{\bar{\rho}\ (\bar{e}+\bar{k})\} + \frac{\partial}{\partial z}\{\bar{\rho}\ \bar{u}\ (\bar{e}+\bar{k}) + \bar{u}\ \bar{p}\} =$$

$$= \bar{Q} - \frac{\partial \bar{q}}{\partial z} + \bar{\Phi} + \bar{u}\ \bar{F}\ . \qquad (3.10)$$

The equation of state, such as eq. (2.19), contains usually the internal energy $\bar{e}$ and not the total internal energy $\bar{e} + \bar{k}$. Therefore, eq. (3.10) is modified to eliminate $\bar{k}$. To this end it is assumed that the approximation (3.8) holds, and eq. (3.7) is subtracted from eq. (3.10). The result is

$$\frac{\partial}{\partial t}(\bar{\rho}\ \bar{e}) + \frac{\partial}{\partial z}(\bar{\rho}\ \bar{u}\ \bar{e}) + \bar{p}\ \frac{\partial \bar{u}}{\partial z} = \bar{Q} - \frac{\partial \bar{q}}{\partial z} + \bar{\Phi}\ . \qquad (3.11)$$

Eq. (3.11) is, of course, the first law of thermodynamics and could have been obtained directly from eq. (2.5) without any assumptions about the kinetic energy.

In order to use eqs. (3.3), (3.7), and (3.11) for computations we need among other data estimates for the forces $\bar{\rho}\ \bar{X}$ and $\bar{T}$ and for the heat dissipation function $\bar{\Phi}$. The latter is often expressed in terms of $\bar{T}$ by the following arguments. (See, e.g., Reference 4, page 39 ff. and 972 ff.)

The last integral on the right hand side of the energy equation (2.6) is according to eq. (2.13)

$$W = \int (\Phi + u_j F_j)dV = \oint \tau_{jk} u_j n_k dS + \int \rho\ u_j X_j dV\ . \qquad (3.12)$$

The surface integral in eq. (3.12) represents the work rate of viscous forces acting on the surface S of the control volume. We subdivide this surface into material boundaries (e.g., duct walls) and flow-through surfaces S_o. The work done on material boundaries is called shaft work. The work by viscous forces on the flow-through boundaries is called shear work. Let the corresponding work rates be W_{shaft} and W_o, respectively. In these terms eq. (3.12) is

$$W = \int (\Phi + u_j F_j)dV = W_{shaft} + W_o + \int \rho\ u_j X_j dV\ . \qquad (3.13)$$

The integral W_o over the open boundaries is usually assumed to be negligible. For example, in case of a tube flow it is argued that integration over the core flow region contributes very little to the integral because τ_{jk} is small in that region. Integration over the boundary layer region also contributes little because the velocity u_j is small in the boundary layer. Hence W_o must be small.

If we carry out the derivation of eq. (3.10) using the relation (3.13) we obtain

$$\frac{\partial}{\partial t}[\bar{\rho}\ (\bar{e}+\bar{k})] + \frac{\partial}{\partial z}[\bar{\rho}\ \bar{u}\ (\bar{e}+\bar{k}) + \bar{u}\ \bar{p}] = \bar{Q} - \frac{\partial \bar{q}}{\partial z} + \bar{W}_{shaft} + \bar{W}_o + \bar{\rho}\ \bar{u}\ \bar{X}\ . \quad (3.14)$$

Combining eqs. (3.14), (3.7), and (3.8) we obtain as the first law of thermodynamics instead of eq. (3.11) the equation

$$\frac{\partial}{\partial t}\ (\bar{\rho}\ \bar{e}) + \frac{\partial}{\partial z}\ (\bar{\rho}\ \bar{u}\ \bar{e}) + \bar{p}\ \frac{\partial \bar{u}}{\partial z} = \bar{Q} - \frac{\partial \bar{q}}{\partial z} + \bar{W}_{shaft} + \bar{W}_o - \bar{u}\ \bar{T}\ . \quad (3.15)$$

In this equation the heat dissipation function Φ is approximated by

$$\bar{\Phi} = \bar{W}_{shaft} + \bar{W}_o - \bar{u}\ \bar{T}. \quad (3.16)$$

At the material boundaries the velocity of the fluid is equal to the velocity of the boundary. Therefore, W_{shaft} is non-zero only if the boundaries are moving. If the tube does not contain moving boundaries and $\bar{W}_o$ is neglected, then eq. (3.16) becomes

$$\bar{\Phi} = -\bar{u}\ \bar{T}, \quad (3.17)$$

which is the usual approximation of $\bar{\Phi}$ for tube flows (Reference 4, page 972 ff.)

In case of two-phase flows, e.g. particles submerged into the fluid, $\bar{W}_{shaft}$ is assumed to be the work of drag forces. Let the average particle velocity be $\bar{u}_{particle}$ and the drag force be $\bar{T}_{drag}$. Then

$$\bar{W}_{shaft} = \bar{u}_{particle}\ \bar{T}_{drag}\ . \quad (3.18)$$

The resultant $\overline{T}$ of the viscous forces is in this case the sum of particle drag and wall friction forces

$$\overline{T} = \overline{T}_{drag} + \overline{T}_{wall} . \tag{3.19}$$

The equation for $\bar{\Phi}$ becomes then

$$\bar{\Phi} = (\overline{u}_{particle} - \overline{u})\, \overline{T}_{drag} - \overline{u}\, \overline{T}_{wall} . \tag{3.20}$$

This equation is sometimes modified by an ad hoc factor, see Reference 6, page 81.

In summary, either eq. (3.17), or eq. (3.20) provides a convenient estimate for $\bar{\Phi}$ i.e., for the right-hand side of the energy equation (3.15). Estimates of $\overline{T}_{wall}$ and $\overline{T}_{drag}$ are also needed to express the terms on the right-hand sides of the momentum equation (3.6). It appears from the derivation that no further estimates of flow properties are needed under quite general conditions.

Some limitations of the approximation (3.17) become obvious if we consider non-steady fluctuating flows. In such flows it is possible that the signs of $\overline{T}$ and $\overline{u}$ are temporarily equal. In these cases $\overline{W}_o$ cannot be neglected, because otherwise we would have a negative heat dissipation function. Thus it seems appropriate to ask how accurate is the energy equation (3.15). Our derivation does not provide any clues to an answer to this question. We will therefore rederive the duct flow equations more carefully in the next section, keeping track of all approximations involved.

4. PRECISE GOVERNING EQUATIONS FOR DUCT FLOWS

In this section we will derive complete one-dimensional governing equations for flows through constant area ducts, including formulas for quantities which were neglected in Section 3. We will then discuss the differences between the complete equations and those of the previous

---

[6] G.B. Wallis, *One-Dimensional Two-Phase Flow*, McGraw-Hill Co, New York, 1969.

section, indicating where the previous derivation of the equations is insufficient.

One-dimensional duct flow equations are relations between average flow properties. The equations depend therefore, among other things, on the definitions of the averages. For steady duct flows certain averages and corresponding governing equations have been discussed by Crocco in Reference 5. Because the averages defined by Crocco cannot be used for non-steady flows, our analysis will be different. The results of our analysis can be applied to steady as well as non-steady flows.

First we consider the continuity equation (2.1). For a control volume which consists of a length of Δx_3 of the duct, eq. (2.1) is

$$\frac{\partial}{\partial t}\int_{x_3}^{x_3+\Delta x_3}\left\{\int_A \rho\, ds\right\}dx_3 + \left[\int_A \rho\, u_3 ds\right]_{x_3}^{x_3+\Delta x_3} = 0 \tag{4.1}$$

The integrals $\int \rho ds$ and $\int \rho u_3 ds$ are functions of x_3. We expand these functions in Taylor series, intechange the order of integration over x_3 and differentiation with respect to t in the first term of eq. (4.1), and apply the mean value theorem to that term. The result is

$$\Delta x_3 \left\{\frac{\partial}{\partial t}\int_A \rho\, ds\right\}_{x_3+\theta\Delta x_3} + \Delta x_3 \left\{\frac{\partial}{\partial x_3}\int_A \rho\, u_3 ds\right\}_{x_3} = 0(\Delta x_3^2)\ , \tag{4.2}$$

with $0 \leq \theta \leq 1$.

At the limit $\Delta x_3 \to 0$ eq. (4.2) yields

$$\frac{\partial}{\partial t}\int_A \rho\, ds + \frac{\partial}{\partial x_3}\int_A \rho\, u_3 ds = 0\ . \tag{4.3}$$

We now define for each cross-section x_3 = const. an average fluid density $\bar{\rho}$ by

$$\bar{\rho} = \frac{1}{A}\int_A \rho\, ds \tag{4.4}$$

and an average fluid velocity $\bar{u}$ by

$$\bar{u} = \frac{1}{A\,\bar{\rho}} \int_A \rho\, u_3 ds \ . \tag{4.5}$$

The continuity eq. (4.3) can then be expressed in terms of the average density and velocity as

$$\frac{\partial \bar{\rho}}{\partial t} + \frac{\partial}{\partial x_3}(\bar{\rho}\ \bar{u}) = 0 \ . \tag{4.6}$$

Eq. (4.6) is identical to the continuity equation (3.3). However, we have now established that the continuity equation is of this form only if the average quantities $\bar{\rho}$ and $\bar{u}$ are defined by eqs. (4.4) and (4.5), respectively. Thus, if we chose an alternate definition of the average velocity, e.g., the simple spatial average

$$\tilde{u} = \frac{1}{A} \int_A u_3 ds \ , \tag{4.7}$$

then the corresponding continuity equation would be

$$\frac{\partial \bar{\rho}}{\partial t} + \frac{\partial}{\partial x_3}\ (\bar{\rho}\ \tilde{u}) = \frac{\partial}{\partial x_3}[\bar{\rho}\ (\tilde{u} - \bar{u})] \ . \tag{4.8}$$

The right-hand side of eq. (4.8) is non-zero in general.

The momentum equation (2.2) yields for a duct flow in analogy to eq. (4.3)

$$\frac{\partial}{\partial t} \int_A \rho\, u_k ds + \frac{\partial}{\partial x_3} \int_A \rho\, u_k u_3 ds + \int_A \frac{\partial p}{\partial x_k} ds = \int_A F_k ds \ . \tag{4.9}$$

If the flow is axially symmetric, then for k = 1 and k = 2 all terms in eq. (4.9) are identically zero. In cases of non-axisymmetric flows all three momentum equations are needed to describe the flow, e.g., in the case of a non-vertical tube in a gravity field. We will consider

for simplicity only the momentum equation in the axial x_3-direction, thus restricting the analysis to axisymmetric flows. Eq. (4.9) thus becomes

$$\frac{\partial}{\partial t}\int_A \rho\, u_3 ds + \frac{\partial}{\partial x_3}\int_A \rho\, u_3^2 ds + \int_A \frac{\partial p}{\partial x_3} ds = \int_A F_3 ds \ . \tag{4.10}$$

Eq. (4.10) contains two new flow variables for which averages have to be defined. We chose the following definitions:

$$\bar{p} = \frac{1}{A}\int_A p\ ds \tag{4.11}$$

and

$$\bar{F}_3 = \frac{1}{A}\int_A F_3 ds \ . \tag{4.12}$$

Expressing the momentum equation (4.10) in terms of average quantities we then obtain

$$\frac{\partial}{\partial t}(\bar{\rho}\ \bar{u}) + \frac{\partial}{\partial x_3}(\bar{\rho}\ \bar{u}^2) + \frac{\partial \bar{p}}{\partial x_3} = \bar{F}_3 + C_m \tag{4.13}$$

with

$$C_m = \frac{\partial}{\partial x_3}\left\{\bar{\rho}\ \bar{u}^2 - \frac{1}{A}\int_A \rho\ u_3^2 ds\right\} =$$

$$= \frac{\partial}{\partial x_3}\left\{\frac{1}{A^2\bar{\rho}}\left[\left(\int_A \rho\ u_3 ds\right)^2 - \int_A \rho\ ds \int_A \rho\ u_3^2 ds\right]\right\} \tag{4.14}$$

Comparing the momentum equation (4.13) with the corresponding eq. (3.5), we see that the latter equation is in error. The reason for this error is that eq. (3.4) should have contained the term $\Delta z \cdot C_m \cdot A$. Eq. (4.14) shows that this term is non-zero in general. The expression in square brackets in eq. (4.14) is negative or zero according to Schwarz's inequality. It is zero if and only if u_3 = const. across the duct. Hence the correction term C_m is zero only in case of a slug flow or if the term is independent of x_3. The latter is the case for steady incompressible flows through constant area ducts. In all more interesting cases C_m is non-zero and its magnitude should be estimated to justify the neglect of C_m, or C_m should be included in the momentum equation.

The average force per unit volume, $\overline{F}_3$, is defined by eq. (4.12). For later reference we note that according to equation (2.11) F_3 is a sum of body forces and viscous forces. We define the corresponding averages by

$$\overline{T}_3 = \frac{1}{A} \int_A \frac{\partial \tau_{3k}}{\partial x_k} \, ds \tag{4.15}$$

and

$$\overline{X}_3 = \frac{1}{\overline{\rho} A} \int_A \rho \, X_3 ds \ . \tag{4.16}$$

With these definitions we have

$$\overline{F}_3 = \overline{\rho} \overline{X}_3 + \overline{T}_3 \ . \tag{4.17}$$

We now consider the first law of thermodynamics, eq. (2.5). First we obtain for the duct flow in analogy to eq. (4.9)

$$\frac{\partial}{\partial t} \int_A \rho \, e \, ds + \frac{\partial}{\partial x_3} \int_A \rho \, e \, u_3 ds + \int_A p \frac{\partial u_k}{\partial x_k} \, ds = \int_A (Q - \frac{\partial q_k}{\partial x_k}) \, ds + \int_A \Phi ds \ . \tag{4.18}$$

In order to express this equation in terms of averages we define

$$\bar{e} = \frac{1}{\overline{\rho} \, A} \int_A \rho \, e \, ds \ , \tag{4.19}$$

$$\overline{\Phi} = \frac{1}{A} \int_A \Phi ds = \frac{1}{A} \int_A \tau_{kj} \frac{\partial u_k}{\partial x_j} \, ds \tag{4.20}$$

and

$$\overline{H} = \frac{1}{A} \int_A \left(Q - \frac{\partial q_k}{\partial x_k} \right) ds \ . \tag{4.21}$$

With these definitions eq. (4.18) becomes

$$\frac{\partial}{\partial t}(\bar{\rho}\ \bar{e}) + \frac{\partial}{\partial x_3}(\bar{\rho}\ \bar{e}\ \bar{u}) + \bar{p}\ \frac{\partial \bar{u}}{\partial x_3} = \bar{H} + \bar{\Phi} + C_{e1} + C_{e2}\ , \qquad (4.22)$$

where

$$C_{e1} = \bar{p}\ \frac{\partial \bar{u}}{\partial x_3} - \frac{1}{A}\int_A p\ \frac{\partial u_k}{\partial x_k}\, ds \qquad (4.23)$$

and

$$C_{e2} = \frac{\partial}{\partial x_3}(\bar{\rho}\ \bar{e}\ \bar{u}) - \frac{1}{A}\frac{\partial}{\partial x_3}\int_A \rho\ e\ u_3 ds\ . \qquad (4.24)$$

The nature of the correction terms C_{e1} and C_{e2} is similar to that of the correction term C_m in the momentum equation. They are zero for slug flow and should be estimated in other cases. If we compare eq. (4.22) with the corresponding eq. (3.11), we see that the latter is in error. The reason for the error is an oversight of a term $\Delta z(C_{e1} + C_{e2})\cdot A$ which should have been introduced in eq. (3.9). The correction terms enter the equations because a product of function averages is in general not equal to the average of the product of the functions. Or, differently expressed, multiplications of functions and averaging of functions are not commutative operations.

We mentioned in Section 3 that $\bar{\Phi}$ is usually approximated by $-\bar{u}\ \bar{T}_3$. It was also shown that such an approximation is based on the assumption that a term W_o can be neglected. We will now investigate the approximation more carefully. By the definition (4.20) we have

$$\bar{\Phi} = \frac{1}{A}\int_A \tau_{kj}\ \frac{\partial u_k}{\partial x_j}\, ds =$$

$$= \frac{1}{A}\int_A \frac{\partial}{\partial x_j}(\tau_{kj} u_k)\, ds - \frac{1}{A}\int_A u_k\ \frac{\partial \tau_{kj}}{\partial x_j}\, ds =$$

$$= \frac{1}{A}\frac{\partial}{\partial x_3}\int_A \tau_{k3} u_3 ds - \frac{1}{A}\int_A u_k T_k ds\ . \qquad (4.25)$$

The first term on the right-hand side of eq. (4.25) we recognize as $\overline{W}_o$, i.e., the average of the gradient of the work rate of viscous forces on the cross section A. The second term may be approximated by $-\bar{u}\,\bar{T}_3$. The final formula for $\overline{\Phi}$ is then

$$\overline{\Phi} = -\,\bar{u}\,\bar{T}_3 + \overline{W}_o + C_\Phi \,, \tag{4.26}$$

where

$$\overline{W}_o = \frac{1}{A}\frac{\partial}{\partial x_3}\int_A \tau_{k3}u_k\,ds \tag{4.27}$$

and

$$C_\Phi = \bar{u}\,\bar{T}_3 - \frac{1}{A}\int_A u_k T_k\,ds \,. \tag{4.28}$$

Combining eqs. (4.26) and (4.22), we obtain for the energy equation (first law of thermodynamics) the expression

$$\frac{\partial}{\partial t}(\bar{\rho}\,\bar{e}) + \frac{\partial}{\partial x_3}(\bar{\rho}\,\bar{e}\,\bar{u}) + \bar{p}\,\frac{\partial \bar{u}}{\partial \bar{x}_3} = \bar{H} - \bar{u}\,\bar{T}_3 + C_{e1} + C_{e2} + \overline{W}_o + C_\Phi \,. \tag{4.29}$$

The first two correction terms, C_{e1} and C_{e2}, appear in eq. (4.29) because of the averaging of some terms. The last two correction terms, $\overline{W}_o$ and C_Φ, are due to the approximation of $\overline{\Phi}$ by $-\bar{u}\,\bar{T}_3$.

The equation for the kinetic energy can be treated formally in the same manner as the equation for the internal energy. One can introduce error terms, corresponding to C_{e1} and C_{e2}, either in the kinetic energy equation or in the equation for the total internal energy. Since typically only the equation for internal energy is needed for computations, the other equations are not formally derived.

In summary, we have shown that the one-dimensional governing equations for average flow properties in duct flows are not the same as equations for locally one-dimensional flows. If the medium is compressible then the additional terms in the governing equations vanish only for slug flow. For other flows the magnitudes of the terms should be estimated for each case to check their significance. Formulas given in this section may be used for that purpose.

If the duct is axially symmetric, it is more convenient to use cylindrical coordinates than the cartesian coordinates of this section. We give therefore in Appendix A all pertinent formulas in cylindrical coordinates.

5. EXAMPLES OF TUBE FLOWS

5.1 Incompressible Steady Flow Through Cylindrical Tubes

In the case of an incompressible steady duct flow the flow velocity is constant along the duct and dependent on the radial coordinate r only. Also, only the axial coordinate u of the velocity is non-zero. Therefore, of all the correction terms given in Appendix A, only C_ϕ can be non-zero in this case. It is given by eq. (A.35), which reduces to

$$C_\phi = \bar{u}\,\bar{T}_z - \frac{2}{R^2}\int_0^R u\,\frac{\partial}{\partial r}(r\,\tau_{rz})\,dr\ . \tag{5.1.1}$$

The shearing stress $\tau_{rz}(r)$ is in the present case a linear function of r. This is a consequence of the second momentum equation (A.13) which reduces to

$$\frac{\partial p}{\partial z} = \frac{1}{r}\frac{\partial}{\partial r}(r\,\tau_{rz})\ . \tag{5.1.2}$$

The left-hand side of eq. (5.1.2) is constant. Therefore, τ_{rz} must be linear in r:

$$\tau_{rz}(r) = \frac{r}{R}\,\tau_{rz}(R)\ . \tag{5.1.3}$$

Substituting eq. (5.1.3) into eq. (5.1.1), we obtain

$$C_\phi = \bar{u}\,\bar{T}_z - \frac{4\tau_{rz}(R)}{R^3}\int_0^R u\,r\,dr =$$

$$= \bar{u}\,\bar{T}_z - 2\,\bar{u}\,\tau_{rz}(R)\,\frac{1}{R}\ . \tag{5.1.4}$$

The average shear stress $\overline{T}_z$ is, according to eqs. (A.17) and (A.21),

$$\overline{T}_z = \frac{2}{R^2} \int_0^R \frac{\partial}{\partial r} (r \tau_{rz}) \, dr = 2\tau_{rz}(R) \frac{1}{R} . \tag{5.1.5}$$

Substituting eq. (5.1.5) into eq. (5.1.4) we see that the correction term C_Φ is zero.

Hence the average flow equations are exact for incompressible steady flows through circular tubes. This is essentially a consequence of eq. (5.1.3) and the result is valid for either turbulent or laminar flows. Also, we have not made use of Stokes equations for the stress tensor, nor made any assumptions about the viscosity of the fluid.

5.2 Lagrange's Interior Ballistics Flow

As an example for non-steady tube flows we consider Lagrange's approximation to interior ballistics flow (Reference 7). The approximation is obtained by postulating that the average axial velocity $\bar{u}$ of the gas in a gun tube is at any time a linear function of the axial distance z, i.e.,

$$\bar{u}(z,t) = \frac{z}{z_p(t)} u_p(t) , \tag{5.2.1}$$

where $z_p(t)$ and $u_p = dz_p/dt$ are the location and velocity of the projectile, respectively. We assume that the local velocity can have axial as well as radial components which may depend on z, t, and on the radial coordinate r.

Some consequences of the assumption (5.2.1) are discussed in the Appendix B. In summary, the discussion shows that this assumption, complemented with a second Lagrange's assumption

$$\rho = \rho_o \frac{z_p(0)}{z_p(t)} , \tag{5.2.2}$$

---

[7] *J. Corner, Theory of the Interior Ballistics of Guns, John Wiley and Sons, New York, 1950.*

is consistent with the average continuity equation (4.6) for the flow. One can also assume that the local velocity vector has the form

$$u^* = \begin{pmatrix} \bar{u}(z,t)\cdot f(r) \\ \bar{v}(z,t)\cdot h(r) \end{pmatrix}. \qquad (5.2.3)$$

For any reasonable functions $\bar{u}(z,t)$ and $f(r)$ one can determine corresponding funtions $\bar{v}(z,t)$ and $h(r)$ such that the local continuity equation is satisfied. (The necessary formulas are given in Appendix B.) However, a flow characterized by eqs. (5.2.1) through (5.2.3) in general does not satisfy the local momentum equations if constant viscosities are assumed. Hence Lagrange's approximation, (5.2.1) and (5.2.2), and a local velocity field of the type (5.2.3) can be consistent only for inhomogeneous media, i.e., media with variable viscosity.

Because an exact solution of the viscous tube flow equations is not available, we cannot obtain exact values for all correction terms. However, the correction term in the momentum equation is independent of the viscosities and can be computed exactly for any flow profile. In contrast, the correction terms in the energy equation can be computed only if additional information is available about the stress tensor and the internal energy profile. These terms we will estimate by computing their values for constant viscosities and for a number of "reasonable" flow profiles. We expect by such calculations to obtain at least order-of-magnitude estimates of the correction terms.

Particularly we will consider flow profiles of two types. First we will assume a flow field which is described by

$$\left.\begin{aligned} u(r,z,t) &= z\,\frac{u_p(t)}{z_p(t)}\,\frac{n+2}{n}\left[1-\left(\frac{r}{R}\right)^{n}\right] \\ \text{and} \\ v(r,z,t) &= -R\,\frac{u_p(t)}{z_p(t)}\,\frac{1}{n}\left(\frac{r}{R}\right)\left[1-\left(\frac{r}{R}\right)^{n}\right]. \end{aligned}\right\} \qquad (5.2.4)$$

This flow field has a Hagen-Poiseuille profile for n = 2. For larger values of n it approximates turbulent flow profiles or profiles with thin boundary layers.

As a second example we will consider a flow profile which approximates the universal profile for steady turbulent tube flow.

The flow field defined by eq. (5.2.4) satisfies the local continuity equation, if the density is given by eq. (5.2.2). We note the interesting fact that local continuity requires the radial flow component to be directed toward the center of the tube. This is due to the higher mass flow rate at the center and due to the assumed increase of the average axial velocity $\bar{u}(z,t)$ with z.

The correction term C_m of the momentum equation is given for our flow by eq. (B.48)

$$C_m = \left\{ 1 - \frac{2}{R^2} \int_0^R f^2 r \, dr \right\} \frac{\partial}{\partial z} (\rho u^2) \, . \tag{5.2.5}$$

Substituting

$$f(r) = \frac{n+2}{n} \left[1 - \left(\frac{r}{R} \right)^n \right] \tag{5.2.6}$$

into eq. (5.2.5), we obtain

$$C_m = - \frac{1}{n+1} \frac{\partial}{\partial z} (\bar{\rho} \, \bar{u}^2) . \tag{5.2.7}$$

The momentum equation is therefore in terms of the average axial velocity

$$\frac{\partial (\bar{\rho} \, \bar{u})}{\partial t} + \left(1 + \frac{1}{n+1} \right) \frac{\partial}{\partial z} (\bar{\rho} \, \bar{u}^2) + \frac{\partial \bar{p}}{\partial z} = \bar{T}_z \, . \tag{5.2.8}$$

Eq. (5.2.8) shows that in the case of a Hagen-Poiseuille profile the momentum transport term in the momentum equation should be increased by about 33%. Even for a rather flat profile with, say, n = 10 the correction term is 9% in this example.

The first correction term C_{e1} of the energy equation is zero in our example because the density ρ is independent of r and z. (See Appendix B for a discussion of this term.)

The second correction term C_{e2} of the energy equation is (see eq. (B.54))

$$C_{e2} = \frac{\partial}{\partial z}\left\{\bar{\rho}\ \bar{u}\ \frac{2}{R^2}\int_0^R (1-f)\ e\ r\ dr\right\} =$$

$$= \rho\ \frac{u_p}{z_p}\ \frac{2}{R^2}\ \frac{\partial}{\partial z}\left\{z\int_0^R (1-f)\ e\ r\ dr\right\}. \qquad (5.2.9)$$

In order to compute this term we would need to make an assumption about the internal energy function e. For the present discussion we will not make any assumptions and leave eq. (5.2.9) unchanged.

The average heat dissipation function $\bar{\Phi}$, which appears on the right hand-side of the energy equation can be computed by the formulas (B.56) and (B.57). The result of the computation is

$$\bar{\Phi} = 2\mu\bar{u}^2\ \frac{1}{R^2}\ \frac{(n+2)^2}{n} + \left(\frac{u_p}{z_p}\right)^2\left(2\mu\ \frac{n+3}{n+1} - \frac{2}{3}\ \mu + \mu'\right). \qquad (5.2.10)$$

The equation for the average internal energy (first law of thermodynamics) is in our case

$$\frac{\partial(\bar{\rho}\ \bar{e})}{\partial t} + \frac{\partial}{\partial z}\ (\bar{\rho}\ \bar{e}\ \bar{u}) + \bar{p}\ \frac{\partial\bar{u}}{\partial z} = \bar{H} + \bar{\Phi} + C_{e2}\ . \qquad (5.2.11)$$

Substituting eqs. (5.2.9) and (5.2.10) into eq. (5.2.11), we obtain

$$\frac{\partial(\bar{\rho}\ \bar{e})}{\partial t} + \frac{\partial}{\partial z}\left[\bar{\rho}\ \bar{e}\ \bar{u}\left(1 - \frac{1}{\bar{e}}\ \frac{2}{R^2}\int_0^R (1-f)\ e\ r\ dr\right)\right] + \bar{p}\ \frac{\partial\bar{u}}{\partial z} =$$

$$= \bar{H} + 2\mu\ \bar{u}^2\ \frac{1}{R^2}\left[\frac{(n+2)^2}{2n} + \left(\frac{2}{3}\ \frac{n+4}{n+1} + \frac{\mu'}{2\mu}\right)\left(\frac{R}{z}\right)^2\right]. \qquad (5.2.12)$$

In eq. (5.2.12) we have included the correction term C_{e2} into the energy flux term on the left hand side. It is readily apparent from the form of the term that the correction is zero, if the specific internal energy e is independent of r.

In the heat dissipation function on the right-hand side of eq. (5.2.12) the term with the factor $(R/z)^2$ can generally be neglected, because $(R/z)^2$ is of the order 10^{-2}. (R/z is large in the vicinity of the breech, where the one-dimensional approximation should not be used anyway.) The other term in the square brackets is usually replaced by $-\bar{u}\,\bar{T}_z$. If this is done, then two additional correction terms should be included in the equation. The general formulas for these terms are given by eqs. (B.61) and (B.62). They are in our case

$$\bar{W}_o = \left(\frac{u_p}{z_p}\right)^2 \left\{\frac{2}{R^2}\int_o^R \mu\left[R^2h'^2 + \left(\frac{R}{r}\right)^2 h^2 + 2f^2\right] r\,dr + \mu' - \frac{2}{3}\mu\right\} =$$

$$= \left(\frac{u_p}{z_p}\right)^2 \cdot \left[2\mu\,\frac{n+5/2}{n+1} + \mu' - \frac{2}{3}\mu\right] \qquad (5.2.13)$$

and

$$C_\Phi = \bar{u}\,\bar{T}_z + \left(\frac{u_p}{z_p}\right)^2 \mu\,\frac{2}{R^2}\int_o^R \left[R^2h'^2 + \left(\frac{R}{z}\right)^2 h^2 + z^2f'^2\right] r\,dr =$$

$$= \bar{u}\,\bar{T}_z + \left(\frac{u_p}{z_p}\right)^2 2\mu\,\frac{1}{2(n+1)} + 2\mu\,\bar{u}^2\,\frac{1}{R^2}\,\frac{(n+2)^2}{2n} \qquad (5.2.14)$$

The first term $\bar{u}\,\bar{T}_z$ in eq. (5.2.14) is according to eq. (B.63)

$$\bar{u}\,\bar{T}_z = 2\mu\,\bar{u}^2\,\frac{1}{R}\,f'(R) = -\,2\mu\,\bar{u}^2\,\frac{1}{R^2}\,(n+2) = -\left(\frac{u_p}{z_p}\right)^2\left(\frac{z}{R}\right)^2 2\mu\,(n+2). \qquad (5.2.15)$$

Comparing eqs. (5.2.13) and (5.2.15), we see that the term $\bar{W}_o$ is indeed small relative to the magnitude of $\bar{u}\,\bar{T}_z$. In Section 3 such a ratio of magnitudes was anticipated based on plausibility arguments.

The total correction is the sum of C_Φ and $\bar{W}_o$. Combining eqs. (5.2.13) through (5.2.15) we obtain for the sum

$$C_\Phi + \bar{W}_o = \bar{u}\,\bar{T}_z\left[\frac{n-2}{2n} - \left(\frac{R}{z}\right)^2 \frac{2}{3}\,\frac{n+4}{(n+1)(n+2)}\left(1 + \frac{\mu'}{\mu}\,\frac{3(n+1)}{4(n+4)}\right)\right] \qquad (5.2.16)$$

The right-hand side of the energy equation (5.2.12) is thus

$$\bar{H} + \bar{\Phi} = \bar{H} - \bar{u}\,\bar{T}_z + C_\Phi + \bar{W}_o = \bar{H} - \overline{u\,T}_z \quad (1\text{-}a) \ , \qquad (5.2.17)$$

where a is a relative correction which is to be applied to $\overline{u\,T}_z$. It is given by

$$a = \frac{n-2}{2n} - \left(\frac{R}{z}\right)^2 \frac{2}{3} \frac{n+4}{(n+1)(n+2)} \left(1 + \frac{\mu'}{\mu} \frac{3}{4} \frac{n+1}{n+4}\right) \ . \qquad (5.2.18)$$

The second term in this formula can in general be neglected, because $(R/z)^2$ is of the order 10^{-2}. The first term is zero only for n = 2, i.e., for a Hagen-Poiseuille flow profile. In this case the shear stress is a linear function of r, which causes certain correction terms to vanish, as shown in Section 5.1. For a flat flow profile with, say, n = 10, the relative correction is a = 0.4. Clearly such a 40% approximation error will be seldom tolerable. Hence for flat flow profiles and constant viscosities the approximation of $\bar{\Phi}$ by $-\overline{u\,T}_z$ is not realistic for calculations in interior ballistics.

The flow profile which is defined by eq. (5.2.4) does not have the characteristic form of a fully developed turbulent flow profile for any n. We may therefore ask whether the correction terms are possibly smaller for such a profile. In order to investigate this question we approximate the universal turbulent profile (see, e.g., Reference 8, page 512) by defining

$$f(r) = 0.456 \left\{1 - \left(\frac{r}{R}\right)^{1.5} + 2\left[1 - \left(\frac{r}{R}\right)^{15}\right]\right\} \ . \qquad (5.2.19)$$

The corresponding function h(r) is

$$h(r) = -\,0.456\,\frac{r}{R}\left\{\frac{1}{3.5}\left[1 - \left(\frac{r}{R}\right)^{1.5}\right] + \frac{2}{17}\left[1 - \left(\frac{r}{R}\right)^{15}\right]\right\} \ . \qquad (5.2.20)$$

---

[8] *H. Schlichting, Boundary Layer Theory, McGraw-Hill, New York (4th Edition), 1960.*

The correction term C_m of the momentum equation can now be computed using eq. (5.2.5). The result is

$$C_m = - 0.090 \frac{\partial}{\partial z} (\bar{\rho}\, \bar{u}^2) \ . \qquad (5.2.21)$$

In analogy to eq. (5.2.8) we conclude from eq. (5.2.21) that the momentum flux term in the momentum equation should be increased by 9% in the present case.

Assuming as before constant viscosities, we obtain for the average heat dissipation function

$$\bar{\Phi} = 2\mu\, \bar{u}^2 \frac{1}{R^2}\, 7.840 + \left(\frac{u_p}{z_p}\right)^2 (1.699\, \mu + \mu') \qquad (5.2.22)$$

For the product $-\bar{u}\, \bar{T}_z$ we obtain

$$- \bar{u}\, \bar{T}_z = - 2\mu\, \bar{u}^2 \frac{1}{R^2} f'(R) = 2\mu\, \bar{u}^2 \frac{1}{R^2} \cdot 14.364 \ . \qquad (5.2.23)$$

The right-hand side of the energy equation (5.2.12) is therefore

$$\bar{H} + \bar{\Phi} = \bar{H} - \bar{u}\, \bar{T}_z \left[1 - 0.454 + \left(\frac{R}{z}\right)^2 0.059 \left(1 + 0.588 \frac{\mu'}{\mu}\right)\right] \qquad (5.2.24)$$

The error which is introduced by replacing $\bar{\Phi}$ by $-\bar{u}\, \bar{T}_z$ is about 45% of $|\bar{u}\, \bar{T}_z|$. As in the previously treated case, such errors will be seldom tolerable.

We may conclude from these examples that the magnitudes of correction terms are essentially the same for flow profiles described by eq. (5.2.4) as for profiles described by eqs. (5.2.19) and (5.2.20). Using conventional tube flow equations, e.g. from Reference 4, for interior ballistics calculations, one introduces errors in the momentum and energy equations which are of the order of 9-50% of several of the terms. The examples indicate that an investigation of magnitudes of the correction terms is necessary whenever average flow equations are used to describe non-steady tube flows.

6. CONCLUSIONS

Tube flow governing equations for average properties differ from one-dimensional flow equations. The differences are caused by the fact that averaging of functions and multiplication of functions are not commutative operations. The magnitudes of the differences depend on the particular problem. If the unsteady tube flow is of a type which is encountered in interior ballistics, then several terms in the equations can be in error by up to 50%.

One consequence of the various correction terms in the equations is that the continuity and momentum equations cannot be combined to yield a simple equation for the average axial velocity component. Instead, the original equation for the average axial momentum component is the simplest form. Correspondingly, the energy equation should be formulated for the internal energy per unit volume instead of using the specific internal energy.

The popular approximation of the heat dissipation function by the product of average velocity and average shear stress is appropriate only in the simplest cases, e.g., for steady flows or flows with a Hagen-Poiseuille velocity profile. In other cases the approximation can be off by up to 50%. In cases of more complicated flows even the sign of the approximation can be wrong. Hence the approximation should not be used unless one can demonstrate its validity in the particular case of application.

Formulas for the correction terms in the governing equations can be derived for other than simple tube flows following the outline of this paper. The derivations which are presented in some engineering textbooks neglect important first-order terms. The apparent success of the inaccurate equations for the treatment of tube flows is probably due to the fact that the neglected terms are small or vanish for steady flows, for which most comparisons between calculation and experiments are made.

REFERENCES

1. H.S. Tsien, "The Equations of Gas Dynamics," in Fundamentals of Gas Dynamics, edited by H.W. Emmons, Princeton University Press, 1958.

2. G.K. Batchelor, An Introduction to Fluid Dynamics, Cambridge University Press, 1967.

3. G. Hamel, Mechanik der Kontinua, R.G. Teubner, Stuttgart, 1956.

4. A.H. Shapiro, The Dynamics and Thermodynamics of Compressible Fluid Flow, Vol. I and II, Roland Press Company, New York, 1954.

5. L. Crocco, "One-Dimensional Treatment of Steady Gas Dynamics" in Fundamentals of Gas Dynamics II, edited by W. Emmons, Princeton University Press, 1958.

6. G.B. Wallis, One-Dimensional Two-Phase Flow, McGraw-Hill Co, New York, 1969.

7. J. Corner, Theory of the Interior Ballistics of Guns, John Wiley and Sons, New York, 1950.

8. H. Schlichting, Boundary Layer Theory, McGraw-Hill, New York (4th Edition), 1960.

APPENDIX A

FORMULAS IN CYLINDRICAL COORDINATES

In Section 2 through 4 a convenient cartesian tensor notation was used to derive all formulas. If the results are to be used for axially symmetric tube flows, then it is more convenient to use cylindrical coordinates. In this appendix we express the important formulas in these coordinates.

Stokes equation for the stress tensor of an isotropic fluid can be expressed in coordinate independent form as follows (Reference 3, page 132; Reference 2, page 144)

$$\tau = 2\mu\,\varepsilon + (\mu' - \tfrac{2}{3}\mu)\,\mathrm{div}\,u^{*}\cdot I\ , \qquad (A.1)$$

where τ is the stress tensor, ε is the strain rate tensor, u^* is the velocity vector of the fluid, and I is the unit tensor. The viscosities μ and μ' in eq. (A.1) need not be constant, i.e., the fluid under consideration need not be homogeneous. However, μ as well as μ' must be positive or zero.

Next we compute the work rate of the viscous forces acting within an arbitrary volume. To this end we compute the inner product of the viscous forces $\nabla\cdot\tau$ with the velocity vector u^* and integrate over the volume. The result can be expressed as follows:

$$\int (u^{*}\cdot(\nabla\cdot\tau))\,dV = \oint (u^{*}\cdot(\tau\cdot n))\,dS - \int \Phi\,dV\ . \qquad (A.2)$$

In eq. (A.2) n is a unit vector, orthogonal to the surface of the volume V and pointing inward, and Φ is the heat dissipation function defined by

$$\Phi = \{\tau\varepsilon\}_{trace} = 2\mu\,\{\varepsilon^2\}_{trace} + (\mu' - \tfrac{2}{3}\mu)\,\mathrm{div}\,u^{*}\,\{\varepsilon\}_{trace}\ . \qquad (A.3)$$

Because $\mathrm{div}\,u^{*} = \{\varepsilon\}_{trace}$, eq. (A.3) can be also expressed as follows:

$$\Phi = 2\mu\,\{\varepsilon^2\}_{trace} + (\mu' - \tfrac{2}{3}\mu)\,\{\varepsilon\}^2_{trace}\ . \qquad (A.4)$$

Eq. (A.4) corresponds to eq. (2.10) in cartesian coordinates. Eq. (A.2) corresponds to eq. (2.13) in cases where the body forces X_j are absent.

We now express the various quantities appearing in the equations using cylindrical coordinates. Let the coordinates be r, ϕ and z. Components of vectors and tensors we denote by attaching corresponding indexes to the quantities. Thus, the velocity vector u* is

$$u^* = (u_r,\ u_\phi,\ u_z)\ . \tag{A.5}$$

The strain rate tensor ε has the following components (Reference 2, Page 602)

$$\left.\begin{aligned}
&\varepsilon_{rr} = \frac{\partial u_r}{\partial r}, \quad \varepsilon_{\phi\phi} = \frac{1}{r}\frac{\partial u_\phi}{\partial \phi} + \frac{1}{r}u_r, \quad \varepsilon_{zz} = \frac{\partial u_z}{\partial z}, \\
&\varepsilon_{r\phi} = \frac{1}{2}\left[r\frac{\partial}{\partial r}\left(\frac{1}{r}u_\phi\right) + \frac{1}{r}\frac{\partial u_r}{\partial \phi}\right], \\
&\varepsilon_{rz} = \frac{1}{2}\left[\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r}\right], \\
&\varepsilon_{\phi z} = \frac{1}{2}\left[\frac{1}{r}\frac{\partial u_z}{\partial \phi} + \frac{\partial u_\phi}{\partial z}\right].
\end{aligned}\right\} \tag{A.6}$$

The vector $\nabla\cdot\tau$ has the components T_r, T_ϕ and T_z, representing the viscous forces acting in the three coordinate directions. The components are

$$(\nabla\cdot\tau)_r = T_r = \frac{\partial \tau_{rr}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{r\phi}}{\partial \phi} + \frac{\partial \tau_{rz}}{\partial z} + \frac{1}{r}(\tau_{rr} - \tau_{\phi\phi})\ ,$$

$$(\nabla\cdot\tau)_\phi = T_\phi = \frac{\partial \tau_{r\phi}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\phi\phi}}{\partial \phi} + \frac{\partial \tau_{\phi z}}{\partial z} + 2\frac{1}{r}\tau_{r\phi}\ ,$$

$$(\nabla\cdot\tau)_z = T_z = \frac{\partial \tau_{rz}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\phi z}}{\partial \phi} + \frac{\partial \tau_{zz}}{\partial z} + \frac{1}{r}\tau_{rz}\ . \tag{A.7}$$

We specialize these equations for the case of an axisymmetric flow without swirl through a circular tube. The flow is then independent of the coordinate ϕ, and the ϕ-component of the velocity, u_ϕ, is zero. In order to simplify the notation we denote the non-zero velocity components as follows:

$$u_z = u(t,z,r) \ ,$$
$$u_r = v(t,z,r) \ . \tag{A.8}$$

Let R be the radius of the tube. The average density is then defined by

$$\bar{\rho}(t,z) = \frac{2}{R^2} \int_0^R \rho(t,z,r)\ r\ dr \ . \tag{A.9}$$

The average axial velocity is

$$\bar{u}(t,z) = \frac{2}{R^2 \bar{\rho}(t,z)} \int_0^R \rho(t,z,r)\ u(t,z,r)\ r\ dr \ . \tag{A.10}$$

The local continuity equation is

$$\frac{\partial \rho}{\partial t} + \frac{1}{r} \frac{\partial}{\partial r} (r\ \rho\ v) + \frac{\partial}{\partial z} (\rho\ u) = 0 \ . \tag{A.11}$$

The corresponding equation for the averages is

$$\frac{\partial \bar{\rho}}{\partial t} + \frac{\partial}{\partial z} (\bar{\rho}\ \bar{u}) = 0 \ . \tag{A.12}$$

The local balance of momentum is expressed by the following two differential equations:

$$\left.\begin{aligned} &\frac{\partial(\rho\ v)}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}(\rho\ r\ v^2) + \frac{\partial}{\partial z}(\rho\ u\ v) + \frac{\partial p}{\partial r} = T_r \\ &\frac{\partial(\rho\ u)}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}(\rho\ r\ u\ v) + \frac{\partial}{\partial z}(\rho\ u^2) + \frac{\partial p}{\partial z} = T_z \end{aligned}\right\} \quad (A.13)$$

The third momentum equation is staisfied identically because of our symmetry assumptions. The right-hand sides of eqs. (A.13) depend on the strain rate tensor ε by eq. (A.7) and (A.1). In our case the strain rate tensor has the following components:

$$\left.\begin{aligned} &\varepsilon_{rr} = \frac{\partial v}{\partial r}\ , \quad \varepsilon_{\phi\phi} = \frac{1}{r}v\ , \quad \varepsilon_{zz} = \frac{\partial u}{\partial z}\ , \\ &\varepsilon_{r\phi} = 0\ , \quad \varepsilon_{\phi z} = 0\ , \\ &\varepsilon_{rz} = \frac{1}{2}\left[\frac{\partial v}{\partial z} + \frac{\partial u}{\partial r}\right]. \end{aligned}\right\} \quad (A.14)$$

The divergence of the velocity vector is

$$\operatorname{div} u^* = \{\varepsilon\}_{trace} = \varepsilon_{rr} + \varepsilon_{\phi\phi} + \varepsilon_{zz} =$$

$$= \frac{1}{r}\frac{\partial}{\partial r}(r\ v) + \frac{\partial u}{\partial z}\ . \quad (A.15)$$

The viscous stress tensor τ has the components

$$\left.\begin{aligned}
\tau_{rr} &= 2\mu \frac{\partial v}{\partial r} + (\mu' - \frac{2}{3}\mu)\ \mathrm{div}\ u^* \ , \\
\tau_{r\phi} &= 0 \ , \\
\tau_{rz} &= \mu \left[\frac{\partial v}{\partial z} + \frac{\partial u}{\partial r}\right] , \\
\tau_{\phi\phi} &= 2\mu \frac{1}{r} v + (\mu' - \frac{2}{3}\mu)\ \mathrm{div}\ u^* \ , \\
\tau_{\phi z} &= 0 \ , \\
\tau_{zz} &= 2\mu \frac{\partial u}{\partial z} + (\mu' - \frac{2}{3}\mu)\ \mathrm{div}\ u^* \ .
\end{aligned}\right\} \quad (A.16)$$

The right-hand side of the local momentum equations (A.13) is

$$\left.\begin{aligned}
(\nabla\cdot\tau)_r &= T_r = \frac{1}{r}\frac{\partial}{\partial r}(r\,\tau_{rr}) + \frac{\partial \tau_{rz}}{\partial z} - \frac{1}{r}\tau_{\phi\phi} \\
(\nabla\cdot\tau)_z &= T_z = \frac{1}{r}\frac{\partial}{\partial r}(r\,\tau_{rz}) + \frac{\partial \tau_{zz}}{\partial z}
\end{aligned}\right\} \quad (A.17)$$

Substituting (A.16) into (A.17), we obtain

$$T_r = \frac{1}{r}\frac{\partial}{\partial r}\left(2\mu\, r\, \frac{\partial v}{\partial r}\right) - 2\mu\, \frac{v}{r^2} + \frac{\partial}{\partial r}\left[\left(\mu' - \frac{2}{3}\mu\right) \operatorname{div} u^*\right] + \frac{\partial}{\partial z}\left[\mu\, \frac{\partial v}{\partial z} + \mu\, \frac{\partial u}{\partial r}\right],$$

$$T_z = \frac{1}{r}\frac{\partial}{\partial r}\left[\mu\, r\, \frac{\partial v}{\partial z} + \mu\, r\, \frac{\partial u}{\partial r}\right] + \frac{\partial}{\partial z}\left[2\mu\, \frac{\partial u}{\partial z} + \left(\mu' - \frac{2}{3}\mu\right) \operatorname{div} u^*\right] \quad (A.18)$$

The momentum equation for the averages is

$$\frac{\partial(\bar{\rho}\,\bar{u})}{\partial t} + \frac{\partial}{\partial z}\left(\bar{\rho}\,\bar{u}^2\right) + \frac{\partial \bar{p}}{\partial z} = \bar{T}_z + C_m\,, \quad (A.19)$$

where the average pressure $\bar{p}$ is defined by

$$\bar{p} = \frac{2}{R^2}\int_0^R p(t,z,r)\, r\, dr \quad (A.20)$$

and the average viscous force $\bar{T}_z$ by

$$\bar{T}_z = \frac{2}{R^2}\int_0^R T_z(t,z,r)\, r\, dr\,. \quad (A.21)$$

The correction term C_m in eq. (A.19) is

$$C_m = \frac{\partial}{\partial z}\left\{\bar{\rho}\,\bar{u}^2 - \frac{2}{R^2}\int_0^R \rho\, u^2\, r\, dr\right\}. \quad (A.22)$$

The equation for the local internal energy is

$$\frac{\partial(\rho\, e)}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}\left(r\, \rho\, e\, v\right) + \frac{\partial}{\partial z}\left(\rho\, e.u\right) + p\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\, v\right) + \frac{\partial u}{\partial z}\right]. \quad (A.23)$$

The heat dissipation function Φ is given by eq. (A.4). Substituting the strain rate tensor components from eq. (A.6) into (A.4), we obtain

$$\Phi = 2\mu \left[\left(\frac{\partial v}{\partial r}\right)^2 + \frac{1}{2}\left(\frac{\partial v}{\partial z} + \frac{\partial u}{\partial r}\right)^2 + \left(\frac{v}{r}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2\right] + \\ + (\mu' - \frac{2}{3}\mu)\left[\frac{1}{r}\frac{\partial}{\partial r}(r\,v) + \frac{\partial u}{\partial z}\right]^2 . \qquad (A.24)$$

The energy equation for the average internal energy is

$$\frac{\partial(\bar{\rho}\,\bar{e})}{\partial t} + \frac{\partial}{\partial z}(\bar{\rho}\,\bar{e}\,\bar{u}) + \bar{p}\frac{\partial \bar{u}}{\partial z} = \bar{H} + \bar{\Phi} + C_{e1} + C_{e2} , \qquad (A.25)$$

where

$$\bar{e} = \frac{1}{\bar{\rho}} \cdot \frac{2}{R^2}\int_0^R \rho\, e\, r\, dr , \qquad (A.26)$$

$$\bar{H} = \frac{2}{R^2}\int_0^R (Q - \operatorname{div} q)\, r\, dr , \qquad (A.27)$$

$$\bar{\Phi} = \frac{2}{R^2}\int_0^R \Phi\, r\, dr , \qquad (A.28)$$

$$C_{e1} = \bar{p}\frac{\partial \bar{u}}{\partial z} - \frac{2}{R^2}\int_0^R p\left[\frac{\partial u}{\partial z}\, r + \frac{\partial}{\partial r}(r\,v)\right] dr \qquad (A.29)$$

and

$$C_{e2} = \frac{\partial}{\partial z}(\bar{\rho}\,\bar{u}\,\bar{e}) - \frac{2}{R^2}\int_0^R \frac{\partial}{\partial z}(\rho\, u\, e)\, r\, dr . \qquad (A.30)$$

The term $-\bar{u}\,\bar{T}_z$ is often used instead of $\bar{\Phi}$ in eq. (A.25). In that case the equation becomes

$$\frac{\partial(\bar{\rho}\,\bar{e})}{\partial t} + \frac{\partial}{\partial z}(\bar{\rho}\,\bar{e}\,\bar{u}) + \bar{p}\,\frac{\partial \bar{u}}{\partial z} = \bar{H} - \bar{u}\,\bar{T}_z + C_{e1} + C_{e2} + \bar{W}_o + C_\phi \,. \qquad (A.31)$$

The additional correction terms $\bar{W}_o$ and C_ϕ are

$$\bar{W}_o = \frac{2}{R^2}\int_o^R \mathrm{div}(\tau\, u^*)\, r\, dr \qquad (A.32)$$

and

$$C_\phi = \bar{u}\,\bar{T}_z - \frac{2}{R^2}\int_o^R u^* \cdot (\nabla\cdot\tau)\, r\, dr \,. \qquad (A.33)$$

In eq. (A.32) we have

$$\tau\, u^* = \begin{pmatrix} \tau_{rr} v + \tau_{rz} u \\ 0 \\ \tau_{rz} v + \tau_{zz} u \end{pmatrix}$$

and

$$\mathrm{div}(\tau\, u^*) = \frac{1}{r}\frac{\partial}{\partial r}\left[r\,(\tau_{rr} v + \tau_{rz} u)\right] + \frac{\partial}{\partial z}\left[\tau_{rz} v + \tau_{zz} u\right] .$$

Therefore

$$\bar{W}_o = \frac{2}{R^2}\int_o^R \frac{\partial}{\partial z}\left[\tau_{rz} v + \tau_{zz} u\right] r\, dr =$$

$$= \frac{2}{R^2}\int_o^R \frac{\partial}{\partial z}\left[\mu\, v\left(\frac{\partial v}{\partial z} + \frac{\partial u}{\partial r}\right) + 2\mu\, u\,\frac{\partial u}{\partial z} + \left(\mu' - \frac{2}{3}\mu\right) u\, \mathrm{div}\, u^*\right] r\, dr \,, \qquad (A.34)$$

where div u* is given by (A.15).

The integrand in (A.33) can be obtained from eq. (A.17). Carrying out the substitutions, we obtain for the second correction term

$$C_\phi = \bar{u}\,\bar{T}_z - \frac{2}{R^2}\int_0^R (v\,T_r + u\,T_z)\,r\,dr =$$

$$= \bar{u}\,\bar{T}_z - \frac{2}{R^2}\int_0^R \left\{ v\left[\frac{1}{r}\frac{\partial}{\partial r}(r\,\tau_{rr}) + \frac{\partial \tau_{rz}}{\partial z} - \frac{1}{r}\tau_{\phi\phi}\right] + \right.$$

$$\left. + u\left[\frac{1}{r}\frac{\partial}{\partial r}(r\,\tau_{rz}) + \frac{\partial \tau_{zz}}{\partial z}\right]\right\} r\,dr\,. \qquad \text{(A.35)}$$

The separate expressions for C_ϕ and $\bar{W}_o$ might be of interest for the discussion of approximations. Usually C_ϕ is neglected completely and $\bar{W}_o$ is assumed to be small by plausibility arguments. The total correction, which is caused by replacement of $\bar{\Phi}$ by $-\bar{u}\,\bar{T}_z$, is the sum of $\bar{W}_o$ and C_ϕ. The sum is, of course,

$$\bar{W}_o + C_\phi = \bar{u}\,\bar{T}_z + \bar{\Phi} =$$

$$= \bar{u}\,\bar{T}_z + \frac{2}{R^2}\int_0^R \left\{ 2\mu\left[\left(\frac{\partial v}{\partial r}\right)^2 + \frac{1}{2}\left(\frac{\partial v}{\partial z} + \frac{\partial u}{\partial r}\right)^2 + \left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{v}{r}\right)^2\right] + \right. \qquad \text{(A.36)}$$

$$\left. + \left(\mu' - \frac{2}{3}\mu\right)(\operatorname{div} u^*)^2\right\} r\,dr\,.$$

Eq. (A.36) may be more advantageous for actual calculations than (A.34) and (A.35) because it does not contain derivatives of the viscosities.

APPENDIX B

LAGRANGE'S APPROXIMATION TO INTERIOR BALLISTICS FLOW

It is plausible to assume that in a gun tube the average axial velocity $\bar{u}(z,t)$ of each cross-section is a linear function of the distance z from the breech of the weapon. Let $z_p(t)$ be the location of the projectile and $u_p(t) = dz_p/dt$ be its velocity. The above-mentioned Lagrange's approximation is then

$$\bar{u}(z,t) = \frac{z}{z_p(t)}\, u_p(t) \ . \tag{B.1}$$

In the classical Lagrange's approximation (B.1) is supplemented with the assumption that the gas density in the tube is a function of time only.

In this appendix we shall investigate some consequences of these assumptions. Particularly we are interested in finding if there is a three-dimensional viscous tube flow which satisfies Lagrange's assumptions.

First we will consider flows in which the gas density is a separable function of z, t, and the radial coordinate r:

$$\rho(r,z,t) = g(r) \cdot P(z) \cdot K(t) \ . \tag{B.2}$$

Later we will specialize our considerations to the classical Lagrange's approximation, where P(z) and g(r) are constants.

We assume that g(r) is non-dimensional and normalized by

$$\frac{2}{R^2} \int_0^R g(r)\ r\ dr = 1 \ . \tag{B.3}$$

The product of the other two functions in eq. (B.2) is then the average density

$$\bar{\rho}(z,t) = P(z)\ K(t) = \frac{2}{R^2} \int_0^R \rho(r,z,t)\ r\ dr \ . \tag{B.4}$$

In eq. (B.1) the variables z and t are already separated. We assume that the dependence of u on r can be separated also, such that

$$u(r,z,t) = f(r) \cdot \bar{u}(z,t) \ . \tag{B.5}$$

It was shown in Section 4 that a reasonable definition of the average axial velocity $\bar{u}$ in terms of the local velocity u is

$$\bar{u} = \frac{1}{\rho} \cdot \frac{2}{R^2} \int_0^R u \rho r \, dr \ . \tag{B.6}$$

With this definition of $\bar{u}$ we have the following relation between the nondimensional functions f(r) and g(r):

$$\frac{2}{R^2} \int_0^R f(r) \, g(r) \, r \, dr = 1 \ . \tag{B.7}$$

The funtions $\bar{u}(z,t)$ and $\bar{\rho}(z,t)$ satisfy the continuity equation (4.6), i.e.,

$$\frac{\partial \bar{\rho}}{\partial t} + \frac{\partial}{\partial z} (\bar{\rho} \, \bar{u}) = 0 \ . \tag{B.8}$$

Substituting the product P·K for $\bar{\rho}$ into eq. (B.8) and the expression (B.1) for $\bar{u}$ we obtain

$$P(z) \cdot K'(t) + K(t) \frac{u_p(t)}{z_p(t)} \frac{d}{dz} \left(z \, P(z) \right) = 0 \ . \tag{B.9}$$

This equation has solutions of the form

$$\bar{\rho} = \rho_o \left(\frac{z}{z_p(t)} \right)^m (m+1) \, \frac{z_p(o)}{z_p(t)} \tag{B.10}$$

with arbitrary m. In eq. (B.10) ρ_o is the average density of the gas in the tube at time t = 0. For m = 0 we obtain the classical Lagrange's

solution. More generally we may assume $\bar{\rho}$ to be, e.g., of the form

$$\bar{\rho} = \frac{\rho_o}{A_o + \frac{1}{m+1} A_m} \left[A_o + A_m \left(\frac{z}{z_p(t)} \right)^m \right] \frac{z_p(o)}{z_p(t)} \tag{B.11}$$

with arbitrary m, A_o and A_m. For physical reasons $m \geq 0$, $A_o \geq 0$ and $A_m \geq -A_o$.

Next we investigate the radial velocity component v(r,z,t). The local continuity equation is

$$\frac{\partial \bar{\rho}}{\partial t} + \frac{1}{r} \frac{\partial}{\partial r}(r \rho v) + \frac{1}{r} \frac{\partial}{\partial \phi} (\rho w) + \frac{\partial}{\partial z} (\rho u) = 0, \tag{B.12}$$

where w is the angular velocity component. Let w = 0 (no swirl) and

$$v = \bar{v}(z,t) \cdot h(r) . \tag{B.13}$$

Eq. (B.12) can then be expressed by

$$\frac{\partial \bar{\rho}}{\partial t} \cdot g(r) + \bar{\rho}\, \bar{v}\, \frac{1}{r} \frac{d}{dr} [r\, g(r)\, h(r)] + g(r)\, f(r) \frac{\partial}{\partial z} [\bar{\rho}\, \bar{u}] = 0 . \tag{B.14}$$

Eliminating $\partial\bar{\rho}/\partial t$ from eq. (B.14) with the aid of eq. (B.8), we obtain

$$[-1+f(r)]\, g(r) \frac{\partial(\bar{\rho}\, \bar{u})}{\partial z} + \bar{\rho}\, \bar{v}\, \frac{1}{r} \frac{d}{dr}[r\, g(r)\, h(r)] = 0 . \tag{B.15}$$

This equation is satisfied by the functions

$$\bar{v}(z,t) = R \frac{1}{\bar{\rho}} \frac{\partial(\bar{\rho}\, \bar{u})}{\partial z} \tag{B.16}$$

and

$$h(r) = \frac{1}{R\, r\, g} \int_o^r (1-f)\, g\, r\, dr . \tag{B.17}$$

Eqs. (B.13), (B.16), and (B.17) give the local radial velocity $v(r,z,t,)$ for any flow profile specified by $\bar{\rho}$, $\bar{u}$, g, and f. Clearly, the factor h(r) is not normalized in the same manner as f(r). Therefore, $\bar{v}(z,t)$ is not an "average" radial velocity. If $\bar{\rho}$ is given by eq. (B.11), then

$$\bar{v}(z,t) = \frac{A_o + (m+1)\left(\frac{z}{z_p(t)}\right)^m A_m}{A_o + \left(\frac{z}{z_p(t)}\right)^m A_m} \frac{u_p(t)}{z_p(t)} \quad . \tag{B.18}$$

It is interesting to note that $\bar{v}$ is not zero for $z = 0$ and $z = z_p$. This is an indication that the assumption (B.5) about separation of variables for the axial velocity is not valid in the vicinities of the breech and the projectile. These regions we will therefore exclude from our considerations.

In summary we have found a flow field in a cylindrical tube which satisfies the local continuity equation and Lagrange's assumption (B.1). The flow field is described by the following functions

$$\left.\begin{aligned} u &= \frac{z}{z_p(t)} u_p(t) \cdot f(r) \\ \rho &= \bar{\rho}(z,t) \cdot g(r) \\ v &= \bar{v}(z,t) \cdot h(r) \end{aligned}\right\} \tag{B.19}$$

If one specifies u, then $\bar{\rho}$ is given by eq. (B.11) and $\bar{v}$ is given by eq. (B.18). The dependence of the flow field on r can be specified by two functions, g(r) and f(r), from which h(r) is then computed by eq. (B.17). The function g(r) has to be positive for $0 \leq r \leq R$ and normalized by eq. (B.3). We assume also that $g'(0) = 0$. The function f(r) has to satisfy the conditions

and

$$\left.\begin{aligned} f'(o) &= 0 \\ f(R) &= 0 \ . \end{aligned}\right\} \tag{B.20}$$

It is normalized by eq. (B.7). Hence we have a total of three conditions which restrict the choice of f(r).

Instead of specifying f(r) we may also specify h(r). The function f(r) is then given by

$$f(r) = 1 - \frac{R}{r\, g(r)} [r\, g(r)\, h(r)]' \tag{B.21}$$

The function h(r) has to satisfy the following four conditions

$$\left.\begin{aligned} h(0) &= 0 \,, \\ h''(0) &= 0 \,, \\ h(R) &= 0 \,, \\ h'(R) &= 1/R \,. \end{aligned}\right\} \tag{B.22}$$

The flow field also has to satisfy the momentum equations. The analysis of these equations is more complicated because it involves, in addition to the velocity and density functions, the pressure function p(r,z,t) and the viscosities μ and μ', which in general are variable. We have tried to restrict our considerations to the special case with constant viscosities and classical Lagrange's approximation (i.e., g(r) ≡ 1). We have found that the flow field, as defined by eq. (B.19), does not satisfy the momentum equations in this special case. The formulas for correction terms, which we shall derive at the end of this appendix, are therefore to be considered as approximations only.

If g(r) ≡ 1, then the flow field is given by

$$u = \bar{u}(z,t)\, f(r) = z\, \frac{u_p(t)}{z_p(t)}\, f(r) \,, \tag{B.23}$$

$$v = R\, \frac{\partial \bar{u}}{\partial z}\, h(r) = R\, \frac{u_p(t)}{z_p(t)}\, h(r) = \frac{u_p}{z_p}\, \frac{1}{r} \int_0^r (1-f)\, r\, dr \,, \tag{B.24}$$

$$\rho = \bar{\rho}(t) = \rho_o\, \frac{z_p(o)}{z_p(t)} \,. \tag{B.25}$$

Note that according to eq. (B.24) the radial velocity component v is independent of z for the classical Lagrange's approximation.

The divergence of the velocity u* of this flow is (see eq. (A.15))

$$\text{div } u^* = \frac{1}{r}\frac{\partial(rv)}{\partial r} + \frac{\partial u}{\partial z} = \frac{\partial \bar{u}}{\partial z} = \frac{u_p(t)}{z_p(t)} . \qquad (B.26)$$

The components of the forces caused by the viscous stress tensor are given by eq. (A.18). In our case, assuming constant viscosities, we obtain for the r-component

$$T_r = 2\mu\,\bar{v}\,\frac{1}{r}\,[(r\,h')' - \frac{1}{r}\,h] + \mu\,\frac{\partial \bar{u}}{\partial z}\,f' =$$

$$= [-2\mu\,f' + \mu\,f']\,\frac{\partial \bar{u}}{\partial z} = -\,\mu\,\frac{u_p}{z_p}\,f' . \qquad (B.27)$$

The z-component of the force is

$$T_z = \frac{1}{r}\,\mu\,\bar{u}\,[r\,f']' = \mu\,z\,\frac{u_p}{z_p}\,\frac{1}{r}\,[r\,f']' . \qquad (B.28)$$

The local momentum equations are according to eqs. (A.13) and (B.12)

$$\rho\,\frac{\partial v}{\partial t} + \rho v\,\frac{\partial v}{\partial r} + \rho u\,\frac{\partial v}{\partial z} + \frac{\partial p}{\partial r} = T_r , \qquad (B.29)$$

$$\rho\,\frac{\partial u}{\partial t} + \rho v\,\frac{\partial u}{\partial r} + \rho u\,\frac{\partial u}{\partial z} + \frac{\partial p}{\partial z} = T_z . \qquad (B.30)$$

For a flow field described by eqs. (B.23) through (B.24) eq. (B.29) is

$$\rho R\,\frac{d}{dt}\left(\frac{u_p}{z_p}\right) h + \rho R^2 \left(\frac{u_p}{z_p}\right)^2 h\,h' + \frac{\partial p}{\partial r} = -\,\mu\,\frac{u_p}{z_p}\,f' . \qquad (B.31)$$

Differentiating eq. (B.31) with respect to z, we obtain

$$\frac{\partial^2 p}{\partial r \partial z} = 0 \ . \qquad \text{(B.32)}$$

The function $p(r,z,t)$ is therefore of the form

$$p(r,z,t) = p_1(r,t) + p_2(z,t) \ . \qquad \text{(B.33)}$$

Eq. (B.31) might be used to determine the function $p_1(r,t)$ if the other terms in the equation are given.

Eq. (B.30) is in the present case, i.e., for the flow described by eqs. (B.23) through (B.25)

$$\rho z \ f \frac{d}{dt}\left(\frac{u_p}{z_p}\right) + \rho R \left(\frac{u_p}{z_p}\right)^2 z \ h \ f' + \rho z \left(\frac{u_p}{z_p}\right)^2 f^2 + \frac{\partial p_2}{\partial z} = \mu z \frac{u_p}{z_p} \frac{1}{r} (r \ f')' \qquad \text{(B.34)}$$

or

$$\rho f \frac{z_p}{u_p} \frac{d}{dt}\left(\frac{u_p}{z_p}\right) (R \ f'h + f^2) + \frac{z_p}{u_p} \frac{1}{z} \frac{\partial p_2}{\partial z} - \mu \frac{1}{2} (r \ f')' = 0 \ . \qquad \text{(B.35)}$$

From eq. (B.35) we can conclude that the expression $(\partial p_2/\partial z)/z$ is independent of z. The various terms in this equation are products of functions of r and t and the equation has the form

$$g_1(t) f_1(r) + g_2(t) f_2(r) + g_3(t) + f_3(r) = 0 \ . \qquad \text{(B.36)}$$

Such an equation can be satisfied identically only if either all $g_i(t)$ are constant or all $f_i(r)$ are constant. The case with all $f_i(r)$ = const. corresponds to a slug flow in which we are not interested. Assuming the time functions $g_i(t)$ to be constant, we obtain first

$$g_2(t) = \rho \frac{u_p}{z_p} = A$$

or

$$\rho_o z_p(o) \frac{dz_p}{dt} \frac{1}{z_p} = A \ . \tag{B.37}$$

Eq. (B.37) can be integrated to yield

$$z_p(t) = z_p(o) \frac{1}{1-At/\rho_o} \tag{B.38}$$

The corresponding velocity of the projectile is

$$u_p(t) = z_p(o) \frac{A}{\rho_o} \frac{1}{(1-At/\rho_o)^2} \tag{B.39}$$

The density as a function of time is

$$\rho(t) = A \frac{z_p}{u_p} = \rho_o \left(1-A \frac{t}{\rho_o}\right) \ . \tag{B.40}$$

The first time function in eq. (B.35) is then

$$g_1(t) = \rho \frac{z_p}{u_p} \frac{d}{dt} \left(\frac{u_p}{z_p}\right) = \rho_o \left(1-A \frac{t}{\rho_o}\right)^2 \frac{d}{dt} \left(\frac{1}{1-At/\rho_o}\right) \equiv A \tag{B.41}$$

Let the value of the third time function be B. We obtain then

$$\frac{\partial p_2}{\partial z} = B \cdot z \frac{u_p}{z_p} = B \ z \frac{A}{\rho_o} \frac{1}{1-At/\rho_o} = A \ B \ \rho_o \frac{z}{z_p(o)} \ z_p(t) \tag{B.42}$$

Eq. (B.35) takes now the form

$$A(f+f^2+R \ h \ f') + B - \mu \frac{1}{r} (r \ f')' = 0 \ . \tag{B.43}$$

The functions f(r) and h(r) are related by eq. (B.21). We can therefore express (B.43) in terms of h(r) only. We also multiply the equation by R^2/μ. The result is

$$A^* \left[1 - \frac{R}{r} (r\,h)' + \left(1 - \frac{R}{2} (r\,h)' \right)^2 - R.h \left(\frac{R}{r} (r\,h)' \right)' \right] +$$

$$+ \frac{R^2}{r} \left\{ r \left[\frac{R}{2} (r\,h)' \right]' \right\}' + B^* = 0 \,, \qquad \text{(B.44)}$$

with the constants

$$A^* = \frac{R^2}{\mu} A = \frac{R\,u_p \rho}{\mu} \cdot \frac{R}{z_p} \qquad \text{(B.45)}$$

and

$$B^* = \frac{R^2}{\mu} B = \left(\frac{\partial p}{\partial z} \cdot \frac{R^2}{\mu\, u_p} \right) \frac{z_p}{z} \,. \qquad \text{(B.46)}$$

The first factor in eq. (B.45) is a Reynolds number of the projectile. It is typically of the order 10^6. The first factor in eq. (B.46) for B* is a Poiseuille number of the flow. Its magnitude is of the order 10^4. The function h(r) has to satisfy the differential equation (B.44) and the four boundary conditions (B.22). Since eq. (B.44) is of third order only, the function h(r) will in general not satisfy all boundary conditions. We conclude, therefore, that Lagrange's approximation is not consistent with a flow field which can be described by separation of variables, eqs. (B.23) through (B.25).

In Section 5.2 we have nevertheless used this flow field to obtain estimates of correction terms because we were not able to find an exact three-dimensional solution of Navier-Stokes equations which is also consistent with Lagrange's approximation.

Next we compute the various correction terms for the average flow equations using the formulas of Appendix A and the flow described by eq. (B.19).

The correction term C_m of the momentum equation is given by eq. (A.22):

$$C_m = \frac{\partial}{\partial z}\left\{\bar{\rho}\ \bar{u}^2 - \frac{2}{R^2}\int_0^R \rho u^2 r\ dr\right\} . \qquad (B.47)$$

Substituting the expressions (B.23) and (B.25) for u and ρ respectively into eq. (B.47), we obtain

$$C_m = \frac{\partial}{\partial z}\left\{\bar{\rho}\ \bar{u}^2 - \bar{\rho}\ \bar{u}^2\ \frac{2}{R^2}\int_0^R f^2(r)\ r\ dr\right\}$$

$$= \left\{1 - \frac{2}{R^2}\int_0^R f^2\ r\ dr\right\}\frac{\partial}{\partial z}\left(\bar{\rho}\ \bar{u}^2\right) . \qquad (B.48)$$

Eq. (B.48) is of course valid for any functions $u = \bar{u}(z,t)\cdot f(r)$ and $\rho = \bar{\rho}(z,t)$. In Lagrange's case $\bar{\rho}$ is independent of z, and $\bar{u}$ is linear in z. We obtain in this case

$$C_m = \left\{1 - \frac{2}{R^2}\int_0^R f^2\ r\ dr\right\} 2z\ \bar{\rho}(t)\ \frac{u_p(t)}{z_p(t)} . \qquad (B.49)$$

The energy equation has several correction terms. First we consider the term C_{el}, given by eq. (A.29):

$$C_{el} = \bar{p}\ \frac{\partial \bar{u}}{\partial z} - \frac{2}{R^2}\int_0^R p\ \left[r\ \frac{\partial u}{\partial z} + \frac{\partial}{\partial r}\left(r\ v\right)\right] dr . \qquad (B.50)$$

We substitute the flow field formulas (B.19) into this equation and obtain

$$C_{el} = \bar{p}\ \frac{\partial \bar{u}}{\partial z} - \frac{2}{R^2}\int_0^R \left[r\ f\ \frac{\partial \bar{u}}{\partial z} + \bar{v}\ (r\ h)'\right] p\ dr =$$

$$= \bar{p}\ \frac{\partial \bar{u}}{\partial z} - \frac{2}{R^2}\int_0^R \left[\frac{\partial \bar{u}}{\partial z}\left(r - \frac{R}{g}\ (r\ h)'\right) + \frac{R}{\bar{\rho}}\ \frac{\partial(\bar{\rho}\ \bar{u})}{\partial z}\ (r\ h)'\right] p\ dr . \qquad (B.51)$$

This expression can be transformed by partial integration and some algebra into

$$C_{e1} = \frac{2}{R}\frac{\partial \bar{u}}{\partial z}\int_0^R \left(\frac{1}{g} - 1\right)(r\,h)'\,p\,dr - \frac{2}{R}\frac{\bar{u}}{\bar{\rho}}\frac{\partial \bar{\rho}}{\partial z}\int_0^R (r\,h)'\,p\,dr\,. \qquad (B.52)$$

Eq. (B.52) shows that the correction term C_{e1} is zero in the following cases:

(a) $g \equiv 1$ and $\frac{\partial \bar{\rho}}{\partial z} = 0$, i.e., the classical Lagrange's case;

(b) $g \equiv 1$ and $\frac{\partial p}{\partial r} = 0$, i.e., ρ and p independent of r;

(c) $h \equiv 0$, or $f \equiv 1$, i.e., slug flow.

The second correction term of the energy equation, C_{e2}, depends on the local internal energy. According to eq. (A.30),

$$C_{e2} = \frac{\partial}{\partial z}\left\{\bar{\rho}\,\bar{u}\,\bar{e} - \frac{2}{R^2}\int_0^R \rho\,u\,e\,r\,dr\right\}. \qquad (B.53)$$

If the flow field is given by eq. (B.19), then

$$C_{e2} = \frac{\partial}{\partial z}\left\{\bar{\rho}\,\bar{u}\,\frac{2}{R^2\rho}\int_0^R \rho\,e\,r\,dr - \frac{2}{R^2}\int_0^R \rho u e r\,dr\right\} =$$

$$= \frac{\partial}{\partial z}\left\{\bar{\rho}\,\bar{u}\,\frac{2}{R^2}\int_0^R g\,(1-f)\,e\,r\,dr\right\}\,. \qquad (B.54)$$

This correction term vanishes if the internal energy e is independent of the radial coordinate r.

The remaining correction terms, $\bar{W}_o$ and C_Φ, in the energy equation are caused by the replacement of the heat dissipation function $\bar{\Phi}$ by the product $-\bar{u}\,\bar{T}_z$. The heat dissipation function Φ is according to eq. (A.24)

$$\Phi = 2\mu\left[\bar{v}^2 h'^2 + \frac{1}{2}\left(\frac{\partial \bar{v}}{\partial z}h + \bar{u}\,f'\right)^2 + \frac{1}{r^2}\bar{v}^2 h^2 + \left(\frac{\partial \bar{u}}{\partial z}\right)^2 f^2\right] +$$

$$+ \left(\mu' - \frac{2}{3}\mu\right)\left[\frac{1}{r}\bar{v}\,h' + \frac{\partial \bar{u}}{\partial z}f\right]^2\,. \qquad (B.55)$$

For the flow field described by eqs. (B.23) through (B.25) we obtain

$$\Phi = 2\mu \left[R^2 h'^2 + \left(\frac{R}{r}\right)^2 h^2 + f^2 + \frac{1}{2} z^2 f'^2\right] \left(\frac{u_p}{z_p}\right)^2 + \left(\mu' - \frac{2}{3}\mu\right) \left(\frac{u_p}{z_p}\right)^2 \qquad (B.56)$$

The average $\overline{\Phi}$ is by definition

$$\overline{\Phi} = \frac{2}{R^2} \int_0^R \Phi \, r \, dr \, . \qquad (B.57)$$

The correction term $\overline{W}_o$ is according to eq. (A.34)

$$\overline{W}_o = \frac{2}{R^2} \int_0^R \left[\mu \frac{\partial \bar{u}}{\partial z} f' \bar{v} \, h + 2\mu \frac{\partial}{\partial z} \left(\bar{u} \frac{\partial \bar{u}}{\partial z}\right) f^2 + (\mu' - \frac{2}{3}\mu) \frac{\partial \bar{u}}{\partial z} \frac{u_p}{z_p} f\right] r \, dr =$$

$$= \frac{2}{R^2} \left(\frac{u_p}{z_p}\right)^2 \int_0^R \left[\mu(f'h + 2f^2) + (\mu' - \frac{2}{3}\mu) f\right] r \, dr \, . \qquad (B.58)$$

The correction term C_ϕ is given by eq. (A.35). In the present case with constant viscosities we obtain by substituting eqs. (B.27) and (B.28) into eq. (A.35)

$$C_\phi = \bar{u} \, \bar{T}_z - \frac{2}{R^2} \int_0^R \left[-\mu \, \bar{v} \, h \frac{u_p}{z_p} f' + \mu \, \bar{u} \, f \, z \frac{u_p}{z_p} \frac{1}{r} (r \, f')'\right] r \, dr =$$

$$= \bar{u} \, \bar{T}_z + \frac{2}{R^2} \int_0^R \left[\mu \, R \, h \, f' + \mu \, z^2 f'^2\right] \left(\frac{u_p}{z_p}\right)^2 r \, dr \, . \qquad (B.59)$$

By partial integration and using eq. (B.21) we can show that

$$\int_0^R R \, h \, f' r \, dr = \int_0^R \left(R^2 h'^2 + \left(\frac{R}{r}\right)^2 h^2\right) r \, dr \, . \qquad (B.60)$$

Making use of eq. (B.60) we can express the correction terms as follows:

$$\overline{W}_o = \left(\frac{u_p}{z_p}\right)^2 \frac{2}{R^2} \int_o^R \left[\mu\, R^2 h'^2 + \mu \left(\frac{R}{r}\right)^2 h^2 + 2\mu\, f^2\right] r\, dr + \left(\frac{u_p}{z_p}\right)^2 \left(\mu' - \frac{2}{3}\mu\right) \tag{B.61}$$

and

$$C_\Phi = \bar{u}\, \overline{T}_z + \left(\frac{u_p}{z_p}\right)^2 \frac{2}{R^2} \int_o^R \left[\mu\, R^2 h'^2 + \mu \left(\frac{R}{r}\right)^2 h^2 + \mu\, z^2 f'^2\right] r\, dr\ . \tag{B.62}$$

The first term in eq. (B.62) is according to eq. (B.28)

$$\bar{u}\, \overline{T}_z = \bar{u}^2 \mu \frac{2}{R^2} \int_o^R (r\, f')'\, dr = 2\mu\, z^2 \left(\frac{u_p}{z_p}\right)^2 \frac{1}{R} f'(R)\ . \tag{B.63}$$

The correction terms $\overline{W}_o$ and C_Φ vanish for a slug flow. However, for such a flow $\bar{u}\, \overline{T}_z$ and $\bar{\Phi}$ are also zero.

FINITE ELEMENT STRESS ANALYSIS OF AXISYMMETRIC BODIES UNDER TORSION

Tien-Yu Tsui
Army Materials & Mechanics Research Center
Watertown, Massachusetts

ABSTRACT. A finite element procedure for linear stress analysis of axisymmetric bodies subjected to torsional loads is developed. The formulation is based on the assumed stress hybrid model. Applications are made to cylinders and cones. Excellent agreements are obtained between the exact solution and the finite element results.

I. INTRODUCTION. The present study is motivated by the consideration of the stress analysis of artillery projectiles. During firing, the projectile is subjected to a combination of various loads which are (1) axial load-due to linear acceleration of the projectile, (2) centrifugal load-due to angular rotation of the projectile, (3) torsional load-due to angular acceleration of the projectile, (4) internal load-due to setback on H.E. and (5) external load-due to gun tube constraint, band pressure and balloting. In view of the complexity of the geometry of the projectile, a finite element analysis must be performed in order to determine the stresses and deformations in the projectile. Since the projectile has an axis of rotational symmetry, it is only logical that an axisymmetric ring element would model it more accurately and efficiently. In a MIT study [1], which was performed for AMMRC under a contract, an axisymmetric ring element was developed based on the assumed-stress hybrid finite element model. However, it can only treat axisymmetric loads of the projectile.

It is the goal of the present analysis to develop an axisymmetric solid of revolution element which can be used to determine the stresses and deformations in axisymmetric structural bodies under torsional loads. The assumed stress-hybrid model is employed to derive the element stiffness matrix such that the results can be combined with that from the MIT study.

Other finite element formulations for solution of axisymmetric structural bodies under torsion can also be made. The axisymmetric quadrilateral element, based on the displacement formulation in the ANSYS finite element program [2], can be used for modeling axisymmetric structures with non-axisymmetric loadings such as bending, shear or torsion. Different finite element formulations have also been developed for the solution of torsion of nonprismatric bars [3,4].

II. DERIVATION OF ELEMENT STIFFNESS MATRIX. The assumed-stress hybrid model is based on a modified complementary energy principle [5]. It assumes compatible displacements along the interelement boundaries and a stress field which satisfies equilibrium within each element.

The total complementary energy of one finite element is given by

$$\pi_c = \frac{1}{2} \int_V S_{ijk\ell}\, \sigma_{ij}\, \sigma_{k\ell}\, dV - \int_A T_i u_i dA \tag{1}$$

where

$S_{ijk\ell}$ = compliance constants of the material

σ_{ij} = stresses

V = volume

A = boundary area

T_i = surface traction

u_i = prescribed boundary displacement

The function π_c is a minimum when the stresses satisfy the equilibrium condition. In deriving the element stiffness matrix, the displacements along each boundary of the finite element are expressed in terms of the nodal displacements $\underset{\sim}{q}$ and certain interpolation functions $\underset{\sim}{L}$, such that the displacement compatibility conditions with the neighboring elements are satisfied.

$$\underset{\sim}{u} = \underset{\sim}{L}\underset{\sim}{q} \tag{2}$$

The element stresses in the interior of the element are then expanded in terms of a finite number of stress parameters $\underset{\sim}{\beta}$

$$\underset{\sim}{\sigma} = \underset{\sim}{P}\underset{\sim}{\beta} \tag{3}$$

where $\underset{\sim}{P}$ is chosen to satisfy the homogeneous equilibrium equations. The surface tractions can be written in the form of

$$\underset{\sim}{T} = \underset{\sim}{R}\underset{\sim}{\beta} \tag{4}$$

where $\underset{\sim}{R}$ is obtained by applying the element boundary conditions to $\underset{\sim}{P}$. Substituting Eqs. (2) to (4) in Eq. (1) one obtains,

$$\pi_c = \frac{1}{2} \underset{\sim}{\beta}^T \underset{\sim}{H}\underset{\sim}{\beta} - \underset{\sim}{\beta}^T \underset{\sim}{G}\underset{\sim}{q} \tag{5}$$

where

$$\underset{\sim}{H} = \int_V \underset{\sim}{P}^T \underset{\sim}{S}\underset{\sim}{P} dV \tag{6}$$

and

$$\underset{\sim}{G} = \int_A \underset{\sim}{R}^T \underset{\sim}{L} dA \tag{7}$$

The best approximate solution for $\underset{\sim}{\beta}$ for the problem is obtained by setting $\partial\pi_c/\partial\underset{\sim}{\beta}$ to zero. The result is

$$\underset{\sim}{H}\,\underset{\sim}{\beta} - \underset{\sim}{G}\,\underset{\sim}{q} = 0 \tag{8}$$

From which,

$$\underset{\sim}{\beta} = \underset{\sim}{H}^{-1}\underset{\sim}{G}\,\underset{\sim}{q} \tag{9}$$

The first term in the expression of π_c (Eq. (1)) represents the total strain energy U in the element. By substituting Eq. (9) into U one obtains:

$$U = \frac{1}{2}\,\underset{\sim}{q}^T\,\underset{\sim}{G}^T\,\underset{\sim}{H}^{-1}\,\underset{\sim}{G}\,\underset{\sim}{q} \tag{10}$$

By definition, the strains energy can be written as follows:

$$U = \frac{1}{2}\,\underset{\sim}{q}^T\,\underset{\sim}{K}\,\underset{\sim}{q} \tag{11}$$

where $\underset{\sim}{K}$ is the element stiffness matrix.

By comparing Eqs. (10) and (11), one obtains the element stiffness matrix for the hybrid stress model:

$$\underset{\sim}{K} = \underset{\sim}{G}^T\,\underset{\sim}{H}^{-1}\,\underset{\sim}{G} \tag{12}$$

III. FORMULATION OF AXISYMMETRIC SOLID OF REVOLUTION ELEMENT BY ASSUMED-STRESS HYBRID MODEL. Since only structural problems in the shape of body of revolution are considered, an axisymmetric solid of revolution element is developed. For convenience, the cylindrical coordinates are used. Let u, v and w denote, respectively, the components of displacement in the radial (r), tangential (θ) and axial (z) directions. The relationships between the components of displacements and the strain components are:

$$\varepsilon_r = \frac{\partial u}{\partial r}$$

$$\varepsilon_\theta = \frac{u}{r} + \frac{\partial v}{r\partial\theta}$$

$$\varepsilon_z = \frac{\partial w}{\partial z} \tag{13}$$

$$\gamma_{r\theta} = \frac{\partial u}{r\partial\theta} + \frac{\partial v}{\partial r} - \frac{v}{r}$$

$$\gamma_{rz} = \frac{\partial u}{\partial z} + \frac{\partial w}{\partial r}$$

$$\gamma_{z\theta} = \frac{\partial v}{\partial z} + \frac{\partial w}{r\partial\theta}$$

The equilibrium equations are:

$$\frac{\partial \sigma_r}{\partial r} + \frac{1}{r}\frac{\partial \tau_{r\theta}}{\partial \theta} + \frac{\partial \tau_{rz}}{\partial z} + \frac{\sigma_r - \sigma_\theta}{r} = 0$$

$$\frac{\partial \tau_{rz}}{\partial r} + \frac{1}{r}\frac{\partial \tau_{\theta\partial}}{\partial \theta} + \frac{\partial \sigma_z}{\partial z} + \frac{\tau_{rz}}{r} = 0 \qquad (14)$$

$$\frac{\partial \tau_{r\theta}}{\partial r} + \frac{1}{r}\frac{\partial \sigma_\theta}{\partial \theta} + \frac{\partial \tau_{\theta z}}{\partial z} + \frac{2\tau_{r\theta}}{r} = 0$$

In the application of these equations to the torsional problem the semi-inverse method may be used and the components of displacements u and w are assumed to be zero. It can be shown that the solution obtained on the basis of such an assumption satisfies all the equations of elasticity and therefore represents the true solution of the problem [6], Substituting in Eq. (13) $u = w = 0$, and making use of the fact that from symmetry the displacement v does not vary with the angle θ, one obtains:

$$\varepsilon_r = \varepsilon_\theta = \varepsilon_z = \gamma_{rz} = 0$$

$$\gamma_{r\theta} = \frac{\partial v}{\partial r} - \frac{v}{r} \qquad (15)$$

$$\gamma_{\theta z} = \frac{\partial v}{\partial z}$$

Equation (15) together with Hooke's law determines that of all the six stress components σ_r, σ_θ, σ_z, τ_{rz}, $\tau_{r\theta}$, $\tau_{\theta z}$ only $\tau_{r\theta}$ and $\tau_{\theta z}$ are different from zero. As a result of this, the first two of Eq. (14) are identically satisfied, and the third of these equations becomes:

$$\frac{\partial \tau_{r\theta}}{\partial r} + \frac{\partial \tau_{\theta z}}{\partial z} + \frac{2\tau_{r\theta}}{r} = 0 \qquad (16)$$

Hence, in the subsequent formulation of an axisymmetric solid of revolution ring element one is only concerned with the displacement v and the equilibrium equation given by Eq. (16). The ring element has four nodes and a general quadrilateral cross-section in the r-z plane (Fig. 1).

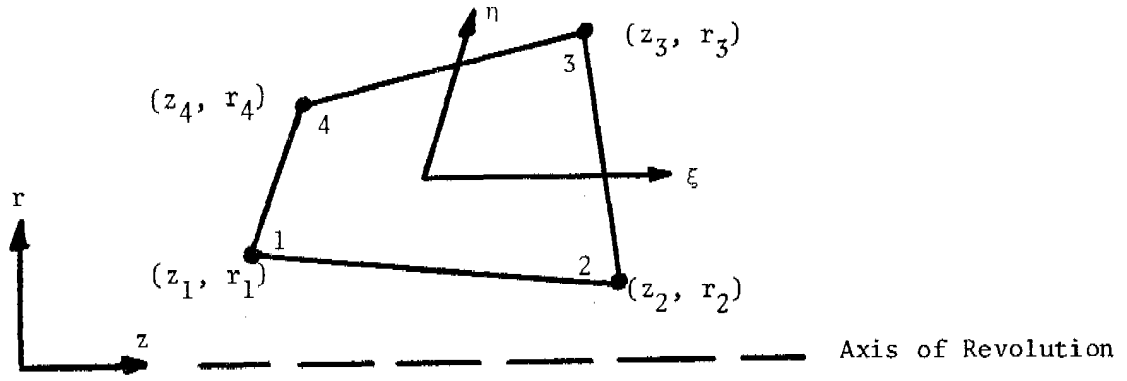


Fig. 1 General Quadrilateral Axisymmetric Solid of Revolution Element in Global System

A. Element Displacement Assumptions

In the assumed-stress hybrid model, the interpolation function must be selected in such a manner that it assures the displacement compatibility between neighboring elements. Thus, for the 4 node 4 degree-of-freedom axisymmetric element the appropriate choice is a linear interpolation. In terms of the nodal quantities it assumes the following form:

$$v = \frac{(1-S)}{2} v_i + \frac{(1+S)}{2} v_{i+1} \tag{17}$$

S is the boundary coordinate which is normalized for each element side such that it varies between -1 and +1. v_i and v_{i+1} are the values of v at the nodes i and i+1 with $v_5 = v_1$. Eq. (17) is the interpolation function for the ith side of the element (i = 1, 2, 3, 4). The value of r and z along the ith side can be related to the nodal coordinate r_i, z_i in the same manner as the displacement interpolation function; i.e.

$$r = \frac{(1-S)}{2} r_i + \frac{(1+S)}{2} r_{i+1}$$

$$z = \frac{(1-S)}{2} z_i + \frac{(1+S)}{2} z_{i+1} \tag{18}$$

When interpolation of the displacements in the interior of an element is required, the following bilinear interpolation function may be used:

$$v = \frac{1}{4} \sum_{i=1}^{4} P_i v_i \tag{19}$$

where v_i is the value of v at the ith node of the element, P_i is defined by the following equation:

$$P_i = (1+\zeta_i \zeta)(1+\eta_i \eta) \tag{20}$$

ζ and η are related to r,z by the following equations

$$z(\zeta,\eta) = \frac{1}{4} \sum_{i=1}^{4} P_i z_i$$

$$r(\zeta,\eta) = \frac{1}{4} \sum_{i=1}^{4} P_i r_i \qquad (21)$$

Equation (21) describes a coordinate transformation between a general quadrilateral in the r-z plane and a square (with side length equal to 2) in the ζ-η plane (Fig. 2).

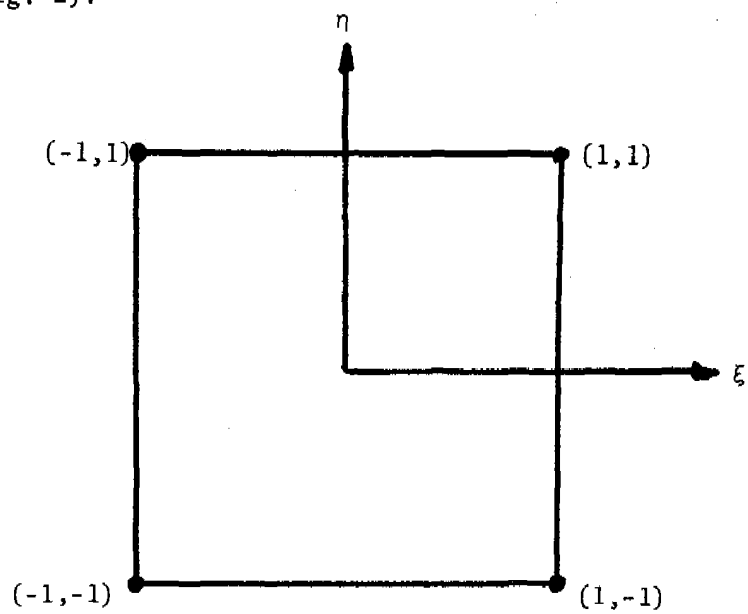


Fig. 2 Solid of Revolution in Transformed ξ-η System

B. Element Stress Assumptions

There is no prior knowledge concerning the selection of proper stress assumptions except they must satisfy the equilibrium equations exactly. In the present analysis, the following stress assumption is chosen:

$$\tau_{r\theta} = \beta_1 + \beta_2 r + \beta_3 z$$

$$\tau_{\theta z} = \beta_4 + \beta_5 r + \beta_6 z \qquad (22)$$

Substituting Eq. (22) into Eq. (16), one obtains

$$\beta_2 + \beta_6 + 2\beta_1 \frac{1}{r} + 2\beta_2 + 2\beta_3 \frac{\sigma}{r} = 0 \tag{23}$$

In order for the assumed stress (Eq. (22)) to satisfy the equilibrium (Eq. (16)) exactly, the following relationships must hold:

$$\begin{aligned} 3\beta_2 + \beta_6 &= 0 \\ 2\beta_1 &= 0 \\ 2\beta_3 &= 0 \end{aligned} \tag{24}$$

As a result, the stress assumptions (Eq. (22)) becomes

$$\begin{aligned} \tau_{r\theta} &= \beta_2 r \\ \tau_{z\theta} &= \beta_4 + \beta_5 r - 3\beta_2 z \end{aligned} \tag{25}$$

By comparing Eqs. (3) and (25), one obtains:

$$\underset{\sim}{\sigma} = \begin{bmatrix} \tau_{r\theta} \\ \tau_{z\theta} \end{bmatrix} \tag{26}$$

$$\underset{\sim}{\beta} = \begin{bmatrix} \beta_2 \\ \beta_4 \\ \beta_5 \end{bmatrix} \tag{27}$$

$$\underset{\sim}{P} = \begin{bmatrix} r & 0 & 0 \\ -3z & 1 & r \end{bmatrix} \tag{28}$$

C. Definition of Compliance Matrix

The matrix $\underset{\sim}{S}$ in Eq. (6) is the compliance matrix that relates stresses $\underset{\sim}{\sigma}$ to strains $\underset{\sim}{\varepsilon}$:

$$\underset{\sim}{\varepsilon} = \underset{\sim}{S}\,\underset{\sim}{\sigma} \tag{29}$$

In the present study (for an isotropic material) S is given as follows:

$$\underset{\sim}{S} = \frac{1}{E} \begin{bmatrix} 2(1+\nu) & 0 \\ 0 & 2(1+\nu) \end{bmatrix} \tag{30}$$

and

$$\underset{\sim}{\varepsilon} = \begin{bmatrix} \gamma_{r\theta} \\ \gamma_{z\theta} \end{bmatrix} \tag{31}$$

D. Calculation of Matrix $\underset{\sim}{H}$

For an axisymmetric element, the matrix $\underset{\sim}{H}$ in Eq. (6) may be written as

$$\underset{\sim}{H} = \int_{V_n} \underset{\sim}{P}^T(r,z)\ \underset{\sim}{S}\ \underset{\sim}{P}\ (r,z)\ r dr d\theta dz \tag{32}$$

Using the coordinates transformation defined in Eq. (21) and performing the θ-integration analytically, the expression for $\underset{\sim}{H}$ becomes

$$\underset{\sim}{H} = 2\pi \int_{-1}^{1} \int_{-1}^{1} \underset{\sim}{P}^T\ (\zeta,\eta)\underset{\sim}{S}\ \underset{\sim}{P}(\zeta,\eta) r(\zeta,\eta) |J| d\zeta d\eta \tag{33}$$

r and z can be written in terms of ζ and η as

$$\begin{aligned} r &= a_0 + a_1\zeta + a_2\eta + a_3\zeta\eta \\ z &= b_0 + b_1\zeta + b_2\eta + b_3\zeta\eta \end{aligned} \tag{34}$$

where

$$\begin{aligned} a_0 &= \frac{1}{4}(r_1 + r_2 + r_3 + r_4) \\ a_1 &= \frac{1}{4}(-r_1 + r_2 + r_3 - r_4) \\ a_2 &= \frac{1}{4}(-r_1 - r_2 + r_3 + r_4) \\ a_3 &= \frac{1}{4}(r_1 - r_2 + r_3 - r_4) \end{aligned} \tag{35}$$

and the b_i's are defined similarly with z_i substituted for r_i. The Jacobian $|J|$ is defined by

$$|J| = \left| \frac{\partial(z,r)}{\partial(\zeta,\eta)} \right| = A_1 + A_2\zeta + A_3\eta \tag{36}$$

where

$$\begin{aligned} A_1 &= a_2b_1 - a_1b_2 \\ A_2 &= a_3b_1 - a_1b_3 \\ A_3 &= a_2b_3 - a_3b_2 \end{aligned} \tag{37}$$

E. Calculations of Matrix G

For the calculation of the matrix $\underset{\sim}{G}$ in Eq. (7), expressions for tractions and displacements along the boundary are required. The traction T_θ for the ith side of an element are given in terms of the stresses by

$$(T_\theta)_i = \tau_{z\theta}(\nu_1)_i + \tau_{r\theta}(\nu_2)_i \tag{38}$$

where

$$\begin{aligned} (\nu_1)_i &= \frac{r_{i+1} - r_i}{\ell_i} \\ (\nu_2)_i &= \frac{z_i - z_{i+1}}{\ell_i} \\ \ell_i &= \sqrt{(r_{i+1} - r_i)^2 + (z_{i+1} - z_i)^2} \end{aligned} \tag{39}$$

Substituting Eq. (25) into Eq. (38), one obtains

$$(T_\theta)_i = [-3z(\nu_1)_i + r(\nu_2)_i]\beta_2 + (\nu_1)_i\beta_4 + r(\nu_2)_i\beta_5 \tag{40}$$

From Eqs. (4), (27) and (40), one obtains

$$\underset{\sim}{T} = [T_\theta] \tag{41}$$

$$\underset{\sim}{R} = [-3z\nu_1 + r\nu_2,\ \nu_1,\ r\nu_2] \tag{42}$$

The boundary displacement are related to the nodal displacements, $\underset{\sim}{q}$, as

$$\underset{\sim}{u} = \underset{\sim}{L}\,\underset{\sim}{q}$$

with

$$\{q_1, q_2, q_3, q_4\} = \{v_1, v_2, v_3, v_4\} \tag{43}$$

Linear boundary displacements are assumed as given by Eqs. (17). By defining

$$\underset{\sim}{L}_1 = \begin{bmatrix} \frac{(1-s)}{2} & 0 \\ 0 & \frac{(1-s)}{2} \end{bmatrix}$$

and

$$\underset{\sim}{L}_2 = \begin{bmatrix} \frac{(1+s)}{2} & 0 \\ 0 & \frac{(1+s)}{2} \end{bmatrix} \tag{44}$$

The matrix $\underset{\sim}{L}$ for each side may be readily written as:

$$\begin{aligned} &\text{Side 1} \quad \underset{\sim}{L} = [\underset{\sim}{L}_1 \vdots \underset{\sim}{L}_2 \vdots 0 \vdots 0] \\ &\text{Side 2} \quad \underset{\sim}{L} = [0 \vdots \underset{\sim}{L}_1 \vdots \underset{\sim}{L}_2 \vdots 0] \\ &\text{Side 3} \quad \underset{\sim}{L} = [0 \vdots 0 \vdots \underset{\sim}{L}_1 \vdots \underset{\sim}{L}_2] \\ &\text{Side 4} \quad \underset{\sim}{L} = [\underset{\sim}{L}_2 \vdots 0 \vdots 0 \vdots \underset{\sim}{L}_1] \end{aligned} \tag{45}$$

The contribution to the matrix $\underset{\sim}{G}$ from each side is now given by

$$\underset{\sim}{G} = 2\pi \int_{-1}^{1} \underset{\sim}{R}^T(S)\, \underset{\sim}{L}(S)\, r(S)\, |J|\, dS \tag{46}$$

The Jacobian, $|J|$, is simply half the length of the side and the total value of $\underset{\sim}{G}$ is obtained by summing the contributions of Eq. (46) from each side. A Gaussian integration rule is used for the evaluation of Eq. (46).

When the matrices $\underset{\sim}{H}$ and $\underset{\sim}{G}$ have been formed, and $\underset{\sim}{H}$ has been inverted (note that $\underset{\sim}{H}$ is symmetric and positive definite), the element stiffness matrix, $\underset{\sim}{k}$, is determined by

$$\underset{\sim}{k} = \underset{\sim}{G}^T \underset{\sim}{H}^{-1} \underset{\sim}{G} \tag{47}$$

IV. NUMERICAL EXAMPLES. In order to examine the adequacy of the element developed in this study, it is employed to analyze two problems with known

analytical solutions. The first problem is the torsion of a solid cylinder (Fig. 3) and the second one is the torsion of a truncated cone (Fig. 4). The exact solution [6] of each problem is given here along with the figure.

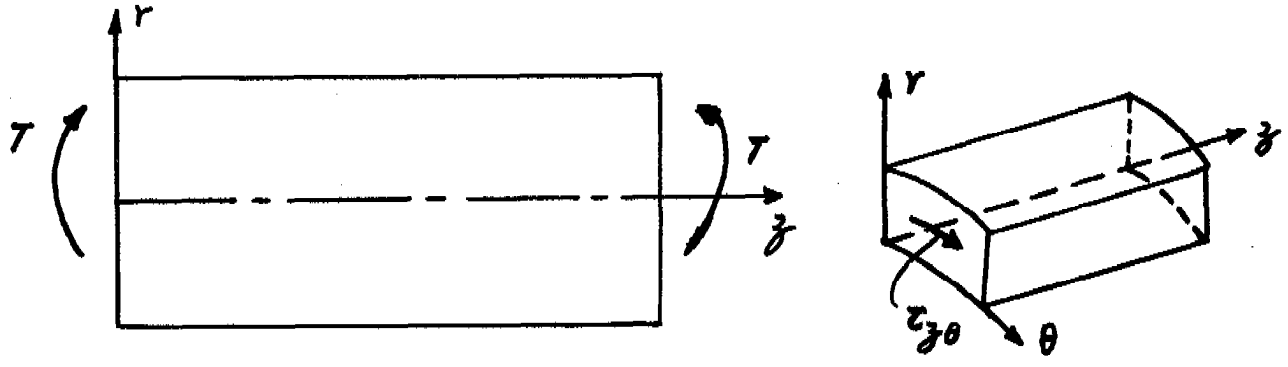


Fig. 3 Torsion of Solid Cylinder

$$\tau_{z\theta} = \frac{Tr}{J} \quad , \quad J = \frac{\pi a^4}{2}$$

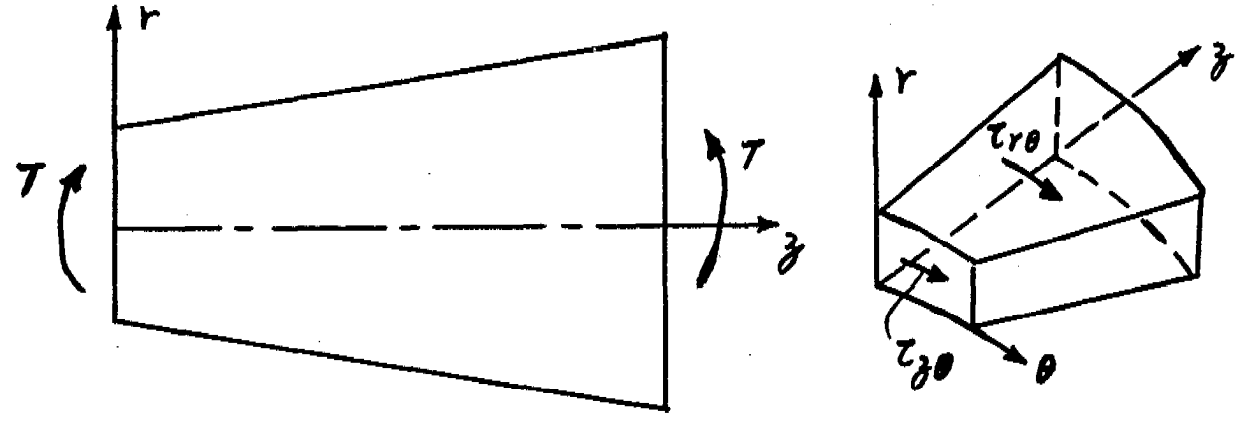


Fig. 4 Torsion of Truncated Cone

$$\tau_{\theta z} = -\frac{crz}{(r^2+z^2)^{5/2}} \quad , \quad \tau_{r\theta} = -\frac{cr^2}{(r^2+z^2)^{5/2}}$$

$$c = -\frac{T}{2\pi(\frac{2}{3} - \cos\alpha + \frac{1}{3}\cos^3\alpha)}$$

The results of the finite element solution of these problems compared with the exact solutions are shown in Tables 1 and 2. Study of the convergence

of the finite element solutions versus the total number of elements used is also indicated in the tables. As it can be seen that an excellent comparison is obtained between the finite element solutions and the exact solutions in the solid cylinder case for the various mesh sizes employed. In the case of truncated cone, excellent agreement is achieved for the $\tau_{z\theta}$ stress in the larger mesh size case.

It should be added here that other forms of stress assumptions were also studied including second and third order terms. However, less than satisfactory results were obtained.

TABLE I

Comparison of Finite Element Solution and Exact Solution (Solid Cylinder)

(Length of Cylinder = 12", Radius = 3.0")

Exact Solution	Location (z) / (r)	96 Elements			24 Elements			8 Elements	
		0.75	1.50	2.25	1.5	3.0	4.5	3.0	6.0
62.88	1.50	62.96	62.89	62.87	62.87	62.91	62.85	62.54	63.00
73.36	1.75	73.38	73.37	73.34					
83.83	2.00	83.87	83.84	83.71	83.71	83.89	83.81		
94.31	2.25	94.33	94.31	94.31				93.69	94.57
104.79	2.50	104.77	104.78	104.80	104.60	104.80	104.81		
115.27	2.75	115.19	115.26	115.28					
125.75	3.00	125.55	125.80	125.72	125.27	125.93	125.66	124.85	126.17

TABLE 2

Comparison of Finite Element Solution and Exact Solution (Truncated Cone)

(Length of Cone = 12", Small End Radius = 1.5", Larger End Radius = 3.0")

Location		$\tau_{z\theta}$			$\tau_{r\theta}$		
z	r	Exact	96 Elements	24 Elements	Exact	96 Elements	24 Elements
1.5	1.687	653.73	564.73	930.52	81.69	95.53	24.71
3.0	1.875	476.70	455.14	384.60	59.59	60.76	70.10
4.5	2.062	358.07	353.93	403.34	41.75	43.13	30.71
6.0	2.25	275.87	276.03	263.42	34.48	32.42	33.26
7.5	2.437	216.94	218.00	225.63	27.11	25.25	21.76
9.0	2.625	173.72	174.74	173.97	21.72	20.16	19.57
10.5	2.813	141.27	141.91	139.62	17.66	16.33	13.17

ACKNOWLEDGEMENT. The author would like to thank Prof. Robert L. Spilker of the University of Illinois at Chicago Circle and Prof. T.H.H. Pian of M.I.T. for their helpful discussions.

V. REFERENCES

1. Spilker, Robert L., "A Study of Elastic-Plastic Analysis by the Assumed-Stress Hybrid Finite Element Model, With Application To Thick Shells of Revolution," AMMRC CTR 74-71 (also MIT ASRL TR 175-1), December, 1974.

2. DeSalvo, Gabriel J. and Swanson, John A., "ANSYS, Engineering Analysis System," Swanson Analysis Systems, Inc., March 1, 1975.

3. Krahula, Joseph L., and Lauterbach, Gerald F., "A Finite Element Solution for Saint-Venant Torsion," AIAA Journal, Vol. 7, No. 12, December, 1969, pp. 2200-2203.

4. Herrmann, L.R., "Elastic Torsional Analysis of Irregular Shapes," Proceedings of the American Society of Civil Engineers, Journal of the Engineering Mechanics Division, Vol. 91, No. EM6, December, 1965, Pt. 1.

5. Pian, T.H.H., "Derivation of Element Stiffness Matrix by Assumed Stress Distributions," AIAA Journal, Vol. 2, No. 7, 1969, pp. 1333-1336.

6. Timoshenko, S. and Goodier, J.N., "Theory of Elasticity," Second Edition, McGraw-Hill, 1951.

THREE-DIMENSIONAL ELASTIC STRESS AND DISPLACEMENT ANALYSIS OF FINITE GEOMETRY SOLIDS CONTAINING CRACKS*

Jonathan Kring, John Gyekenyesi, and Alexander Mendelson
Lewis Research Center and
U.S. Army Air Mobility R&D Laboratory
Cleveland, Ohio 44135

ABSTRACT. The line method of analysis is applied to the Navier-Cauchy equations of elastic equilibrium to calculate the displacement fields in finite geometry bars containing central, surface, and double-edge cracks under extensionally applied uniform loading. The application of this method to these equations leads to coupled sets of simultaneous ordinary differential equations whose solutions are obtained along sets of lines in a discretized region. Normal stresses and the stress intensity factor variation along the crack periphery are calculated using the obtained displacement field. The reported results demonstrate the usefulness of this method in calculating stress intensity factors for commonly encountered crack geometries in finite solids.

INTRODUCTION. The main goal of fracture mechanics is the prediction of the load at which a structure weakened by a crack will fail. Knowledge of the stress and displacement distributions near the crack tip is of fundamental importance in evaluating this load at failure. During the early development of crack mechanics most of the effort was focused on through-thickness cracks which could be characterized as two-dimensional. However, part-through cracks are the most common type of crack defect found in actual service conditions (ref. 1).

Because of the geometric singularity associated with any crack type problem, only limited analytical work has been done in the past on these problems. Early theoretical solutions for three-dimensional flaw configurations usually involved the discussion of cracks in infinite or semi-infinite solids (refs. 2 to 8). For this reason, results for finite geometry stress intensity factors are usually given in terms of magnification factors applied to some convenient reference solution. In addition, considerable scatter exists in the reported results as obtained by different investigators (ref. 9). In our work these difficulties are avoided by solving the finite dimensional problems directly.

Recently, approximate solutions of the finite geometry surface crack problem were obtained by the boundary integral equation method (ref. 10) and the finite element method (ref. 11). An alternate semi-analytical method suitable for the elastic solution of crack problems is the line method of analysis. Successful application of this method to finite geometry solids containing cracks has been demonstrated by Gyekenyesi and Mendelson (ref. 12). Although the concept of the line method for solving partial differential equations is not new (ref. 13), its application in the past has been limited to simple examples. The basis of this technique is the substitution of finite differences for the derivatives with respect to all the independent variables except one for which

*This article has been issued a NASA Technical Memorandum No. 73717.

the derivatives are retained. This approach replaces a given partial differential equation with a system of simultaneous ordinary differential equations whose solutions can then be obtained in closed form. These equations describe the dependent variable along lines which are parallel to the coordinate in whose direction the derivatives were retained. Application of the line method is most useful when the resulting ordinary differential equations are linear and have constant coefficients.

An inherent advantage of the line method over other numerical methods is that good results are obtained from the use of relatively coarse grids. This use of a coarse grid is permissible because parts of the solutions are obtained in terms of continuous functions. Additional accuracy in normal stress distributions is derived from the fact that they are expressed as first-order derivatives of the displacements and these derivatives can be analytically evaluated. Inherently inaccurate numerical differentiation is required only for evaluating the shear stresses, but this presents no important loss of accuracy since they are an order of magnitude smaller than the normal stresses. For problems with geometric singularities, additional accuracy is derived from using a displacement formulation since the resulting deformations are not singular.

It is the purpose of this report to present a simple and systematic approach to the elastic analysis of three-dimensional, finite geometry solids containing traction-free cracks. The need for these specific solutions has existed for a number of years in fracture toughness testing.

REDUCTION OF THE NAVIER-CAUCHY EQUATIONS TO SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS. Within the framework of linearized elasticity theory, the equations of elastic equilibrium in terms of displacements are

$$(\lambda + G)\frac{\partial e}{\partial x} + G\nabla^2 u = 0 \tag{1}$$

$$(\lambda + G)\frac{\partial e}{\partial y} + G\nabla^2 v = 0 \tag{2}$$

$$(\lambda + G)\frac{\partial e}{\partial z} + G\nabla^2 w = 0 \tag{3}$$

where the body forces are assumed to be zero and the dilatation is

$$e = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \tag{4}$$

For a finite geometry solid with rectangular boundaries, we construct three sets of parallel lines (fig. 1(a)). Each set of lines is parallel to one of the coordinate axes and thus perpendicular to the corresponding coordinate plane. An approximate solution of equation (1) can be obtained by developing solutions of ordinary differential equations along the x-directional lines. As seen in the figure, ther are a total of $\iota = NY \times NZ$ such lines where NY is the number of lines along the y-direction and NZ is the number of lines along the z-direction in a given plane, respectively. We define the displacements along these lines as $u_1, u_2, \ldots, u_\iota$. The derivatives of the y-directional displacements on these lines with respect to y are defined as $v'|_1, v'|_2, \ldots, v'|_\iota$, and the derivatives of the z-directional displacements with respect to z are defined as $w'|_1, w'|_2, \ldots, w'|_\iota$. These displacements and derivatives can than be regarded as functions of x only since they are variables on x-directional lines. When these definitions are used, the ordinary differential equation along a generic line ij (a double subscript is used here for simplicity of writing) in figure 1(b) may be written as

$$\frac{d^2u_{ij}}{dx^2} + \frac{(1 - 2\nu)}{2(1 - \nu)} \left[-\left(\frac{2}{h_y^2} + \frac{2}{h_z^2} \right) u_{ij} + (u_{i+1,j} + u_{i-1,j}) \frac{1}{h_y^2} + \frac{1}{h_z^2} (u_{i,j+1} + u_{i,j-1}) \right] + \frac{f_{ij}(x)}{2(1 - \nu)} = 0 \qquad (5)$$

where

$$f_{ij}(x) = \left. \frac{dv'}{dx} \right|_{ij} + \left. \frac{dw'}{dx} \right|_{ij} \qquad (6)$$

and

$$v' = \frac{dv}{dy}, \quad w' = \frac{dw}{dz}, \quad \nu = \frac{\lambda}{2(G + \lambda)} .$$

Similar differential equations are obtained along the other x-directional lines. Since each equation has the terms of the displacements on the surrounding lines, these equations constitute a system of ordinary differential equations for the displacements $u_1, u_2, \ldots, u_\iota$.

The set of ι second order differential equations represented by (5) can be reduced to a set of 2ι first order differential equations by treating the derivatives of the u's as an additional set of ι unknowns, i.e., defining

$$u_{\iota+1} = \frac{du_1}{dx}, \quad u_{\iota+2} = \frac{du_2}{dx}, \quad \text{etc.} \tag{7}$$

The resulting 2ι equations can now be written as a single first order matrix differential equation

$$\frac{du}{dx} = A_1 U + R(x) \tag{8}$$

where U and R are column matrices of 2ι elements each and A_1 is a $2\iota \times 2\iota$ matrix of the constant coefficients appearing in equations (5) and (7).

In a similar manner, to solve equations (2) and (3), ordinary differential equations are constructed along the y- and z-directional lines respectively. These equations are also expressed in an analogous form to equations (8); they are

$$\frac{dV}{dy} = A_2 V + S(y) \tag{9}$$

$$\frac{dw}{dz} = A_3 W + T(z) \tag{10}$$

Equations (8) to (10) are linear first-order ordinary matrix differential equations. They are, however, not independent, but are coupled through the vectors, R, S and T whose components are given by equations similar to (6). The elements of the coefficient matrices A_1, A_2, and A_3 are all constants, being functions of the mesh spacing and Poisson's ratio only.

Noting that a second-order differential equation can satisfy only a total of two boundary conditions and since three-dimensional elasticity problems have three boundary conditions at every point of the bounding surface, some of the boundary data must be incorporated into the surface line differential equations. Hence, conditions of normal stress and displacement are enforced through the constants of the homogeneous solutions while shear stress boundary data must be incorporated into the differential equations of the surface lines. The application of the specified shear conditions permits the use of central difference approximations when surface line differential equations are constructed. The details of constructing these equations are found in reference 14.

SOLUTION OF THE SYSTEMS OF ORDINARY DIFFERENTIAL EQUATIONS. The systems of ordinary differential equations (8) to (10) can be solved by any of a number of standard techniques. The method used herein was basically the matrizant or Peano-Baker method of integration (ref. 15). For equation (8) the solution can be written as

$$U(x) = e^{A_1 x} U(0) + e^{A_1 x} \int_0^x e^{-A_1 \eta} R(\eta)\, d\eta \qquad (11)$$

with similar solutions for equations (9) and (10). U(0) is the initial value vector, determined from the boundary conditions. The conversion of given boundary data into required initial values is discussed in more detail in reference 14.

The matrizant $e^{A_1 x}$ is generally evaluated by its matrix series. For larger values of x, when convergence becomes slow, additive formulas may be used. In addition, similarity transformations can be used to diagonalize the matrix A_1. These various techniques for improving the accuracy are discussed in detail in reference 14.

Since equations (8) to (10) and their boundary conditions are highly coupled, it is generally impossible to directly evaluate their solutions. Thus, a successive approximation procedure must be employed where assumed values must be used initially for the required unknowns. The cyclic resubstitution of the obtained solutions into the coupling vectors and the boundary conditions will usually converge to the correct solution, depending mainly on the accuracy to which the required matrizant can be evaluated.

Once the successive approximation procedure has converged and the displacement field in the body has been calculated, the normal stress distributions can be obtained directly by using the stress-displacement equations. The shear stresses, however, can be evaluated only through finite difference approximations for the required displacement gradients.

STRESS INTENSITY FACTOR. The stress intensity factor K_I was at first obtained from the calculated stresses and displacements by extending the usual definition

$$K_I = \lim_{R \to 0} \sigma_y (2\pi R)^n \qquad (12)$$

to discrete data, where R is measured from and is normal to the crack front and n is the singularity. It was found, however, that due to the coarseness of the grid used, the usual plotting and extrapolating techniques gave results that were erratic and of questionable accuracy. This was compounded by the fact that the precise crack tip location is not really known except that it is approximately midway between two lines, one of which has zero displacement specified in the crack plane and one of which has zero stress specified. It was found, however, that by using two terms in the stress and displacement series expansions around the crack tip, good results could be obtained even with the coarse grid used. Furthermore, this also permitted us to determine the actual crack tip location from the computed results. The method utilized is as follows. We take

$$v\Big|_{y=0} = \alpha K_I \left[\sqrt{\frac{R + r}{2\pi}} + \frac{L_I}{K_I} \sqrt{(R + r)^3} \right] \tag{13}$$

$$\sigma_y\Big|_{y=0} = K_I \left[\frac{1}{\sqrt{2\pi(R - r)}} + \frac{L_I}{K_I} \sqrt{R - r} \right] \tag{14}$$

where α is a function of Poisson's ratio, n was assumed to be -1/2 and r is the crack edge position correction measured from the originally assumed midpoint position. Using displacement data from three adjacent nodes to the crack edge in equation (13), values of αK_I, L_I/K_I, and r are calculated for each value of z, with R also measured from the half-way point between nodes specifying boundary stresses and displacements, respectively. Substituting values of L_I/K_I and r into equation (14), we can calculate K_I as a function of the corrected crack edge distance, $\rho = R - r$. A plot of $\ln K_I$ versus $\sqrt{\rho}$ as $\sqrt{\rho} \to 0$ can then be used to obtain K_I. In a similar manner, α can now be calculated from equation (13), where the corrected crack edge distance with the displacement data is $\rho = R + r$.

APPLICATION TO TENSILE FRACTURE SPECIMENS CONTAINING CRACKS. A great amount of experimental work has been done in fracture mechanics (ref. 16) through the use of crack-notched specimens. In the past, many different types of specimens have been used to determine a material's fracture toughness. The most common early specimens employed in these tests were the center-cracked and double-edge-notched bar specimens. Figures 2(a) and 3(a) show the finite rectangular bars with through-thickness, traction-free central and double-edge cracks, respectively. Because of the symmetric geometry and loading, only one-eighth of each bar has to be discretized as shown in figures 2(b) and 3(b).

NUMERICAL RESULTS. - Center-Cracked Tensile Fracture Specimen. The solution of this problem was obtained by using two different sets of lines along the coordinate axes so that the convergence of the finite difference approximations could be checked. In a given direction, uniform line spacing was used in all computations with no other restriction being placed on the selection of the grid size. The crack edge location with respect to the imposed grid was initially assumed to be halfway between nodes specifying normal stress and displacement boundary conditions, respectively. Subsequently, using the obtained near crack tip stresses and displacements, a more accurate crack edge location was established for calculating the stress intensity factor. This approach was considered acceptable since the results from the two sets of lines at corresponding points did not change, although the crack edge to node distance was considerably decreased for the finer mesh. The successive approximation procedure required for decoupling the three sets of ordinary differential equations was terminated when the difference between successively calculated nondimensionalized displacements, which are of the order of unity, at every point was less than a present value (10^{-6}). As expected, the convergence rate of this

successive approximation procedure was greatly dependent on the initial guess for the required unknowns in the coupling vectors and boundary conditions. For maximum computer efficiency, displacement data obtained from the use of coarse grids was interpolated to obtain improved starting values for the computations involving the final spacing of lines. The required initial quantities for the preliminary coarse grid calcualtions were taken to be zero in our work. All calculations were performed on a UNIVAC 1100/40 computer, using double precision arithmetic.

For the selected geometry, the crack opening displacements and normal stresses from our analysis and those from Raju's finite element method (ref. 17) are compared in Table I. Although in our solution of the gross displacement and stress fields, the minimum corrected crack edge distance is $\rho = 0.042c$, crack opening displacements and stresses can be calculated from equations (13) and (14) at any value of ρ for which these equations are assumed to be accurate ($\rho = 0.40c$ or less in this problem). As seen from Table I, there is good agreement in most displacements, with the normal stress at the surface showing the greatest difference.

The dimensionless crack opening displacement is shown in figure 4. Agreement with the finite element results is seen to be very good. It is noteworthy that the results correspond to elliptical crack profiles in all cases.

An indication of the accuracy of our technique for computing the stress intensity factor is seen in figure 5. This figure shows the stress intensity factor variation across the bar thickness. The results obtained in reference 17 using a finite elements method for a geometry almost identical to one of the geometries in this paper is also shown. It is seen that very good agreement is obtained between two completely different methods. In addition, Isida's plane solution (ref. 18), corrected for finite width and length, is also shown in this figure for comparison. Note that these results indicate a small increas in K_I at the surface with increasing bar thickness. Interestingly, for bars with $t > 3c$, K_I increases gradually with z, reaching a maximum near $z = 0.85t$, and then decreases rapidly to its surface value.

Surface Crack Tensile Fracture Specimen. Figure 6 shows a finite geometry bar containing a traction free rectangular surface crack. Because of the symmetric geometry and loading, only one-fourth of the bar has to be discretized as shown in figure 6(b).

Selected results of the dimensionless surface crack opening displacements are shown in figure 7. Note that the crack opening increases rapidly with crack depth for $0.21 \leq a/t \leq 0.87$, slightly exceeding even the surface crack displacement of a through-thickness crack at $a/t = 0.87$. The plane strain solution for a finite width center cracked bar is also shown in figure 7 for reference. Final displacement values in this report were obtained from a set of 100, 140, and 140 x-, y-, and z-directional differential equations, respectively. A typical computer run for this system of equations takes approximately 30 to 40 minutes of CPU time and 720 K bytes of storage.

In order to show the singularity of the stresses, the y-directional normal stress in the crack plane is plotted in figure 8 for $a/t = 0.536$. The results clearly indicate the singular nature of σ_y along the crack periphery.

Double Edge Crack Tensile Fracture Specimen. Our last example is the finite bar with double edge cracks. The crack opening displacements for this problem are presented in figure 9. The stress intensity factor variation as a function of bar thickness is shown in figure 10. In both, results from the finite element method are shown for comparison. Agreement is again excellent.

CONCLUSIONS. The line method of analysis presented affords a practical way for analysis of three-dimensional crack problems, at least for bodies with reasonably regular boundaries. Because parts of the solution are obtained as continuous functions along the lines chosen, relatively good accuracy can be obtained with coarse grids. Results of the analysis include the displacements and normal stresses at every node inside the body from which the stress intensity factor variations were easily calculated. In addition it should be noted that the common semi-elliptical surface crack problem could also be analyzed by merely changing the boundary conditions at certain nodes in the crack plane. Introduction of plasticity into the analysis could also be accomplished by changing the coupling terms in equations (8) to (10). Since these have to be determined by an interative process in any case, it would seem possible to solve the elastoplastic problem by a simple extension of the present method. Whether this approach is practical requires further investigation.

REFERENCES

1. Srawley, John E.; and Esgar, Jack B.: Investigation of Hydrotest Failure of Thiokol Chemical Corporation 260-Inch-Diameter SL-1 Motor Case. NASA TM X-1194, 1966.

2. Irwin, G. R.: Crack-Extension-Force for a Part-Through Crack in a Plate. J. Appl. Mech. vol. 29, no. 4, Dec. 1962, pp. 651-654.

3. Smith, Fred. W.: Stresses Near a Semi-Circular Edge Crack. Ph.D. Thesis, Univ. Washington, 1966.

4. Smith, F. W.; and Alavi, M. J.: Stress Intensity Factors for a Part-Circular Surface Flaw. Proceedings of the First International Pressure Vessel Conference, Holland, 1969.

5. Thresher, R. W.; and Smith, F. W.: Stress-Intensity Factors for a Surface Crack in a Finite Solid. J. Appl. Mech., vol. 39, no. 1, Mar. 1972, pp. 195-200.

6. Shah, R. C.; and Kobayashi, A. S.: Stress Intensity Factor for an Elliptical Crack Under Arbitrary Normal Loading. J. Eng. Fract. Mech., vol. 3, July 1971, pp. 71-96.

7. Shah, R. C.; and Kobayashi, A. S.: On the Surface Flaw Problem. The Surface Crack: Physical Problem and Computational Solutions. Am. Soc. Mech. Eng., 1972, pp. 79-124.

8. Hartranft, R. J.; and Sih, G. C.: Alternating Method Applied to Edge and Surface Crack Problems. Methods of Analysis and Solutions of Crack Problems: Recent Developments in Fracture Mechanics; Theory and Methods of Solving Crack Problems. Leiden, Noordhoff International Publ., 1973, pp. 179-238.

9. Keays, R. H.: A Review of Stress Intensity Factors for Surface and Internal Cracks. ARL/SM-Rept-343, Aeronautical Res. Labs., 1973.

10. Cruse, T. A.: Numerical Evaluation of Elastic Stress Intensity Factors by the Boundary-Integral Equation Method. The Surface Crack: Physical Problems and Computational Solutions. Am. Soc. Mech. Eng., 1972, pp. 153-170.

11. Marcal, P. V.: Three-Dimensional Finite Element Analysis for Fracture Mechanics. The Surface Crack: Physical Problems and Computational Solutions. Am. Soc. Mech. Eng., 1972, pp. 187-202.

12. Gyekenyesi, J. P.; and Mendelson, A.: Three-Dimensional Elastic Stress and Displacement Analysis of Finite Geometry Solids Containing Cracks. Int. J. Fract., vol. II, no. 3, June 1975, pp. 409-429.

13. Jones, D. J.; South, J. C.; and Klunker, E. B.: On the Numerical Solution of Elliptic Partial Differential Equations by the Method of Lines. J. Comput. Phys., vol. 9, 1972, pp. 496-527.

14. Gyekenyesi, J. P.: Solution of Some Mixed Boundary Value Problems of Three-Dimensional Elasticity by the Method of Lines. Ph.D. Thesis, Michigan State Univ., 1972.

15. Frazer, R. A.; Duncan, W. J.; and Collar, A. R.: Elementary Matrices and Some Applications to Dynamic and Differential Equations. Cambridge Univ. Press, 1938.

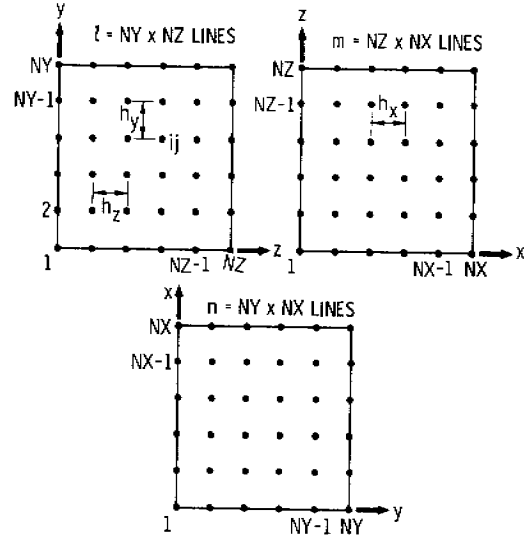
16. Sih, G. C.: Handbook of Stress Intensity Factors. Inst. Fract. Solid Mech., Lehigh Univ., 1973.

17. Raju, I. S.; and Newman, J. C., Jr.: Three-Dimensional Finite-Element Analysis of Finite Thickness Fracture Specimens. NASA TN D-8414, 1977.

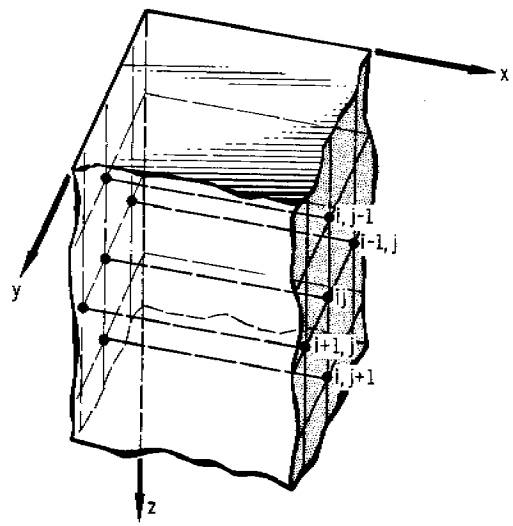
18. Isida, M.: Effect of Width and Length on Stress Intensity Factors of Internally Cracked Plates Under Various Boundary Conditions. Int. J. Fract. Mech., vol. 7, Sept. 1971, pp. 301-316.

TABLE I. - DIMENSIONLESS CRACK OPENING DISPLACEMENTS $Ev/\sigma_0 c$ AND DIMENSIONLESS y-DIRECTIONAL STRESSES σ_y/σ_0 FOR A RECTANGULAR BAR UNDER UNIFORM TENSION CONTAINING A THROUGH-THICKNESS CENTRAL CRACK [ALL DATA IN y = 0 PLANE WITH ρ = CORRECTED CRACK EDGE DISTANCE.]

	FINITE ELEMENT METHOD - REFERENCE 10 c = 1.0; W = 2.0; L = 1.75; t = 3.0; ν = 1/3			LINE METHOD c = 1.0; W = 1.92; L = 1.69; t = 2.90; ν = 1/3		
z	ρ	$\frac{Ev}{\sigma_0 c}$	$\frac{\sigma_y}{\sigma_0}$	ρ	$\frac{Ev}{\sigma_0 c}$	$\frac{\sigma_y}{\sigma_0}$
0	0.0132	0.3991	9.126	0.0132	0.384	8.975
	.0264	.5626	6.650	.0264	.541	6.552
	.040	.6910	5.381	.040	.666	5.599
	.080	.9698	3.852	.080	.937	3.996
	.120	1.1758	3.245	.120	1.142	3.330
	.160	1.3475	2.736	.160	1.314	2.912
	.200	1.4928	2.434	.200	1.463	2.602
	.400	1.9935	1.703	.400	2.023	1.888
	.600	2.3120	1.295	.600	2.390	1.320
	.920	------	-----	.920	2.640	.532
	1.000	2.5347	.531	1.000	2.651	-----
$\frac{t}{2}$	0.0132	0.4293	7.748	0.0132	0.441	8.463
	.0264	.6102	5.259	.0264	.622	6.121
	.040	.7552	4.046	.040	.766	5.030
	.080	1.0801	2.821	.080	1.081	3.667
	.120	1.3282	2.426	.120	1.321	3.048
	.160	1.5391	2.100	.160	1.524	2.664
	.200	1.7197	1.989	.200	1.679	2.392
	.400	2.3433	1.533	.400	2.385	1.678
	.600	2.7412	1.253	.600	2.830	1.210
	.920	------	-----	.920	3.140	.656
	1.000	3.0167	.703	1.000	3.157	-----

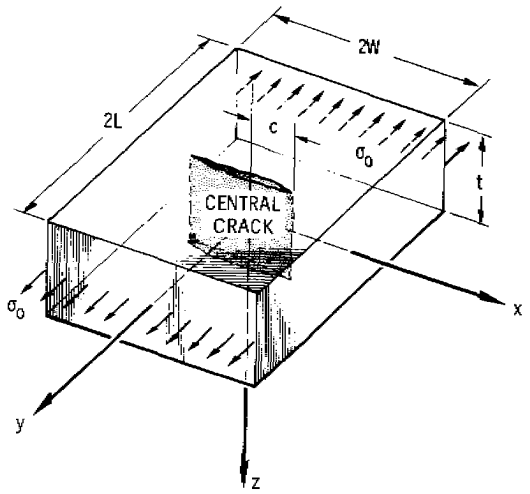


(a) THREE SETS OF LINES PARALLEL TO x-, y-, AND z-COORDINATES AND PERPENDICULAR TO CORRESPONDING COORDINATE PLANES.

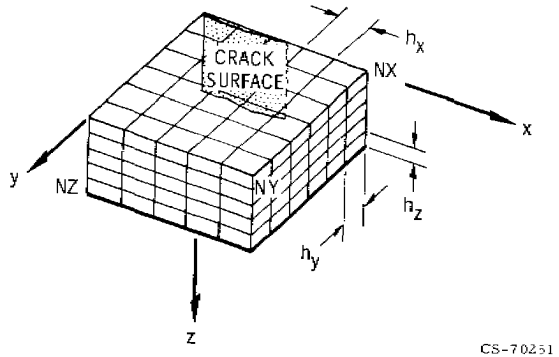


(b) SET OF INTERIOR LINES PARALLEL TO x-COORDINATE.

Figure 1. - Sets of lines parallel to Cartesian coordinates.

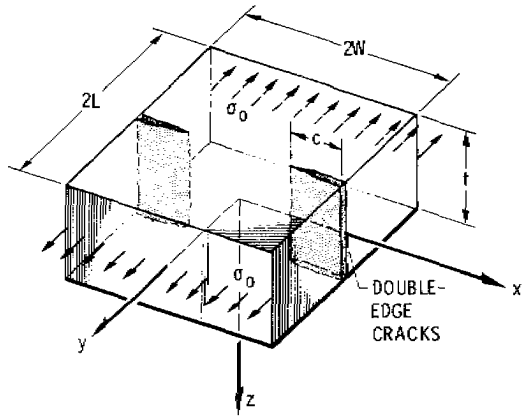


(a) RECTANGULAR BAR WITH THROUGH-THICKNESS CENTRAL CRACK.

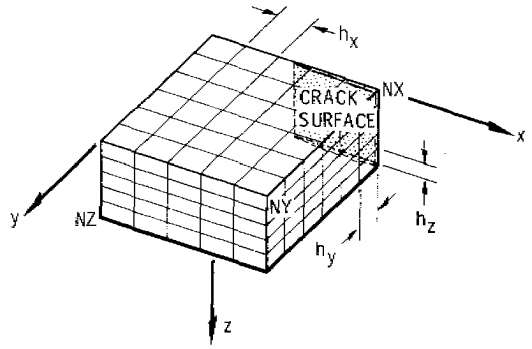


(b) DISCRETIZED REGION OF RECTANGULAR BAR WITH THROUGH-THICKNESS CENTRAL CRACK.

Figure 2. - Rectangular bar with through-thickness central crack under uniform tension.



(a) RECTANGULAR BAR WITH THROUGH-THICKNESS DOUBLE EDGE CRACKS.



(b) DISCRETIZED REGION OF RECTANGULAR BAR WITH DOUBLE-EDGE CRACKS.

Figure 3. - Rectangular bar with through-thickness double-edge cracks under uniform tension.

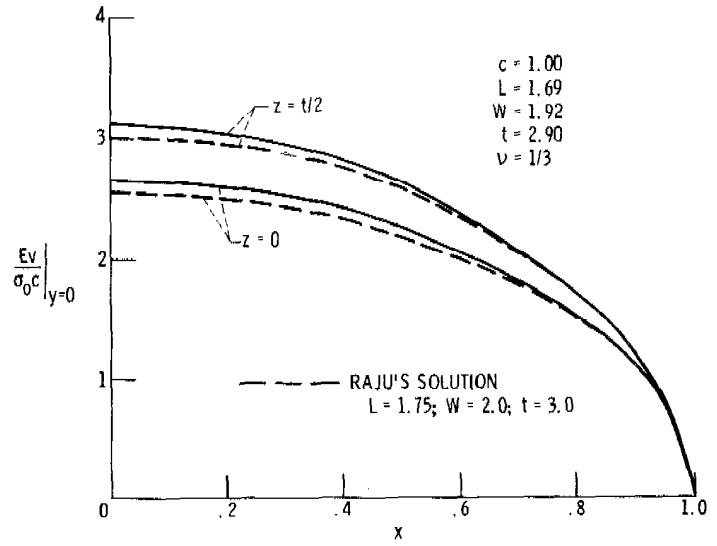


Figure 4. - Crack opening displacement for center-cracked bar under uniform tension.

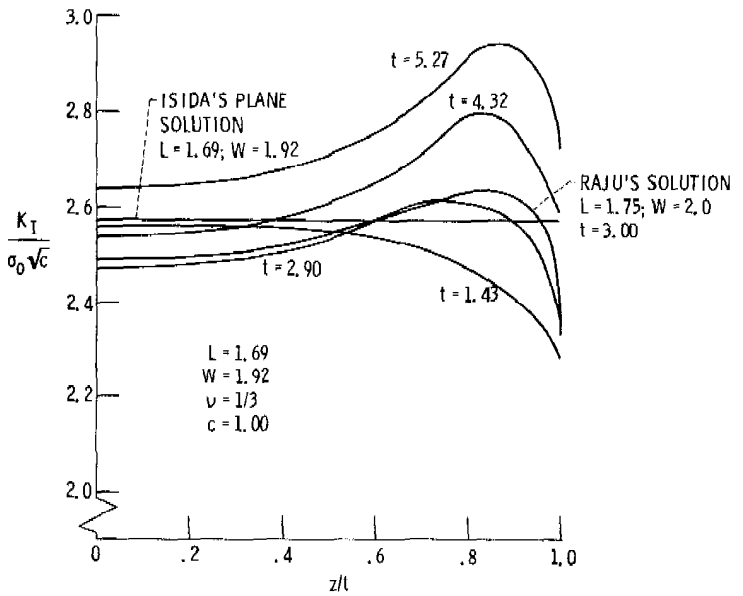
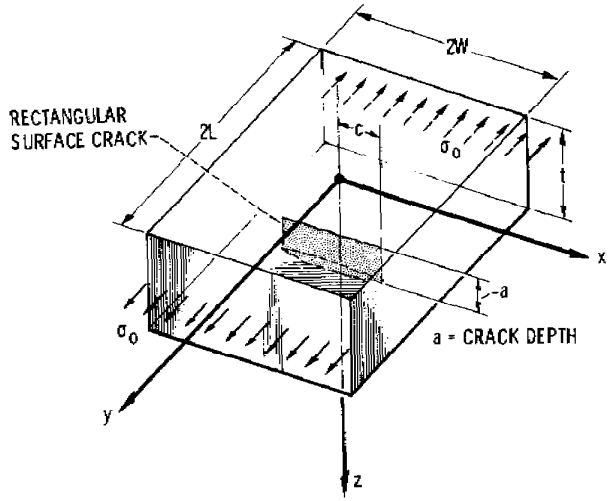
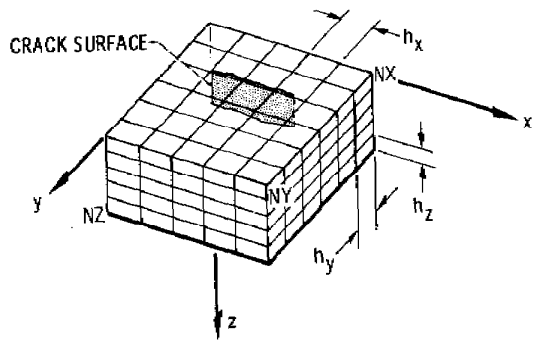


Figure 5. - Stress-intensity factor variation as a function of bar thickness for a center-cracked rectangular bar under uniform tension.



(a) BAR WITH RECTANGULAR SURFACE CRACK.



(b) DISCRETIZED REGION OF BAR WITH RECTANGULAR SURFACE CRACK.

Figure 6. - Bar with rectangular surface crack under uniform tension.

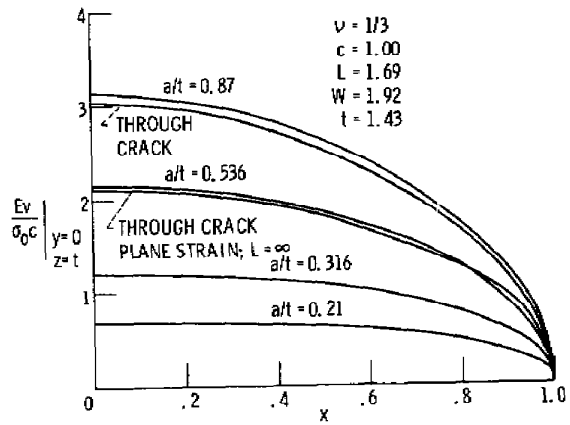
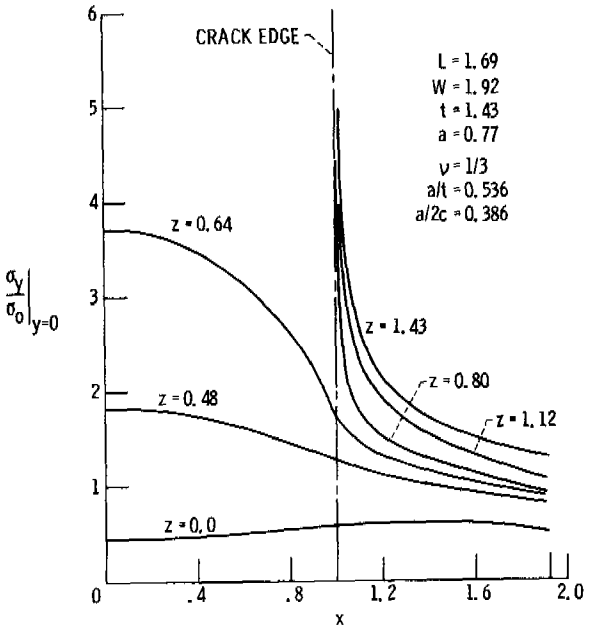
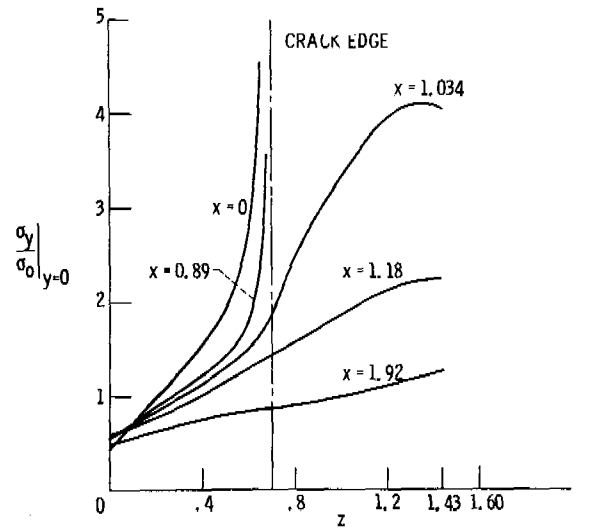


Figure 7. - Surface crack opening displacement variation as a function of crack depth for a rectangular bar under uniform tension.



(a) DIMENSIONLESS y-DIRECTIONAL NORMAL STRESS VARIATION ALONG BAR WIDTH.



(b) DIMENSIONLESS y-DIRECTIONAL NORMAL STRESS VARIATION ACROSS BAR THICKNESS.

Figure 8. - Dimensionless y-directional normal stress distribution in the crack plane for a bar under uniform tension containing a rectangular surface crack.

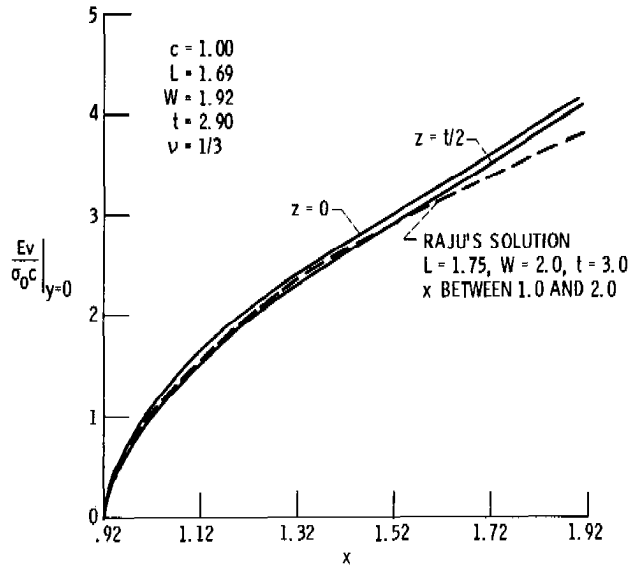


Figure 9. - Crack opening displacement for rectangular bar under uniform tension containing through-thickness double-edge cracks.

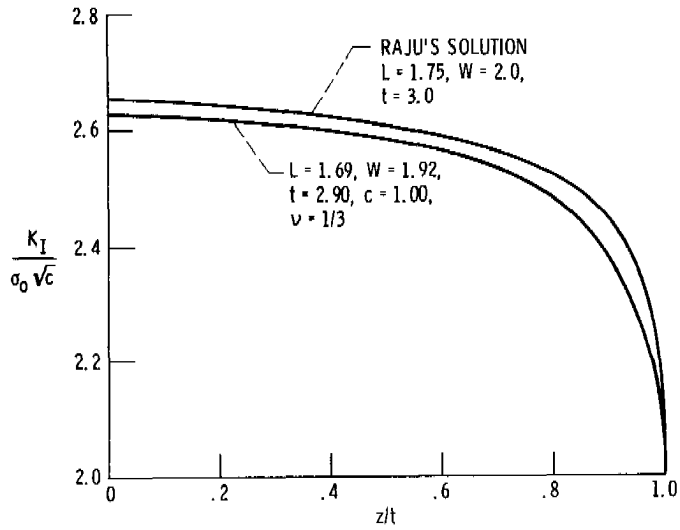


Figure 10. - Stress intensity factor variation as function of bar thickness for rectangular bar under uniform tension containing double-edge cracks.

FULLY PLASTIC DEFORMATION IN ANISOTROPIC ANNULAR PLATES UNDER INTERNAL PRESSURE

P. C. T. Chen
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. The problem considered is an elastic-plastic, annular plate radially stressed by uniform internal pressure. The partially plastic deformation problem is extended to the fully plastic case. The plate is elastically as well as plastically orthotropic but isotropic in its plane. The exact solution is obtained on the basis of the J_2 deformation theory, the Hill's yield criterion and a modified Ramberg-Osgood Law. Numerical results for the stresses, strains and displacement will be presented and discussed.

1. INTRODUCTION. The problem considered is an elastic-plastic annular plate radially stressed by uniform internal pressure. For ideally plastic materials, the stress solution for this statically determinate problem was first obtained by Mises [1] and the corresponding two strain solutions were obtained by the present author on the basis of both J_2 deformation and flow theories [2]. For elastic-plastic strain-hardening materials, an exact solution for the partially plastic deformation problem was recently reported in [3] for the isotropic case and in [4] for the anisotropic case. Analytical expressions were derived but only the effect of geometric ratio on the stresses was briefly discussed in that technical note.

In the present paper, the partially plastic deformation problem is extended to the fully plastic case. A unified treatment is given here for both cases. The plate is elastically as well as plastically orthotropic but isotropic in its plane. The material model is assumed to obey the J_2 deformation theory, the Hill's yield criterion and a modified Ramberg-Osgood law [5]. The exact solution will be presented and the effect of strain hardening on the stresses, strains and displacement for the complete range of loading will be discussed.

2. BASIC EQUATIONS. Assuming small strains and neglecting inertia forces in the axisymmetric state of plane stress, the radial and tangential stresses, σ_r and σ_θ, must satisfy the equilibrium equation,

$$\sigma_\theta = (d/dr)(r\sigma_r) \quad ; \tag{1}$$

and the corresponding strains, ε_r and ε_θ , are given in terms of the radial displacement, u , by

$$\varepsilon_r = du/dr \quad , \quad \varepsilon_\theta = u/r \ . \tag{2}$$

We shall assume that the plate is elastically as well as plastically orthotropic but isotropic in its plane, $E = E_r = E_\theta$, $\nu = \nu_{r\theta} = \nu_{\theta r}$, $\nu_1 = \nu_{rz} = \nu_{\theta z}$ are the elastic constants, and R is the plastic strain ratio related to the yield stress ratio ω, plastic Poisson's ratio μ_p by $1 - (2\omega^2)^{-1} = \mu_p = R/(1+R)$. According to the simple deformation theory, the strains are related to the stresses by

$$\begin{aligned} \varepsilon_r &= E^{-1}(\sigma_r - \nu\sigma_\theta) + (E_s^{-1} - E^{-1})(\sigma_r - \mu_p\sigma_\theta) \ , \\ \varepsilon_\theta &= E^{-1}(\sigma_\theta - \nu\sigma_r) + (E_s^{-1} - E^{-1})(\sigma_\theta - \mu_p\sigma_r) \ , \end{aligned} \tag{3}$$

where E_s is the secant modulus on the effective stress-strain curve with $E_s = \sigma/\varepsilon$ and

$$\sigma = (\sigma_r^2 + \sigma_\theta^2 - 2\mu_p\sigma_r\sigma_\theta)^{1/2} \ . \tag{4}$$

If a modified uniaxial relation of the Ramberg-Osgood type is assumed [5], we have

$$E_s^{-1} = E^{-1} \text{ for } \sigma \leq \sigma_y \quad ; \quad E_s^{-1} = E^{-1}(\sigma/\sigma_y)^{n-1} \text{ for } \sigma \geq \sigma_y \tag{5}$$

and the initial yield surface is defined by the ellipse $\sigma = \sigma_y$.

Since the compressibility of the material is taken into account, the longitudinal strain ε_z can be determined by

$$\varepsilon_r + \varepsilon_\theta + \varepsilon_z = E^{-1}(1 - \nu - \nu_1)(\sigma_r + \sigma_\theta) \ , \tag{6}$$

which holds in the elastic as well as plastic region.

The boundary conditions on the problem are

$$\sigma_r(a, t) = -P \ , \ \sigma_r(b, t) = 0 \ , \tag{7}$$

Where a, b and P are the inner, outer radii and internal pressure, respectively, and t denotes some monotonic parameter such as P or the elastic-plastic boundary ρ. In addition, all stresses, strains and displacement must be continuous throughout the entire region.

In the following, the solutions will be presented in terms of nondimensional quantities defined by

$$\alpha = a/b, \quad \xi = r/b, \quad \beta = \rho/b, \quad \lambda = P/\sigma_y ,$$

$$S_r = \sigma_r/\sigma_y, \quad S_\theta = \sigma_\theta/\sigma_y, \quad S = \sigma/\sigma_y ,$$

$$e_r = E\epsilon_r/\sigma_y, \quad e_\theta = E\epsilon_\theta/\sigma_y, \quad e_z = E\epsilon_z/\sigma_y . \tag{8}$$

3. SOLUTION IN THE PLASTIC REGION ($\alpha \leq \xi \leq \beta$). Following Nadai for isotropic problems [6], we introduce the parametric representation ($0 \leq \phi \leq \pi/2$)

$$S_r = -S\cos\phi/\sin2\delta ,$$

$$S_\theta = -S\cos(\phi + 2\delta)/\sin2\delta , \tag{9}$$

which satisfies equation (4) identically and leads to the following equation upon substituting into the equation of equilibrium,

$$\xi^{-1}d\xi = [\sin2\delta(\tan\delta + \tan\phi)]^{-1}(\tan\phi d\phi - S^{-1}dS) . \tag{10}$$

By the extended Michell theorem [7], the stress solution for the present problem is independent of ν. So choose $\nu = \mu_p$ in (3) and make use of (2), (5), (8) and (10), we have

$$\xi^{-1}d\xi = [-\sin2\delta(\cot\delta + \cot\phi)]^{-1}(\cot\phi d\phi + nS^{-1}dS) . \tag{11}$$

The above two ordinary differential equations form a system and can be integrated exactly if we know the boundary values for S and ϕ at $\xi = \alpha$ or β. These values at the two boundaries of the plastic region will be determined in the next section. Since S and ϕ are functions of ξ and β, the notation $S_{\xi\beta} = S(\xi,\beta)$, $\phi_{\xi\beta} = \phi(\epsilon,\beta)$ are introduced. After some manipulation, the results are presented in the following form:

$$S_{\xi\beta}/S_{\beta\beta} = G(\phi_{\xi\beta}, \phi_{\beta\beta}) ,$$

$$(\beta/\xi)^2 = F(\phi_{\xi\beta}, \phi_{\beta\beta}) , \tag{12}$$

where

$$G(\phi_{\xi\beta}, \phi_{\beta\beta}) = \left[\frac{n\sin\phi_{\beta\beta} - \cot\delta\cos\phi_{\beta\beta}}{n\sin\phi_{\xi\beta} - \cot\delta\cos\phi_{\xi\beta}}\right]^{\mu} \exp\left[\frac{(n-1)\cot\delta}{n^2+\cot^2\delta}(\phi_{\beta\beta}-\phi_{\xi\beta})\right] , \tag{13}$$

$$F(\phi_{\xi\beta}, \phi_{\beta\beta}) = \frac{\sin(\phi_{\xi\beta}+\delta)}{\sin(\phi_{\beta\beta}+\delta)} \times \left[\frac{n \sin\phi_{\beta\beta} - \cot\delta \cos\phi_{\beta\beta}}{n \sin\phi_{\xi\beta} - \cot\delta \cos\phi_{\xi\beta}}\right]^{\frac{n\csc^2\delta}{n^2+\cot^2\delta}}$$

$$\times \exp\left[\frac{(n^2-1)\cot\delta}{n^2+\cot^2\delta} (\phi_{\beta\beta} - \phi_{\xi\beta})\right] , \qquad (14)$$

$\mu = (n + \cot^2\delta)/(n^2 + \cot^2\delta)$,

and δ is the anisotropic parameter defined in the first quadrant by

$$\tan^2\delta = (1-\mu_p)/(1 + \mu_p) = 1/(1 + 2R) = 1/(4\omega^2-1) . \qquad (15)$$

The solution for the strains in the plastic region ($\alpha \le \xi \le \beta$) of an elastic-plastic plate with finite n can be obtained from (3) and (6), using (5), (8) and the above stress solution. After some manipulation, the equations for the dimensionless strains can be written as

$$e_r = -S^n_{\xi\beta} \sin(\phi_{\xi\beta}+2\delta) - S_{\xi\beta}\cos(\phi_{\xi\beta}+2\delta)(\cos2\delta-\nu)/\sin2\delta ,$$

$$e_\theta = S^n_{\xi\beta} \sin\phi_{\xi\beta} - S_{\xi\beta} \cos\phi_{\xi\beta}(\cos2\delta-\nu)/\sin2\delta .$$

$$e_z = [(2\sin\delta)S^n_{\xi\beta} - (2\sin\delta-\nu_1/\sin\delta)S_{\xi\beta}] \cos(\phi_{\xi\beta}+\delta) . \qquad (16)$$

4. DETERMINATION OF THE BOUNDARY VALUES. For small internal pressure ($\lambda \le \lambda^*$), the plate will be elastic throughout ($\alpha \le \xi \le 1$) and the solution is omitted here. The critical value λ^* to cause incipient plastic deformation is

$$\lambda^* = (1-\alpha^2) [2 (1 + \mu_p) + 2 (1 - \mu_p)\alpha^4]^{-1/2} . \qquad (17)$$

For values of p larger than p*, the plate becomes plastic in the inner region.

The foregoing solution in the plastic region ($\alpha \le \xi \le \beta$) has been derived on the assumption that we know the boundary values for S and ϕ at $\xi = \beta$. These values at the two boundaries at $\xi = \alpha$ and β will be determined in this section for the partially-plastic as well as fully-plastic case.

A. Partially-Plastic Deformation ($\lambda^* \leq \lambda \leq \lambda^{**}$). In this case, the plate will be plastic in the inner region ($\alpha \leq \xi \leq \beta$) and still elastic in the outer region ($\beta \leq \xi \leq 1$). In the outer elastic region, the equations for the dimensionless stresses and strains are

$$\left.\begin{matrix} S_r \\ S_\theta \end{matrix}\right\} = \frac{1}{2}\,(1 \mp \xi^{-2})/(\sin^2\delta + \beta^{-4}\cos^2\delta)^{1/2} ,$$

$$\left.\begin{matrix} e_r \\ e_\theta \end{matrix}\right\} = \frac{1}{2}\,[(1-\nu) \mp (1+\nu)\xi^{-2}]/(\sin^2\delta + \beta^{-4}\cos^2\delta)^{1/2} ,$$

$$e_z = -\nu_1/(\sin^2\delta + \beta^{-4}\cos^2\delta)^{1/2} . \qquad (18)$$

The above equations together with the continuity requirement of the stresses, strains and displacement at the elastic-plastic interface lead to

$$S_{\beta\beta} = 1 \text{ and} \quad , \qquad (19a)$$

$$\tan\phi_{\beta\beta} = (\beta^2\tan\delta + \cot\delta)/(1-\beta^2) . \qquad (19b)$$

At the inside surface, $\xi = \alpha$, $\phi = \phi_{\alpha\beta}$, equations (12) with the boundary condition (7) reduce to

$$\lambda = S_{\alpha\beta}\cos\phi_{\alpha\beta}/\sin 2\delta , \qquad (19c)$$

$$(\beta/\alpha)^2 = F(\phi_{\alpha\beta}, \phi_{\beta\beta}) , \qquad (19d)$$

and

$$S_{\alpha\beta}/S_{\beta\beta} = G(\phi_{\alpha\beta}, \phi_{\beta\beta}) . \qquad (19e)$$

The five equations (19a to e) are sufficient to determine the five unknowns β, $\phi_{\beta\beta}$, $S_{\beta\beta}$, $\phi_{\alpha\beta}$, $S_{\alpha\beta}$ in terms of λ. Alternatively, we can determine λ, $\phi_{\beta\beta}$, $S_{\beta\beta}$, $\phi_{\alpha\beta}$, $S_{\alpha\beta}$ in terms of β. The latter approach has been chosen in obtaining numerical results. The partially-plastic deformation reaches its upper limit when β increases to 1. The corresponding pressure factor is λ^{**}.

B. Fully-Plastic Deformation ($\lambda \geq \lambda^{**}$). In this case, the plate will be plastic in the entire region ($\alpha \leq \xi \leq \beta$) and $\beta = 1$. Equations (9) and (12) with the boundary conditions (7) reduce to

$$\phi_{11} = \pi/2 , \qquad (20a)$$

$$\lambda = S_{\alpha 1}\,\phi_{\alpha 1}/\sin 2\delta , \qquad (20b)$$

$$(1/\alpha)^2 = F(\phi_{\alpha 1}, \phi_{11}) , \tag{20c}$$

and

$$S_{\alpha 1}/S_{11} = G(\phi_{\alpha 1}, \phi_{11}) . \tag{20d}$$

The four equations (20a to d) are sufficient to determine $\phi_{\alpha 1}$, $S_{\alpha 1}$, ϕ_{11}, S_{11} in terms of λ.

Once the boundary values for S and ϕ at $\xi = \alpha$ and β are determined, the solution for the stresses, strains and displacement in the plastic region can be calculated as shown in the preceding section.

5. DISCUSSIONS OF RESULTS. Since the deformation theory is used, the validity of the above solution should be assessed by applying Budiansky's criterion [8] which requires the following inequality to be satisfied,

$$[(nS_{\xi\beta}^{n-1}-1)/(S_{\xi\beta}^{n-1}-1)]^{1/2} \geq (n \tan\phi_{\xi\beta} \tan\delta-1)/(\tan\phi_{\xi\beta} + \tan\delta). \tag{21}$$

It has been verified numerically that all the values of $S_{\xi\beta}$ and $\phi_{\xi\beta}$ reported here satisfy the inequality. We may discuss the effects of elastic constants (E, ν, ν_1), plastic material constants (σ_y, n, R) and geometric ratio (a/b) on the stresses, strains and displacement for various values of pressure factor (P/σ_y). Some typical results for the distribution of stresses and strains in the 2219-T87 aluminum plate with $E = 10.5 \times 10^6$ psi, $\nu = \nu_1 = .3$, $\sigma_y = 5.5 \times 10^4$ psi, $n = 9$, $R = 1$, $b/a = 3$ were presented in [3]. The effects of geometric ratio on the stresses at the inside surface of several partly-plastic plates with $\nu = \nu_1 = .3$, $n = 9$, $R = 1$ and $b/a = 2,3,4,10$ were reported in [4]. Neither the numerical results for the strains and displacement nor the effect of strain hardening and anisotropy have been discussed. In the following additional results will be given.

The emphasis herein is on the effect of strain hardening in a partly-plastic as well as fully-plastic annular plate radially stressed under uniform internal pressure. In order to compare with the elastic-perfectly-plastic solution given in [2], we choose $b/a = 2$, $\nu = \nu_1 = 0.3$, $R = 1$ and $n = 3,9,15, \infty$. The numerical results for the stresses, strains and displacement are presented in graphical form for the complete range of loading. For the partially plastic case, all the results can be expressed as functions of radius of elastic-plastic boundary ρ. These are shown in Figures 1-4. The effect of strain hardening (n) on the radial stress or pressure at the inside surface for various sizes of plastic zone is shown in Figure 1 and the corresponding effect on the circumferential stress at the inside surface, in Figure 2. The effect of strain hardening on the stresses is quite significant as shown in the above figures. The differences become larger as the size of plastic zone expands. It is interesting to point out that the stresses at the inside surface corresponding to initial yielding is independent of strain hardening but depend on geometric ratio and anisotropic parameter as

shown in [4]. The numerical results for the radial and axial strains at the inside surface for various sizes of plastic zone are presented in Figure 3 and the corresponding results for the radial displacement or circumferential strain at the inside surface, in Figure 4. The differences for all strain components become larger as plastic zone expands, however, those differences for circumferential strain are very small. In Figure 4, we also present the results for the residual circumferential stress at the inside surface as a function of radius of elastic-plastic boundary. The results were obtained by assuming elastic unloading when the pressure corresponding to any size of plastic zone was removed completely. As shown in Figure 4, the effect of increased strain hardening (i.e., a smaller value of n) is to decrease the magnitude of the compressive residual circumferential stress. Since the smaller value of this magnitude is less favorable and larger pressure is required to reach it, we conclude from this point of view that the strain hardening effect is undesirable. Other considerations such as reduced ductility and reduced fatigue resistance also support this conclusion. The best candidate is a high strength material with little strain hardening.

For the fully plastic case, the results are no longer functions of radius of elastic-plastic boundary ρ but they can be expressed as functions of internal pressure or radial expansion as presented in Figures 5 to 8. In all these figures, the results for the complete range of loading are given with broken lines for elastic ranges, solid curves for partially-plastic ranges and dotted curves for fully plastic ranges. The radial and circumferential stresses at the inside surface for several values of n are shown in Figure 5. The effect of strain hardening (n) on the stresses is quite significant as shown here. For large stresses, the curves for finite n in this figure approach radially. The numerical results for the radial and axial strains at the inside surface are presented in Figure 6 and the corresponding results for the radial displacement or circumferential strain at the inside surface, in Figure 7. The effect of strain hardening (n) on the strains and displacement can be seen from these two figures. We also present the results for the residual circumferential stress at the inside surface as a function of inside radial expansion in Figure 8. Again, the assumption of elastic unloading is used after removing the pressure completely. It should also be noted that the present solution has been obtained on the basis of small strain assumption. All the strain results reported here are indeed small since they must be multiplied by the yield strain, σ_Y/E, which can take on a very small value. However, for a given material with a given yield strain the validity of the small strain approximation may be violated. The large strain approach to this problem should be examined.

REFERENCES

1. Mises, R. V., "Three Remarks on the Theory of the Ideal Plastic Body," Reissner Anniversary Volume, 1949, pp. 415-429.

2. Chen, P. C. T., "A Comparison of Flow and Deformation Theories in a Radially Stressed Annular Plate," Journal of Applied Mechanics, Vol. 40, No. 1, Trans, ASME, Vol. 95, 1973, pp. 283-287.

3. Chen, P. C. T., "An Exact Solution to an Elastic-Plastic Deformation Problem in a Radially-Stressed Annular Plate," Transaction of the Twenty-Second Conference of Army Mathematics, ARO Report 77-1, 1977, pp. 227-238.

4. Chen, P. C. T., "Elastic-Plastic Analysis of a Radially-Stressed Annular Plane," Journal of Applied Mechanics, Vol. 44, 1977, pp. 167-169.

5. Budiansky, B., "An Exact Solution to an Elastic-Plastic Stress Concentration Problem," Prikladnaya Mathematika i Physik, Vol. 35, No. 1, 1971, pp. 40-48.

6. Nadai, A., Theory of Flow and Fracture of Solids, McGraw-Hill, New York, Vol. 1, 1950, Chapter 33.

7. Budiansky, B., "Extension of Michell's Theorem to Problems of Plasticity and Creep," Quarterly of Applied Mathematics, Vol. 16, 1958, pp. 307-309.

8. Budiansky, B., "A Reassessment of Deformation Theories of Plasticity," Journal of Applied Mechanics, Vol. 26, Trans. ASME, Vol. 81, 1959, pp. 259-264.

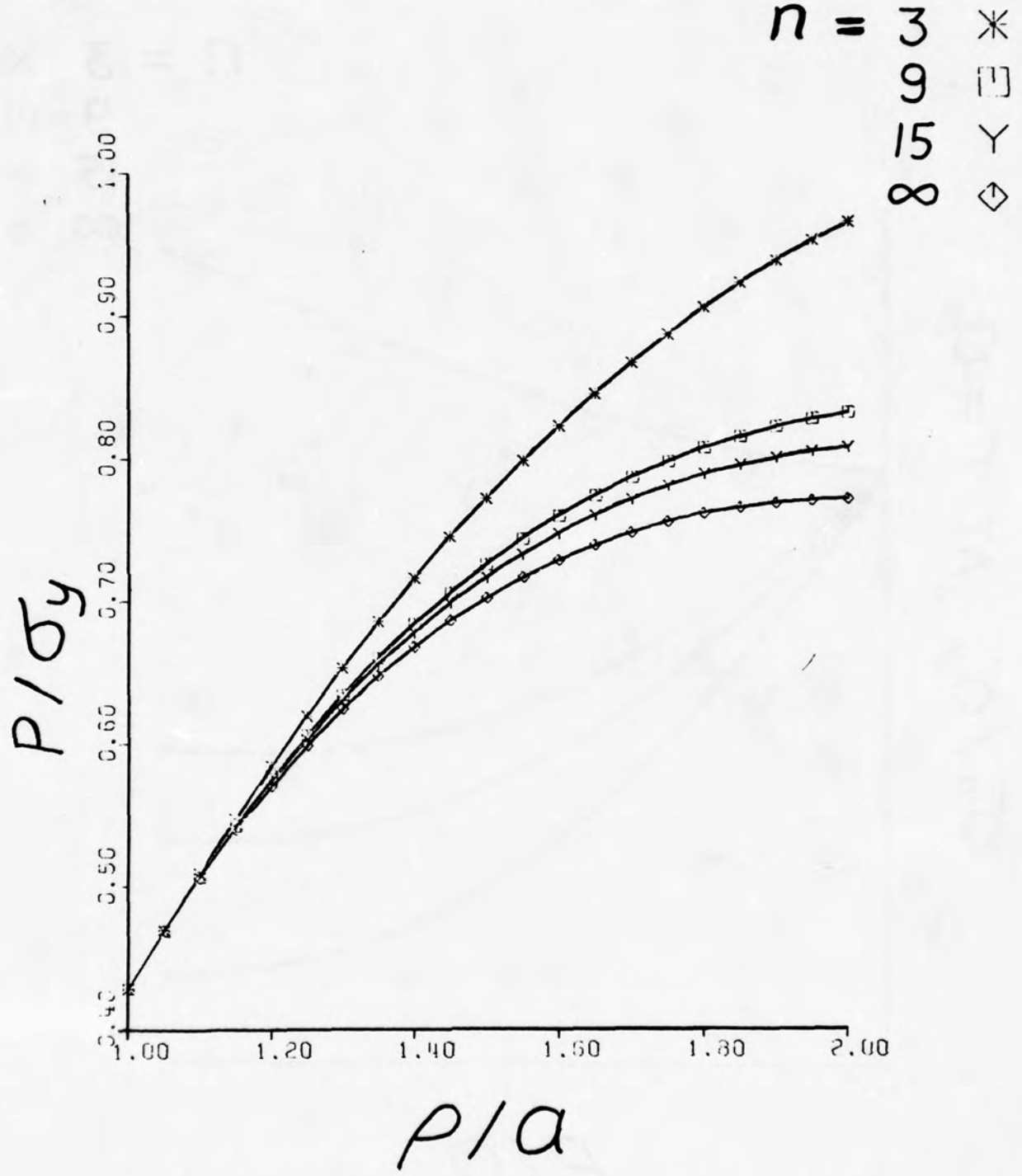


Fig. 1 Relation between pressure and location of elastic-plastic boundary (b/a = 2, $\nu = \nu_1 = .3$, R = 1).

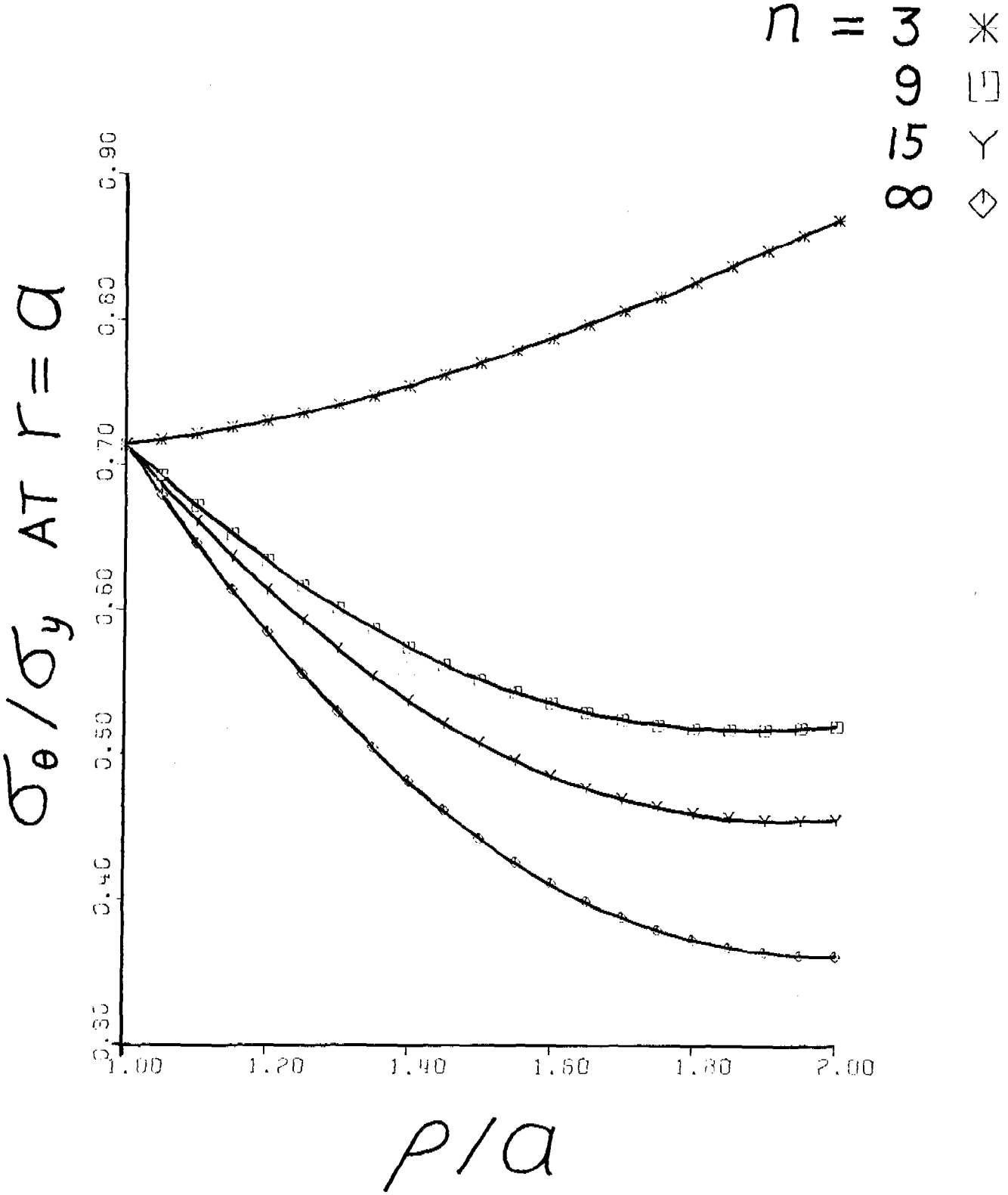


Fig. 2 Relation between inside tangential stress and location of elastic-plastic boundary (b/a = 2, $\nu = \nu_1 = .3$, R = 1).

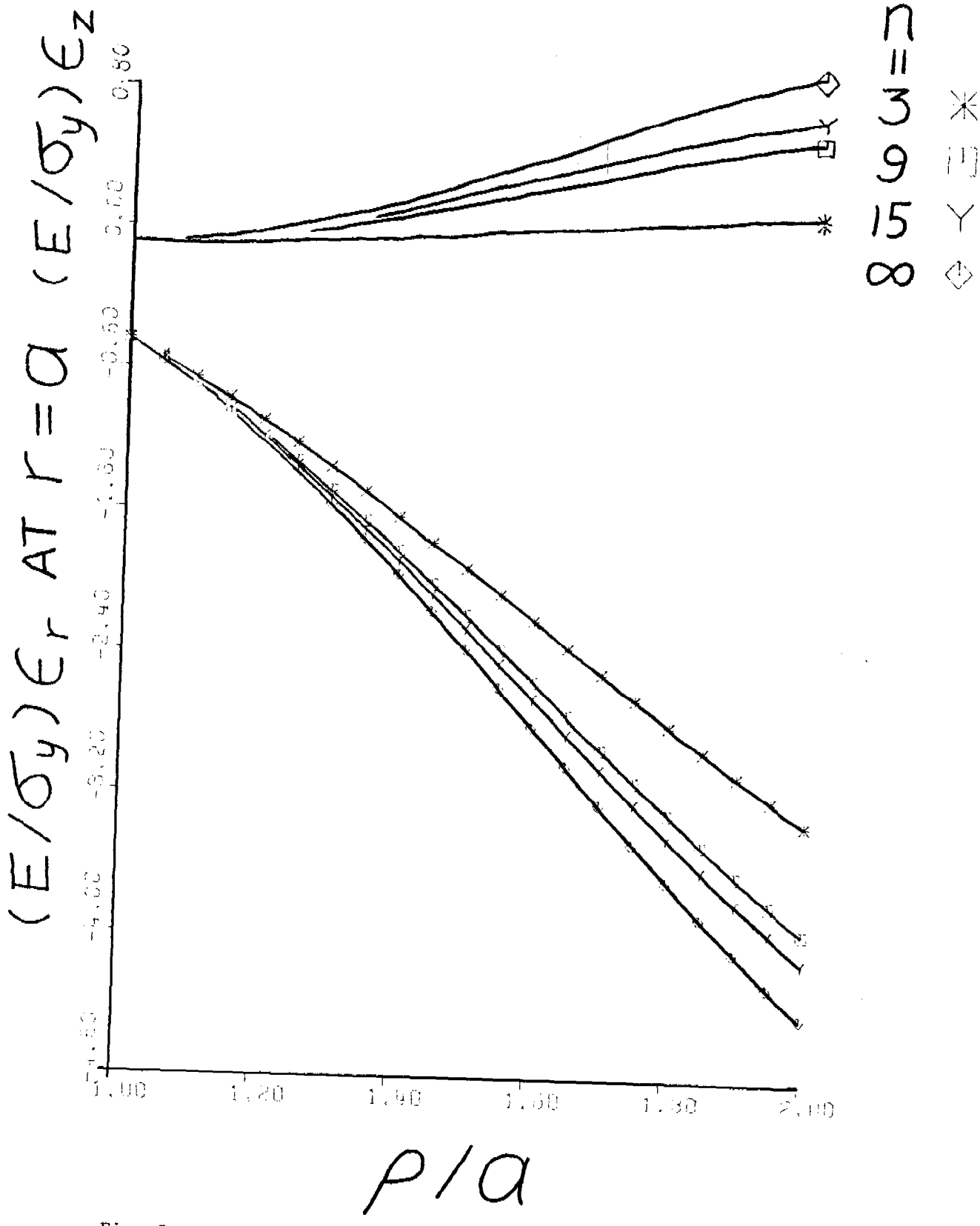


Fig. 3 Radial and axial strains at the inside surface as functions of radius of elastic-plastic boundary ($b/a = 2$, $\nu = \nu_1 = .3$, $R = 1$).

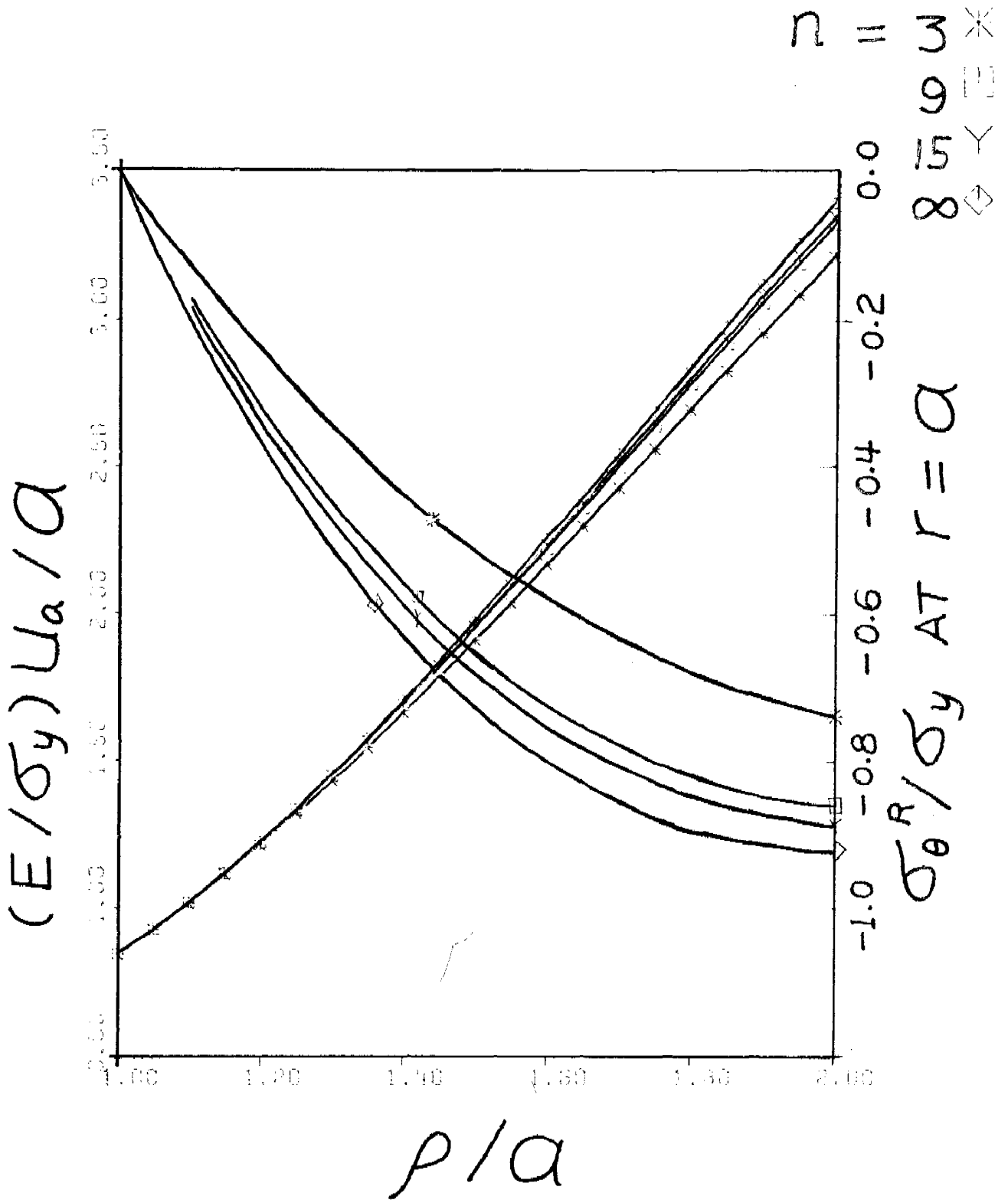


Fig. 4 Radial displacement and residual circumferential stress at the inside surface as function of elastic-plastic boundary (b/a = 2, $\nu = \nu_1 = .3$, R = 1).

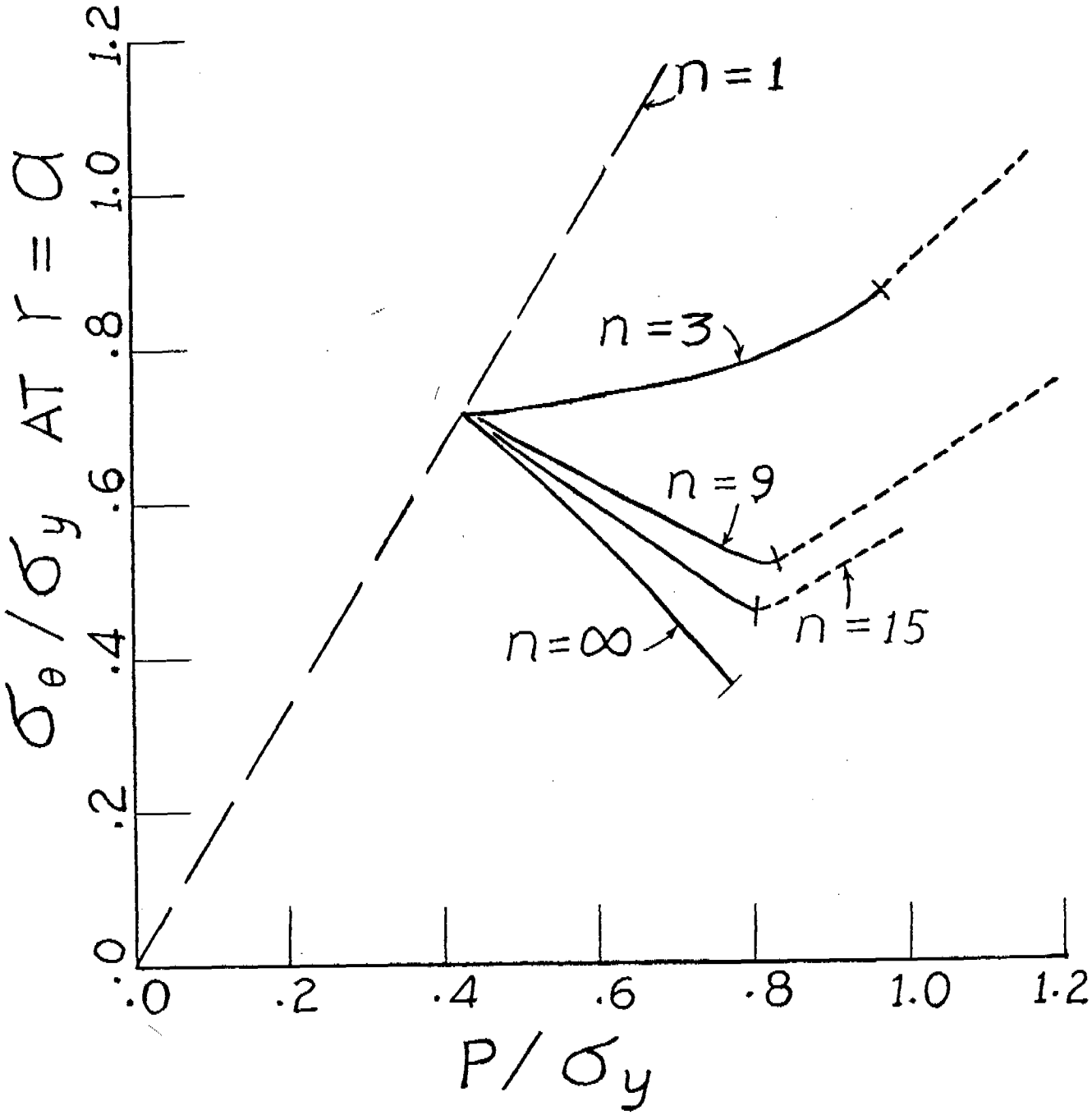


Fig. 5 Circumferential stress at the inner boundary as function of internal pressure (b/a = 2, $\nu = \nu_1 = .3$, R = 1).

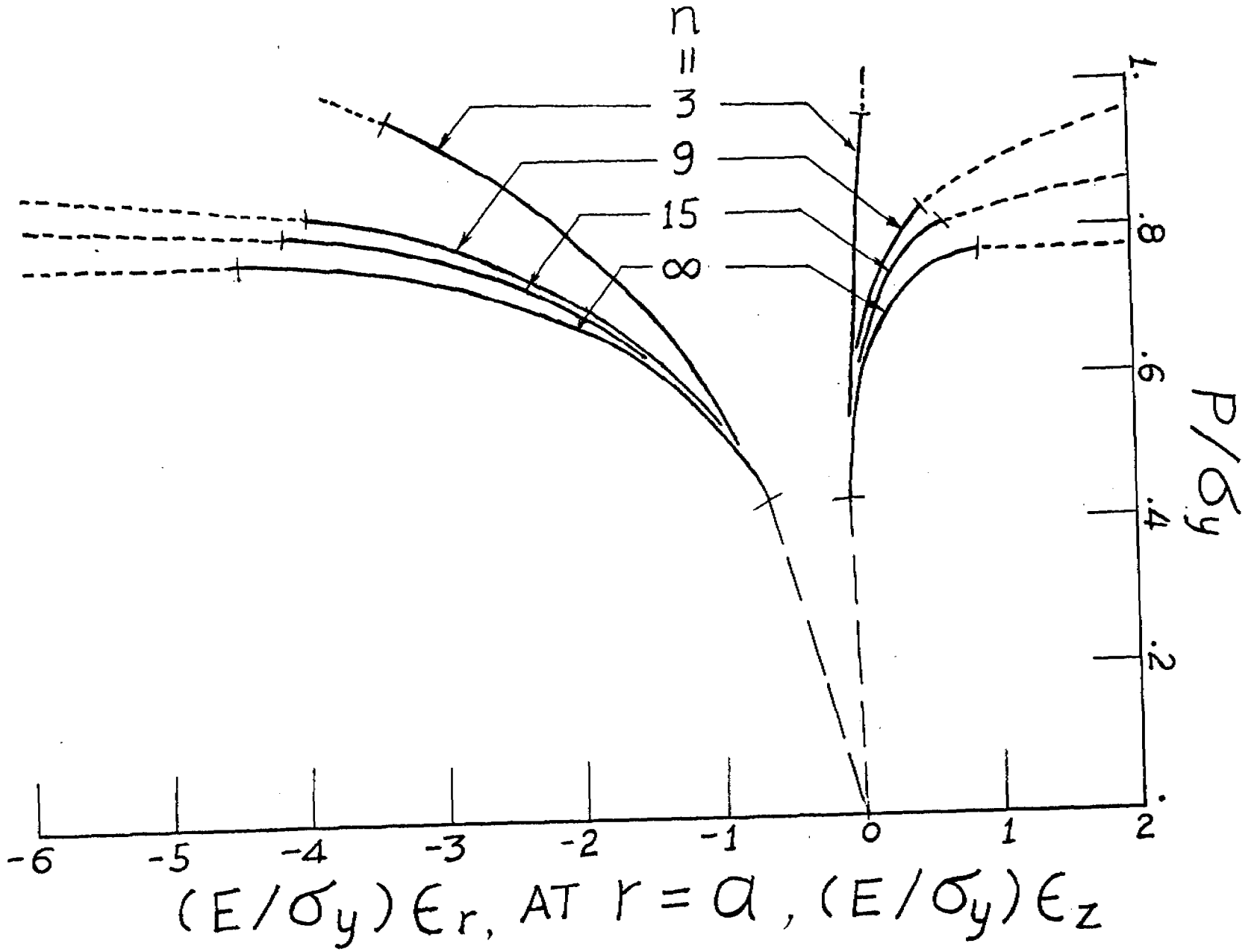


Fig. 6 Radial and axial strains at the inside surface as functions of internal pressure ($b/a = 2$, $\nu = \nu_1 = .3$, $R = 1$).

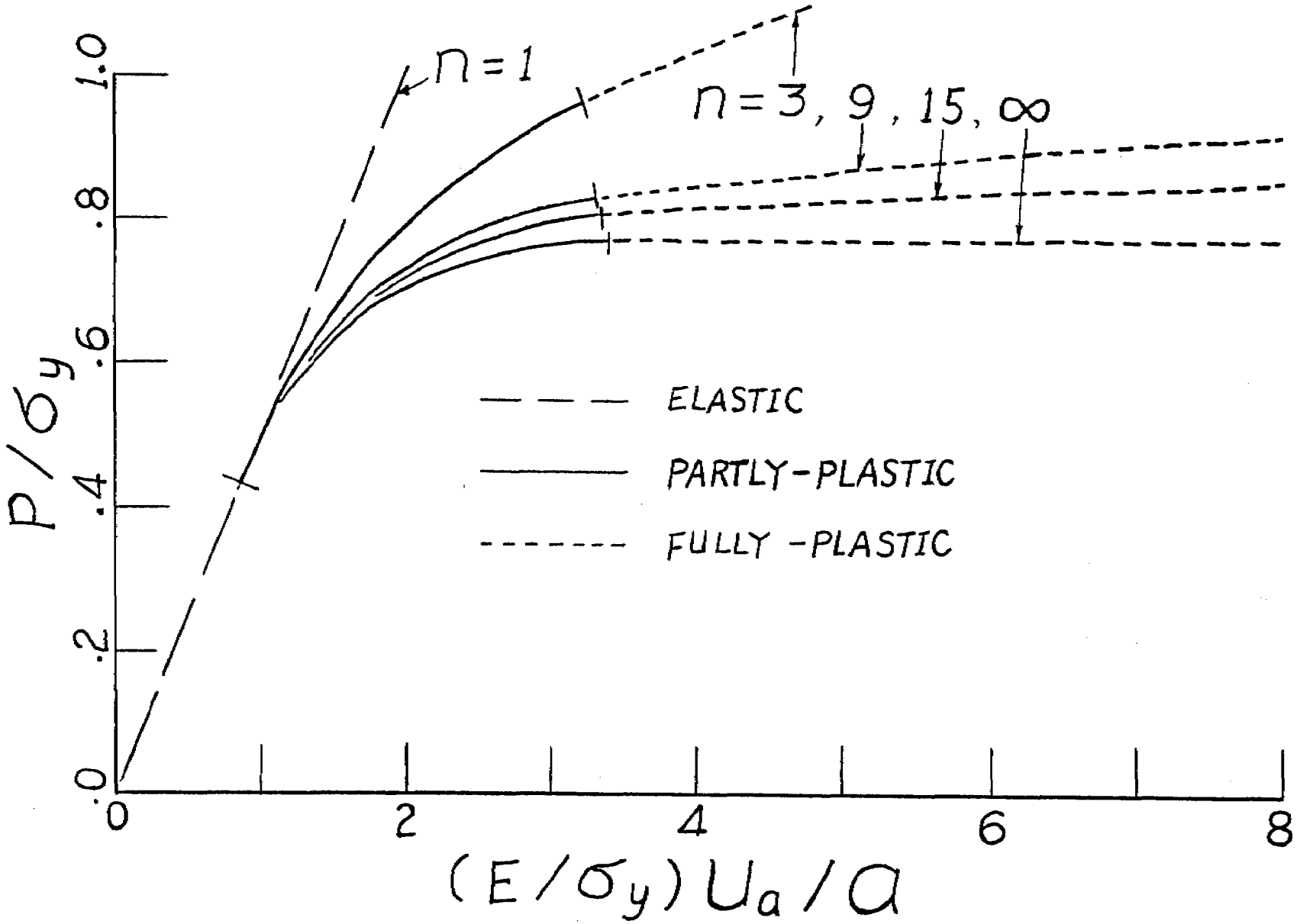


Fig. 7 Pressure factor as functions of inside radial displacement (b/a = 2, $\nu = \nu_1 = .3$, R = 1).

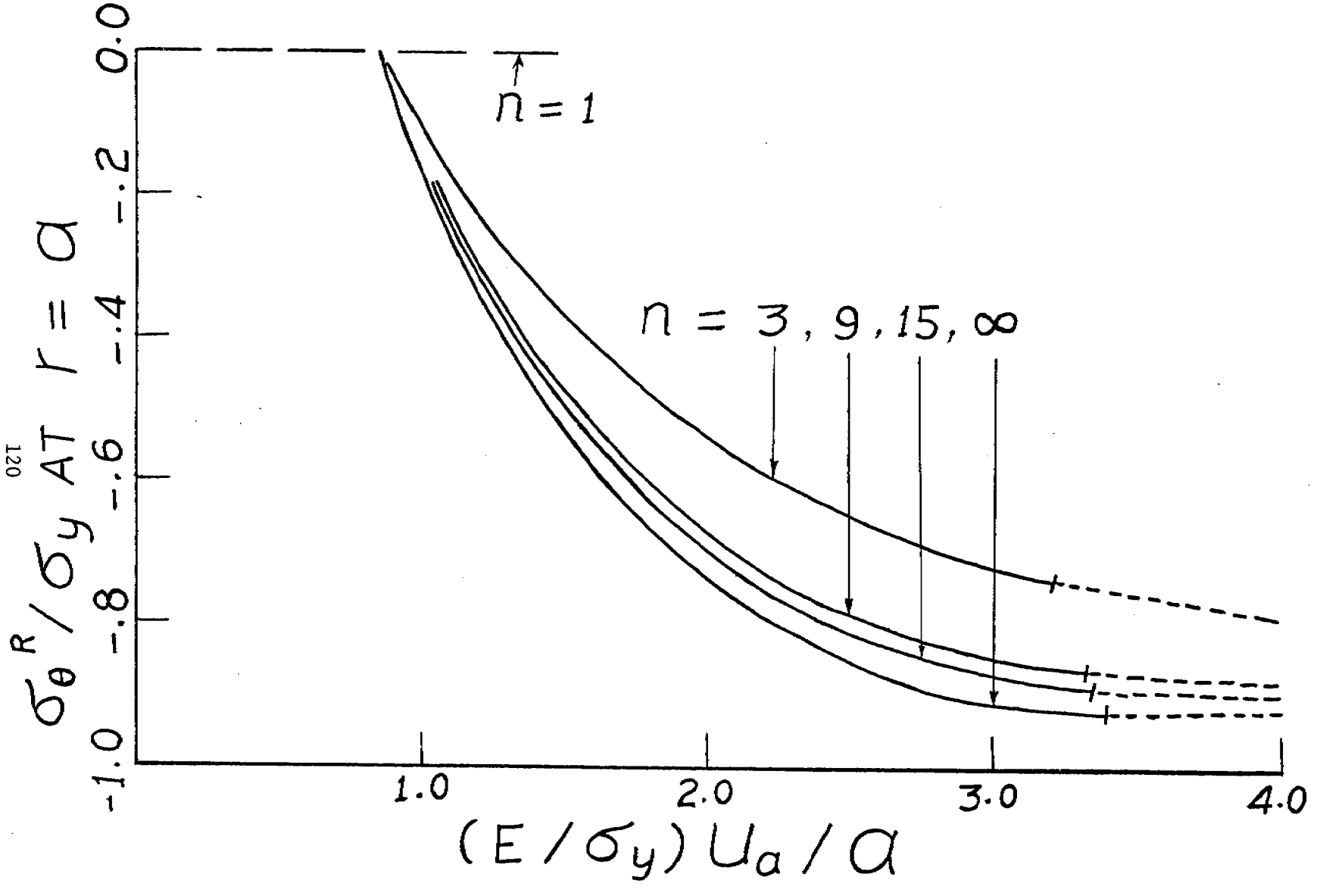


Fig. 8 Residual circumferential stress at the inside surface as functions of radial expansion (b/a = 2, $\nu = \nu_1 = .3$, R = 1).

COMPUTER SIMULATION OF SHOCK PROPAGATION IN THE ONE-DIMENSIONAL LATTICE*

John D. Powell and Jad H. Batteh
Physical Sciences Branch
Ballistic Modeling Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

ABSTRACT. The equations of motion for the atoms in a one-dimensional lattice subjected to steady shock compression are solved numerically. The atoms are assumed to interact through a nearest-neighbor, Morse-type potential. The effect of the initial state of the lattice upon the shock profile is studied by considering two sets of initial conditions. In the first, the atoms are at rest in their equilibrium positions prior to compression by the shock wave; in the second, the lattice is initially in thermal equilibrium at approximately room temperature. The lattice is found to support the propagation of well-defined, stable pulses (solitons) and the physical implications of these pulses are discussed.

I. INTRODUCTION. In the past decade or so, computer-molecular-dynamic techniques have been used by a number of different investigators to study shock propagation in discrete crystal lattices. The work has been motivated by a belief that the usual continuum approximation may, at least in certain cases, fail to describe the effect adequately. These investigations have, in our opinion, raised but failed to answer some important questions regarding the existence of steady state and the approach to thermal equilibrium behind the shock front. For this reason, and because of the importance of shock propagation to detonation theory, we have in the last few months initiated a program in computer molecular dynamics. As an initial effort, we developed in-house a computer code which solves the atomic equations of motion for a one-dimensional lattice subjected to shock compression. The results for the one-dimensional case are of some interest in themselves, but this case really represents only an initial effort in the development of a full three-dimensional code. We now describe the results of the calculations, emphasizing particularly how and why they differ from the continuum results.

II. MODEL AND EQUATIONS. The model employed is shown in Fig. 1. It consists of a one-dimensional, monatomic, chain of N atoms, each having mass m, which interact through some interatomic potential. At time T=0, the first particle is subjected to steady compression at a nondimensional velocity of unity - it is unity because we normalize all the velocities in the calculation to the compression velocity - and this compression produces a shock wave which propagates through the crystal.

* For more detailed discussion, see John D. Powell and Jad H. Batteh, "Shock Propagation in the One-Dimensional Lattice", BRL Report (to be published).

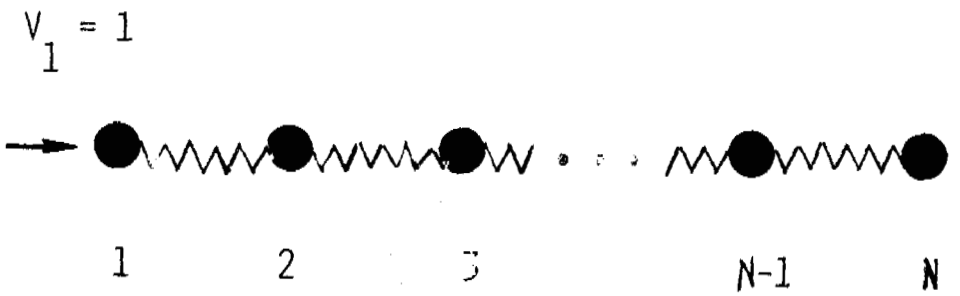


Figure 1. Model for simulating shock propagation in a one-dimensional, discrete lattice.

The classical equations of motion of each atom describing the response of the lattice to the shock wave are then solved numerically. We have considered several forms of interatomic potential but, for purposes of discussion, will consider only the Morse interaction here.

The differential equation of motion for the j^{th} particle in the lattice is just given by Newton's second law, namely,

$$\frac{d^2S_j}{dT^2} = \frac{1}{4A}\left\{\exp\left[-2A(S_j-S_{j-1})\right] - \exp\left[-A(S_j-S_{j-1})\right] - \exp\left[-2A(S_{j+1}-S_j)\right] + \exp\left[-A(S_{j+1}-S_j)\right]\right\}.$$

In this expression, S_j is the nondimensional displacement of the j^{th} particle, T the nondimensional time, and A a parameter which represents the nonlinearity or anharmonicity of the lattice. Hereafter, it will be referred to simply as the nonlinearity parameter. In this equation, we assume that only nearest-neighbor interactions are important - i.e., that only the j-1 st and j+1 st particles exert an appreciable force on the j^{th}.

There is, of course, an equation such as this for <u>every</u> particle (several hundred in our calculation), and the equations are coupled because the force exerted on a given particle depends upon the position of those adjacent to it.

<u>III. THE INITIALLY QUIESCENT LATTICE.</u> In our initial calculations, we assumed for simplification that all atoms in the lattice were initially at rest in their equilibrium positions prior to being excited by the shock front. The results of the calculations can be understood most easily by comparing the velocity-time trajectories for several particles in the lattice and determining what aspects of the general shock profile can be understood therefrom. Results are for the case A = 1.0.

Therefore, in Fig. 2, we consider initially the velocity-time trajectory of the 25^{th} particle in the lattice subsequent to its excitation by the shock. In all graphs we have arbitrarily readjusted the time axis so that, at time T=0, the particle first feels the effect of the shock. Prior to that time, it is at rest in its equilibrium position. Just behind the front, we find that the velocity of the particle varies along rather well-defined pulses whose amplitude decreases with increasing time. Although it is not shown in the graph, the amplitude of the velocity oscillations eventually approaches some asymptotic value which increases with increasing A. If we view a particular particle, then, these pulses appear to propagate into the lattice from the end at which compression is occurring and it is found that the speed with which they propagate decreases with decreasing amplitude. Thus, they tend to spread apart as they propagate. By the time the shock has reached the 500^{th} particle in the lattice, shown on the lower half of the figure, we find extremely well-defined pulses, having reached an amplitude of about 2,

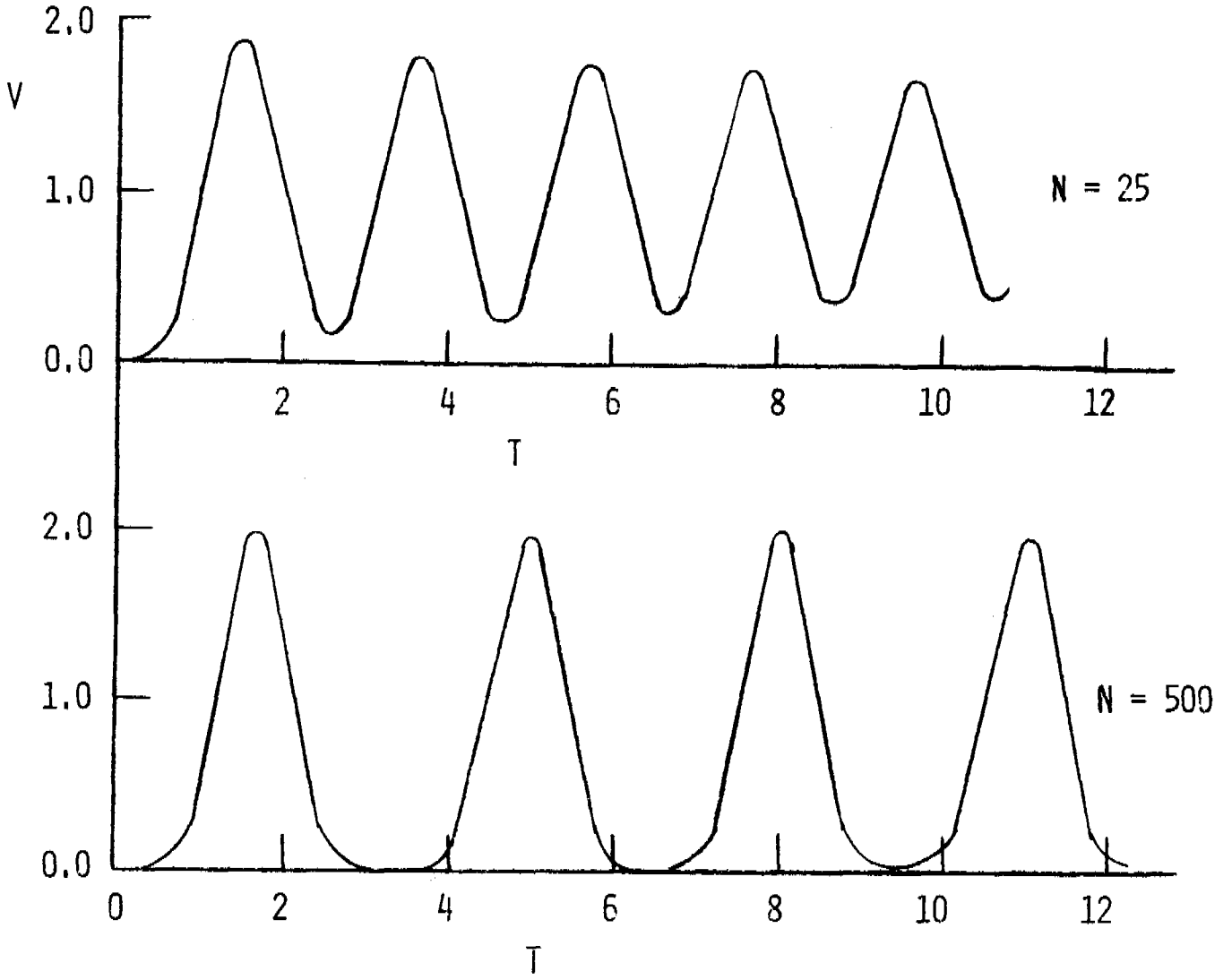


Figure 2. Velocity-time trajectories for the initially quiescent, Morse-potential lattice for the case A=1.0.

in the vicinity of the front. The pulses are more distantly spaced in time than when the shock front was at the 25th particle owing to the spreading effect just mentioned. These well-defined pulses have been observed to propagate in other nonlinear, dispersive media; they maintain their shapes upon collision with one another (demonstrated in the next section) and are referred to as solitons. Physically, they represent a balance between the nonlinearity which tends to steepen the pulse and dispersion which tends to spread it out. Dispersion is simply a characteristic property of discrete lattices in which the higher-frequency components in the pulse propagate more slowly than the lower-frequency components.

The conclusion for this case, then, is that, unlike as is assumed in continuum theories, the profile is not steady because of the spreading effect and the lattice does not approach thermal equilibrium at a higher temperature behind the shock front because of the propagation of well-defined pulses. Essentially, all the energy deposited into the lattice by the shock wave is propagated in the form of solitary waves and no energy remains for thermalization. Results similar to this have been obtained by Tasi* for the case of a slightly nonlinear, cubic interatomic potential for which he solved the equations of motion using an elegant perturbation technique.

IV. THE LATTICE AT NONZERO INITIAL TEMPERATURE. The calculations of the preceding section are somewhat of a compromise to physical reality because, as was stated previously, it was assumed that initially each atom was at rest in its equilibrium position. It was therefore desirable to next determine the effect of nonzero ambient temperature upon the propagation of the solitary waves. For this reason, we modified our code to account for a nonzero initial temperature. Specifically, we randomly assigned, according to a Maxwellian distribution, velocities to all the particles in the lattice. The initial energy of the lattice was equivalent to that in a corresponding lattice in thermal equilibrium at roughly room temperature. The lattice was then allowed to oscillate freely and we employed several checks to ensure that it was in a state of thermal equilibrium. We then subjected this lattice to shock compression in the same manner as before.

In Fig. 3 we show the velocity-time trajectory of the 145th particle in the lattice just after the shock front has reached it. Again the shock front arrives at time T=0 on this arbitrary time scale. There is some suggestion of solitary waves propagating in the lattice but, owing to the thermal background, the shape of the pulses is not so well defined as in the previous case. In fact, though it is perhaps not evident in this figure, in some cases we studied the shapes of the pulses were so poorly defined that we were unable to determine whether they were actually solitons or simply large perhaps unstable variations in the thermal background. In an effort to resolve this point, we redid the calculation and instantaneously stopped the compression when the shock wave was at

* J. Tasi, J. Appl. Phys. 43, 4016 (1972); 44, 4569 (1973); 44, 2245 (1973).

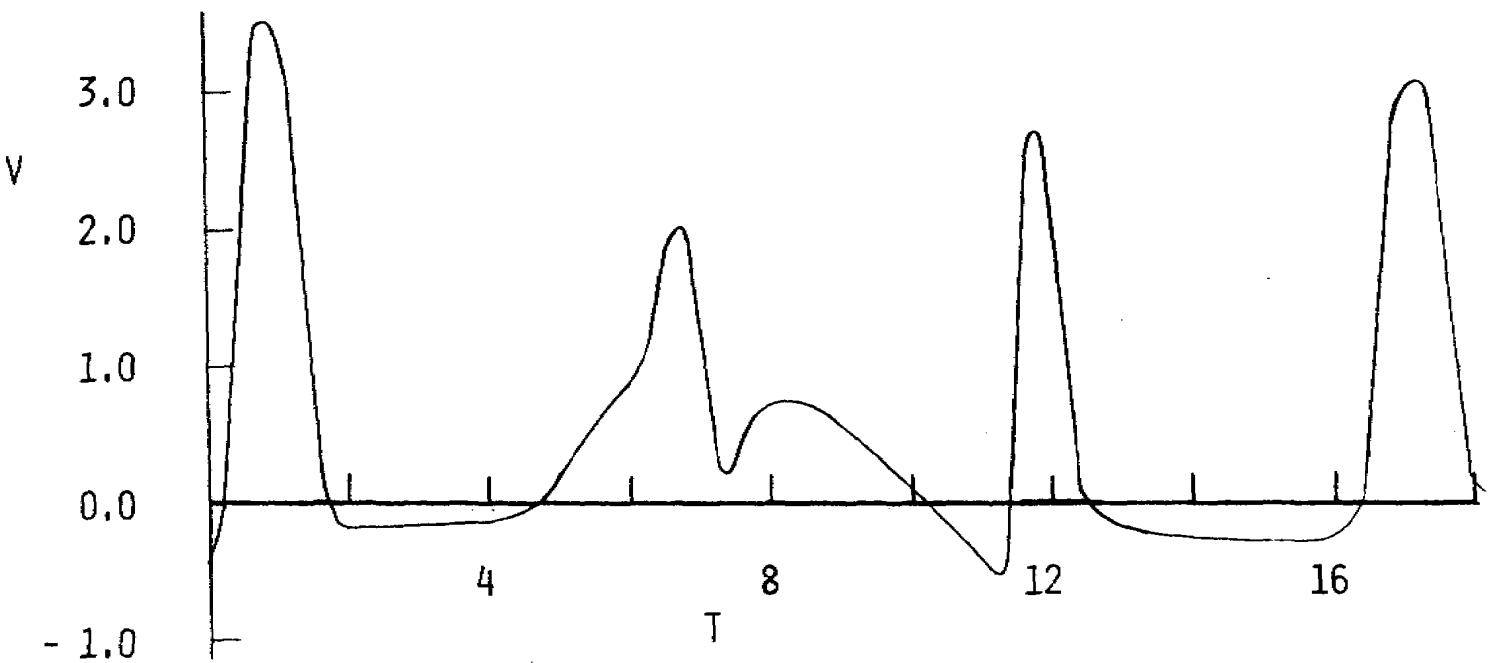


Figure 3. Velocity-time trajectory for the 145^{th} particle in the lattice at nonzero initial temperature.

the 140th particle. We then set every atom in the lattice beyond the 140th at rest in its equilibrium position so that we essentially had the shock-compressed lattice next to a cold lattice. Our belief was that the thermal background would proceed slowly into the cold lattice, whereas the high-amplitude solitons should propagate in rapidly and we could therefore separate the two effects.

The result of the calculation is shown in Fig. 4 in which is plotted the velocity-time trajectory of the 145th particle after the compression had been stopped at the 140th. Note that we indeed find solitons, of varying amplitude this time but, near the front, the amplitudes are higher than in the case for the initially quiescent lattice. If we wait a time much longer than is shown on the graph, we will eventually see the thermal background from the hot shock-compressed lattice propagating into the cold lattice. As before, we found that the pulses propagate at a rate which increases with increasing amplitude.

Since the solitons have velocities that vary with amplitude, one can easily calculate at what point in the lattice two solitons should be coincident and interact. In Fig. 5 we show such an interaction. At the 170th particle the solitons are still separated and propagating to the left. The higher-amplitude, faster-moving pulse is behind the slower. By the time the solitons have propagated to the 185th particle, a nonlinear interaction is occurring and finally, at the 200th particle, the pulses have separated and assumed their original shapes. The resultant disturbance in the region of interaction, of course, is not a simple linear superposition of separate amplitudes because the differential equations are nonlinear. Consequently, we see that the solitons are extremely stable entities that do not scatter irreversibly even when they collide with one another.

From the calculation, we can conclude that owing to the spreading effect of solitons of different propagation velocities, the profile is again nonsteady in time. Despite the fact that solitons propagate in the vicinity of the front, it is interesting to ask if any thermalization of the energy deposited by the shock wave occurs. Unfortunately, at this point we have no complete answer to this question but the evidence seems to suggest that it does. Specifically we have calculated the velocity distribution function for atoms which were well behind the shock front at several different times and found that it is approximately Maxwell-Boltzmann and remains essentially constant in time. The distribution function consistently corresponds to a temperature that is about three times the ambient temperature of the lattice. Our tentative speculation then - and it is only speculation at this point - is that the perturbation induced by the initial thermal oscillations will prevent the formation of, or perhaps lead to the decay of, some of the solitons that might otherwise form and that this energy eventually becomes thermalized. Additional study of the effect will be necessary before firm conclusions can be drawn.

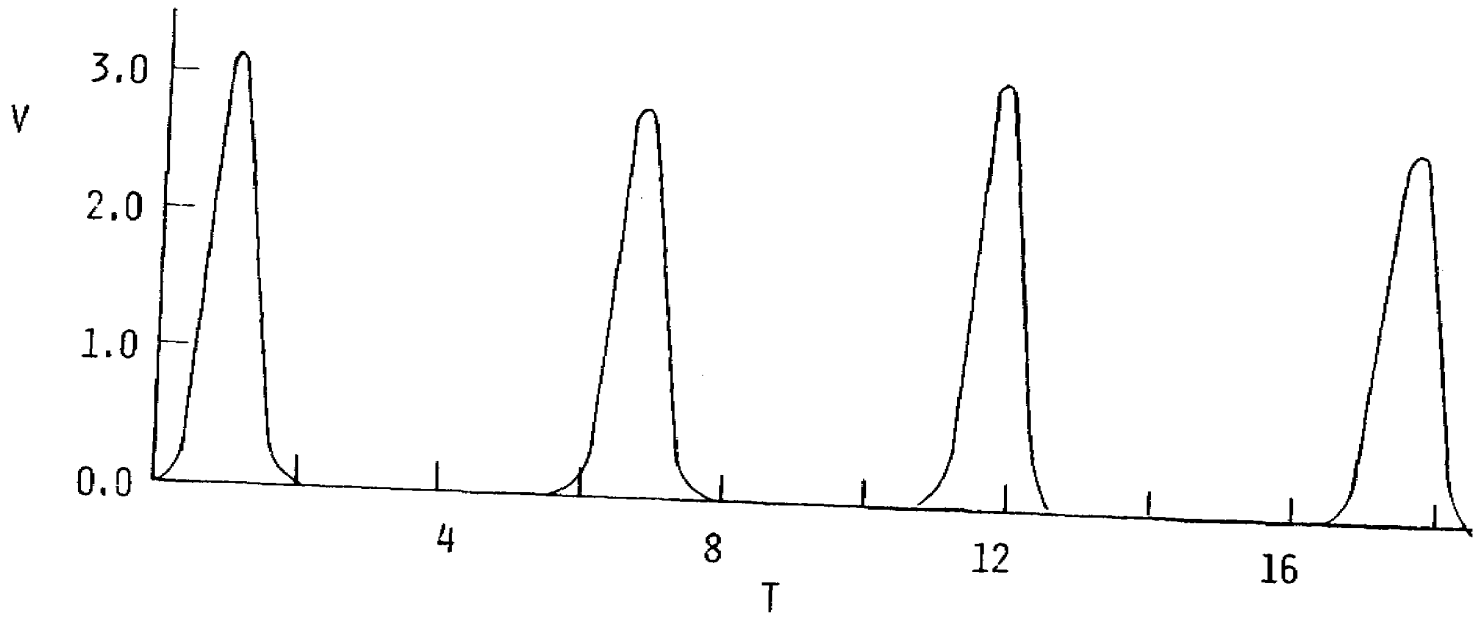


Figure 4. Propagation of solitons into the cold lattice.

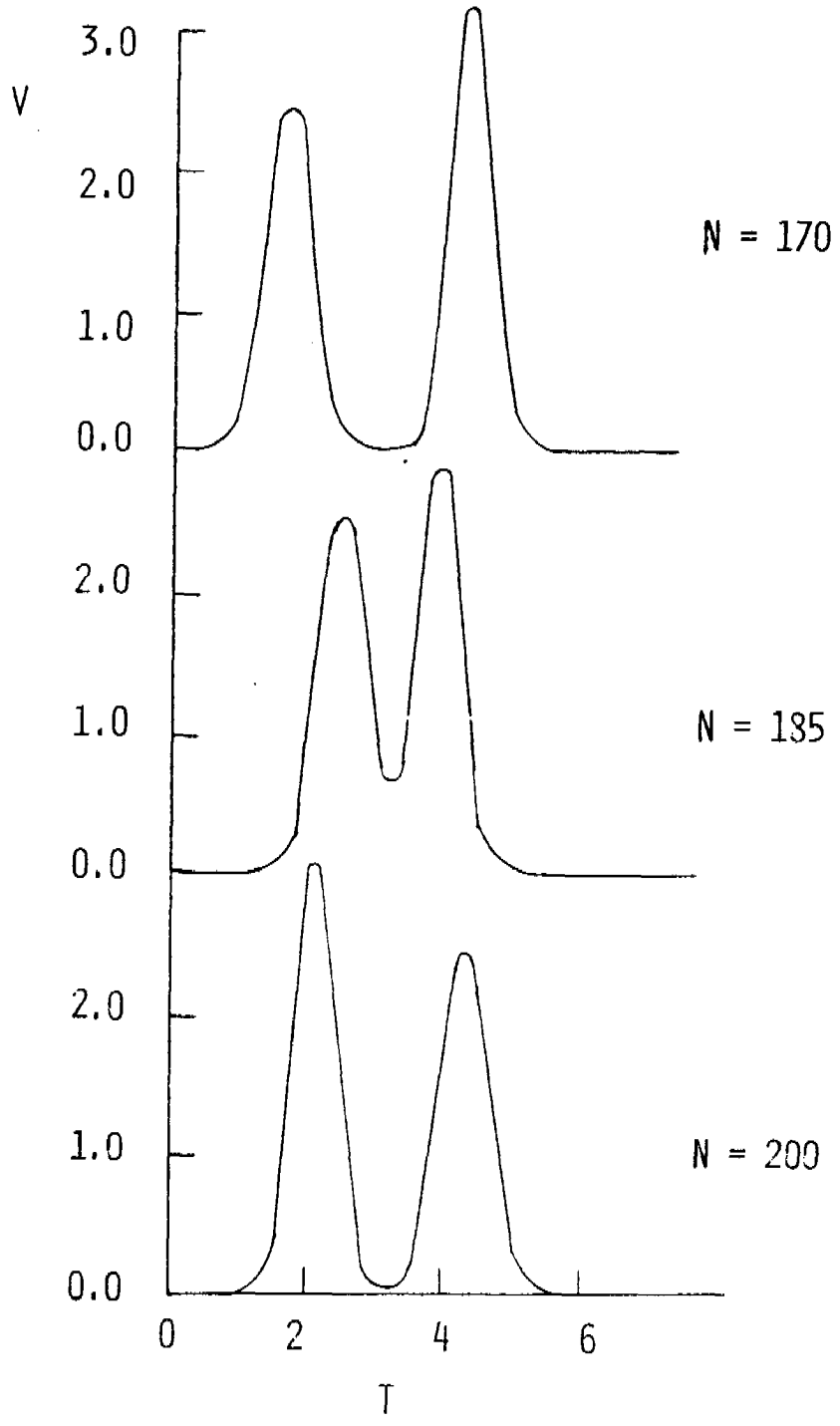


Figure 5. Velocity-time trajectories for three particles illustrating a soliton collision. The time scale is arbitrary.

V. SUMMARY AND CONCLUSIONS. To summarize, we have found that in the one-dimensional case the propagation of solitons tends to prevent the rapid establishment of thermal equilibrium behind the shock front, and the spreading effect prevents the shock profile from approaching a steady state as is generally assumed in continuum theories of shock propagation. Consequently, the transition region between the equilibrated region of the crystal ahead of the front and that behind the front grows as the shock propagates into the crystal.

In the near future, we intend to extend the calculations to three dimensions and to more realistic, perhaps impure, crystals to see if similar effects persist in that case. If so, we believe that these effects may be significant in the study of shock-induced detonations. Most current theories assume that the transition region can be ignored and that the only effect which the shock has is to raise the temperature, pressure, and density of the crystal to values higher than the ambient values. Chemical reactions then occur in a thermally equilibrated part of the crystal. It would appear, however, that since the transition region grows, chemical reactions may occur in a region characterized by extreme nonequilibrium and it is important to assess the effects of this environment upon chemical - reaction rates. It is unlikely, for instance, that such rates can be characterized by an Arrhenius - type relation whose validity is dependent upon the existence of thermal equilibrium. We hope to investigate this problem in greater detail once the calculations have been extended to three dimensions.

A PERTURBATION EXPANSION OF THE NAVIER-STOKES EQUATIONS FOR SHOCK WAVES*

Jad H. Batteh and John D. Powell
Physical Sciences Branch
Ballistic Modeling Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005

ABSTRACT. A perturbation expansion is developed for the steady-state Navier-Stokes equations describing one-dimensional shock propagation in an ideal gas. The temperature dependence of the viscosity and thermal conductivity is accounted for, though the specific heat and Prandtl number are assumed constant. It is shown that if the first n-1 terms of the expansion are known, the solution for the nth term can be reduced to a quadrature. The expansion is evaluated explicitly to second order in the shock strength. A comparison of the second-order approximation with a special-case, exact solution indicates good agreement even for rather strong shock waves. The perturbation solution provides a simple, analytic technique for determining the effect of temperature-dependent transport coefficients on the structure of weak shock waves.

I. INTRODUCTION. For sufficiently weak shock waves, the variation of the flow variables through the front can be determined from the Navier-Stokes equations. In general, these equations are too complicated to solve exactly and one must usually resort to numerical techniques. It is desirable to supplement the numerical solutions with approximate, analytic solutions wherever possible since they permit simpler calculations for the range in which they are valid and generally lead to greater physical insight. In this paper, we will describe a perturbation expansion we have developed for solving the steady-state Navier-Stokes equations describing the propagation of a planar shock wave in an ideal gas. The approximate solution we obtain from the expansion offers a simple, analytic technique for evaluating the effect of temperature-dependent transport coefficients on the structure of weak shock waves.

Figure 1 shows a typical shock profile in the co-ordinate frame in which the shock wave is stationary. In the laboratory frame, the shock actually propagates into a stationary gas with a speed u_i in the negative x-direction. As it propagates, the shock wave compresses and heats the gas and sets it in motion so that the density and

*A more detailed treatment of this presentation can be found in Reference (1).

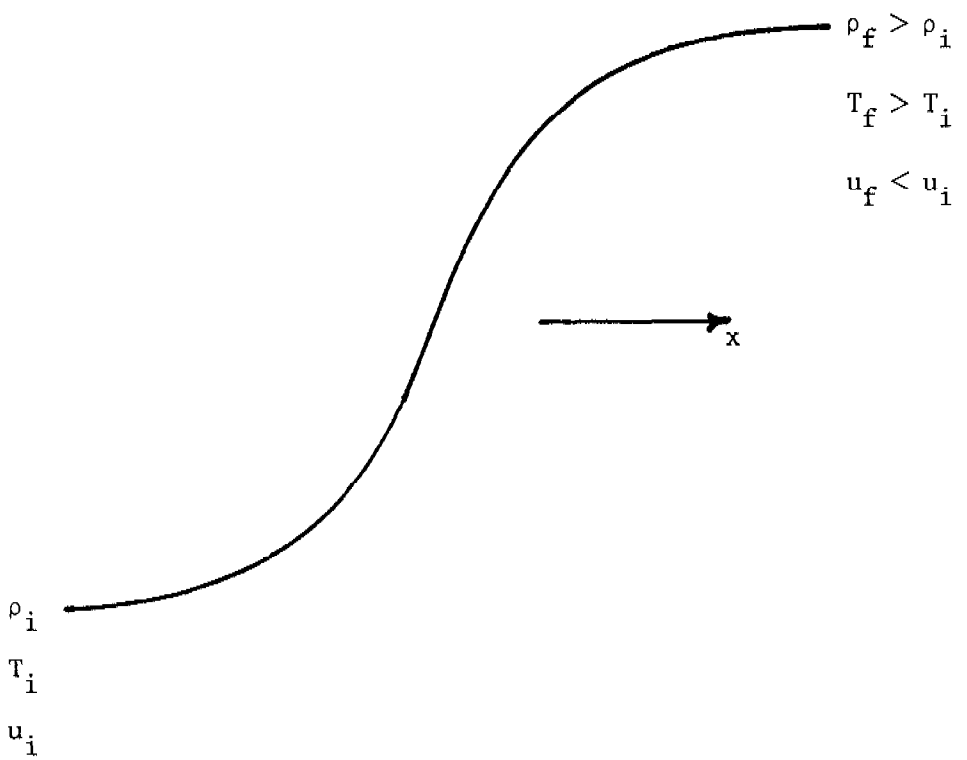


Figure 1. Steady-state shock profile in the stationary frame.

temperature behind the front, denoted by ρ_f and T_f, respectively, are greater than the corresponding values, ρ_i and T_i, ahead of the front. In the stationary frame, the flow velocity behind the front, u_f, is less than that ahead of the shock wave. The width of the shock front is typically several molecular mean free paths and decreases as the strength of the shock increases.

II. THE FLOW EQUATIONS. The shock profile in the stationary frame is determined by the steady-state hydrodynamic equations which express the conservation of mass, momentum and energy:

$$\frac{d}{dx}(\rho u) = 0 \tag{1}$$

$$\rho u \frac{du}{dx} + \frac{dP_{xx}}{dx} = 0 \tag{2}$$

$$\rho u \frac{de}{dx} + \frac{dQ}{dx} + P_{xx} \frac{du}{dx} = 0 \ . \tag{3}$$

In these equations, e is the specific internal energy of the gas, Q is the energy dissipated due to heat conduction and P_{xx} is the appropriate element of the pressure tensor.

Equations (1) - (3) are general and exact, but they contain insufficient information to determine the flow variables. To close the set of equations, it is necessary to specify a form for the thermal conduction and the pressure element, as well as an equation of state. For sufficiently weak shock waves, where the gradients in the flow variables are small, the thermal conduction is adequately represented by Fourier's Law

$$Q = - \kappa \frac{dT}{dx} \ , \tag{4}$$

where κ is the thermal conductivity and the pressure term by Stokes' hypothesis

$$P_{xx} = P - 4/3 \ \mu \frac{du}{dx} \ , \tag{5}$$

where P is the pressure and μ is the viscosity. The conservation equations, together with Fourier's Law and Stokes' hypothesis, are generally referred to as the Navier-Stokes equations. Finally, we assume for this analysis a perfect gas so that the equation of state

can be written as

$$p = \frac{\gamma-1}{\gamma} c_p \rho T \tag{6}$$

where c_p is the specific heat at constant pressure and γ is the ratio of that specific heat to the one at constant volume. The values of c_p and γ are assumed to remain constant during the shock compression.

It should be emphasized that the Navier-Stokes equations are valid only for relatively weak shock waves. For strong shocks, the gradients of the flow variables become large and this fact precludes the use of Fourier's Law and Stokes' hypotehsis. One must then resort to more complicated analyses based on the Boltzmann equation.

It is convenient to define a shock strength parameter, ε, according to

$$\varepsilon = \frac{M^2-1}{\gamma M^2+1} \tag{7}$$

where M is the Mach number given by the ratio of the shock speed u_i to the sound speed in the gas, c. The Mach number approaches unity for weak shock waves and increases without bound with increasing shock strength. Consequently, ε approaches zero for weak shock waves and has a maximum value of $1/\gamma$ which is less than unity.

We choose to nondimensionalize the flow velocity and the temperature according to

$$\eta = u/u_o \tag{8}$$

$$t = (1-\gamma\varepsilon^2)\ T/T_o \tag{9}$$

where u_o is the average velocity across the shock wave and T_o is the average temperature. These average quantities can be determined directly from the Rankine-Hugoniot conditions (2) which relate the final values of the flow variables behind the front to the initial values ahead of the shock. In terms of the shock strength parameter, the averages are given by

$$T_o = \frac{T_i+T_f}{2} = \frac{1-\gamma\varepsilon^2}{(1-\gamma\varepsilon)(1+\varepsilon)}\ T_i \tag{10}$$

$$u_o = \frac{u_i+u_f}{2} = \frac{u_i}{1+\varepsilon}\ . \tag{11}$$

The transport coefficients, μ and κ, are allowed to vary with temperature, but we take advantage of the fact that their functional dependence is similar in the temperature range of interest. Consequently, the thermal conductivity and viscosity are represented by

$$\kappa = \kappa_o f(t) \tag{12}$$

$$\mu = \mu_o f(t) \tag{13}$$

where κ_o and μ_o are constants and $f(1) = 1$.

Finally, we define a nondimensional co-ordinate

$$y = \frac{\rho_i u_i}{\mu_o} x \ . \tag{14}$$

For weak shock waves, the length scale $\mu_o/\rho_i u_i$ is on the order of a molecular mean free path.

Since Equation (1) representing conservation of mass can be integrated directly, the Navier-Stokes equations can be reduced to a pair of coupled, nonlinear differential equations. The two remaining equations can be written in terms of the nondimensional variables as

$$\frac{4}{3}\gamma f(t)\eta \ \frac{d\eta}{dy} + \gamma\eta(1-\eta) + \eta - t = 0 \tag{15}$$

$$\frac{1}{Pr}\frac{dt}{dy} - \frac{4}{3}\ \eta\frac{d\eta}{dy} + \frac{(1+\gamma)}{2f(t)}\left[\ (\eta-1)^2 - \varepsilon^2\ \right] = 0 \ . \tag{16}$$

In Equations (15) and (16), $f(t)$ represents the temperature dependence of the transport coefficients and

$$Pr = \frac{\mu c_p}{\kappa} \tag{17}$$

is the Prandtl number which is of order unity for gases. For this analysis, Pr is a constant since c_p is assumed to be a constant and μ and κ have the same dependence on temperature.

Equations (15) and (16) are to be solved subject to the boundary conditions specified by the Rankine-Hugoniot relations, namely

$$\eta(-\infty) = 1+\varepsilon \quad ; \quad \eta(\infty) = 1-\varepsilon \tag{18}$$

$$t(-\infty) = (1-\gamma\varepsilon)(1+\varepsilon) \; ; \; t(\infty) = (1+\gamma\varepsilon)(1-\varepsilon). \tag{19}$$

III. PERTURBATION EXPANSION AND FORMAL SOLUTION. We now proceed to solve Equations (15) and (16) by a perturbation expansion. In determining the form of the expansion, we make use of two observations. First, the boundary conditions, Equations (18) and (19), indicate that the variation of η and t across the shock is of order ε. Since ε approaches zero as the shock strength decreases, it emerges as a natural choice for the perturbation parameter. Second, both theory and experiment indicate that the shock thickness varies as $1/(M-1)$ for weak shocks, which in turn varies as $1/\varepsilon$. Therefore, we expect that dt/dy and $d\eta/dy$ will be proportional to $\varepsilon\,\Delta t$ and $\varepsilon\,\Delta\eta$, respectively. Actually, this judgement does not have to be made a priori but emerges during the perturbation expansion procedure as the only consistent ordering of the derivatives.

These two observations suggest expansions for η and t of the form*

$$\eta = \sum_j \varepsilon^j \; \phi_j(\varepsilon y) \tag{20}$$

$$t = \sum_j \varepsilon^j \; \theta_j(\varepsilon y) \tag{21}$$

where $\phi_o = \theta_o = 1$. The series given by Equations (20) and (21) correspond to expanding the velocity about its midvalue and the temperature about the value $T_o(1-\gamma\varepsilon^2)$ which approaches the midvalue as the shock strength diminishes.

In order to determine the functions ϕ_j and θ_j, the expansion is substituted into Equations (15) and (16) and the resulting equations solved separately to each order in ε. For the sake of brevity the details of the expansion procedure will be omitted. A more detailed treatment of the perturbation expansion and resulting solution can be found in Reference (1).

Selecting the lowest order terms in the two equations results in an ordinary differential equation for ϕ_1 which can be solved to yield

$$\phi_1 = -\tanh(\lambda r) \tag{22}$$

where

$$r = \varepsilon y$$

*All sums extend from 0 to ∞ unless otherwise noted.

and a relation between θ_1 and ϕ_1,

$$\theta_1 = (1-\gamma)\phi_1. \tag{23}$$

The constant λ in Equation (22) is given by

$$\lambda = \frac{3(\gamma+1)Pr}{8Pr+6(\gamma-1)} \quad . \tag{24}$$

Furthermore, the constant of integration has been chosen so that y=0 represents the location of the midvalue of the velocity profile.

For $n \geqslant 2$, solving the system of equations for order ε^n results in a differential equation for ϕ_n which can be written as

$$\frac{d\phi_n}{dr} - 2\lambda\phi_1\,\phi_n = T_{n-1} \tag{25}$$

where T_{n-1} is a rather complicated function of the n-1 preceding ϕ_j and θ_j. Since T_{n-1} is not a function of ϕ_n or θ_n, Equation (25) is a linear, ordinary differential equation which has the formal solution

$$\phi_n = \operatorname{sech}^2(\lambda r)\left[\int \cosh^2(\lambda r) T_{n-1}\, dr\right]. \tag{26}$$

Again, θ_n can be related algebraically to ϕ_n and the preceding ϕ_j and θ_j.

Consequently, it is possible to obtain, at least formally, as many terms in the expansion as one wishes by repeatedly evaluating the integral in Equation (26). In practice, this soon becomes difficult to do analytically since T_{n-1} becomes progressively more complicated as n increases. However, we were able to evaluate the integral explicitly to obtain the second-order term, that is the term in the solution of order ε^2. The resulting second-order approximations to the velocity and temperature profiles are given by

$$\eta_2 = \left(\frac{u}{u_o}\right)_2 = 1-\varepsilon \tanh(\lambda r) + \varepsilon^2 A \operatorname{sech}^2(\lambda r) \ell n \operatorname{sech}(\lambda r) \tag{27}$$

$$\begin{aligned} t_2 = (1-\gamma\varepsilon^2)\left(\frac{T}{T_o}\right)_2 &= 1 + \varepsilon(\gamma-1)\tanh(\lambda r) \\ &+ \varepsilon^2\left\{-A(\gamma-1)\operatorname{sech}^2(\lambda r)\,\ell n \operatorname{sech}(\lambda r)\right. \\ &\left. + \gamma\left[(1-4\lambda/3)\operatorname{sech}^2(\lambda r) - 1\right]\right\} \end{aligned} \tag{28}$$

where

$$A = 1 - \frac{3(\gamma^2-1)(4Pr-3)}{[4Pr+3(\gamma-1)]^2} - (\gamma-1)\left(\frac{df}{dt}\right)_{t=1} \qquad (29)$$

IV. DISCUSSION. It is apparent that as $r \to \pm \infty$ the second-order solutions, given by Equations (27) and (28), are in identical agreement with the Rankine-Hugoniot relations, expressed in the form of Equations (18) and (19), for all shock strengths. Of course, agreement at the endpoints does not guarantee that the approximate solutions adequately represent the entire profile. In an attempt to estimate the range of validity of the approximate solutions, we compared them with a special-case calculation which is known to have an analytic solution.

As first shown by Becker (3), the steady-state Navier-Stokes equations can be solved exactly for the case where μ and κ are temperature independent and $Pr = 3/4$. We have compared our approximate solutions with the exact solution for that case with γ equal to 5/3. In Figure (2), we show the first- and second-order approximations to the velocity profile for a Mach number equal to 2. This value represents approximately the maximum Mach number for which the Navier-Stokes equations themselves are valid. For this case, the exact solution is indistinguishable from the second-order approximation, the difference being less than 1% throughout the entire range of r. Although the results are not reproduced here, the second-order approximation to the temperature ratio was also found to be within 1% of the exact solution at M=2.

The agreement was even better for smaller values of M, as would be expected. As M increased, the deviation between the exact and second-order profiles increased; but even for a Mach number of 10, the difference was less than 15%. Of course, the results at large Mach numbers are only of academic interest since the original equations are then no longer valid.

These results are somewhat surprising since other expansions using M-1 as a perturbation parameter are valid over a much smaller range. It appears that in this case the choice of an expansion parameter significantly affects the accuracy of the approximation.

There are, in fact, two indications that suggest that ε is a more appropriate expansion parameter than M-1. First, ε never exceeds unity as the Mach number increases, whereas M-1 increases without bound as $M \to \infty$. Second, with our nondimensionalization we were able to satisfy the boundary conditions exactly for all Mach numbers with, at most, a second-order approximation, whereas all the terms in the series are required if M-1 is used as an expansion parameter.

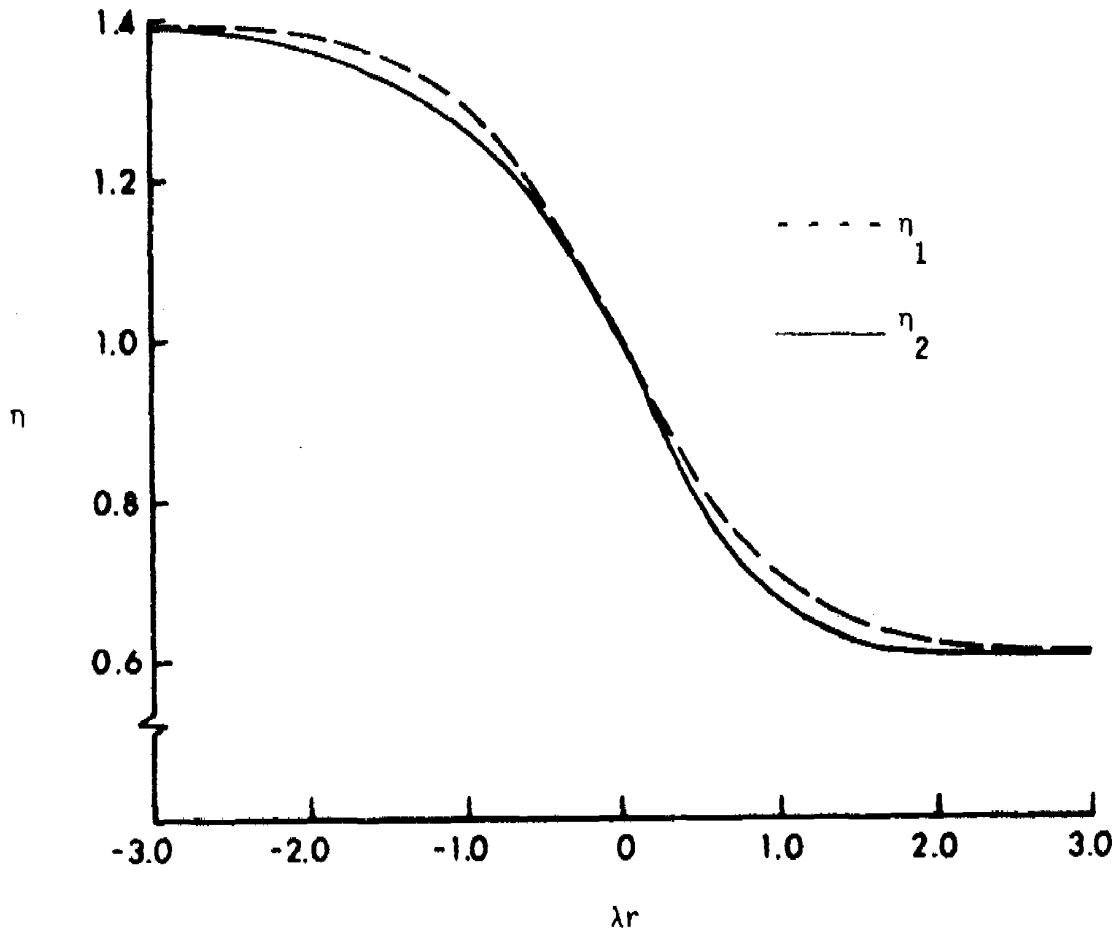


Figure 2. Comparison of first- and second-order solutions for temperature-independent transport coefficients. Becker's solution for η is not shown since it was found to be coincident with η_2. The results are plotted for the following values of the parameters: M=2, Pr=3/4, γ=5/3.

The second-order approximation can be used to examine the effect of temperature-dependent transport coefficients on the shock profile since the terms of order ε^2 in Equations (27) and (28) depend on df/dt at t=1. For example, simple kinetic theory (4) predicts that for a hard-sphere gas the Prandtl number is equal to 2/3 while the viscosity and thermal conductivity are given by

$$\mu = \mu_o t^{\frac{1}{2}} \tag{30}$$

$$\kappa = \kappa_o t^{\frac{1}{2}} . \tag{31}$$

In Figure 3, we have plotted η_2 as calculated from Equation (27) for the case in which M=2, Pr=2/3, γ=5/3, and μ and κ are given by Equations (30) and (31). The profile is compared with the solution for η_2 obtained by holding μ and κ constant at their upstream values. Both profiles are plotted as a function of the dimensionless parameter $z = \lambda\varepsilon\rho_i u_i x/\mu_i$. It is apparent from the figure that including the temperature-dependence of the transport coefficients changes the value of η_2 by as much as 12% at this Mach number. The broader front in the temperature-dependent case is to be expected since the dissipation is enhanced by allowing μ and κ to increase with temperature.

V. SUMMARY. We have developed a perturbation expansion for the Navier-Stokes equations describing steady-state, one-dimensional shock propagation in an ideal gas. Formally, the expansion can be evaluated to any order by quadrature. The second-order solution has been determined explicitly and this solution appears to accurately represent the shock profile for the range of Mach numbers in which the Navier-Stokes equations themselves are valid. The expansion provides a simple, analytic method for investigating the effect of temperature-dependent transport coefficients on the shock profile.

REFERENCES

1. J.D. Powell and J.H. Batteh, "Perturbation Expansion of the Navier-Stokes Equations for Shock Waves," Phys. Fluids 20, 734 (1977).

2. H.W. Liepmann and A. Roshko, Elements of Gasdynamics (Wiley, New York, 1957), Ch. 4.

3. R. Becker, "Stosswelle und Detonation," Z. Physik 8, 321 (1922).

4. J.O. Hirschfelder, C.F. Curtiss and R.B. Bird, Molecular Theory of Gases and Liquids (Wiley, New York, 1954), Ch. 1.

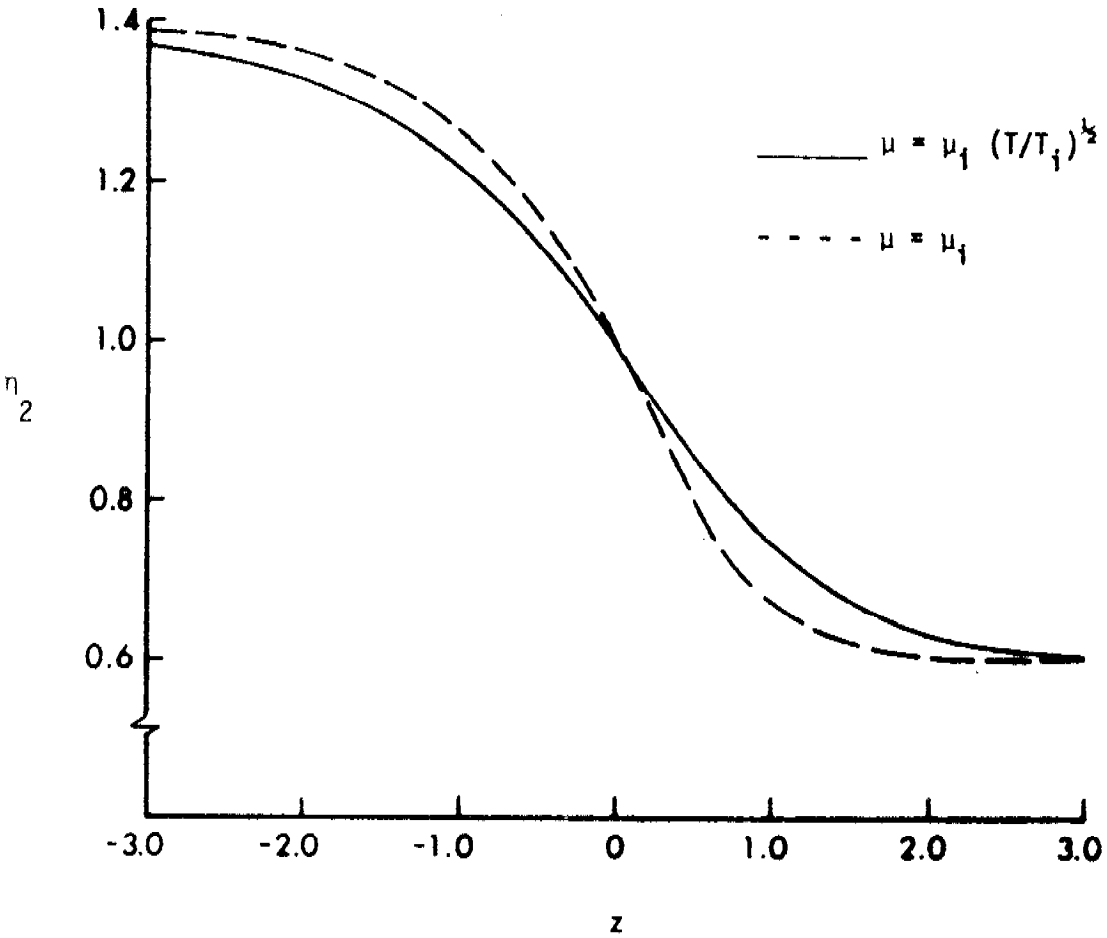


Figure 3. Effect of temperature-dependent transport coefficients on velocity profile. Results are plotted for M=2, Pr=2/3, and γ=5/3.

A GENERALIZED COMPARISON PRINCIPLE AND MONOTONE METHOD FOR SECOND ORDER BOUNDARY VALUE PROBLEMS IN BANACH SPACES*

S. R. Bernfeld[1], V. Lakshmikantham[1], S. Leela[1,2]

Abstract

A generalized comparison principle for second order differential inequalities is established in a Banach space where the inequalities are given relative to an arbitrary cone. In case the Banach space is the real line the comparison principle reduces to the classical maximum principle.

The comparison principle is then used to develop a monotone method to generate two-sided bounds on solutions of nonlinear boundary value problems for ordinary differential equations in a Banach space.

1. *Introduction*

Monotone methods have been used to generate maximal and minimal solutions of nonlinear boundary value problems for both ordinary and partial differential equations. This study was originally motivated by the problem of extending the chord method as used by Keller [8] and Sattinger [11], who considered nonlinear partial differential equations containing no gradient term. The inclusion of a gradient term was first introduced by Chandra and Davis [5] who considered the ordinary boundary value problem

$$u'' = f(t,u,u') \tag{1.1}$$

$$B^i u = \alpha_i u(i) + (-1)^{i+1}\beta_i u'(i) = b_i, \quad i = 0,1. \tag{1.2}$$

Here $f \in C[[0,1] \times R \times R, R]$, $\alpha_0,\alpha_1 \geq 0$, $\beta_0,\beta_1 > 0$.
They assumed that f depended linearly on u'. This restriction

*Research partially supported by U. S. Army Research Grant DAAG29-77-G0062.

[1]Mathematics Department, University of Texas at Arlington, Arlington, Texas 76019

[2]Mathematics Department, State University of New York, College at Geneseo, Geneseo, New York 14454

on u' was eliminated by Bernfeld and Chandra [2]. The approach that is used in [5] and [2] is to first ascertain the existence of a lower solution $u_0(t)$ and upper solution $v_0(t)$ such that $u_0(t) \leq v_0(t)$. By assuming a Nagumo Condition [3] on $f(t,u,u')$ one is able by standard methods to obtain a uniform estimate on the derivative of any solution $z(t)$ of (1.1) such that $u_0(t) \leq z(t) \leq v_0(t)$ $t \in [0,1]$. These conditions imply the existence of solutions of (1.1) and (1.2). In order to obtain the monotone iterations, one linearizes $f(t,u,u')$ in the variable u about any element $z(t)$ such that $u_0(t) \leq z(t) \leq v_0(t)$. More precisely consider

$$u'' = F(t,u,u',z), \tag{1.3}$$

where $F(t,u,u',z) = f(t,z,u') + \delta u - \delta z$, where $u_0(t) \leq z(t) \leq v_0(t)$ and δ is an upper bound on $|f_u(t,u,u')|$ for $t \in [0,1]$, $u_0(t) \leq u(t) \leq v_0(t)$ and $|u'| \leq N$, where N is obtained from the Nagumo Condition. The existence of solutions of (1.3) and (1.2) follows as in the problem (1.1),(1.2). By using the classical maximum principle one obtains for each $z(t)$ such that $u_0(t) \leq z(t) \leq v_0(t)$ a unique solution $w(t)$ of (1.3), (1.2) such that $u_0(t) \leq w(t) \leq v_0(t)$. If we define the mapping A to be governed by the rule $w = Az$, then by the maximum principle one has that $Au_0 \geq u_0$, $Av_0 \leq v_0$ and A is monotone on $< u_0, v_0 >$, that is, $z_1 \leq z_2 \Rightarrow Az_1 \leq Az_2$. Finally by defining the sequences $u_n = Au_{n-1}$, $v_n = Av_{n-1}$ one shows that $\{u_n\}$ and $\{v_n\}$ converge uniformly and monotonically to the minimal and maximal solutions respectively of the BVP(1.1), (1.2).

The extension of these results to R^n and infinite dimensional systems of the type (1.1) is important, for parabolic equations of the form

$$u_{xx} = f(t,x,u,u_t,u_x) \tag{1.4}$$

can be approximated, using the method of lines, by n-dimensional and infinite dimensional systems of type (1.1) (see Liskovets [9] and Thompson [13]). Chandra, Lakshmikantham, and Leela [6] extended the development of the monotone method to the Banach space E of bounded sequences $x = \{\xi_i\}$, $i \in Z^+$. In particular they considered the BVP(1.1), (1.2) in which $f \in C\,[[0,1] \times E \times E, E]$,

α_i, β_i scalars, $\alpha_0, \alpha_1 \geq 0$, $\beta_0, \beta_1 > 0$, $b_0, b_1 \in E$. Their results of course included R^n as we can identify R^n as a subspace of E.

One of the biggest problems in extending the scalar case is the lack of a suitable maximum principle in higher dimensions. In [6] a comparison principle in E was developed which played the same role as the classical maximum principle for the scalar problem. All inequalities are assumed to be componentwise. An appropriate existence theorem in Banach spaces is used in which the modified function approach is the principle tool. In particular appropriate compactness type conditions are imposed on f in terms of the measure of noncompactness (see J. Chandra, V. Lakshmikantham, A. R. Mitchell [7]). The comparison result allows us to assert that the solution of (1.1), (1.2) lies between $U_0(t)$ and $V_0(t)$, the lower and upper solutions respectively. (See also R. Thompson [14] for a good discussion of existence in the space E.) The comparison principle thus becomes the main tool in the development of the monotone method. Conditions such as quasimonotonicity on f and the restriction that $f_i(t,u,u') \equiv f_i(t,u,u_i')$ are necessary in order to successfully develop the monotone method. Recall, for example, that in R^n, $n > 1$, the only known methods showing that $U_0(t) \leq V_0(t)$, where $U_0(t)$ and $V_0(t)$ are lower and upper solutions in the usual sense

$$U_0'' \geq f\big(t, U_0(t), U_0'(t)\big),$$
$$V_0'' \leq f\big(t, V_0(t), V_0'(t)\big),$$

are to require that $f_i(t,x,x') \equiv f_i(t,x,x_i')$. One can require more conditions on U_0 and V_0, thus eliminating the restrictions on f, but this seems to be incompatible with the development of the monotone method.

As in scalar case the comparison principle yields the uniqueness of the solutions of system (1.3) and leads to the subsequent development of the monotone operator A (see Section IV).

In this note we shall extend the comparison theorem to arbitrary Banach spaces E. We shall deal with the second order differential inequalities in which the inequality relation is induced by a cone K in E. We shall discuss the natural development of the monotone method but not present the details here. Our results

will include those in [5], [2], [6] and complement the results of R. Thompson [15]. We will be more precise about this later.

2. Preliminaries

Let B be a Banach space with norm $|\cdot|$ and let B^* denote the set of continuous linear functionals. Let K be a cone in B which induces a partial ordering $\leq$ as follows: $x \leq y$ if and only if $y - x \in K$.

A linear functional $\phi \in B^*$ is called a positive linear functional if $\phi(x) \geq 0$ whenever $x \in K$. Let K^* denote the set of positive linear functionals. Notice then that K is contained in the closed halfspace $C_\phi = \{x \in B \mid \phi(x) \geq 0, \ \phi$ a positive linear functional$\}$. Thus the positive linear functionals are support functionals and since K is a cone in B, then K is the intersection of all the closed half-spaces which support it. If $K_G \subseteq K^*$ and $K = \cap \{C_\phi \mid \phi \in K_G\}$ then we say K_G generates K. Let $K_u = \{\phi \in K_G \mid \|\phi\| = 1\}$, and $\overline{K_u}$ the closure of K_u in the weak star topology.

Denote by $int\ K$, the interior of K and if $int\ K \neq 0$, then K is called a solid cone. It is interesting to note that $x \in int\ K$ if and only if there exists $\varepsilon > 0$ such that $\phi(x) \geq \varepsilon$ for all $\phi \in K_u$.

3. Comparison Result

As indicated in the introduction an essential feature of the monotone method for (1) (2) is the application of a comparison principle, which in one dimension, is the classical maximum principle. We will state the comparison result and indicate its implications.

We consider the boundary value (1.1), (1.2), where $f : I \times B \times B \to B$ is continuous with $I = [0,1]$. Let K be a cone in B generated by K_u.

Definition: We say $f(t,x,x')$ is quasimonotone nonincreasing in x with respect to K if $x \leq y$, $\phi(x) = \phi(y)$, $\phi(x') = \phi(y')$ implies $\phi\big(f(t,x,x')\big) \geq \phi\big(f(t,y,y')\big)$ for all $\phi \in K_u$.

For the case in which $B = R^n$, $1 \leq n \leq \infty$, where K is the cone generated by the positive projections, this condition on f implies that $f_i(t,x,x') \equiv f_i(t,x,x_i')$. This can be seen by letting $x = y$, and thus obtaining $f_i(t,x,x') = f_i(t,x,y')$ whenever $x_i' = y_i'$. If for example $f(t,x,x')$ is linear in x and x' then $f(t,x,x') = P(t)x' + Q(t)x$, where $P(t)$ is a diagonal matrix and $Q(t)$ is a matrix in which $q_{ij}(t) \leq 0$ $i \neq j$.

We consider a family of functions $\{Z_\lambda(t)\}$, $\lambda > 0$, $t \in [0,1]$, and we shall write $Z(\lambda,t) \equiv Z_\lambda(t)$. We say this family is admissible if $Z(\lambda,\cdot)$ is $\in C^2[I,B]$ for each λ and $Z(\cdot,t)$ is continuous in λ for each t and $Z(0,t) = 0$ for $0 \leq t \leq 1$. The family is said to satisfy a uniformity condition at $\lambda = 0$ ($\lambda = \infty$) if for each $\phi \in K_u$ $\phi\big(Z_\lambda(t)\big) \to 0$ as $\lambda \to 0$ uniformly for $t \in [0,1]$ $\big(\phi\big(Z_\lambda(t)\big) \to \infty$ as $\lambda \to \infty$ uniformly for $t \in [0,1]\big)$. An example of such a family is $\{\lambda\, Z(t)\}$, $\lambda > 0$, $Z(t)$ continuous, $\phi\big(Z(t)\big) > 0$ for each $t \in [0,1]$ and for each $\phi \in K_u$.

We are now ready to state our main result of this section.

Theorem 3.1 Let B be a Banach space and K a cone in B. Assume $f(t,x,x')$ is continuous and quasimonotone nonincreasing in x with respect to K. Suppose V and W are lower and upper solutions of (1.1) and (1.2) respectively; that is, $V,W \in C^2[I,B]$ and for $t \in I$, $i = 0,1$

$$\begin{aligned} V''(t) &\geq f\big(t,V(t),V'(t)\big), \quad B^i V \leq b_i \\ W''(t) &\leq f\big(t,W(t),W'(t)\big), \quad B^i W \geq b_i. \end{aligned} \tag{3.1}$$

Let $\{Z_\lambda(t)\}$ be an admissible family satisfying a uniformity condition at $\lambda = 0$ and $\lambda = \infty$ such that for each $t \in [0,1]$ and for each $\phi \in K_u$

$$\begin{aligned} \phi\big(Z_\lambda''(t)\big) < \phi\big(f\big(t,W(t) + Z_\lambda(t),W'(t) + Z_\lambda'(t)\big)\big) \\ - \phi\big(f\big(t,W(t),W'(t)\big)\big) \end{aligned} \tag{3.2}$$

and $B^0 Z_\lambda(0) > 0$ $B^1 Z_\lambda(1) > 0$. Then $V(t) \leq W(t)$ on I.

This theorem includes a result of Schroder [12, references therein] who considered the case in which $B = R$. Moreover our result includes the comparison theorem in [6] which was concerned with the case $B = R^n$ $1 \leq n \leq \infty$. In that case K is the set of vectors all of whose components are greater than or equal to zero and K_u are the positive projections.

We now apply Theorem 3.1 to the case in which f is linear, that is, $f(t,x,x') = Q(t)x + P(t)x'$, where $P(t)$ and $Q(t)$ map $B \to B$ linearly. Then the quasimonotonicity of f with respect to K implies that $x \leq y$, $\phi(x) = \phi(y)$ then $\phi\big(Q(t)x\big) \geq \phi\big(Q(t)y\big)$. Thus $Q(t)x$ is quasimonotone nonincreasing in x with respect to K. Moreover we find that $\phi\big(P(t)x'\big) = \phi\big(P(t)y'\big)$ when $\phi(x') = \phi(y')$ (in case $B = R^n$ $1 \leq n \leq \infty$, this implies that $P(t)$ is a diagonal matrix, where K is the cone of vectors whose components are nonnegative). We thus arrive at the following corollary assuming all the hypotheses of Theorem 3.1 in the special case that f is linear.

Corollary 3.1 Suppose that

(i) $W \in C^2[I,B]$ and for $t \in I$

$$W'' \leq P(t)W' + Q(t)W, \quad B^i W \geq 0, \quad i = 0,1,$$

where $Q(t)$ is a quasimonotone nonincreasing linear mapping with respect to K and $P(t)$ has the property that $\phi\big(P(t)x'\big) = \phi\big(P(t)y'\big)$ whenever $\phi(x') = \phi(y')$ for each $\phi \in K_u$.

(ii) let $\{Z_\lambda(t)\}$ be an admissible family satisfying a uniformity condition at $\lambda = 0$, $\lambda = \infty$ such that for each $t \in [0,1]$ and for each $\phi \in K_u$

$$\phi\big((Z_\lambda'')(t)\big) < \phi\big(Q(t)Z_\lambda(t)\big) + \phi\big(P(t)Z_\lambda'(t)\big) \quad \text{and} \quad B^0\big(Z_\lambda(0)\big) > 0,$$
$$B^1\big(Z_\lambda(1)\big) > 0.$$

Then $W(t) \geq 0$.

The proof of this corollary follows from Theorem 3.1 by letting $V \equiv 0$. In the case $B = R$ this corollary corresponds to the generalized maximum principle [10, p. 8].

We observe that the generalized comparison theorem has much flexibility because for a given $f(t,x,x')$ one may have many cones in which f is quasimonotone. Moreover we do not require the cones have an interior, as is often assumed, for the strict inequalities used here are in terms of linear functionals. Thus although $x > 0$ usually implies that x is in the interior of a cone, we only require $\phi(x) > 0$ for each $\phi \in K_u$. This may occur in cones with no interior such as in l^p.

4. Monotone Method

We shall briefly discuss the monotone method here.

The first step is to obtain existence of solutions of (1.1) and (1.2). We do this by using the modified function approach [3] combined with a Nagumo condition, assuming the existence of lower and upper solutions $V(t)$, $W(t)$ respectively with $V(t) \leq W(t)$. The system (1.3) is developed as follows: assume, as before, $f(t,x,y)$ is quasimonotone in x with respect to K. Let $f(t,x,y)$ be continuously differentiable in x and y, and for $t \in I$, $V(t) \leq x \leq W(t)$ and $|y| \leq N$ (obtained using Nagumo condition, see [3]), the Frechet derivative of f with respect to x, $D_x f$, satisfies for each $z \in B$

$$D_x f(t,x,y)[z] \leq Q(z), \tag{4.1}$$

where $Q(z)$ is quasimonotone nonincreasing in z with respect to K. We then define for any element $n(t)$ such that $V(t) \leq n(t) \leq W(t)$

$$\begin{aligned} F(t,x,x') &\equiv F(t,x,x',n) \\ &= F\big(t,n(t),x'\big) + Q\big(x - n(t)\big). \end{aligned}$$

Using the above arguments it can be shown that the boundary value problem (1.2) and

$$x'' = F(t,x,x') \tag{4.2}$$

has a unique solution $x(t)$ such that $V(t) \leq x(t) \leq W(t)$ for each $n(t)$ such that $V(t) \leq n(t) \leq W(t)$.

For each $n \in C(I,B)$ such that $V(t) \leq n(t) \leq W(t)$ on I, define the mapping A by

$$An = x$$

where x is the unique solution of the boundary value problem (4.2), (1.2). Using Theorem 3.1 one can establish that A is a monotone operator on the segment $\langle V,W\rangle = \{u \in E : V(t) \leq u \leq W(t),\ t \in I\}$. If we define the sequences

$$V_n = AV_{n-1}, \quad W_n = AW_{n-1}, \tag{4.3}$$

where $V_0 = V$ and $W_0 = W$ we can state our main result of this section:

Theorem 4.1 Assume $f : I \times B \times B \to B$ is completely continuous and that the hypotheses of Theorem 3.1 is satisfied. Assume the operator Q defined in (4.1) is completely continuous and that

$F(t,x,x')$ satisfies a Nagumo condition. Then the sequences $\{V_n\}$, $\{W_n\}$ defined by (4.3) converge uniformly and monotonically to the minimal and maximal solutions V_{min}, W_{max} respectively of the boundary value problem (1.1), (1.2) on $\langle V,W\rangle$.

If there exists a unique solution of (1.1), (1.2) then the sequences V_n, W_n converge uniformly to the solution, and we thus have a constructive technique to obtain solutions.

One may weaken the assumption of complete continuity by assuming conditions on f in terms of the measure of noncompactness [7].

One application of our work arises in the study of stochastic differential equations. For example, recent work in [4] suggests that an appropriate model for the birefringence of a solution of proteins polymerizing under an electric field is the system (1.1), (1.2). Here the Banach space B would be the space of distribution functions over $(0,\infty)$ since at any given time the length of the polymers form a certain type of distribution which varies with time.

Finally there has been other work concerned with the problem with which we have addressed ourselves. In particular Amann and Crandall [1] have obtained a monotone method for nonlinear semilinear elliptic boundary value problems.

REFERENCES

[1] Amann, H. and Crandall, M., On some existence theorems for semi-linear elliptic equations, (to appear).

[2] Bernfeld, S. and Chandra, J., Minimal and maximal solutions of nonlinear boundary value problems, *Pacific J. of Math*, (to appear).

[3] Bernfeld, S. and Lakshmikantham, V., *An Introduction to Nonlinear Boundary Value Problems*, Academic Press, New York, 1974.

[4] Bernfeld, S. and Judy, M., On the characterization of macromolecular length distributions by analysis of electrical birefringence decay, *Proceedings of an International Symposium on Nonlinear Analysis and Applications*, edited by V. Lakshmikantham, Academic Press, New York, 1977.

[5] Chandra, J. and Davis, P., A monotone method for quasilinear boundary value problems, *Arch. Rat. Mech. Anal.*, 54 (1974), 257-266.

[6] Chandra, J., Lakshmikantham, V., Leela, S., A monotone method for infinite system of nonlinear boundary value problems, *Arch. Rat. Mech. Anal.*, (to appear).

[7] Chandra, J., Lakshmikantham, V., Mitchell, A., Existence of solutions of boundary value problems for nonlinear second order systems in a Banach Space, (to appear).

[8] Keller, H., Elliptic boundary value problems suggested by nonlinear diffusion processes, *Arch. Rat. Mech. Anal.*, 35 (1969), 363-381.

[9] Liskovets, O., The method of lines, *Diff. Eq.*, 1(1965), 1308-1323.

[10] Protter, M. and Weinberger, H., *Maximum Principles in Differential Equations*, Prentice-Hall, Inc., New Jersey, 1967.

[11] Sattinger, D., Monotone methods in nonlinear elliptic and parabolic boundary value problems, Indiana Univ. Math. J., 21(1972), 979-1000.

[12] Schroder, J., Inverse-monotone nonlinear differential operators of the second order, (to appear).

[13] Thompson, R., Convergence and error estimates for the method of lines for certain nonlinear elliptic and elliptic-parabolic equations, *SIAM J. Num. Anal.*, 13(1976), 27-43.

References, continued

[14] Thompson, R., Differential inequalities for infinite second order systems and an application to the method of lines, *J. Diff. Eq.*, 17(1975), 421-434.

[15] Thompson, R., An invariance property of solutions to second order differential inequalities in ordered Banach spaces, *SIAM J. Math. Anal.*, (to appear).

COMPARISON THEOREMS FOR SECOND-ORDER LINEAR DIFFERENTIAL EQUATIONS

Leon Kotin
Communications/Automatic Data Processing Laboratory
U. S. Army Electronics Command, Fort Monmouth, New Jersey 07703

1. Introduction. We consider the equation

$$Lu \equiv u'' + pu' - qu = 0 \tag{1}$$

and solutions $u = u(x)$ which are positive on an interval

$$I = [x_0, x_1), \quad x_1 \leq \infty. \tag{2}$$

Adopting an idea suggested by Garrett Birkhoff, we use Sturmian-type arguments to determine how majorizing $p = p(x)$ or $q = q(x)$ -- i. e., replacing p or q with a larger function -- affects the magnitude of the positive solution. This leads to some interesting corollaries; e. g., if u is a positive solution of $(ru')' - qu = 0$ on I with $0 < rq \leq c^2$ on I and $u' \leq cu/r$ at x_0, then $u \leq u_0 \exp(c\int_{x_0}^{x} dt/r(t))$. These results appear to be new even though the technique is classical.

Throughout, we shall let $u_0 \equiv u(x_0)$, $v_0 \equiv v(x_0)$.

2. Majorizing the coefficient p. Typical is our first result, in which we compare the damping coefficients p in two equations of the form (1).

Lemma 1. Let u be a positive solution of (1) and $v = v(x)$ a positive solution of

$$v'' + p_1 v' - qv = 0 \tag{3}$$

on the interval I.

(i) If $p \leqq p_1$ and $v' \geqq 0$ on I, and $u'v-uv' \geqq 0$ at x_0, then $u/u_0 \geqq v/v_0$ on I.

(ii) If $p \leqq p_1$ and $v' \leqq 0$ on I, and $u'v-uv' \leqq 0$ at x_0, then $u/u_0 \leqq v/v_0$ on I.

(iii) If $p \geqq p_1$ and $v' \geqq 0$ on I, and $u'v-uv' \leqq 0$ at x_0, then $u/u_0 \leqq v/v_0$ on I.

(iv) If $p \geqq p_1$ and $v' \leqq 0$ on I, and $u'v-uv' \geqq 0$ at x_0, then $u/u_0 \geqq v/v_0$ on I.

Proof. Multiply (1) by v and (3) by u and subtract, getting $w' + pw = (p_1-p)uv'$, where $w = u'v-uv'$. Thus $[\exp(\int_{x_0}^{x} p\,dt)w]' = (p_1-p)\exp(\int_{x_0}^{x} p\,dt)uv'$ whence, for case (i), $\exp(\int_{x_0}^{x} p\,dt)w \geqq w(x_0) \geqq 0$. Thus $w = u'v-uv' \geqq 0$. Dividing by $uv > 0$ and integrating complete the proof of the first case. The other cases are proved similarly.

Example 1. Consider the functions $v = \exp(\pm cx^2/2)$, with $0 < c = \text{const.}$, which are linearly independent solutions of $v'' - v'/x - c^2x^2v = 0$. Then when $Lu = 0$ and $x_0 > 0$,

(i) If $p \leqq -1/x$ and $u'(x_0) \geqq cx_0u_0$, then $u \geqq u_0\exp[c(x^2-x_0^2)/2]$.

(ii) If $p \leqq -1/x$ and $u'(x_0) \leqq -cx_0u_0$, then $u \leqq u_0\exp[-c(x^2-x_0^2)/2]$.

(iii) Reverse all inequalities in (i).

(iv) Reverse all inequalities in (ii).

Now assume $0 < q \in C^1(I)$ and consider

$$v'' + (2aq^{\frac{1}{2}} - q'/2q)v' - qv = 0, \tag{4}$$

with solutions

(5) $$v = \exp\left[(-a\pm\sqrt{a^2+1})\int_{x_0}^{x} q^{\frac{1}{2}}dt\right]$$

on I. Then Lemma 1 immediately yields the following result.

Theorem 1. Let u be a positive solution of $Lu = 0$ on I with $0 < q \in C^1(I)$. Then with a = const. and $b_\pm \equiv -a\pm(a^2+1)^{\frac{1}{2}}$, we have

(i) If $p \leqq 2aq^{\frac{1}{2}} - q'/2q$ on I and $u' \geqq b_+q^{\frac{1}{2}}u$ at x_0, then $u \geqq u_0\exp\left[b_+\int_{x_0}^{x} q^{\frac{1}{2}}\,dt\right]$ on I.

(ii) If $p \leqq 2aq^{\frac{1}{2}} - q'/2q$ on I and $u' \leqq b_-q^{\frac{1}{2}}u$ at x_0, then $u \leqq u_0\exp\left[b_-\int_{x_0}^{x} q^{\frac{1}{2}}\,dt\right]$ on I.

(iii) Reverse all inequalities in(i).

(iv) Reverse all inequalities in (ii).

The special cases $a = 0$ (whence $b_\pm = \pm 1$) and $p \equiv 0$ are particularly interesting. Taking the extreme case that both $a = 0$ and $p \equiv 0$, we have the

Corollary. Let u be a positive solution of $u'' - qu = 0$ on I, with $0 < q \in C^1(I)$. Then

(i) If $q' \leqq 0$ on I and $u' \geqq q^{\frac{1}{2}}u$ at x_0, then $u \geqq u_0\exp(\int_{x_0}^{x} q^{\frac{1}{2}}dt)$ on I.

(ii) If $q' \leqq 0$ on I and $u' \leqq -q^{\frac{1}{2}}u$ at x_0, then $u \leqq u_0\exp(-\int_{x_0}^{x} q^{\frac{1}{2}}dt)$ on I.

(iii) Reverse all inequalities in (i).

(iv) Reverse all inequalities in (ii).

3. Majorizing the coefficient q. Now we compare the restoring coefficients q in two equations of the form (1).

Theorem 2. Let u and v be positive solutions of (1) and

(6) $$v'' + pv' - q_1v = 0,$$

respectively, on I. If $q \leqq q_1$ on I and $u'v-uv' \leqq 0$ at x_0, then $u/u_0 \leqq v/v_0$ on I.

Proof. Multiply (1) by v and (6) by u and subtract, getting $w' + pw = (q-q_1)uv$, where $w \equiv u'v-uv'$. Then

$$[\exp(\int_{x_0}^{x} p\ dt)w]' = (q-q_1)uv \exp(\int_{x_0}^{x} p\ dt).$$

We complete the proof by integrating and considering the signs of u, v, $q-q_1$ and $w(x_0)$.

The conclusion of Theorem 2 is clearly valid, verbatim, when u and v are positive solutions of the self-adjoint equations $(ru')' - qu = 0$ and $(rv')' - q_1v = 0$ on I, with $r > 0$. Now consider in particular the self-adjoint equation

(7) $\qquad (rv')' - (c^2/r)v = 0, \quad r = r(x) > 0, \quad c = \text{const.} \neq 0,$

with a solution

(8) $$v = u_0 \exp(c\int_{x_0}^{x} dt/r(t)).$$

Then we conclude

Corollary 1. Let u be a positive solution of $(ru')' - qu = 0$ on I, with $r > 0$ and $q > 0$. (i) If $rq \leqq c^2$ on I and $u' \leqq cu/r$ at x_0, then $u \leqq u_0 \exp(c\int_{x_0}^{x} dt/r(t))$. (ii) Reverse all inequalities in (i).

Another application of Theorem 2 is

Corollary 2. Let u be a positive solution on I of $u'' + 2fu' - qu = 0$, $f \in C^1(I)$. Then for any c = const., we have (i) If $q \leqq -f^2-f'+c^2$ on I and $u' \leqq (c-f)u$ at x_0, then

$u \leqq u_0 \exp\left[c(x-x_0) - \int_{x_0}^{x} f(t)dt\right]$ on I; (ii) Reverse all inequalities in (i).

Proof. Apply Theorem 2 to the equation

$$v'' + 2fv' + (f^2+f'-c^2)v = 0,$$

with a solution $v = \exp(cx - \int_{x_0}^{x} f(t)dt)$.

Example 2. If we now consider the exact equation $v'' + pv' + p'v = 0$ and apply Theorem 2, we find that if u is a positive solution of $Lu = 0$ and if $q \leqq -p'$, then

$$u \leqq \exp(-\int_{x_0}^{x} p\ dt)\left[u_0 + a\int_{x_0}^{x}\exp(\int_{x_0}^{t} p\ ds)dt\right],$$

where $a = (u'+pu)\big|_{x_0}$.

Similarly, by considering the equation $v'' + pv' = 0$, we obtain

Corollary 3. Let u be a positive solution of $Lu = 0$, and suppose that $u_0 + u'(x_0)\int_{x_0}^{x}\exp(-\int_{x_0}^{t} p\ ds)dt > 0$ on I. Then (i) If $q \geqq 0$, then $u \geqq u_0 + u'(x_0)\int_{x_0}^{x}\exp(-\int_{x_0}^{t} p\ ds)dt$; and (ii) Reverse all inequalities in (i).

4. The differential inequality $Lv \geqq 0$. Similar techniques permit us to derive a result reminiscent of Čaplygin's([1, p. 139], [3, p. 15]), which requires that $q \geqq 0$. The following theorem eliminates that condition at the expense of requiring that $u > 0$ and $v > 0$. While this result is known [2, pp. 27-28], we include it here because it is an almost trivial corollary of Theorem 2.

Theorem 3. Let v be a given positive function $\in C^2(I)$ and

u be any positive solution of (1) on I. Then (i) If $Lv \geqq 0$ on I and $u'v - uv' \leqq 0$ at x_0, then $u/u_0 \leqq v/v_0$ on I; and (ii) Reverse all inequalities in (i).

Proof. The given function v satisfies (6) with $q_1 \equiv (v''+pv')/v$. Since $q_1-q = L(v)/v$, the conclusion is an immediate consequence of Theorem 2.

Now by considering convenient comparison functions v and applying Theorem 3, we obtain the following result, which is independent of the damping coefficient p.

Corollary 1. Suppose $Lu = 0$ on I. (i) Let $q \geqq 0$; if $u_0 > 0$ and $u'(x_0) \geqq 0$, then $u \geqq u_0$ on I. (ii) Let $q \leqq 0$; if $u > 0$ on I and $u'(x_0) \leqq 0$, then $u \leqq u_0$ on I.

Proof. Consider the constant function $v \equiv u_0$ and apply Theorem 3 for the cases $q \geqq 0$ and $q \leqq 0$.

Similarly, by considering the function $v = e^{cx}$, we obtain

Corollary 2. If u is a positive solution of (1) on I and $p^2+4q \leqq 0$, then $u \leqq u_0 \exp\left[(u'/u)\Big|_{x_0}(x-x_0)\right]$ on I. On the other hand, the last inequality is reversed if $c^2+pc-q \leqq 0$ on I, where $c \equiv u'(x_0)/u_0$.

REFERENCES

1. E. F. Beckenbach and R. Bellman, "Inequalities," Springer, New York, 1965.
2. R. Bellman, "Methods of Nonlinear Analysis II," Academic Press, New York, 1973.
3. C. A. Swanson, "Comparison and Oscillation Theory of Linear Differential Equations," Academic Press, New York, 1968.

THE COLLAPSED CUBIC ISOPARAMETRIC ELEMENT AS A SINGULAR ELEMENT FOR CRACK PROBLEMS

S. L. Pu, M. A. Hussain and W. E. Lorensen
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. For the 12-node quadrilateral isoparametric elements, it is shown that the inverse square root singularity of the strain field at the crack tip can be obtained by the simple technique of collapsing the quadrilateral elements into triangular elements around the crack tip and placing the two mid-side nodes of each side of the triangles at 1/9 and 4/9 of the length of the side from the tip. This is analogous to placing the mid-side nodes at quarter points in the vicinity of the crack tip for the quadratic isoparametric elements.

The advantages of this method are that the displacement compatibility is satisfied throughout the region and that there is no need of special crack tip elements. The stress intensity factors can be accurately obtained by using general purpose programs having isoparametric elements such as NASTRAN.

1. INTRODUCTION. The direct application of the finite element method to crack problems was studied by a number of investigators [1-3]. No special attention was given to the singular nature of stress and strain at the crack tip. Because of the large strain gradients in the vicinity of a crack tip, it requires the use of an extremely fine element grid near the crack tip. By comparing the finite element result of displacement components or stress components at a nodal point with the corresponding asymptotic result of displacement or stress components at that node, the stress intensity factor could be estimated. The estimated value of stress intensity factor varies over a considerable range, depending on which node is taken for computation. This results in poor estimates if displacements are taken at nodal points either very close to or far away from the crack tip.

An improved finite element technique was developed by Wilson [4]. It combined the asymptotic expansion of displacements in a small circular core region surrounding a crack tip and the finite element approximation outside a polygon approximating the circular arc of the core region. The displacement fields obtained from these two approximations are not, in general, continuous along the asymptotic expansion-finite element interface except at discrete nodal points.

An alternative finite element approach to crack problems is the use of special elements in the region of the crack tip, e.g. [5-7]. In [5] Tracey employs quadrilateral isoparametric elements which become triangular around the crack tip. The displacement functions of the two types of elements are selected such that displacements are continuous everywhere and the near tip displacements are proportional to the square root of the distance from the crack tip.

Henshall and Shaw [8] and Barsoum [9] showed that special crack tip elements were unnecessary. For two-dimensional 8-node quadrilateral elements, the inverse square root singularity of the strain field at the crack tip is obtained by collapsing quadrilateral elements into triangular elements and placing the mid-side nodes at quarter-points from the tip. The quarter-point quadratic isoparametric elements as singular elements for crack problems have been implemented in NASTRAN by Hussain et al [10].

In order to reduce the computer core requirement and to simplify the modeling of a structure, better known but lower order finite elements have been abandoned in favor of cubic 12-node isoparametric quadrilateral elements as described by Zienkiewicz [11]. In this paper, the concept of quarter-point quadratic isoparametric element is extended to 12-node cubic isoparametric elements. The correct order of strain singularity at the crack tip is achieved in a simple manner by collapsing the quadrilateral elements into triangular elements and by placing the two middle nodes of a side at 1/9 and 4/9 of the length of the side from the tip. The 12-node isoparametric elements have been implemented in NASTRAN. Both mode I and mixed mode crack problems are computed by NASTRAN using the collapsed elements to assess the accuracy. The stability of results is discussed when the collapsed triangular elements are used.

2. THE 12-NODE QUADRILATERAL ISOPARAMETRIC ELEMENT. A typical 12-node quadrilateral element in Cartesian coordinates (x,y) which is mapped to a square in the curvilinear space (ξ,η) with vertices at $(\pm 1, \pm 1)$ is shown in Figure 1. The assumption for displacement components takes the form:

$$\left.\begin{aligned} u &= \sum_{i=1}^{12} N_i(\xi,\eta)u_i \\ v &= \sum_{i=1}^{12} N_i(\xi,\eta)v_i \end{aligned}\right\} \qquad (1)$$

where u,v are x,y components of displacement of a point whose natural coordinates are ξ,η; u_i,v_i are displacement components of node i and $N_i(\xi,\eta)$ is the shape function which is given by [11]

$$\begin{aligned} N_i(\xi,\eta) = {} & \frac{1}{256}(1+\xi\xi_i)(1+\eta\eta_i)[-10+9(\xi^2+\eta^2)][-10+9(\xi_i^2+\eta_i^2)] \\ & + \frac{81}{256}(1+\xi\xi_i)(1+9\eta\eta_i)(1-\eta^2)(1-\eta_i^2) \\ & + \frac{81}{256}(1+\eta\eta_i)(1+9\xi\xi_i)(1-\xi^2)(1-\xi_i^2) \end{aligned} \qquad (2)$$

for node i whose Cartesian and curvilinear coordinates are (x_i, y_i) and (ξ_i, η_i) respectively. The details of the shape functions and the numbering sequence are given in Figure 1.

The same shape functions are used for the transformation of coordinates, hence the name isoparametric,

$$\left.\begin{aligned} x &= \sum_{i=1}^{12} N_i(\xi,\eta)x_i \\ y &= \sum_{i=1}^{12} N_i(\xi,\eta)y_i \end{aligned}\right\} \qquad (3)$$

The element stiffness matrix is found in the usual way and is given by [9;10]

$$[K] = \int_{-1}^{1}\int_{-1}^{1} [B]^T[D][B] \det |J| d\xi d\eta \qquad (4)$$

where [B] is a matrix relating joint displacements to strain field

$$[B] = [\ldots B_i \ldots .], \qquad [B_i] = \begin{bmatrix} \frac{\partial N_i}{\partial x} & 0 \\ 0 & \frac{\partial N_i}{\partial y} \\ \frac{\partial N_i}{\partial y} & \frac{\partial N_i}{\partial x} \end{bmatrix} \qquad (5a)$$

and [D] is the material stiffness matrix and is given for the case of plane stress by

$$[D] = \frac{E}{1-\nu^2} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & (1-\nu)/2 \end{bmatrix} \qquad (5b)$$

in which E is Young's modulus and ν is Poisson's ratio.

The Jacobian matrix [J] is given by

$$[J] = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix} = \begin{bmatrix} \cdots & \frac{\partial N_i}{\partial \xi} & \cdots \\ \cdots & \frac{\partial N_i}{\partial \eta} & \cdots \end{bmatrix} \begin{bmatrix} \cdot & \cdot \\ \cdot & \cdot \\ x_i & y_i \\ \cdot & \cdot \\ \cdot & \cdot \end{bmatrix} \tag{6}$$

whenever the determinant of [J] is zero, the stresses and strains become singular [8-10]. The derivatives of shape functions are

$$\frac{\partial N_i}{\partial \xi} = \frac{1}{256}(1 + \eta\eta_i)[-10 + 9(\xi_i^2 + \eta_i^2)](-10\xi_i + 18\xi + 27\xi_i\xi^2 + 9\xi_i\eta^2)$$

$$+ \frac{81}{256}\xi_i(1 + 9\eta\eta_i)(1 - \eta^2)(1 - \eta_i^2)$$

$$+ \frac{81}{256}(1 + \eta\eta_i)(1 - \xi_i^2)(9\xi_i - 2\xi - 27\xi_i\xi^2) \tag{7a}$$

$$\frac{\partial N_i}{\partial \eta} = \frac{1}{256}(1 + \xi\xi_i)[-10 + 9(\xi^2 + \eta^2)](-10\eta_i + 18\eta + 27\eta_i\eta^2 + 9\eta_i\xi^2)$$

$$+ \frac{81}{256}\eta_i(1 + 9\xi\xi_i)(1 - \xi^2)(1 - \xi_i^2)$$

$$+ \frac{81}{256}(1 + \xi\xi_i)(1 - \eta_i^2)(9\eta_i - 2\eta - 27\eta_i\eta^2) \tag{7b}$$

3. THE CRACK TIP ELEMENT. In an 8-node quadratic isoparametric element, Henshell and Shaw [8] and Barsoum [9] found independently that the strain became singular at the corner node if the mid-side nodes were placed at the quarter points of the sides from the corner node. This singularity is achieved in a similar way for a 12-node isoparametric element by placing the two middle nodes at the 1/9 and 4/9 of the length of the sides from the common node of two sides.

For simplicity, let us consider the singularity along the side $\eta = -1$ of Figure 1. In general the cubic mapping functions are

$$x = a_0 + a_1\xi + a_2\xi^2 + a_3\xi^3 \tag{8}$$

$$u = b_0 + b_1\xi + b_2\xi^2 + b_3\xi^3 \tag{9}$$

For $\xi = -1, -1/3, 1/3$ and 1, the corresponding values of x and u are

$$x = 0, \alpha\ell, \beta\ell, \ell$$

$$u = u_1, u_2, u_3, u_4$$

The constants a's and b's in terms of these values of x and u are

$$\left.\begin{aligned} a_0 &= \frac{\ell}{16}(-1 + 9\alpha + 9\beta), \quad a_1 = \frac{\ell}{16}(-1 - 27\alpha + 27\beta) \\ a_2 &= \frac{9\ell}{16}(1 - \alpha - \beta), \quad a_3 = \frac{9\ell}{16}(1 + 3\alpha - 3\beta) \end{aligned}\right\} \tag{10}$$

$$\left.\begin{aligned} b_0 &= \frac{1}{16}(-u_1 + 9u_2 + 9u_3 - u_4), \quad b_1 = \frac{1}{16}(u_1 - 27u_2 + 27u_3 - u_4) \\ b_2 &= \frac{9}{16}(u_1 - u_2 - u_3 + u_4), \quad b_3 = \frac{9}{16}(-u_1 + 3u_2 - 3u_3 + u_4) \end{aligned}\right\} \tag{11}$$

To have singular strain at $x = 0$ ($\xi = -1$), the reduced Jacobian, $\frac{dx}{d\xi}$, must vanish at $\xi = -1$. From (8) we have

$$\frac{dx}{d\xi} = a_1 + 2a_2\xi + 3a_3\xi^2 \tag{12}$$

For $\xi = -1$, $\frac{dx}{d\xi} = 0$ leads to the equation

$$\beta = 2\alpha + \frac{2}{9} \tag{13}$$

In order to have the inverse square root singularity for $\frac{du}{dx}$,

$$\frac{du}{dx} = \frac{du}{d\xi}\frac{d\xi}{dx} = (b_1 + 2b_2\xi + 3b_3\xi^2)/\frac{dx}{d\xi}$$

x must be a quadratic function of ξ so that the inverse gives ξ as a function of $x^{1/2}$. This leads to $a_3 = 0$ or

$$1 + 3\alpha - 3\beta = 0 \tag{14}$$

The solution of (13) and (14) gives

$$\alpha = 1/9 \quad \text{and} \quad \beta = 4/9 \tag{15}$$

Equations (8) and (9) become

$$x = \frac{\ell}{4}(1+\xi)^2 \quad \text{or} \quad \xi = -1 + 2\sqrt{\frac{x}{\ell}} \tag{16}$$

$$u = u_1 + \frac{1}{2}(-11u_1 + 18u_2 - 9u_3 + 2u_4)\sqrt{\frac{x}{\ell}} + \frac{9}{2}(2u_1 - 5u_2 + 4u_3 - u_4)\frac{x}{\ell}$$

$$+ \frac{9}{2}(-u_1 + 3u_2 - 3u_3 + u_4)\left(\frac{x}{\ell}\right)^{3/2} \tag{17}$$

From (17) it is clear $\frac{du}{dx}$ has singularity of the order $\frac{1}{\sqrt{x}}$ at $x = 0$.

The inverse square root singularity at $x = 0$ along any other ray emanating from node 1 can be achieved by degenerating the quadrilateral element into a triangular element with the side 10, 11, 12, 1 collapsed to a point at the crack tip and placing grid points 2, 9 at $\ell/9$ and 3, 8 at $4\ell/9$ from the tip, Figure 2, where ℓ is the length of the sides corresponding to $\eta = \pm 1$. For simplicity and without loss of generality we take $\beta = 0$ and the Cartesian coordinates of nodal points as follows:

Node	1	2	3	4	5	6	7	8	9
x/ℓ	0	$\frac{1}{9}$	$\frac{4}{9}$	1	$\frac{2+\cos\alpha}{3}$	$\frac{1+\cos\alpha}{3}$	$\cos\alpha$	$\frac{4\cos\alpha}{9}$	$\frac{\cos\alpha}{9}$
y/ℓ	0	0	0	0	$\frac{\sin\alpha}{3}$	$\frac{2\sin\alpha}{3}$	$\sin\alpha$	$\frac{4\sin\alpha}{9}$	$\frac{\sin\alpha}{9}$

Using (3),

$$\begin{aligned} x/\ell &= N_A + N_B \cos\alpha \\ y/\ell &= N_B \sin\alpha \end{aligned} \tag{18}$$

where

$$\begin{aligned} N_A &= \frac{1}{9}N_2 + \frac{4}{9}N_3 + N_4 + \frac{2}{3}N_5 + \frac{1}{3}N_6 = \frac{1}{8}(1+\xi)^2(1-\eta) \\ N_B &= \frac{1}{3}N_5 + \frac{2}{3}N_6 + N_7 + \frac{4}{9}N_8 + \frac{1}{9}N_9 = \frac{1}{8}(1+\xi)^2(1+\eta) \end{aligned} \tag{19}$$

The determinant of Jacobian is

$$|J| = \begin{vmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{vmatrix} = \ell^2 \left(\frac{\partial N_A}{\partial \xi}\frac{\partial N_B}{\partial \eta} - \frac{\partial N_A}{\partial \eta}\frac{\partial N_B}{\partial \xi}\right)\sin\alpha = \frac{\ell^2}{16}(1+\xi)^3\sin\alpha \qquad (20)$$

This shows the strain is singular at $x = 0$ ($\xi = -1$) along any ray from $x = 0$ since $|J| = 0$ at $\xi = -1$ for all η. From (19) and (18) and using polar coordinates $x/\ell = \rho\cos\Theta$, $y/\ell = \rho\sin\Theta$, where $\rho = r/\ell$, we obtain

$$\left.\begin{aligned} \xi &= -1 + 2\rho^{\frac{1}{2}}\left[\cos\left(\Theta - \tfrac{\alpha}{2}\right)/\cos\left(\tfrac{\alpha}{2}\right)\right]^{1/2} \\ \eta &= \tan\left(\Theta - \tfrac{\alpha}{2}\right)/\tan\left(\tfrac{\alpha}{2}\right) \end{aligned}\right\} \qquad (21)$$

For a more general collapsed triangular element in Figure 2 when $\beta \neq 0$, these two equations take the forms

$$\left.\begin{aligned} \xi &= -1 + 2R^{\frac{1}{2}}, \quad R^{\frac{1}{2}} = \rho^{\frac{1}{2}}\left[\cos\left(\Theta - \tfrac{\alpha+\beta}{2}\right)/\cos\left(\tfrac{\alpha-\beta}{2}\right)\right]^{\frac{1}{2}} \\ \eta &= \tan\left(\Theta - \tfrac{\alpha+\beta}{2}\right)/\tan\left(\tfrac{\alpha-\beta}{2}\right) \end{aligned}\right\} \qquad (22)$$

The displacements components u,v at a point (ξ,η) of the triangular element of Figure 2 are

$$\left.\begin{aligned} u &= \sum_{i=1}^{12} N_i(\xi,\eta)u_i = A_0(\eta,u_i) + A_1(\eta,u_i)(1+\xi) \\ &\quad + A_2(\eta,u_i)(1+\xi)^2 + A_3(\eta,u_i)(1+\xi)^3 \\ v &= \sum_{i=1}^{12} N_i(\xi,\eta)v_i = A_0(\eta,v_i) + A_1(\eta,v_i)(1+\xi) \\ &\quad + A_2(\eta,v_i)(1+\xi)^2 + A_3(\eta,v_i)(1+\xi)^3 \end{aligned}\right\} \qquad (23)$$

where

$$A_0(\eta,u_i) = \{2(-1+9\eta^2)[(1-\eta)u_1 + (1+\eta)u_{10}] + 18(1-\eta^2)[(1+3\eta)u_{11} + (1-3\eta)u_{12}]\}/32$$

The displacement derivatives are

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial x} = \frac{1}{(1+\xi)^2}\left(-\frac{4}{\ell}\right)(1+\eta)\frac{\partial A_0(\eta,u_i)}{\partial \eta}$$

$$+ \frac{1}{(1+\xi)}\frac{2}{\ell}\left[A_1 - 2(1+\eta)\frac{\partial A_1}{\partial \eta}\right] + \ldots$$

$$\frac{\partial u}{\partial y} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial y} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial y} = \frac{1}{(1+\xi)^2}\frac{4[(1+\cos\alpha) - \eta(1-\cos\alpha)]}{\ell\sin\alpha}$$

$$\frac{\partial A_0(\eta,u_i)}{\partial \eta} + \frac{1}{(1+\xi)} \ldots . + \ldots .$$

where

$$\frac{\partial A_0(\eta,u_i)}{\partial \eta} = \{(2 + 36\eta - 54\eta^2)u_1 - (2 - 36\eta - 54\eta^2)u_{10}$$

$$+ 18(3 - 2\eta - 9\eta^2)u_{11} - 18(3 + 2\eta - 9\eta^2)u_{12}\}/32$$

Similar expressions for $\partial v/\partial x$ and $\partial v/\partial y$ with u_i replaced by v_i.

It can be seen that both $\partial A_0(\eta,u_i)/\partial\eta$ and $\partial A_0(\eta,v_i)/\partial\eta$ vanish for all η if

$$u_1 = u_{10} = u_{11} = u_{12} \quad \text{and} \quad v_1 = v_{10} = v_{11} = v_{12} \tag{24}$$

Hence for the strain field to have the inverse square root of singularity at $r = 0$, the nodes 1, 10, 11, 12 which are collapsed into one point must be tied together. This is analogous to the constraints given in [12] for quadratic isoparametric elements. Using multiple constraint conditions, equations (24), the displacement components at (ξ,η) relative to the tip may be written in the form

$$u = \frac{1}{16}R^{\frac{1}{2}}\left[36F_1(\eta,u_i) + F_3(\eta,u_i) + 36\{F_2(\eta,u_i) - F_1(\eta,u_i)\}\right.$$

$$\left. R^{\frac{1}{2}} - 36F_2(\eta,u_i)R\right] \tag{25}$$

$$v = \frac{1}{16}R^{\frac{1}{2}}\left[36F_1(\eta,v_i) + F_3(\eta,v_i) + 36\{F_2(\eta,v_i) - F_1(\eta,v_i)\}\right.$$

$$\left. R^{\frac{1}{2}} - 36F_2(\eta,v_i)R\right] \tag{26}$$

where

$$F_1(\eta,u_i) = (1-\eta)(2u_2 - u_3) + (1+\eta)(2u_9 - u_8)$$

$$F_2(\eta,u_i) = (1-\eta)(-3u_2 + 3u_3 - u_4) + (1+\eta)(-3u_9 + 3u_8 - u_7)$$

$$F_3(\eta,u_i) = 9(1-\eta^2)[(1-3\eta)u_5 + (1+3\eta)u_6]$$

$$- (1-9\eta^2)[(1-\eta)u_4 + (1+\eta)u_7] \qquad (27)$$

and $F_1(\eta,v_i)$ etc. are obtained by replacing u_i by v_i.

4. DETERMINATION OF STRESS INTENSITY FACTORS. The collapsed triangular elements around the crack tip have the correct order of singularity at the tip. The continuity of displacement components is insured throughout the region and because of the use of higher order polynomial for the displacement field, the nodal displacements obtained from the finite element method should be quite accurate. If nodal displacements are substituted into the left hand side of the well known near crack tip displacement formula [13].

$$2Gu(\Theta) = \sum_{n=1,2,\ldots} \{(-1)^{n-1} r^{n-1/2}[d_{2n-1} D_{u1}(n,\Theta) + a_{2n-1} A_{u1}(n,\Theta)]$$

$$+ (-1)^n r^n [d_{2n} D_{u2}(n,\Theta) + a_{2n} A_{u2}(n,\Theta)]\} \qquad (28)$$

$$2Gv(\Theta) = \sum_{n=1,2,\ldots} \{(-1)^{n-1} r^{n-1/2}[d_{2n-1} D_{v1}(n,\Theta) + a_{2n-1} A_{v1}(n,\Theta)]$$

$$+ (-1)^n r^n [d_{2n} D_{v2}(n,\Theta) + a_{2n} A_{v2}(n,\Theta)]\} \qquad (29)$$

the coefficients d's and a's can be approximately determined by a finite number of terms on the right hand sides of (28) and (29). The stress intensity factors K_1 and K_2 are related to d_1 and a_1 by

$$K_1 = -d_1\sqrt{2\pi} \quad , \quad K_2 = -a_1\sqrt{2\pi} \qquad (30)$$

In (28) and (29),

$$
\left.
\begin{aligned}
D_{u1}(n,\theta) &= (n - 1/2)\cos(n - \tfrac{5}{2})\theta - (\kappa + n - \tfrac{3}{2})\cos(n - \tfrac{1}{2})\theta \\
D_{u2}(n,\theta) &= n\cos(n - 2)\theta - (\kappa + n + 1)\cos n\theta \\
A_{u1}(n,\theta) &= (n - 1/2)\sin(n - \tfrac{5}{2})\theta - (\kappa + n + 1/2)\sin(n - 1/2)\theta \\
A_{u2}(n,\theta) &= n\sin(n - 2)\theta - (\kappa + n - 1)\sin\theta \\
D_{v1}(n,\theta) &= -(n - 1/2)\sin(n - \tfrac{5}{2})\theta - (\kappa - n + \tfrac{3}{2})\sin(n - \tfrac{1}{2})\theta \\
D_{v2}(n,\theta) &= -n\sin(n - 2)\theta - (\kappa - n - 1)\sin n\theta \\
A_{v1}(n,\theta) &= (n - \tfrac{1}{2})\cos(n - \tfrac{5}{2})\theta + (\kappa - n - 1/2)\cos(n - 1/2)\theta \\
A_{v2}(n,\theta) &= n\cos(n - 2)\theta + (\kappa - n + 1)\cos n\theta
\end{aligned}
\right\} \qquad (31)
$$

where

$$
\kappa = \begin{cases} (3 - \nu)/(1 + \nu) & \text{for plane stress} \\ 3 - 4\nu & \text{for plane strain} \end{cases} \qquad (32)
$$

There are a number of ways to estimate the stress intensity factors from the finite element displacements near a crack tip. On the right hand side of (28) and (29), we may retain only $r^{1/2}$ term or $r^{1/2}$, r,... up to r^p terms, and on the left hand side we may use u and v actually obtained from (25) and (26) or only the part of u and v which correspond to $r^{1/2}$ term in (25), (26) [14,15]. Detailed discussions are given in [16]. The simple, yet accurate way to obtain the stress intensity factors is the use of $v(r_o,\pi)$ and $v(r_o, -\pi)$ for K_1 and $u(r_o,\pi)$ and $u(r_o, -\pi)$ for K_2 [16].

$$
K_1(\pi) = \frac{\sqrt{2\pi}\ 2G\ v(r_o,\pi)}{r_o^{1/2}(\kappa + 1)}, \quad K_1(-\pi) = -\frac{\sqrt{2\pi}\ 2G\ v(r_o,-\pi)}{r_o^{1/2}(\kappa + 1)},
$$

$$
K_1 = \frac{K_1(\pi) + K_1(-\pi)}{2} \qquad (33)
$$

$$K_2(\pi) = \frac{\sqrt{2\pi}\ 2G\ u(r_o,\pi)}{r_o^{1/2}(\kappa + 1)} \quad , \quad K_2(-\pi) = \frac{-\sqrt{2\pi}\ 2G\ u(r_o,-\pi)}{r^{1/2}(\kappa + 1)} \quad ,$$

$$K_2 = \frac{K_2(\pi) + K_2(-\pi)}{2} \tag{34}$$

where $u(r_o,\pi)$, $v(r_o,\pi)$ are rectangular components of displacement of the node at $r = r_o$, $\theta = \pi$ relative to the node at the crack tip referring to the local coordinates with crack tip as origin and the crack on the negative x-axis. This technique gives good results if 1% of the crack length is used for r_o and $\ell = 9r_o$. For a mode I crack, K_1 is given by either $K_1(\pi)$ or $K_1(-\pi)$, K_2 is zero. For a mode II crack, K_1 is zero and K_2 is given by either $K_2(\pi)$ or $K_2(-\pi)$.

5. NASTRAN IMPLEMENTATION. The NASTRAN implementation of the 12-node quadrilateral follows that of the 8-node quadrilateral as described in [10]. The dummy user element facility of NASTRAN is used. This requires coding routines to calculate element stiffness matrices and stress recovery computations. Modifications to existing NASTRAN source code are made to provide proper output formats for the element. Stress intensity factors for mode I and II are calculated using equations (33) and (34). *All stiffness computations are performed in double precision while stress recovery is performed in single precision. Element stiffness matrix computation requires 10 seconds/element on an IBM 360/44.

6. NUMERICAL RESULTS. Three mode I and one mixed mode crack problems are chosen for numerical computation of stress intensity factors. The geometries and loads of mode I tension test specimens are given in Fig. 3. The idealization of a half of the single edge crack is shown in Fig. 4. A similar idealization is used for a quadrant of a center crack or a double edge crack. Three collapsed triangular elements surrounding a mode I crack tip are shown in Fig. 5. Nodes 1 through 10 are numbered counterclockwise similar to nodes 19 through 28 but they are coincide with the crack tip. At the crack tip, the multiple constraint conditions given by equations (24) are either applied or not applied. The multiple constraint has little effect on numerical results of stress intensity factors for the test problems studied here (see table 1). This is probably because the differences among u_1, u_2,...u_{10} and among v_1, v_2...,v_{10} are very small in the elastic range when the nodes 1, 2,...,10 are not tied together.

For each specimen a reference value of K_1 is used for normalization. For a central crack with a/b = 0.4, K_1 = 1.966 [17] is taken as exact. The exact value for the double edge crack is K_1 = 2.00 [18], and for the single edge crack is K_1 = 3.728 [19, using 1 st formula for F(a/b) on page 2.11]. Table 1 gives ratios of K_1 obtained from cubic isoparametric elements to the corresponding exact value of K_1 for various values of r_o/a where r_o is the distance between the crack tip and the nearest node and a is the crack length.

*Three-point Gaussian quadrature is normally used to calculate each partial integration of the double integral (4). As an option, four-point Gaussian quadrature may be used instead.

TABLE 1. K_1 (NASTRAN)/K_1 (EXACT)

r_o/a	0.01		0.015		0.02	
Multiple Constraint	No	Yes	No	Yes	No	Yes
Center Crack a/b = 0.4, H/b = 4.0 Exact K_1 = 1.966	0.981	1.013	0.982	0.999	0.983	0.994
Double Edge Crack a/b = 0.4, H/b = 4.0 Exact K_1 = 2.00	1.000	1.021	0.998	1.007	0.999	1.002
Single Edge Crack a/b = 0.4, H/b = 4.0 Exact K_1 = 3.728	0.980	1.003	0.980	0.991	0.982	0.988

An obliqued edge crack in a rectangular panel under uniform tension is solved by Freese using modified mapping collocation method [20]. The NASTRAN program is used to solve the combined mode I and mode II crack problem using six collapsed triangular elements around the crack tip as shown in Figure 6. For a 45° edge crack with a/b = 0.4, H/b = 2.0, K_1 and K_2, from readings of Bowie's graphs (Figure 1 - 16(a) and (b) of [20]), are approximately 1.86 and 0.88. The idealization and boundary conditions of the slant edge cracked panel are shown in Figure 7. Numerical results of K_1 and K_2 are tabulated in Table 2 for r_o/a = 0.01 and for various other conditions. Again the multiple constraint conditions, namely $u_1 = u_2 = \ldots = u_{19}$ and $v_1 = v_2 = \ldots = v_{19}$, give little effect on values of K_1 and K_2. Numerical results using cubic isoparametric elements for the test problems can also be found in [16,21].

TABLE 2. K_1 AND K_2 FOR 45° EDGE CRACK BY NASTRAN

B.C.	INTEGRATION	MULTIPLE CONSTRAINT	K_1	K_2
1	3 x 3	No	1.89	0.95
1	3 x 3	Yes	1.89	0.96
1	4 x 4	No	1.83	0.92
2	4 x 4	Yes	1.84	0.93

7. THE STABILITY OF COLLAPSED TRIANGULAR ELEMENTS. In a recent report by Hussain and Lorensen [22], it was found that a slight perturbation in placing the mid-side node opposite to the crack tip for a collapsed 8-node quadrilateral element led to unstable results in stress intensity factors. This unstability can be shown in the collapsed 12-node quadrilateral element if one or both middle nodes of the side opposite to the crack tip been slightly perturbed from their nominal positions.

Let node 5 be perturbed as shown in Figure 8. Denoting the perturbed quantities with an asterick we have

$$x_5^*/\ell = \frac{2 + \cos\alpha}{3} + \varepsilon$$
$$y_5^*/\ell = \frac{\sin\alpha}{3} + \varepsilon' \tag{38}$$

A general point (x,y) given by equation (18) will be displaced at

$$x^*/\ell = \frac{1}{8}(1+\xi)^2[(1-\eta) + (1+\eta)\cos\alpha] + \varepsilon\frac{9}{32}(1+\xi)(1-\eta^2)(1-3\eta) \tag{39}$$

$$y^*/\ell = \frac{1}{8}(1+\xi)^2(1+\eta)\sin\alpha + \varepsilon'\frac{9}{32}(1+\xi)(1-\eta^2)(1-3\eta) \tag{40}$$

Along the line $\eta = -1/3$, and replace y^* by $r\sin\Theta$ in (40) we have

$$1 + \xi = \frac{3\varepsilon'}{\sin\alpha}\left[-1 + \sqrt{1 + \frac{4\sin\Theta\sin\alpha}{3\varepsilon'^2}\cdot\frac{r}{\ell}}\right] \tag{41}$$

Since $(1 + \xi)$ is a common factor in displacement components, equations (25) and (26), it is seen that the singularity required, for the crack problems disappears along at least the ray $\eta = -1/3$ in the collapsed triangular case.

As a numerical example, the central crack tension specimen of Figure 3 (a) is again used. If the idealization remained the same as shown in Figure 4 except that the collapsed elements of Figure 5 were replaced by those of Figure 8b (where nodal points 20, 21, 23, 24, 26 and 27 are on a circular arc), the computed stress intensity factor changed from its almost exact value $K_1 = 1.962$ to $K_1 = 1.421$ (nearly 30% error). If only nodal points 26 and 27 were perturbed to their new locations of Figure 8b, the stress intensity factor would become $K_1 = 1.457$ (a 26% error).

8. CONCLUSIONS. By a simple manner, the 12-node isoparametric elements can be used to form a singular element for two-dimensional elastic fracture mechanics analysis. The elements have been successfully implemented in NASTRAN which can now be more efficiently used for more accurate prediction of stress intensity factors of complicated crack problems. The middle nodes of the side opposite to a crack tip in a collapsed triangular element should be accurately located to avoid unstable results. The extension of collapsed triangular elements as singular elements to three-dimensional brick elements can be easily done as in [9,10].

REFERENCES

1. Swedlow, J. L., Williams, M. L., and Yang, W. H., "Elasto-Plastic Stresses and Strains in Cracked Plates," Proceedings First International Conference on Fracture, 1, p. 259, 1966.

2. Kobayashi, A. S., Maiden, D. E. and Simon, B. J., "Application of the Method of Finite Element Analysis to Two-Dimensional Problems in Fracture Mechanics," ASME 69-WA/PVP-12 (1969).

3. Chan, S. K., Tuba, I. S. and Wilson, W. K., "On Finite Element Method in Linear Fracture Mechanics," Engineering Fracture Mechanics, 2, p. 1, 1970.

4. Wilson, W. K., "Combined Mode Fracture Mechanics," Ph.D. Dissertation, University of Pittsburgh, 1969.

5. Tracey, D. M., "Finite Elements for Determination of Crack Tip Elastic Stress Intensity Factors," Engineering Fracture Mechanics, Vol. 3, 1971.

6. Blackburn, W. S., "Calculation of Stress Intensity Factors at Crack Tips Using Special Finite Elements," The Mathematics of Finite Elements and Applications, Brunel University, 1973.

7. Benzley, S. E. and Beisinger, A. E., "Chiles - A Finite Element Computer Program That Calculates the Intensities of Linear Elastic Singularities," Sandia Laboratories, Technical Report SLA-73-0894, 1973.

8. Henshell, R. D., and Shaw, K. G., "Crack Tip Finite Elements Are Unnecessary," International Journal for Numerical Methods in Engineering, Vol. 9, 1975.

9. Barsoum, R. S., "On the Use of Isoparametric Finite Elements in Linear Fracture Mechanics," International Journal for Numerical Methods in Engineering, Vol. 10, 1976.

10. Hussain, M. A., Lorensen, W. E., and Pflegl, G., "The Quarter-Point Quadratic Isoparametric Element As a Singular Element for Crack Problems," NASA TM-X-3428, 1976, p. 419.

11. Zienkiewicz, O. O., The Finite Element Method in Engineering Science, McGraw Hill, London, 1971.

12. Barsoum, R. S., "Triangular Quarter-Point Elements As Elastic and Perfectly-Plastic Crack Tip Elements," International Journal for Numerical Methods in Engineering, Vol. 11, 1977.

13. Williams, M. L., "On the Stress Distribution at the Base of a Stationary Crack," Journal of Applied Mechanics, Vol. 24, 1957.

14. Tracey, D. M., "Discussion of 'On the Use of Isoparametric Finite Elements In Linear Fracture Mechanics' by R. S. Barsoum", Int. Journal for Numerical Methods in Engineering, Vol. 11, 1977.

15. Barsoum, R. S., "Author's Reply to the Discussion by Tracey," Int. Journal for Numerical Methods in Engineering, Vol. 11, 1977.

16. Pu, S. L., Hussain, M. A., and Lorensen, W. E., "The Collapsed Cubic Isoparametric Element as a Singular Element for Crack Problems," Watervliet Arsenal Technical Report in preparation.

17. Isida, M., "Analysis of Stress Intensity Factors for the Tension of a Centrally Cracked Strip with Stiffened Edges," Engineering Fracture Mechanics, Vol. 5,1973.

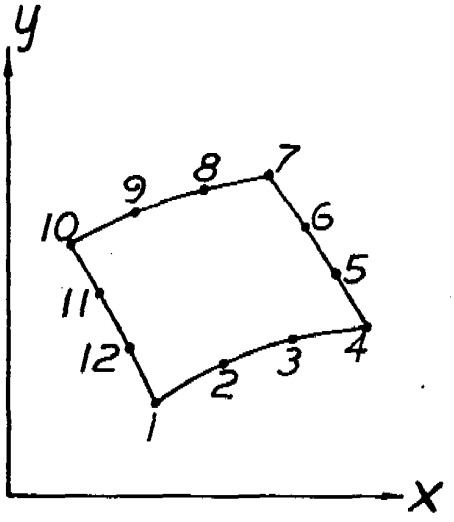
18. Brown, W. F., and Srawley, J. E., "Plane Strain Crack Toughness Testing of High Strength Metallic Materials," ASTM STP-410, 1966.

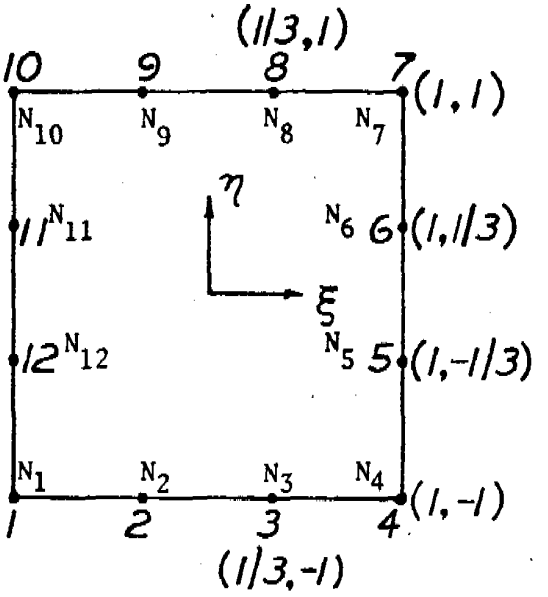
19. Tada, H., Paris, P. and Irwin, G., The Stress Analysis of Cracks Handbook, Del Research Corp., 1973.

20. Bowie, O. L., "Solutions of Plane Crack Problems by Mapping Technique," in Mechanics of Fracture, 1, Edited by G. C. Sih, Noordhoff International Publishing, Leyden, The Netherlands, 1973.

21. Gifford, L. N., "Apes - Second Generation Two-Dimensional Fracture Mechanics and Stress Analysis by Finite Elements," Naval Ship Research and Development Center, Report 4799, December 1975.

22. Hussain, M. A. and Lorensen, W. E.,"Isoparametric Elements As Singular Elements for Crack Problems," Watervliet Technical Report under preparation.





$$N_1 = \frac{1}{32}(1-\eta)(1-\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_2 = \frac{9}{32}(1-\eta)(1-\xi^2)(1-3\xi)$$

$$N_3 = \frac{9}{32}(1-\eta)(1-\xi^2)(1+3\xi)$$

$$N_4 = \frac{1}{32}(1-\eta)(1+\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_5 = \frac{9}{32}(1+\xi)(1-\eta^2)(1-3\eta)$$

$$N_6 = \frac{9}{32}(1+\xi)(1-\eta^2)(1+3\eta)$$

$$N_7 = \frac{1}{32}(1+\eta)(1+\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_8 = \frac{9}{32}(1+\eta)(1-\xi^2)(1+3\xi)$$

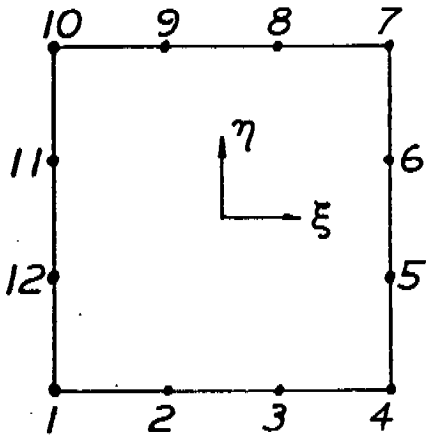
$$N_9 = \frac{9}{32}(1+\eta)(1-\xi^2)(1-3\xi)$$

$$N_{10} = \frac{1}{32}(1+\eta)(1-\xi)[-10+9(\xi^2+\eta^2)]$$

$$N_{11} = \frac{9}{32}(1-\xi)(1-\eta^2)(1+3\eta)$$

$$N_{12} = \frac{9}{32}(1-\xi)(1-\eta^2)(1-3\eta)$$

Figure 1. Shape Functions and Numbering Sequence For a 12-Node Quadrilateral Element.



NODE	x/ℓ	y/ℓ
1	0	0
2	cosβ/9	sinβ/9
3	4cosβ/9	4sinβ/9
4	cosβ	sinβ
5	(2cosβ+cosα)/3	(2sinβ+sinα)/3
6	(cosβ+2cosα)/3	(sinβ+2sinα)/3
7	cosα	sinα
8	4cosα/9	4sinα/9
9	cosα/9	sinα/9
10	0	0
11	0	0
12	0	0

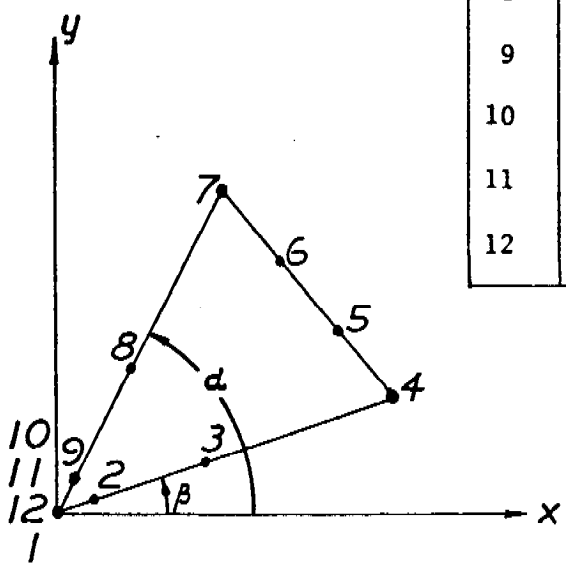
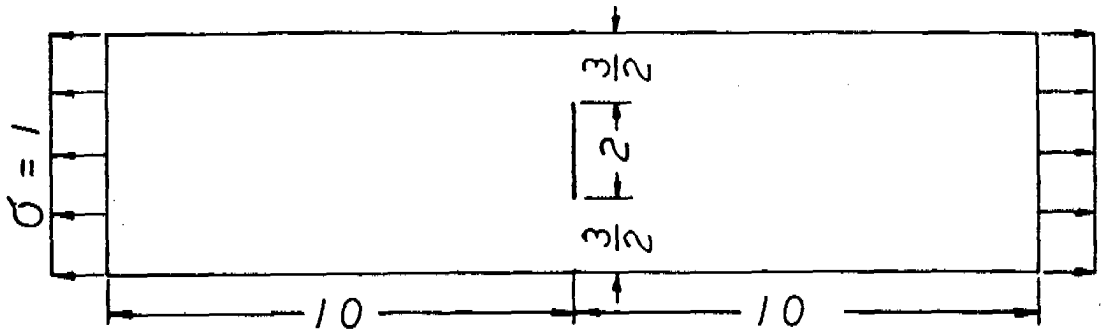
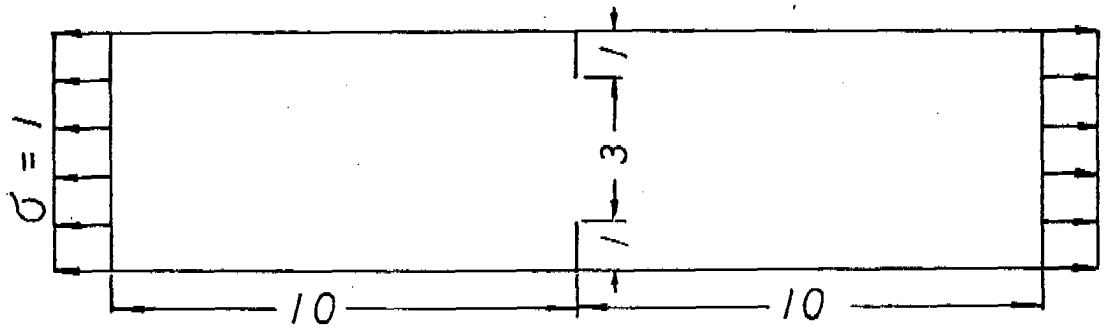


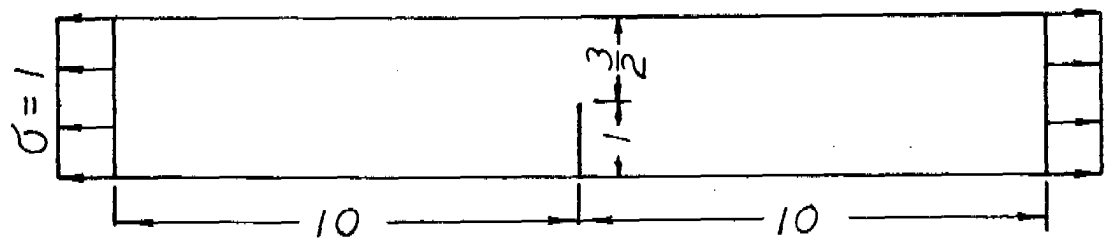
Figure 2. A Normalized Square in (ξ,η) Plane Mapped Into a Collapsed Triangular Element in (x,y) Plane with the side ξ = -1 Degenerated into a Point at the Crack Tip.



(a) Center Crack



(b) Double-Edge Crack



(c) Single-Edge Crack

Figure 3. Three Tension Test Specimens.

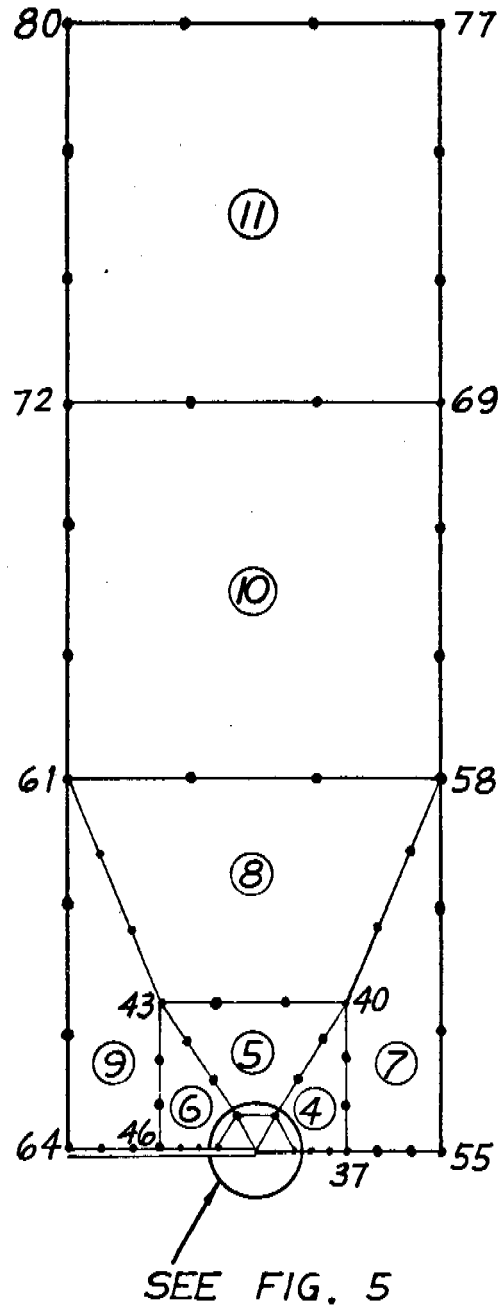


Figure 4. Idealization of a Half of the Single-Edge Cracked Tension Specimen.

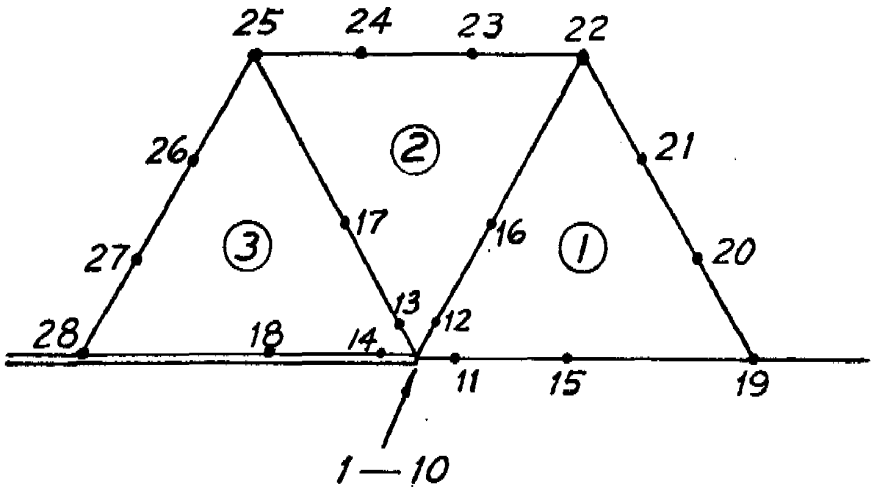


Figure 5. Three Collapsed Triangular Elements Surrounding a Mode I Crack Tip.

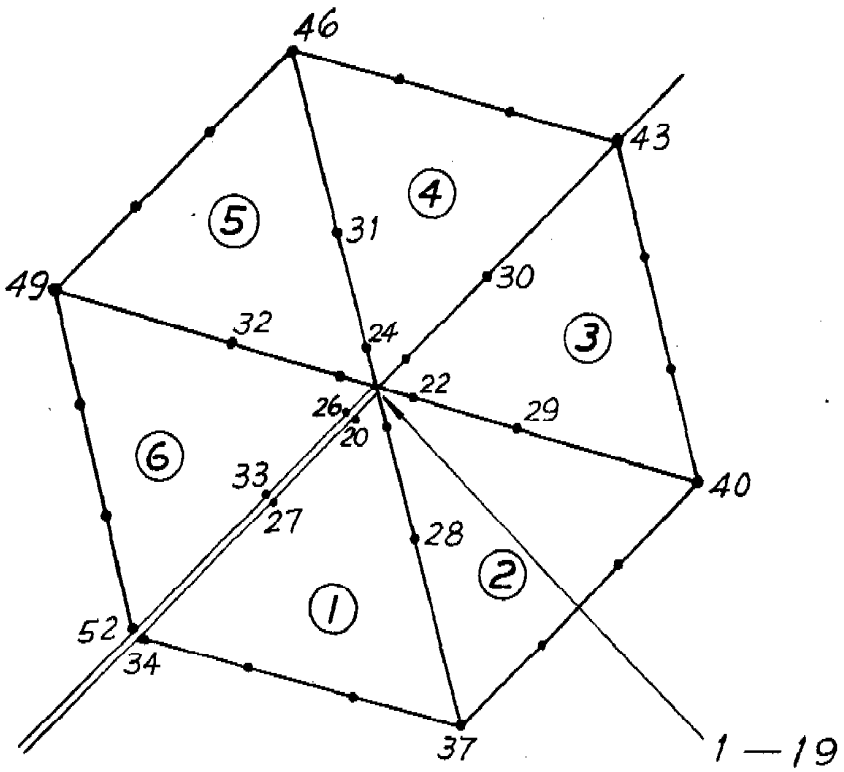


Figure 6. Six Collapsed Triangular Elements Surrounding a Mixed Mode Crack Tip.

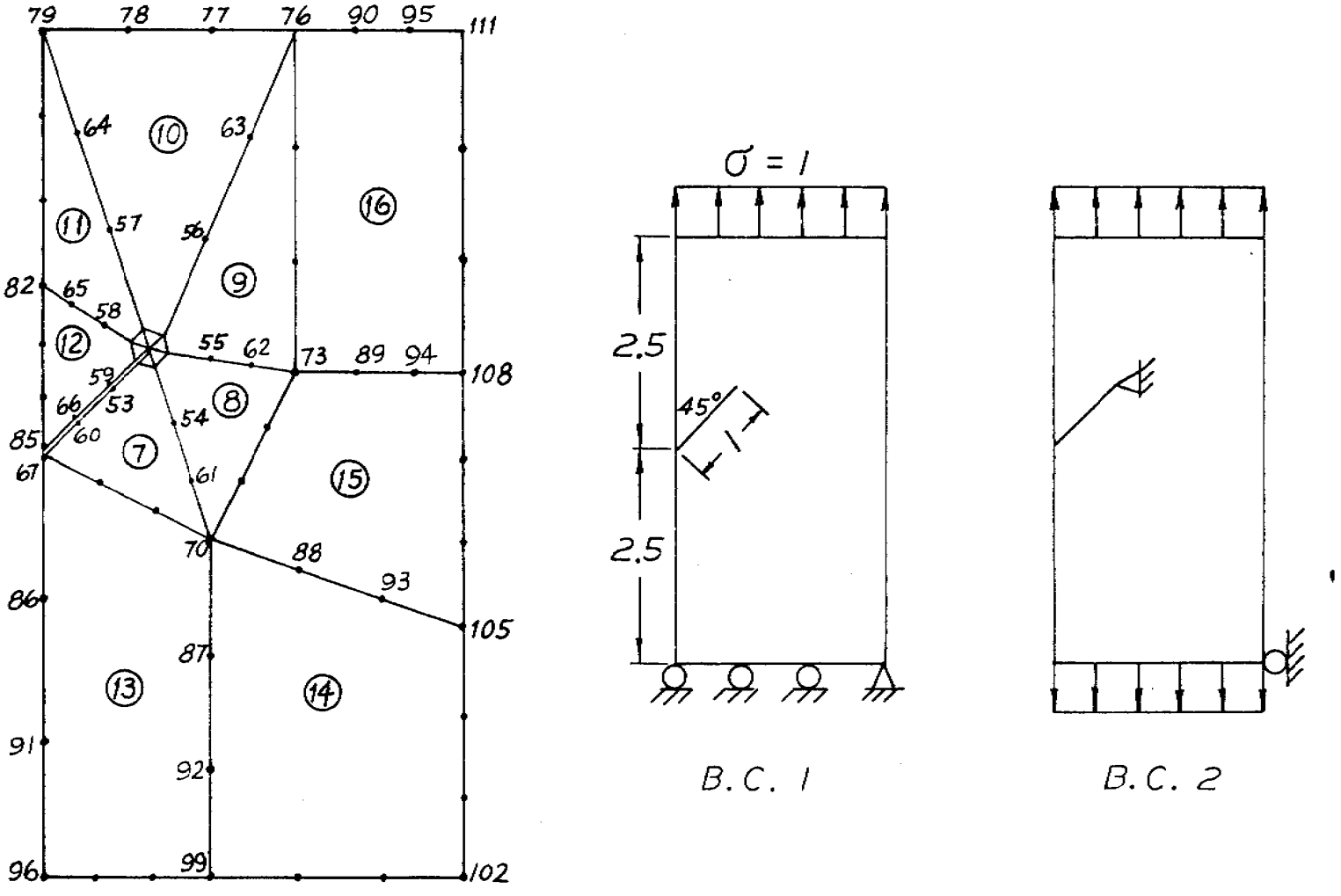


Figure 7. Idealization of a 45-Degree Slant Edge Cracked Panel in Tension.

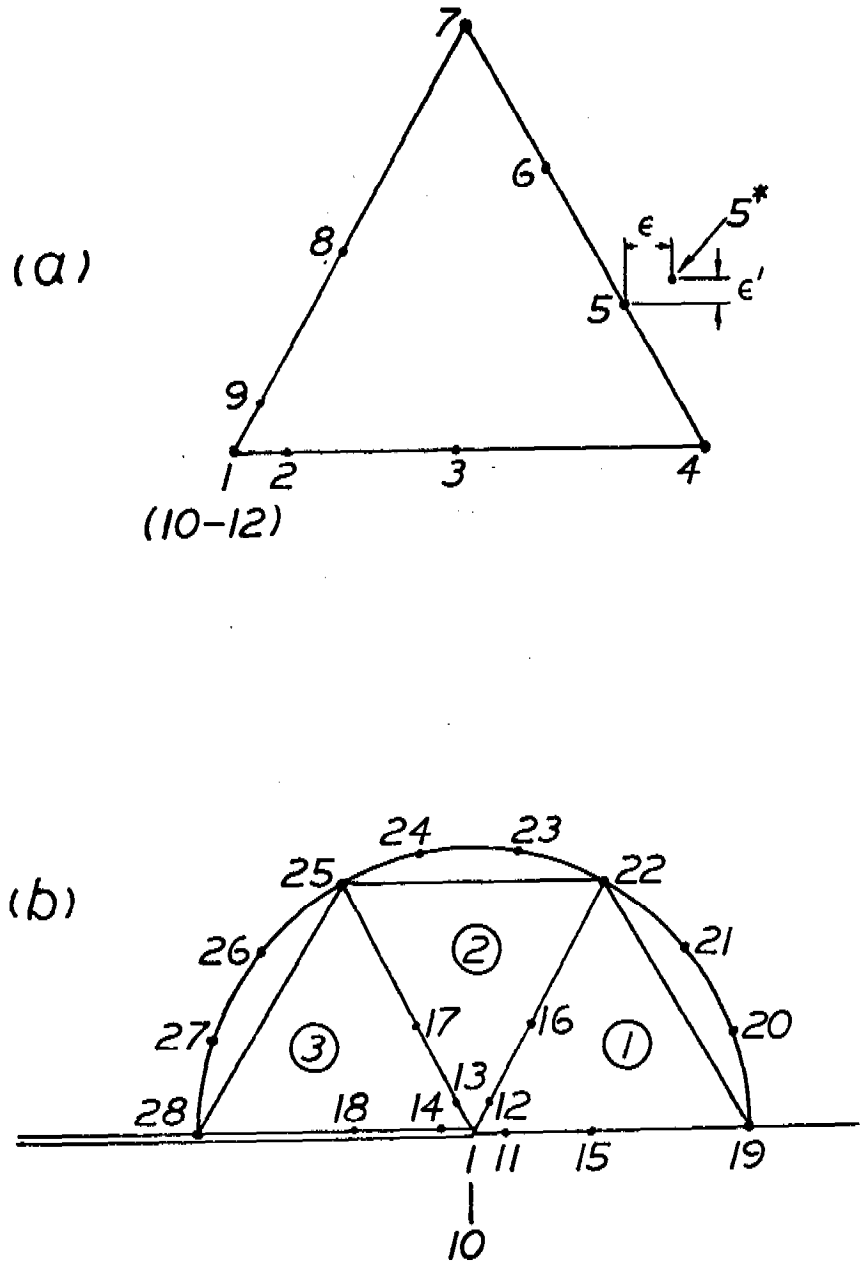


Figure 8(a). Node 5 Perturbed to 5*.
(b). Nodes 20, 21, 23, 24, 26, 27 Perturbed From Their Nominal Positions.

BIVARIATIONAL BOUNDS

Peter D. Robinson*

Mathematics Research Center
University of Wisconsin
Madison, Wisconsin 53706

and

Department of Applied Mathematics
University of Manitoba
Winnipeg, Manitoba, Canada R3T 2N2

ABSTRACT. Let H be a real Hilbert space with symmetric inner product $\langle\ ,\ \rangle$, and let $g \in H$ be an arbitrary vector. Upper and lower bounds of variational type are presented on the quantity $\langle g,\phi\rangle$ associated with the solution $\phi \in H$ of an arbitrary equation $F\phi = 0$. The bounds are based on a bivariational approximation to $\langle g,\phi\rangle$, namely

$$J(\Psi,\Phi) = -\langle\Psi,F\Phi\rangle + \langle g,\phi\rangle\ ,$$

and do not depend on any decomposition of the operator F. Applications to both linear and non-linear problems are indicated.

I. INTRODUCTION. Let A be a self-adjoint operator in a real Hilbert space H with inner product $\langle\ ,\ \rangle$. Associated with the linear equation

$$A\phi = f \tag{1}$$

in H is the well-known Rayleigh-Ritz variational functional

$$R(\Phi) = -\langle\Phi,A\Phi\rangle + 2\langle f,\Phi\rangle,\quad \Phi \in H\ , \tag{2}$$

which is stationary about

$$R(\phi) = \langle f,\phi\rangle \tag{3}$$

for variations in Φ around ϕ, the solution of (1). If A is strictly positive, so that for all $\Phi \in H$

$$\alpha\langle\Phi,\Phi\rangle \geq \langle\Phi,A\Phi\rangle \geq \beta\langle\Phi,\Phi\rangle,\quad \beta > 0\ , \tag{4}$$

then variational bounds

$$R(\Phi) + \frac{1}{\alpha}\left\|A\Phi - f\right\|^2 \leq \langle f,\phi\rangle \leq R(\Phi) + \frac{1}{\beta}\left\|A\Phi - f\right\|^2 \tag{5}$$

---

*On leave from Bradford University, England.

Sponsored by the United States Army under Contract DAAG29-75-C-0024.

are available ($\|\Phi\|$ denotes the Hilbert space norm $\langle\Phi,\Phi\rangle^{1/2}$). If A is indefinite, but nevertheless satisfies for all $\Phi \in H$

$$\|A\Phi\| \geq \gamma\|\Phi\|, \quad \gamma > 0 , \tag{6}$$

then instead of (5) we have the variational bounds

$$R(\Phi) - \frac{1}{\gamma}\|A\Phi - f\|^2 \leq \langle f,\phi\rangle \leq R(\Phi) + \frac{1}{\gamma}\|A\Phi - f\|^2 . \tag{7}$$

Results like (5) and (7) are useful when ϕ cannot be found exactly, but upper and lower bounds on the quantity $\langle f,\phi\rangle$ are required. Other variational bounds can be derived which depend on decomposing all or part of the linear operator A in the form T^*T [1 - 3], but such decompositions are not always available, or convenient (e.g. A might represent an integral operator, or an unwieldy differential operator; or equation (1) might stand for a batch of simultaneous equations). For simplicity and generality we do not consider them here.

The question arises: can we find corresponding bounds on $\langle g,\phi\rangle$ for arbitrary $g \in H$?

II. LINEAR PROBLEMS.

(i) A self-adjoint

Consider the pair of equations in H

$$A\phi = f, \quad A\psi = g \tag{8}$$

for which

$$\langle g,\phi\rangle = \langle f,\psi\rangle . \tag{9}$$

Since for any parameters s and t we have the identity

$$\langle sf + tg, s\phi + t\psi\rangle - \langle sf - tg, s\phi - t\psi\rangle = 4st\langle g,\phi\rangle \tag{10}$$

appropriate subtraction of two pairs of bounds like those in (5) for each of the combined equations

$$A(s\phi \pm t\psi) = sf \pm tg \tag{11}$$

leads to bounds on $\langle g,\phi\rangle$. If A satisfies (4), and the ratio $s:t$ is optimized, we obtain the result

$$J + \xi S - \eta C \leq \langle g,\phi\rangle \leq J + \xi S + \eta C \tag{12}$$

with

$$J(\Psi,\Phi) = -\langle\Psi,A\Phi\rangle + \langle\Psi,f\rangle + \langle g,\Phi\rangle = \langle g,\phi\rangle - \langle\delta\psi,A\delta\phi\rangle \ , \tag{13}$$

$$S(\Psi,\Phi) = \langle A\Phi - f, A\Psi - g\rangle = \langle A\delta\phi, A\delta\psi\rangle \ , \tag{14}$$

$$C(\Psi,\Phi) = \|A\Phi - f\| \, \|A\Psi - g\| = \|A\delta\phi\| \, \|A\delta\psi\| \ , \tag{15}$$

$$\delta\phi = \Phi - \phi, \quad \delta\psi = \Psi - \psi \ , \tag{16}$$

and

$$\xi = \frac{1}{2}\left(\frac{1}{\beta} + \frac{1}{\alpha}\right), \quad \eta = \frac{1}{2}\left(\frac{1}{\beta} - \frac{1}{\alpha}\right) . \tag{17}$$

The bounds in (12) are bivariational in character, and reduce to (5) when $\Psi = \Phi$ and $f = g$. If A satisfies (6) rather than (4), we take $\alpha = -\gamma$, $\beta = +\gamma$ in the foregoing, to generalize (7).

The functional $J(\Psi,\Phi)$ is a bivariational approximation to $\langle g,\phi\rangle$, and it generalizes the Rayleigh-Ritz variational functional $R(\Phi)$.

(ii) A not self-adjoint

When A is not self-adjoint, the bivariational approximation $J(\Psi,\Phi)$ is still available, but is now associated with the pair of equations

$$A\phi = f, \quad A^*\psi = g \ , \tag{18}$$

A^* denoting the Hilbert space adjoint of A. If A satisfies condition (6), it can be shown [4] that the bivariational bounds

$$J - \frac{1}{\gamma}\tilde{C} \leq \langle g,\phi\rangle \leq J + \frac{1}{\gamma}\tilde{C} \tag{19}$$

hold, with

$$\tilde{C}(\Psi,\Phi) = \|A\Phi - f\| \, \|A^*\Psi - g\| = \|A\delta\phi\| \, \|A^*\delta\psi\| \ . \tag{20}$$

Rather better bounds can be derived whenever

$$\frac{1}{2}\langle\Phi,(A + A^*)\Phi\rangle \geq \delta\langle\Phi,\Phi\rangle, \quad \delta > 0, \quad \text{for all } \Phi \in H \ , \tag{21}$$

i.e., whenever the self-adjoint part of A is strictly positive. Then it can be shown [5] that

$$J - \frac{1}{2\delta}(\tilde{C} - \tilde{S}) \leq \langle g,\phi\rangle \leq J + \frac{1}{2\delta}(\tilde{C} + \tilde{S}) \tag{22}$$

with

$$\tilde{S}(\Psi,\Phi) = \langle A\Phi - f, A^*\Psi - g\rangle \ . \tag{23}$$

The different bounds in (12), (19) and (22) are all obtained by adding to or subtracting from the bivariational functional J appropriate "correcting" functionals, which are themselves bivariational approximations to zero. Other techniques [5 - 7] depend on constraining the vectors Φ and Ψ so that the second order term $\langle \delta\psi, A\delta\phi \rangle$ in J takes a definite sign. Decompositions of the operator A are usually involved.

III. NONLINEAR PROBLEMS

The idea of "correcting" a bivariational approximation to $\langle g,\phi \rangle$ works for arbitrary nonlinear problems of form

$$F\phi = 0 \tag{24}$$

in H, provided that F satisfies reasonable conditions. The functional

$$J(\Psi,\Phi) = -\langle \Psi, F\Phi \rangle + \langle g,\phi \rangle \tag{25}$$

generalizes that in (13), and is associated with the pair of equations

$$F\phi = 0, \quad F'(\phi)^*\psi = g \ , \tag{26}$$

$F'(\phi)$ being the linear Gâteaux derivative of F at ϕ. Regarding $J(\Psi,\Phi)$ as a mapping from $H \times H$ into the reals, its Gâteaux derivative at (Ψ,Φ) is described by the mosaic

$$J'\left(\begin{bmatrix}\Psi\\ \Phi\end{bmatrix}\right)\begin{bmatrix}\cdot\\ \cdot\end{bmatrix} = \left\langle \begin{bmatrix} -F\Phi \\ g - F'(\Phi)^*\Psi \end{bmatrix}, \begin{bmatrix}\cdot\\ \cdot\end{bmatrix} \right\rangle \tag{27}$$

and thus will be the null operator when $(\Psi,\Phi) = (\psi,\phi)$, the solutions of (26). The true variational character of $J(\Psi,\Phi)$ is thus established. The original problem (24) is embedded into the larger problem (26), with Ψ playing the role of a sort of Lagrange multiplier. Full details of the analysis are given in [8].

It is interesting to note that, if Ψ can be chosen in terms of Φ so that

$$F'(\Phi)^*\Psi = g \ , \tag{28}$$

then formally

$$J = \langle g, \{\phi - F'(\Phi)^{-1}F\phi\} \rangle = \langle g, N\Phi \rangle \ , \tag{29}$$

$N\Phi$ being Newton's approximation to ϕ. The functional J thus generalizes Newton's approximation in a certain sense, and it does not demand knowledge of unpleasant inverses.

To establish bivariational bounds on $\langle g,\phi \rangle$ in the form

$$J - C \leq \langle g,\phi \rangle \leq J + C \ , \tag{30}$$

where C is a positive bivariational approximation to zero, we assume that

$$\|F\Phi_1 - F\Phi_2\| \geq c\|\Phi_1 - \Phi_2\|, \quad c > 0 , \tag{31}$$

and

$$|\langle \Psi, F\Phi_1 - F\Phi_2 - F'(\Phi_1)[\Phi_1 - \Phi_2]\rangle| \leq \frac{1}{2} K(\Psi) \|\Phi_1 - \Phi_2\|^2 , \quad K(\Psi) \geq 0 , \tag{32}$$

for all

$$\Phi_1, \Phi_2 \in S \subset D(F) \subset H \quad \text{and} \quad \Psi \in \mathfrak{J} \subset D(K) \cap D(F'^*) . \tag{33}$$

Then it is straightforward (see [8]) to show that a suitable 'correcting' functional is

$$C(\Psi, \Phi) = \frac{1}{c} \|F\Phi\| \, \|F'(\Phi)^*\Psi - g\| + \frac{1}{2c^2} K(\Psi) \|F\Phi\|^2 . \tag{34}$$

We note that condition (31) is required even if F is linear (cf. (6)). It implies uniqueness of ϕ in S, and holds for example when $F\phi \equiv G\phi - \phi$ where G is a contraction operator, or when $F\phi \equiv A\phi + f(\phi)$ with A self-adjoint and bounded as in (4), and $df/d\phi$ suitably bounded. Condition (32) indicates that the nonlinearity of F is not too fierce. It is satisfied if either

$$\|F\Phi_1 - F\Phi_2 - F'(\Phi_1)[\Phi_1 - \Phi_2]\| \leq \frac{1}{2} k \|\Phi_1 - \Phi_2\|^2 \quad (\text{with } K(\Psi) = k\|\Psi\|) \tag{35}$$

or (for function-spaces with suitable innerproducts)

$$|F\Phi_1 - F\Phi_2 - F'(\Phi_1)[\Phi_1 - \Phi_2]| \leq \frac{1}{2} k^\dagger |\Phi_1 - \Phi_2|^2 \quad (\text{with } K(\Psi) = k^\dagger |\Psi|_{\sup}) . \tag{36}$$

In a practical situation, the cirtical task can be to find a suitable subset S of the domain of F for which the conditions (31) and (32) hold good.

IV. ILLUSTRATIVE EXAMPLES

Reactor problems

The different bounds for linear problems can be illustrated by reference to neutron diffusion in a reactor, described by the equations

$$\left.\begin{aligned} &(\rho - \sigma^2\nabla^2)\phi(\underline{r},t) + \frac{\partial\phi}{\partial t} = f(\underline{r},t), \quad 0 \leq t \leq T, \quad \underline{r} \in V , \\ &\phi(\underline{r},t) = 0 \quad \text{on} \quad \partial V \quad \text{for all} \quad t, \quad \phi(\underline{r},0) = 0 \quad \text{for all} \quad \underline{r} . \end{aligned}\right\} \tag{37}$$

The total number of neutrons absorbed in time T is proportional to $\langle 1, \phi \rangle$ where

$$\langle \Phi_1, \Phi_2 \rangle = \int_0^T \int_V \Phi_1(\underline{r},t)\Phi_2(\underline{r},t) d\underline{r}\, dt . \tag{38}$$

In a steady-state situation, bounds on $\langle 1,\phi \rangle$ can be found from the simple Rayleigh-Ritz approach in I if $f(\underline{r})$ is constant (cf. [2]); otherwise the theory of II(i) can be used when $f(\underline{r})$ is varying. In a time-dependent situation, the theory of II(ii) applies with (38) as inner product (see [4]).

Pointwise bounds on solutions

By taking g as a suitable kernel function, it is often possible to use bivariational theory to obtain pointwise bounds on solutions. For example, if equation (1) represents the Fredholm integral equation

$$\phi(x) + \lambda \int_a^b k(x,y)\phi(y)dy = f(x) \ , \tag{39}$$

and if $g(x) = k(x',x)$, bounds on $\langle k,\phi \rangle$ lead to bounds on $\phi(x')$ [4,9]. If k is symmetric, the auxiliary equation in (8) is actually the one specifying the reciprocal kernel, and interesting theoretical developments ensue. The bivariational method is a simple yet powerful practical tool in this situation [9].

The same kind of approach can be used for differential equations, with g as the Green's Function for a suitably simple part of the differential operator [10].

Nonlinear problems

The nonlinear diffusion equation

$$\left.\begin{aligned} -\frac{d^2\phi}{dx^2} + (1 + \phi) + \frac{1}{4}(1 + \phi)^2 = 0, \quad -1 \leq x \leq 1, \\ \phi(-1) = 0, \quad \phi(1) = 0 \end{aligned}\right\} \tag{40}$$

is discussed in [8], and bounds on $\langle 1,\phi \rangle$ are evaluated. This quantity could be, for example, the amount of heat stored in a metal bar with ends kept at temperature zero which was suffering a nonlinear heat-loss to the surroundings at temperature -1. A monotonicity theorem can be employed to determine a suitable subset S for the determination of the critical constants in the 'correcting' functional C.

The nonlinear integral equation

$$\int_0^{\pi/2} \frac{\sin(x - y)}{\pi(x - y)} \phi(y)dy = \frac{1}{\phi(x)} \tag{41}$$

occurs in communication theory, and the cosine transform $\langle \cos px,\phi(x) \rangle$ of $\phi(x)$ is proportional to the signal strength. Bivariational bounds on this quantity are reported in [8].

Applications to wide classes of problems are evidently feasible, and much development remains to be done.

V. SOME GENERALIZATIONS. Instead of $\langle g,\phi \rangle$ in the foregoing, we can work with a more general inner product $g(\phi)$. The auxiliary equation in (26) must be replaced by

$$F'(\phi)^*\psi = g'(\phi) \ . \tag{42}$$

It is possible to construct a suitable correcting functional when $g(\phi)$ has a bounded second derivative.

Adaptations can be made to the theory to take explicit account of boundary terms occurring in the specification of adjoint operators (cf. [1-3]). Generalization to complex spaces is not difficult (see [7, 11]), and the theory can be set in more abstract spaces which need not even be normed (see [12]). The bivariational method has proved efficient in calculating dynamic polarizabilities for two-electron atoms at arbitrary complex frequencies ([13]).

REFERENCES

1. B. Noble and M. J. Sewell, 1972, J. Inst. Maths. Applics. 9, 123-193.

2. A. M. Arthurs, 1970, Complementary variational principles, Oxford - Clarendon Press.

3. P. D. Robinson, 1971, in Nonlinear functional analysis and applications, ed. L. B. Rall, Academic Press, 507-576.

4. M. F. Barnsley and P. D. Robinson, 1976, Proc. Roy. Soc. Edinb. 75A, 9, 109-118.

5. M. J. Sewell and B. Noble, 1976, General estimates for linear functionals in nonlinear problems, MRC Report #1703.

6. W. D. Collins, 1976, Dual extremum principles for dissipative systems, MRC Report #1624.

7. P. D. Robinson, 1977, New variational bounds on generalized polarizabilities, MRC Report #1725.

8. M. F. Barnsley and P. D. Robinson, 1977, Bivariational bounds for nonlinear problems, J. Inst. Maths. Applics., in press.

9. P. D. Robinson, 1977, Pointwise bounds on solutions of Fredholm integral equations, MRC Report #1767.

10. M. F. Barnsley and P. D. Robinson, 1976, J. Math. Anal. Applics. 56, 172-84.

11. M. F. Barnsley and G. A. Baker, Jr., 1976, J. Math. Phys. 17, 1019-1027.

12. I. Herrera and M. J. Sewell, 1977, Dual extremum principles for non-negative unsymmetric operators, MRC Report #1743.

13. F. Weinhold and P. D. Robinson, 1977, Bivariational calculations of bounds for complex-frequency polarizabilities, submitted for publication.

SOME GENERIC PROPERTIES OF A LOGIC MODEL FOR ANALYZING HARDWARE MAINTENANCE AND DESIGN CONCEPTS*

James T. Wong, Mathematician
William L. Andre, Research Engineer
U.S. Army Air Mobility Research and Development Laboratory
Ames Research Center
Moffett Field, California

ABSTRACT

A mathematical structure for diagnostic logic modeling was formulated, which allows the intrinsic properties of a complex Logic Model to be studied in an abstract setting. As a result, it was found that a loop-free Logic Model is a partially ordered set and that every permutation of the elements in the terminal set of a finite partially ordered set S partitions S into disjoint subsets. Based on these results, it was deduced that the minimum number of test points required for conclusive detection of malfunctioning components for a loop-free system is equal to the number of elements in the terminal set; this set constitutes the optimal choice for test points. Also, it was established, for each permutation of the elements in the terminal set, a relative failure diagnostic strategy was defined in accordance with Bellman's Principle of Optimality. Finally, for the purpose of illustration, some examples are given.

* Published in Proceedings of the Symposium on "Applications of Decision Theory to Problems of Diagnosis and Repair," June 1976.

COMPUTER GRAPHICS IN A PRODUCTION ENVIRONMENT

William D. Johnston
Analysis and Computation Division
National Range Operations Directorate
US Army White Sands Missile Range, NM

Abstract

This paper examines the mathematical and software techniques required for an efficient, high-production graphics system. It is shown how appropriate design of both software and mathematical procedures will lessen the workload of the customer and reduce throughput time, while satisfying all of the production requirements. Such procedures are designed to isolate the causes of irregularities that may arise during the computer solution, and give the user explicit directives as to corrective action. Also discussed are the software and operational techniques required for information security purposes within the graphics system, as well as the methods of software acquisition for specific areas of the system.

Although primary emphasis is on mathematics and software, as applied to production graphics systems, this paper also examines the following areas:

a. The classification of production vs. demand graphics.

b. User requirements and demands.

c. System operational procedures.

d. Hardware.

e. The evolution of the White Sands production graphics system--an existing high-production system that produces 15,000 plots per month.

I. Production vs. Demand Graphics

The need for graphics falls generally within these two categories, with each type fulfilling specific requirements. Since the needs of the organization will dictate whether a production system is desirable, it is important to distinguish between the two systems and define their characteristics.

The demand (or interactive) graphics system has in recent years become an extremely powerful and highly developed tool in which the user normally accesses a real-time video terminal from which he can produce hard copy prints with an appropriate plug-in device. When equipped with or attached to a processor (micro or otherwise), the images may be edited, rotated, or otherwise modified through keyboard or light-pen entry. Demand graphics systems find their utility in research and development, interactive graphical analysis, and in any situation where the user requires instant presentations or prints in relatively small quantities.

Production graphics systems on the other hand are allied with the need for systems which can generate large quantities of plots, where the plot itself is not a tool in the development process, but a graphical overview of some operation or event. To that end, the plot is not intended to be the source of specific samples of data, and it is neither expected nor intended that the analyst will apply a ruler to the plot to extract specific and precise data from it. For that purpose there exists the detailed point-by-point data listing generated by the data processing equipment. The plot, meanwhile, provides the analyst with an overall picture of the event, and quickly reveals trends, disruptions or deviations, and the probable outcome of the event being plotted.

The need for a true production graphics system exists anywhere there is a large volume of data being generated and where there is a need for graphics to either supplement, or, to a limited extent, stand in for the actual data listings. In these cases, it is desirable that the plots be generated through the use of standardized procedures requiring little action on the part of the initiating party. Furthermore, the graphics system itself should draw all titles, numerical annotation, and other legends on the plots so that there is no need for a clerical post-processing step. In other words, when the plot comes out of the graphics system, it should be ready for immediate insertion in the final data report.

The important factor in deciding whether a production system will be beneficial to a particular organization is the volume of graphics to be produced, including the production of graphics that are desired, but not possible or practical under the existing system. If the operation requires relatively few plots (on the order, say, of one or two hundred per day), or if there is a need for interactive analysis, a small demand system will probably be quite effective. Conversely, an organization that reduces large volumes of data and produces numerous reports containing graphics will find the efficiency of a production system very attractive.

II. User Requirements and Demands

The customer's demands often exceed his requirements, and it is the responsibility of the production manager to decide which services can and should be provided. These decisions affect both hardware and software, and can be intelligently made only after close consultation with the users themselves. Obviously, the requirements may vary widely from one installation to the next, and specific recommendations for one operation could well be partially or wholly invalid for another. Nevertheless, there are several general points that must be reconciled between the production manager and the graphics customers.

The first and most important of these is turn-around time. The user invariably wants fast turn-around, and indeed, it may be possible to provide finished plots in a matter of hours, or even minutes. But this might well require keeping an equipment operator on duty at all times, when it would actually be more efficient to group the graphics jobs so that the actual plotting is done only at certain regular times. Such a system generally guarantees turn-around on some basis such as four, eight, twelve, or twenty-four hours.

In actual practice, there are very few cases which require turn-around in less than twenty-four hours, through the user may think otherwise. The primary method to reduce the customer pressure on turn-around it to provide him with quick-look printer plots on his computer listing. To be effective, however, the printer plots must be of sufficient usefulness to satisfy his curiosity and his anxiety for the next few hours, until the actual finished plots are returned. Whether this very important requirement is met depends upon the design of the printer plot program, which in turn will be discussed in more detail in the section dealing with software.

Another frequent demand is for larger plots in order to obtain increased resolution. As was mentioned previously, however, the objective of production graphics is not to provide a source for specific precise data elements. Furthermore, increased plot size entails increased capital outlay for equipment, as well as increased periodic expenditures for consumables--not to mention the logistical annoyance of finding a place to store the larger sheets of paper. Almost all production graphics requirements can be met with plots (or series of plots) on standard 8 1/2 x 11 inch paper, which permit direct inclusion in the published report. The greatest problem here is in convincing the customer of this fact, especially if he has been accustomed to a larger size in the past. Experience shows that this is nothing more than a matter of conditioning, and if the user can be convinced to at least <u>try</u> a standard size plot for a period of time, he will come to find it satisfactory in every respect.

The desire for multicolor graphics is another aspect that is more related to conditioning than to actual need. There are certainly many cases where it is useful to have two or more plot traces on the same grid, and there must be some way to uniquely identify each one. This may be done by changing the color of the individual trace, or by changing its mode (i.e.,

by drawing a point or symbol plot, as opposed to a line plot). While the multicolor plot is pleasing from an aesthetic point of view, it does leave the door open for error on the part of the equipment operator, and in the long run it may turn out to be completely useless if it must be reproduced in black and white within a printed report. Consequently, the multicolor capability is not necessarily as useful as it would first appear, whereas the single-color multi-mode plot can not only fully satisfy the trace identification requirement, but is actually superior for reproduction purposes.

The final customer requirement to be mentioned here is that of system capacity, or the total number of plots that the system is capable of producing in a given period of time. The demands on graphics systems are similar to the demands on all data processing equipment in the real world, in that any increase in capacity will automatically be met (or exceeded) by an increase in demand. In other words, capacity breeds demand. This is not meant as a humorous aside, but as a statement of a very real fact that computer systems personnel have had to deal with for many years. And this is not in itself all bad either, since increased graphics capability can often speed up the work of the analysts and data report users by eliminating some of the need to pore over long data listings. In any case, as long as the minimum essential capacity for the operation is met, the only major concern beyond that is to ensure that any excess capacity of the system is not so great that it results in purely wasteful expansion of usage, manpower, and materials.

III. Operational Procedures

Basic operational procedures in a production graphics system must satisfy a number of important goals. First, of course they must satisfy the plotting requirements of the customer, but they must do so in an efficient manner if high production is to be achieved. Some of the areas that must be encompassed are the issues of on-line/off-line operation, security, scheduling, and reliability.

The choice between on-line or off-line operation of graphics hardware, as an operational matter, often goes hand-in-hand with the customer's interest in turn-around time. The customer, of course, visualizes greatly improved turn-around by operating the equipment on-line. Without a doubt, this mode has its advantages--generation of the plots at the time of the original computer reduction, elimination of the need for an off line plot tape, and even increased speed of operation of the graphics hardware in many cases. All of this, but at what cost? The most damaging aspect is that the most expensive component of all--the computer itself--is bound by the mechanical speed of the graphics hardware (even to a certain extent in a multiprocessing system). Additionally, one or more I/O ports may be continuously tied up, not to mention the fact that an operator must be on call at all times.

By operating the graphics equipment off-line, the data transfer rate is at full tape speed, and the tapes themselves can be grouped so that operating personnel for the graphics equipment need to be available only during certain specific time periods. Consequently, operating personnel requirements are reduced somewhat, and the computer mainframe is not hobbled by the comparatively slow speed of the plotter hardware. Turn-around can still be provided on some satisfactory pre-arranged basis, such as four, eight, twelve, or twenty-four hours. The optimum scheduling depends on the volume of graphics generated, the operating characteristics of the hardware, and the shift hours of the operating personnel; but this usually has to be modified somewhat to take into account customer requirements and other influencing factors. Some experimentation is to be expected when establishing a new production system.

Probably the most neglected aspect of graphics operations is that of physical and information security, and graphics security often takes second place to other areas of data processing security. Possibly this is due to the fact that the investment in graphics hardware is usually small in comparison to the remainder of the ADP equipment, and improper use of materials or services is not as likely. Nevertheless, physical security affects information security, and classified graphics must be just as closely guarded as any other security sensitive information.

The most straightforward method of handling the problem is to locate the graphics equipment in a physically secure area. This could well be in the same room as the computer itself, though some types of graphics hardware emit fumes or dust which would preclude such an arrangement.

The customer, on the other hand, is sure to want hands-on access to the equipment. This must be carefully avoided in order to prevent interference with the operating personnel, and also to prevent possible breaches of security. His desire to be present during the plotting process will be greatly diminished if he is supplied with useful quick-look printer plots, and if the turn-around time is reasonable and dependable.

Once physical security is achieved, the job of information security becomes a matter of establishing an awareness of the importance of the security of classified graphics, and of course strict accountability and absolute adherence to those procedures established by the installation security office.

Though there are many additional operational questions that must be resolved within any given system, the last matter to be addressed here is that of reliability. This refers not only to hardware reliability which is a maintenance responsibility, but to operational reliability. In other words, the customer must be able to reasonably expect that his plots will be delivered to him within the established schedule. The greatest pitfall is failure to maintain appropriate communications with the hardware maintenance personnel. It is the operator who will discover malfunctions, and he should immediately report all details of the problem to the maintenance section. A foolish, but very common occurence is to shut the machine off and expect the maintenance people to find the trouble during their regular calibration or PM checks. The odds are that they won't find it, and no amount of shifting the blame will get the customer's plots to him any faster.

In an off-line system where tapes are grouped and the actual plotting done at specific times, there is a temptation during slow periods, when only a few tapes are generated, to hold them over until the next scheduled plotting session. This may at first glance appear to improve efficiency, but it is more likely the case that the operator who would have had only a little bit to do will then simply have nothing to do. Then the customer will come knocking on the door at the regular time, only to find out that the turn-around (which he probably thought was too long in the first place) has been increased not by a heavy workload but, incomprehensibly, by a light workload.

On the other hand, total inflexibility can be just as disastrous. If the operations manager is aware of an upcoming increase in production, it is his duty to prepare in advance by scheduling additional periods of operating time.

IV. Software

A. General Requirements

Software is a critical component in the production graphics systems, wherein the mathematician/programmer has an opportunity to develop software which will not only satisfy all of the production and user requirements, but will also reduce throughput time and lessen the customer's workload. One must not make the mistake, however, of assuming that all software must be developed in-house to obtain maximum efficiency. There are, for example, certain basic routines which perform such mundane tasks as buffering and writing the output tape, drawing alphanumeric characters, etc., which are normally supplied by the hardware manufacturer at minimal (or zero) cost. These routines are generally quite versatile and efficient as supplied, and it is simply not possible to develop substitutes in-house at lower cost.

It must also be realized that not all types of graphics software are either necessary or economically justifiable within a given system. Probably the best example of this is in the area of three-dimensional or perspective graphics. Most every organization has an occasional application where 3-D capability would be nice to have, but the acquisition or development of such routines is costly enough that the economic justification in terms of potential benefit should be carefully scrutinized before commiting funds or resources.

The major portion of the software for a production graphics system must, however, be developed in house, and though it must in certain ways be tailored to the specific needs and requirements of the particular installation, there are many features which are essential to any high-efficiency, high-production system. Software design is heavily influenced by the fact that often the customer is an analyst/technician with little or no background in mathematics or programming, and as such he has a tendency to suspect the operation of the plot program before suspecting the data in the event of a malfunction.

In actual practice, most abnormalities are in the data itself, and well-designed graphics software will reveal this to the user on the listing generated during the plot run on the computer. Typical examples are: (a) the analyst used the wrong data tape. (b) He used the right tape, but was looking at the wrong listing. (c) He plotted the wrong parameters. (d) The data was in different units from those stated. (e) The input data file was blank.

The software, then, must first of all minimize the effort required by the user to initiate a run. This is accomplished by providing him with clear and complete documentation beforehand, and by incorporating features in the program that minimize his workload and chance for error. For example, where there are a number of possible selections for a given plot option, the most commonly used selection should be used under default conditions. In other words, the mere act of not specifying any selection should be understood

to mean that the "standard" selection is to be used. If the various options are to be specified on cards, the layout should, of course, be orderly and neat.

Without delving into the innumerable graphics options concerning form and style, we can look specifically at those features whose availability can either diminish the possibility for error initially, or isolate the cause of an irregularity once it does occur.

The first of these is the matter of the selection of the minimum and maximum plotting limits. The specific techniques are described in detail in a following section, but stated briefly, there must be a system which will automatically select limits and increments that are pleasing to the eye, contain all of the data, and do not extend unnecessarily far beyond the limits of the actual data. The implementation of this single option eliminates the source of many errors by doing away with manual scanning of data and manual calculation of appropriate round limits and increments. There must, of course, exist the ability to easily override the option in order to insert specific limits when desired.

Graphics labeling and numerical annotation of axes must be done by the plotter as a part of the plotting process. Any clerical post-processing step for this purpose is prone to error, and in any case will increase the turn-around time dramatically. Security classification labeling, where appropriate, is also handled in the same manner. When classified plots are properly marked at the time they are generated, their susceptibility to mishandling is lessened, and the overall security posture is improved.

The method of specification by the user of headings and legends may vary, but a catalog or table of standard headings might be maintained where there is frequent repetition. The appropriate headings would then be introduced upon recognition of a unique project number of other code. Numerical annotation, on the other hand, should always be generated automatically on the basis of the minimum and maximum plotting limits in use, with no additional action or entries required on the part of the initiator.

There must be a degree of flexibility in the handling of input data if the system is not to break down under varying mission requirements. usually in a production grphics system there will be some standard data format that will be used for the bulk of the processing, but there remain the odd cases that appear from time to time in totally foreign data formats that require some form of graphics processing. One approach is to read the data, rewrite it in standard format, then read it again. Much of this I/O can be avoided, though, if the graphics software has some provision for direct entry of data. This might be through a subroutine call, or through a common storage block. With such a provision, it is only necessary to provide a small routine which is capable of reading or receiving the nonstandard data, and the remainder of the data transfer is accomplished through what is effectively core-to-core movement.

Many graphics malfunctions results from the fact that the data to be plotted is not in the units of measure that the customer believes them to be.

In some cases this will result in total failure of the job to produce the required plots, whereas other cases may result in a built-in bias throughout the plot. The first situation is not particularly dangerous since it is quite clear that some specification must have been in error. On the other hand, the second case generates an insidious error that may go entirely undetected (e.g., plotting yards instead of meters). Unfortunately, there is not a lot that we can do to prevent the customer's carelessness in this matter, but it is possible to incorporate into the graphics software a procedure for units conversion of at least the more common measurement units. Conversion factors would normally be stored table-wise and referenced by an appropriate code during initialization. Digressing somewhat to a point that should be kept in mind, is that although most conversions simply require a multiplicative operation with a single factor, there are a few cases that require an additional operation with an additive factor (degrees Farenheit to degrees Centigrade, for example). The implication is that it may be necessary to maintain two tables--one for each type of factor--and this may be a consideration in terms of core conservation if a great number of conversions are to be made available. Another alternative is to simply make provisions to allow the user to enter the specific factor(s) himself, doing away with the need for conversion factor tables within the graphics software.

It has already been stated that the provision of quick-look printer plots is essential in a production graphics operation. There are a number of good reaons for inclusion of the capability, not the least of which is to reduce customer pressure on turn-around time by temporarily satisfying his curiosity. This in turn makes it easier to deny him physical access to the plotter equipment, where we want to maintain high security. In order for the printer plots to accomplish this, they must be informative, complete, and easy to read. Also, one should not overlook the fact that there are many occasions where a high-resolution plot is not needed, and a printer plot is all that the customer desires. Consequently, the option should exist to generate only the printer plots when so requested, without having to unnecessarily generate the regular plots.

If the printer plot is to satisfy the user, it must incorporate a number of features that allow it to convey as much information as possible at the least possible cost in terms of processing and core storage. The details of these features are further described in Section IV.C.

As to be expected, irregularities will occur from time to time in the processing of graphics data, and though the graphics software cannot prevent them, it can in many cases provide the means to quickly isolate the cause of the problem. One of the most valuable techniques in this respect is the accurate accounting of all data points processed. This means nothing more than accounting for all points examined, and classifying them as to whether they were actually plotted, or were blank, off-scale, or otherwise unusable. A summary of the accounting for each plot is then printed on the customer's computer listing, or on the plots themselves if appropriate. In the event of a failure, the analyst will know exactly how many data points were provided to the plot program, and what their disposition was. This information will often point directly to the cause of an unexpected malfunction. The

beauty of this feature is its simplicity; very little processing time or storage is required. A sample algorithm is included in a later section.

It has not been the intent here to enumerate all conceivable options in a graphics program, but instead to describe a few of the more important features which can reduce user effort, minimize errors, improve throughput, and aid in troubleshooting those problems which are inevitable. Nevertheless, all plot options specified by the user should be carefully checked for errors by the graphics software during its initialization phase, and substitutions made for erroneous entries where practical. In all cases, explicit informative diagnostics should be given.

B. Automatic Selection of Minimum and Maximum Plotting Limits

The automated selection of plotting limits can save the analyst considerable time and effort that would otherwise be spent carrying out this task manually, and it is not particularly difficult to design a routine which can efficiently select limits which are aesthetically pleasing, which contain all of the data, and which do not extend unnecessarily far beyond the limits of the actual data.

To begin with, the raw data contains a range of values, and there will exist some absolute minimum and maximum value for both the ordinate and abscissa. When it comes time to plot the data, we must have some idea of its bounds if the plot is to be meaningful. The retrieval of the absolute min and max values can be achieved by guess, by manual scanning of the data, or by some automated process accomplished by the software. Once obtained, these absolute limits are usually of such a nature that they must be rounded in order for the plot to be easily readable.

The first decision to be made is the manner in which the absolute limits are to be retrieved. Manual processing is out of the question in a production system, but there are still several automated processes to choose from. The first of these is to place the entire data set in core and scan it for the min and max values. Obviously, this system is practical only for those trivial cases where the data set consists of no more than a few thousand pairs of coordinates. Since no production graphics systems can tolerate this type of constraint, it must be discarded. The second method is to read through the entire set of data while it resides on tape or mass storage, saving and updating the minimum and maximum values in the process. Once accomplished, the graphics program will have to reread all of the data to carry out the actual plotting. This is a functional method which may be the only alternative in some cases, but it is wasteful of I/O effort since all of the data must be read twice.

The ideal method of handling the problem is to have the program which initially generates the data save and update the min and max values as each data point is generated. This requires a degree of coordination between the parties responsible for the data reduction software and those responsible for the graphics software, but if accomplished, rereading of the entire data set becomes unnecessary and the savings in I/O processing time are significant.

Once the raw or absolute minimum and maximum values are known, appropriate round limits can be selected which will make the plot more readable. Naturally, there must be a simple method to permit the user to insert any limits he desires (he may want to plot only a particular segment of the data, for example), but the graphics software must also have within it some method of intelligently altering the raw min and max values such that round numbers will appear at the extremities of the axes, as well as at the intermediate increments. In other words, not only must the endpoints have round numbers associated with them, but the increment per unit of axis length must be an appropriate round number also.

The process of selecting the round limits and increment must be optimized, for if the resulting range is unnecessarily broad the plot will contain a great deal of blank space while the plot trace itself is overly compressed. Needless to say, limits which fail to include all of the available data are totally unacceptable. Finally, the routine devised for the task should not be so zealous as to attempt to improve upon limits which are already optimum, as frequently occurs when data is generated artificially during simulations.

The general form of an appropriate algorithm is shown in Figure 1, and a Fortran IV subroutine which incorporates all of the desired features and is usable without modification, is also included. Some samples of limits generated by the routine are included immediately following the program listing.

Automated selection of plotting limits can be the salvation of the harried data analyst, but there are a few minor pitfalls to be guarded against. The most obvious of these is that auto limit selection can cause smooth data to appear rough. For example, a plot depicting a sequence of very small radar range errors would automatically be expanded to fill the whole page. While the data is displayed accurately, the visual impression is that the errors are quite large. In cases like these, it is better to enter standard limits so that all plots of the same type will be uniform.

A second pitfall is that a single piece of erroneous data, if it is itself a minimum or maximum, can upset the selection process. The graphics software has no way of detecting such an occurrence, and must assume that it has been provided with accurate data. Most data reduction software provides some degree of editing and filtering, however, so the frequency of this type of problem is minimal. When it does occur, the analyst has no recourse but to manually select more appropriate limits and rerun the plot.

Another minor problem arises when the plot is to have a logarithmic scale, and either negative data or negative limits exist. Whenever log plots are used and a negative limit is generated, care must be taken not to attempt to take the log of that limit. Also, one can see that if the round limits are generated on the basis of the raw data, the resulting log scale usually will not meet the requirements originally set forth. The most practical procedure is to first check the raw limits, substituting zero for a negative one when necessary. Then take the logs of these raw limits, with the computed log values in turn being supplied as input to the round limits subroutine. This gives the log scale on the axis a more acceptable appearance, and simultaneously avoids an unresolvable condition within the log function.

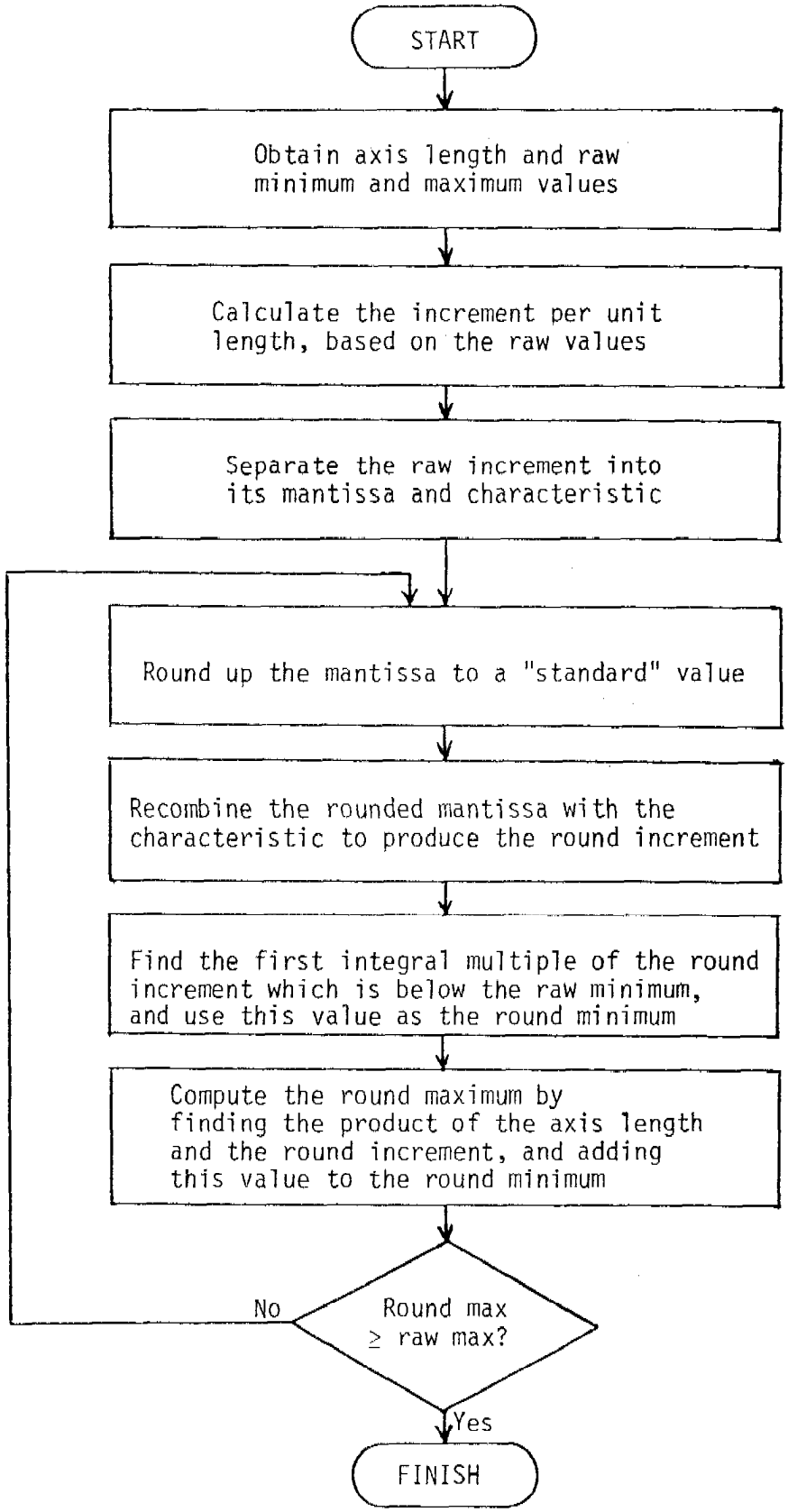

START
Obtain axis length and raw minimum and maximum values
Calculate the increment per unit length, based on the raw values
Separate the raw increment into its mantissa and characteristic
Round up the mantissa to a "standard" value
Recombine the rounded mantissa with the characteristic to produce the round increment
Find the first integral multiple of the round increment which is below the raw minimum, and use this value as the round minimum
Compute the round maximum by finding the product of the axis length and the round increment, and adding this value to the round minimum
Round max ≥ raw max?
No
Yes
FINISH

Figure 1

```
SUBROUTINE RNDLIM    ENTRY POINT 000250

STORAGE USED: CODE(1) 000307; DATA(0) 000050; BLANK COMMON(2) 000000

EXTERNAL REFERENCES (BLOCK, NAME)

 0003   ALOG10
 0004   XPRI
 0005   NERR3$

STORAGE ASSIGNMENT  (BLOCK, TYPE, RELATIVE LOCATION, NAME)

 0001   000070 130G      0001   000031 300L      0001   000056 350L      0001   000103 500L      0001   000105 550L
 0001   000117 600L      0001   000162 700L      0001   000170 750L      0001   000177 800L      0001   000211 830L
 0001   000230 850L      0001   000234 900L      0000 R 000023 FMNTIS    0000 R 000020 GINC      0000 R 000016 GMAX
 0000 R 000015 GMIN      0000 I 000024 I         0000   000035 INJPS     0000 I 000025 K         0000 I 000026 MINT
 0000 I 000014 NMANTS    0000 I 000022 NPWR      0000 R 000030 PDIFF     0000 R 000021 PWR       0000 R 000017 RANGE
 0000 R 000000 STDINC    0000 R 000027 TEMP      0000 R 000031 TMAX
```

```
00101      1*            SUBROUTINE RNDLIM (GIVMIN, GIVMAX, GLENTH, RNDMIN, RNDMAX, RNDINC)RND00001    000002
00101      2*      C                                                                       RND00002    000002
00101      3*      C                                                                       RND00003    000002
00101      4*      C                                                                       RND00004    000002
00101      5*      C       ROUND LIMITS ROUTINE FOR GRAPHICS, DECEMBER 1976.               RND00005    000002
00101      6*      C       WILLIAM D. JOHNSTON, NR-AD-S, WHITE SANDS MISSILE RANGE,        RND00006    000002
00101      7*      C       NEW MEXICO  88002.                                              RND00007    000002
00101      8*      C                                                                       RND00008    000002
00101      9*      C                                                                       RND00009    000002
00101     10*      C       THIS ROUTINE GENERATES OPTIMIZED ROUND LIMITS AND INCREMENTS     RND00010    000002
00101     11*      C       FOR PLOTTING PURPOSES, BASED ON GIVEN RAW MINIMUM AND MAXIMUM   RND00011    000002
00101     12*      C       VALUES AND A PREDETERMINED AXIS LENGTH.                         RND00012    000002
00101     13*      C                                                                       RND00013    000002
00101     14*      C       FOR THE PURPOSES OF THIS ROUTINE, 'MINIMUM' AND 'MAXIMUM'       RND00014    000002
00101     15*      C       SIMPLY REFER TO THE NUMERICAL VALUES AT THE TWO EXTREMITIES     RND00015    000002
00101     16*      C       OF A GIVEN AXIS.  THERE ARE NO RESTRICTIONS ON THEIR VALUES,    RND00016    000002
00101     17*      C       HOWEVER, AND EITHER MAY BE ALGEBRAICALLY GREATER OR LESS        RND00017    000002
00101     18*      C       THAN THE OTHER.                                                 RND00018    000002
00101     19*      C                                                                       RND00019    000002
00101     20*      C       THE ROUTINE IS DESIGNED FOR APPLICATIONS WHERE THE NUMERICAL    RND00020    000002
00101     21*      C       ANNOTATION ON THE PLOT WILL BE AT INTEGRAL MULTIPLES OF THE     RND00021    000002
00101     22*      C       BASIC UNITS OF AXIS LENGTH (INCHES, CENTIMETERS, ETC.).         RND00022    000002
00101     23*      C                                                                       RND00023    000002
00101     24*      C                                                                       RND00024    000002
00101     25*      C           GIVMIN  IS THE RAW GIVEN 'MINIMUM' SUPPLIED BY THE          RND00025    000002
00101     26*      C                   CALLING ROUTINE.                                    RND00026    000002
00101     27*      C                                                                       RND00027    000002
00101     28*      C           GIVMAX  IS THE RAW GIVEN 'MAXIMUM' SUPPLIED BY THE          RND00028    000002
```

```
00101    29*    C                    CALLING ROUTINE.                                          RND00029   000002
00101    30*    C                                                                              RND00030   000002
00101    31*    C            GLENTH  IS THE GIVEN LENGTH, IN ANY UNITS OF MEASURE,             RND00031   000002
00101    32*    C                    SUPPLIED BY THE CALLING ROUTINE.                          RND00032   000002
00101    33*    C                                                                              RND00033   000002
00101    34*    C            RNDMIN  IS THE COMPUTED ROUND 'MINIMUM' VALUE RETURNED BY         RND00034   000002
00101    35*    C                    THE RNDLIM ROUTINE.                                       RND00035   000002
00101    36*    C                                                                              RND00036   000002
00101    37*    C            RNDMAX  IS THE COMPUTED ROUND 'MAXIMUM' VALUE RETURNED BY         RND00037   000002
00101    38*    C                    THE RNDLIM ROUTINE.                                       RND00038   000002
00101    39*    C                                                                              RND00039   000002
00101    40*    C            RNDINC  IS THE COMPUTED ROUND INCREMENT (PER UNIT OF              RND00040   000002
00101    41*    C                    AXIS LENGTH) RETURNED BY THE RNDLIM ROUTINE.              RND00041   000002
00101    42*    C                                                                              RND00042   000002
00101    43*    C                                                                              RND00043   000002
00101    44*    C                                                                              RND00044   000002
00101    45*    C                                                                              RND00045   000002
00101    46*    C                                                                              RND00046   000002
00101    47*    C                                                                              RND00047   000002
00103    48*          DIMENSION  STDINC(12)                                                    RND00048   000002
00103    49*    C         THE VALUE OF NMANTS MUST BE EQUAL TO THE DIMENSION OF STDINC.      RND00049   000002
00104    50*          DATA  NMANTS /12/                                                        RND00050   000002
00104    51*    C         THE TABLE OF STANDARD INCREMENT MANTISSAS, WHICH FOLLOWS, MAY      RND00051   000002
00104    52*    C         CONTAIN ANY NUMBER OF STANDARD INCREMENTS (BUT SEE NOTE            RND00052   000002
00104    53*    C         ABOVE), AND MAY CONTAIN ANY DESIRED VALUES FOR THESE               RND00053   000002
00104    54*    C         INCREMENTS.  HOWEVER, THE STANDARD INCREMENTS IN THE TABLE         RND00054   000002
00104    55*    C         MUST BE LISTED IN ORDER BY INCREASING VALUE, AND THE LAST          RND00055   000002
00104    56*    C         ENTRY IN THE TABLE MUST BE EQUAL TO TEN TIMES THE VALUE OF         RND00056   000002
00104    57*    C         THE FIRST ENTRY IN THE TABLE.                                      RND00057   000002
00106    58*          DATA  STDINC /1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0,      RND00058   000002
00106    59*         1             9.0, 10.0/                                                RND00059   000002
00106    60*    C         SAVE THE GIVEN MINIMUM AND MAXIMUM VALUES.                         RND00060   000002
00110    61*          GMIN = GIVMIN                                                          RND00061   000002
00111    62*          GMAX = GIVMAX                                                          RND00062   000004
00111    63*    C         COMPUTE THE RANGE BETWEEN THE GIVEN MINIMUM AND GIVEN              RND00063   000004
00111    64*    C         MAXIMUM VALUES.  THE SIGN OF THE COMPUTED RANGE WILL INDICATE      RND00064   000004
00111    65*    C         DIRECTION.  (POSITIVE IS INCREASING LEFT TO RIGHT OR               RND00065   000004
00111    66*    C         BOTTOM TO TOP.  NEGATIVE IS THE REVERSE).                          RND00066   000004
00112    67*          RANGE = GMAX - GMIN                                                    RND00067   000006
00112    68*    C         IF THE GIVEN MINIMUM AND MAXIMUM VALUES WERE EQUAL TO EACH         RND00068   000006
00112    69*    C         OTHER, OR IF THE GIVEN AXIS LENGTH IS ZERO, SET ALL RETURNED       RND00069   000006
00112    70*    C         VALUES TO ZERO AND RETURN TO THE CALLING PROGRAM, AS NOTHING       RND00070   000006
00112    71*    C         CAN BE DONE.                                                       RND00071   000006
00113    72*          IF ((RANGE .NE. 0.0) .AND. (GLENTH .GT. 0.0))  GO TO 300               RND00072   000010
00115    73*          RNDMIN = 0.0                                                           RND00073   000022
00116    74*          RNDMAX = 0.0                                                           RND00074   000023
00117    75*          RNDINC = 0.0                                                           RND00075   000024
00120    76*          RETURN                                                                 RND00076   000025
00120    77*    C         COMPUTE THE ABSOLUTE VALUE OF THE GIVEN INCREMENT.  THIS IS        RND00077   000025
00120    78*    C         THE RAW INCREMENT AS COMPUTED FROM THE GIVEN MINIMUM AND           RND00078   000025
00120    79*    C         MAXIMUM VALUES, AND THE LENGTH OF THE AXIS.                        RND00079   000025
00121    80*      300 GINC = ABS(RANGE / GLENTH)                                             RND00080   000031
00121    81*    C         COMPUTE THE POWER (THE CHARACTERISTIC) OF THE GIVEN INCREMENT.     RND00081   000031
00122    82*          PWR = ALOG10(GINC)                                                     RND00082   000034
00123    83*          IF (PWR .LT. 0.0)  PWR = PWR - 1.0                                     RND00083   000040
00125    84*          NPWR = PWR                                                             RND00084   000046
00125    85*    C         COMPUTE THE MANTISSA OF THE GIVEN INCREMENT.                       RND00085   000046
```

```
00126    86*       350 FMNTIS = GINC / 10.0**NPWR                                    RND00086    000056
00126    87*     C       FIND THE FIRST STANDARD INCREMENT MANTISSA (IN THE TABLE)   RND00087    000056
00126    88*     C       WHICH IS EQUAL TO OR GREATER THAN THE COMPUTED RAW MANTISSA. RND00088    000056
00127    89*           DO 500  I=1,NMANTS                                            RND00089    000070
00132    90*           IF (FMNTIS .GT. STDINC(I))  GO TO 500                         RND00090    000070
00132    91*     C       COMBINE THE TABLE VALUE MANTISSA AND THE COMPUTED POWER     RND00091    000070
00132    92*     C       (CHARACTERISTIC) TO GET THE COMPUTED ROUND INCREMENT.       RND00092    000070
00134    93*           RNDINC = STDINC(I) * (10.0**NPWR)                             RND00093    000074
00135    94*           K = I                                                         RND00094    000077
00136    95*           GO TO 600                                                     RND00095    000101
00137    96*       500 CONTINUE                                                      RND00096    000105
00141    97*       550 K = K + 1                                                     RND00097    000105
00141    98*     C       COMBINE THE TABLE VALUE MANTISSA AND THE COMPUTED POWER     RND00098    000105
00141    99*     C       (CHARACTERISTIC) TO GET THE COMPUTED ROUND INCREMENT.       RND00099    000105
00142   100*           RNDINC = STDINC(K) * (10.0**NPWR)                             RND00100    000107
00142   101*     C       THE STATEMENTS FROM HERE THROUGH STATEMENT 750 SELECT A NEW RND00101    000107
00142   102*     C       MINIMUM VALUE, WHICH IS AN INTEGRAL MULTIPLE OF THE ROUND   RND00102    000107
00142   103*     C       INCREMENT AND IS THE FIRST SUCH VALUE EQUAL TO OR LESS THAN RND00103    000107
00142   104*     C       THE GIVEN MINIMUM (OR THE FIRST VALUE GREATER THAN THE GIVEN RND00104   000107
00142   105*     C       MINIMUM IF THE GIVEN MINIMUM IS GREATER THAN THE GIVEN      RND00105    000107
00142   106*     C       MAXIMUM).  THE SIGN OF THE INCREMENT IS CHANGED IF NECESSARY RND00106   000107
00142   107*     C       TO SHOW PROPER DIRECTION.                                   RND00107    000107
00143   108*       600 MINT = GMIN / RNDINC                                          RND00108    000117
00144   109*           TEMP = MINT                                                   RND00109    000127
00145   110*           RNDMIN = TEMP * RNDINC                                        RND00110    000132
00146   111*           PDIFF = ABS((RNDMIN - GMIN) / RANGE)                          RND00111    000135
00147   112*           IF (GMIN .LT. 0.0)  GO TO 700                                 RND00112    000141
00151   113*           IF (RANGE .GT. 0.0)  GO TO 800                                RND00113    000144
00153   114*           IF (PDIFF .GT. 0.0001)  RNDMIN = RNDMIN + RNDINC              RND00114    000147
00155   115*           RNDINC = -RNDINC                                              RND00115    000156
00156   116*           GO TO 800                                                     RND00116    000160
00157   117*       700 IF (RANGE .GT. 0.0)  GO TO 750                                RND00117    000162
00161   118*           RNDINC = -RNDINC                                              RND00118    000164
00162   119*           GO TO 800                                                     RND00119    000166
00163   120*       750 IF (PDIFF .GT. 0.0001)  RNDMIN = RNDMIN - RNDINC              RND00120    000170
00163   121*     C       THE STATEMENTS FROM HERE THROUGH STATEMENT 900 SELECT A NEW RND00121    000170
00163   122*     C       MAXIMUM VALUE WHICH IS AN INTEGRAL MULTIPLE OF THE ROUND    RND00122    000170
00163   123*     C       INCREMENT AND IS THE FIRST SUCH VALUE GREATER THAN THE GIVEN RND00123    000170
00163   124*     C       MAXIMUM (OR THE FIRST VALUE LESS THAN THE GIVEN MAXIMUM IF  RND00124    000170
00163   125*     C       THE GIVEN MAXIMUM IS LESS THAN THE GIVEN MINIMUM).          RND00125    000170
00163   126*     C                                                                   RND00126    000170
00163   127*     C       IN A FEW CASES, THE PROCEDURE WILL RESULT IN ROUND MINIMUM  RND00127    000170
00163   128*     C       AND MAXIMUM VALUES WHICH DO NOT INCLUDE THE GIVEN MAXIMUM   RND00128    000170
00163   129*     C       VALUE.  IN SUCH INSTANCES, TRANSFER IS MADE TO              RND00129    000170
00163   130*     C       STATEMENT 550 TO SELECT THE NEXT HIGHER STANDARD INCREMENT  RND00130    000170
00163   131*     C       MANTISSA, AND THE PROCESS IS REPEATED WITH A NEW ROUND      RND00131    000170
00163   132*     C       INCREMENT TO GET NEW ROUND MINIMUM AND MAXIMUM VALUES.      RND00132    000170
00163   133*     C       IF THE TABLE OF STANDARD INCREMENT MANTISSAS HAS BEEN       RND00133    000170
00163   134*     C       EXHAUSTED, THE CHARACTERISTIC MUST BE INCREASED BY 1, AND   RND00134    000170
00163   135*     C       THE SCAN OF STANDARD INCREMENTS WILL START OVER AT THE      RND00135    000170
00163   136*     C       BEGINNING OF THE TABLE.                                     RND00136    000170
00165   137*       800 TMAX = RNDMIN  +  (RNDINC * GLENTH)                           RND00137    000177
00166   138*           IF (RANGE .LT. 0.0)  GO TO 850                                RND00138    000202
00170   139*           IF (TMAX .GE. GMAX)  GO TO 900                                RND00139    000205
00172   140*       830 IF (ABS((GMAX - TMAX) / RANGE) .LE. 0.0001)  GO TO 900        RND00140    000211
00174   141*           IF (K .LT. NMANTS)  GO TO 550                                 RND00141    000217
00174   142*     C       RAISE THE CHARACTERISTIC BY 1 AND GO BACK TO RECOMPUTE A NEW RND00142    000217
```

```
00174    143*     C          ROUND INCREMENT.
00176    144*              NPWR = NPWR + 1
00177    145*              GO TO 350
00200    146*        850 IF (TMAX .GT. GMAX)  GO TO 830
00202    147*        900 RNDMAX = TMAX
00203    148*              RETURN
00204    149*              END
```

RND00143	000217
RND00144	000223
RND00145	000226
RND00146	000230
RND00147	000234
RND00148	000235
RND00149	000300

GIVEN LIMITS: .2063310+03 .2220620+03 GIVEN AXIS LENGTH: 8.00
COMPUTED ROUND LIMITS: .2050000+03 .2250000+03 COMPUTED ROUND INCREMENT: .2500000+01

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.2050000+03 .2075000+03 .2100000+03 .2125000+03 .2150000+03 .2175000+03 .2200000+03 .2225000+03
.2250000+03

GIVEN LIMITS: .1310000+06 -.2600000+05 GIVEN AXIS LENGTH: 11.00
COMPUTED ROUND LIMITS: .1350000+06 -.3000000+05 COMPUTED ROUND INCREMENT: -.1500000+05

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.1350000+06 .1200000+06 .1050000+06 .9000000+05 .7500000+05 .6000000+05 .4500000+05 .3000000+05
.1500000+05 .0000000 -.1500000+05 -.3000000+05

GIVEN LIMITS: -.8300000-02 -.5560000-04 GIVEN AXIS LENGTH: 4.60
COMPUTED ROUND LIMITS: -.1000000-01 .1500000-02 COMPUTED ROUND INCREMENT: .2500000-02

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

-.1000000-01 -.7500000-02 -.5000000-02 -.2500000-02 .0000000

GIVEN LIMITS: .4961121+05 .5007621+05 GIVEN AXIS LENGTH: 15.00
COMPUTED ROUND LIMITS: .4960000+05 .5020000+05 COMPUTED ROUND INCREMENT: .4000000+02

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.4960000+05 .4964000+05 .4968000+05 .4972000+05 .4976000+05 .4980000+05 .4984000+05 .4988000+05
.4992000+05 .4996000+05 .5000000+05 .5004000+05 .5008000+05 .5012000+05 .5016000+05 .5020000+05

GIVEN LIMITS: .2237670+04 .1403355+03 GIVEN AXIS LENGTH: 6.00
COMPUTED ROUND LIMITS: .2400000+04 .0000000 COMPUTED ROUND INCREMENT: -.4000000+03

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.2400000+04 .2000000+04 .1600000+04 .1200000+04 .8000000+03 .4000000+03 .0000000

GIVEN LIMITS: .5631400+03 .7714500+03 GIVEN AXIS LENGTH: 8.20
COMPUTED ROUND LIMITS: .5400000+03 .7860000+03 COMPUTED ROUND INCREMENT: .3000000+02

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.5400000+03 .5700000+03 .6000000+03 .6300000+03 .6600000+03 .6900000+03 .7200000+03 .7500000+03
.7800000+03

GIVEN LIMITS: .5631400+03 .7714500+03 GIVEN AXIS LENGTH: 9.00
COMPUTED ROUND LIMITS: .5500000+03 .7750000+03 COMPUTED ROUND INCREMENT: .2500000+02

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.5500000+03 .5750000+03 .6000000+03 .6250000+03 .6500000+03 .6750000+03 .7000000+03 .7250000+03
.7500000+03 .7750000+03

GIVEN LIMITS: .5631400+03 .7714500+03 GIVEN AXIS LENGTH: 10.00
COMPUTED ROUND LIMITS: .5500000+03 .8000000+03 COMPUTED ROUND INCREMENT: .2500000+02

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.5500000+03 .5750000+03 .6000000+03 .6250000+03 .6500000+03 .6750000+03 .7000000+03 .7250000+03
.7500000+03 .7750000+03 .8000000+03

GIVEN LIMITS: .5631400+03 .7714500+03 GIVEN AXIS LENGTH: 11.00
COMPUTED ROUND LIMITS: .5600000+03 .7800000+03 COMPUTED ROUND INCREMENT: .2000000+02

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.5600000+03 .5800000+03 .6000000+03 .6200000+03 .6400000+03 .6600000+03 .6800000+03 .7000000+03
.7200000+03 .7400000+03 .7600000+03 .7800000+03

GIVEN LIMITS: .5631400+03 .7714500+03 GIVEN AXIS LENGTH: 12.00
COMPUTED ROUND LIMITS: .5600000+03 .8000000+03 COMPUTED ROUND INCREMENT: .2000000+02

RESULTING AXIS ANNOTATION AT INTEGRAL MULTIPLES OF THE BASIC UNITS OF AXIS LENGTH, STARTING AT ZERO LENGTH:

.5600000+03 .5800000+03 .6000000+03 .6200000+03 .6400000+03 .6600000+03 .6800000+03 .7000000+03
.7200000+03 .7400000+03 .7600000+03 .7800000+03 .8000000+03

C. Quick-look Printer Plots

Though printer lots are of low resolution by nature, they are intended not to take the place of high-resolution plots, but to give the data analyst something to look at while the final plots are being processed. They can serve to reduce customer pressure on turn-around time, and reduce customer desire to be physically present in the plotting room. But if they are to fully satisfy the user, the printer plots must be the same in every detail possible as the final plot. That is, the printer plot must be fully labeled, must have some form of grid, and must be numerically annotated using the same limits.

One exception to this is that in a production system, the physical size of the printer plots remains constant (usually being designed to fill one full page), regardless of variations in size of the high-resolution plots. The reasons are two-fold: First, any reduction in size below one page will further reduce the resolution to an unacceptable level, and second, what little advantage might be gained by expansion is more than offset by the increased complexity of the program. The only significant drawback to constant-size printer plot is an aesthetic one: Intermediate numerical annotation points do not always fall at the same increments as on the high-resolution plot. Remember, though, that the information conveyed is the same, and the printer plot's purpose is only to serve the user until the final plots are delivered a few hours later.

There are a great many printer plot styles in existence, serving a variety of special purposes, but generally speaking, the production graphics system should confine itself to a straightforward rectangular format. The data itself is then represented by appropriately placed symbols within the printed grid. Both the utility and simplicity of such a system are highly attractive.

It should become clear as one undertakes the design of printer plot software that no part of the plot can be printed out until all of the data has been processed, since there is no way of knowing until completion whether a given line of print contains all of its required data points. (There are exceptions, but the program must be able to plot all data--not just the exceptions.) In other words, the entire plot must be stored somewhere for the duration of the plotting process. This does not mean that all of the data must be stored, but that an array of print characters must be maintained and updated as the plotting progresses.

As an example, assume that the printer plot is to consist of a rectangle that measures 50 by 100 increments. This will require 51 lines by 101 print positions, for a total array of 5151 characters. The simplistic approach would be to set aside an array of 5151 words and update the word corresponding to the plotted point. But this is needlessly wasteful of core storage since each word has the potential of holding several characters. If, for example, each word can hold six characters, then each line of 101 characters can be contained in 17 words, and the 51 lines then occupy a total of only 867 words--a net savings of 4284 words.

This method does require character manipulation and masking operations by the software, but they are both simple and straightforward, lending themselves to either Boolean functions, or better, to the FLD bit manipulation function where available. For that matter, they can even be accomplished by integer division/multiplication processes, though not nearly so conveniently.

A representative printer plot program which incorporates all of the desired features is included at the end of this section, along with some samples of its output. While some may find application for the program in its present form, the actual purpose of its inclusion here is simply to demonstrate a working program which has all of the features required in a production graphics operation. This sample routine was prepared for use on a UNIVAC 1108 with 36-bit words and UNIVAC character codes, so use on machines with different word lengths or different character codes would require appropriate changes.

```
SUBROUTINE PRNPLT    ENTRY POINT 001344
           PLOTPT    ENTRY POINT 001373
           PRINTP    ENTRY POINT 001402

STORAGE USED: CODE(1) 001405; DATA(0) 002056; BLANK COMMON(2) 000000

EXTERNAL REFERENCES (BLOCK, NAME)

 0003   XPII
 0004   NWDUS
 0005   NIO2S
 0006   NIO1S
 0007   NERR2S
 0010   NIO3S
 0011   NERR3S

STORAGE ASSIGNMENT  (BLOCK, TYPE, RELATIVE LOCATION, NAME)

 0001   000047 161G      0001   000070 172G      0001   000111 203G      0001   000160 224G      0001   000172 233G
 0001   000174 236G      0001   000243 256G      0001   000244 261G      0001   000256 270G      0001   000257 273G
 0001   000300 305G      0001   000332 314G      0001   000363 325G      0001   000377 335G      0001   000412 341G
 0001   000571 405G      0001   000604 414G      0001   000605 416G      0001   000655 441G      0001   000671 451G
 0001   000705 461G      0001   000741 477G      0001   000761 506G      0001   001015 524G      0001   001030 536G
 0001   001051 546G      0001   001106 565G      0001   001134 603G      0001   000007 6120L     0001   000015 6130L
 0001   000023 6140L     0001   000030 6200L     0001   001221 632G      0001   000063 6350L     0001   001235 641G
 0001   000104 6450L     0001   001251 650G      0001   001275 662G      0001   001276 664G      0001   000125 7000L
 0001   000153 7030L     0001   000166 7050L     0001   000241 7100L     0001   000422 7390L     0000   001715 7701F
 0000   001716 7731F     0000   001727 7741F     0001   000622 7750L     0001   000635 7760L     0000   001730 7761F
 0001   000646 7770L     0000   001732 7771F     0001   000662 7780L     0000   001735 7781F     0001   000676 7790L
 0000   001740 7791F     0001   000711 7795L     0001   000716 7800L     0001   000735 7820L     0000   001743 7831F
 0001   000766 7840L     0000   001746 7851F     0001   001021 7860L     0000   001752 7881F     0001   001056 7890L
 0000   001755 7901F     0001   001112 7910L     0001   001121 7920L     0000   001762 7935F     0001   001201 7940L
 0001   001212 7945L     0001   001226 7950L     0001   001242 7955L     0001   001255 7960L     0000 I 001653 I
 0000 I 001670 IA        0000 I 001677 IAA       0000 I 001671 IB        0000 I 001543 IBLANK    0000 I 001700 IC
 0000 I 001660 ICLASS    0000 I 001655 IFIVEM    0000 I 001657 ILIST     0000 I 001557 IMINUS    0000   002014 INJPS
 0000 I 001551 IPLUS     0000 I 001654 ISIXBK    0000 I 001656 ISIXM     0000 I 001565 ISTAR     0000 I 001674 ITEMP
 0000 I 001663 ITW       0000 I 000000 IWORD     0000 I 001664 IXW       0000 I 001665 IYW       0000 I 001676 J
 0000 I 001704 JA        0000 I 001705 JB        0000 I 001573 JCLASS    0000 I 001675 JTEMP     0000 I 001623 JYLABL
 0000 I 001701 K         0000 I 001603 KCLASS    0000 L 001661 LMIRX     0000 L 001662 LMIRY     0000 I 001702 MASK
 0000 I 001703 MASKBK    0000 I 001672 N         0000 I 001673 NA        0000 R 001615 XANNOT    0000 R 001714 XINCMP
 0000 R 001713 XINCT     0000 R 001666 XPRFAC    0000 R 001712 XPRINC    0000 R 001711 YANNOT    0000 R 001710 YINCMP
 0000 R 001707 YINCT     0000 R 001667 YPRFAC    0000 R 001706 YPRINC

00101     1*          SUBROUTINE PRNPLT (XMIN, XMAX, YMIN, YMAX, ITITLE, IXLABL,      PRN00001   00000
00101     2*        1                    IYLABL, MIRX, MIRY, KLASS)                   PRN00002   00000
00101     3*     C                                                                    PRN00003   00000
00101     4*     C                                                                    PRN00004   00000
```

214

```
00101     5*    C                                                                    PRN00005   000000
00101     6*    C     PRINTER PLOT ROUTINE FOR PRODUCTION GRAPHICS, MAY 1969.        PRN00006   000000
00101     7*    C     WILLIAM D. JOHNSTON, NR-AD-S, WHITE SANDS MISSILE RANGE,       PRN00007   000000
00101     8*    C     NEW MEXICO  88002                                              PRN00008   000000
00101     9*    C                                                                    PRN00009   000000
00101    10*    C                                                                    PRN00010   000000
00101    11*    C     THE PLOTS GENERATED BY THIS ROUTINE ARE FULLY LABELED          PRN00011   000000
00101    12*    C     AND NUMERICALLY ANNOTATED.  LABELS ARE PROPERLY CENTERED.      PRN00012   000000
00101    13*    C                                                                    PRN00013   000000
00101    14*    C                                                                    PRN00014   000000
00101    15*    C                                                                    PRN00015   000000
00101    16*    C                                                                    PRN00016   000000
00101    17*    C                                                                    PRN00017   000000
00101    18*    C     TO INITIALIZE THE PRNPLT PROGRAM, A CALL IS MADE AS            PRN00018   000000
00101    19*    C     FOLLOWS:                                                       PRN00019   000000
00101    20*    C                                                                    PRN00020   000000
00101    21*    C                                                                    PRN00021   000000
00101    22*    C     CALL PRNPLT (XMIN, XMAX, YMIN, YMAX, ITITLE, IXLABL,           PRN00022   000000
00101    23*    C                  IYLABL, MIRX, MIRY, KLASS)                        PRN00023   000000
00101    24*    C                                                                    PRN00024   000000
00101    25*    C                                                                    PRN00025   000000
00101    26*    C     WHERE:                                                         PRN00026   000000
00101    27*    C                                                                    PRN00027   000000
00101    28*    C           XMIN      IS THE VALUE OF THE X-AXIS AT THE LEFT         PRN00028   000000
00101    29*    C                     END.                                           PRN00029   000000
00101    30*    C                                                                    PRN00030   000000
00101    31*    C           XMAX      IS THE VALUE OF THE X-AXIS AT THE RIGHT        PRN00031   000000
00101    32*    C                     END.                                           PRN00032   000000
00101    33*    C                                                                    PRN00033   000000
00101    34*    C           YMIN      IS THE VALUE OF THE Y-AXIS AT THE BOTTOM       PRN00034   000000
00101    35*    C                     END.                                           PRN00035   000000
00101    36*    C                                                                    PRN00036   000000
00101    37*    C           YMAX      IS THE VALUE OF THE Y-AXIS AT THE TOP          PRN00037   000000
00101    38*    C                     END.                                           PRN00038   000000
00101    39*    C                                                                    PRN00039   000000
00101    40*    C                     NOTE:  XMIN MAY BE LARGER THAN XMAX AND        PRN00040   000000
00101    41*    C                            YMIN MAY BE LARGER THAN YMAX.           PRN00041   000000
00101    42*    C                            THESE ARE SIMPLY THE VALUES AT          PRN00042   000000
00101    43*    C                            THE END POINTS OF THE AXES AND          PRN00043   000000
00101    44*    C                            THERE ARE NO RESTRICTIONS ON            PRN00044   000000
00101    45*    C                            THEIR VALUES.                           PRN00045   000000
00101    46*    C                                                                    PRN00046   000000
00101    47*    C           ITITLE    IS A 4-WORD ARRAY CONTAINING THE TITLE         PRN00047   000000
00101    48*    C                     TO BE WRITTEN AT THE TOP OF THE PLOT.          PRN00048   000000
00101    49*    C                     THE TITLE MUST BE LEFT-ADJUSTED IN             PRN00049   000000
00101    50*    C                     THE ARRAY, AND UNUSED CHARACTERS TO THE        PRN00050   000000
00101    51*    C                     RIGHT MUST BE BLANK-FILLED.                    PRN00051   000000
00101    52*    C                                                                    PRN00052   000000
00101    53*    C           IXLABL    IS A 4-WORD ARRAY CONTAINING THE LABEL         PRN00053   000000
00101    54*    C                     TO BE WRITTEN ALONG THE X-AXIS OF THE          PRN00054   000000
00101    55*    C                     PLOT.  THE LABEL MUST BE LEFT-ADJUSTED         PRN00055   000000
00101    56*    C                     IN THE ARRAY, AND UNUSED CHARACTERS TO         PRN00056   000000
00101    57*    C                     THE RIGHT MUST BE BLANK-FILLED.                PRN00057   000000
00101    58*    C                                                                    PRN00058   000000
00101    59*    C           IYLABL    IS A 4-WORD ARRAY CONTAINING THE LABEL         PRN00059   000000
00101    60*    C                     TO BE WRITTEN ALONG THE Y-AXIS OF THE          PRN00060   000000
00101    61*    C                     PLOT.  THE LABEL MUST BE LEFT-ADJUSTED         PRN00061   000000
```

```
00101    02*    C                    IN THE ARRAY, AND UNUSED CHARACTERS TO      PRN00062  000000
00101    03*    C                    THE RIGHT MUST BE BLANK-FILLED.             PRN00063  000000
00101    04*    C                                                                PRN00064  000000
00101    05*    C          MIRX      IS AN INTEGER FLAG TO INDICATE WHETHER      PRN00065  000000
00101    06*    C                    OR NOT TO PLOT THE MIRROR IMAGE OF THE      PRN00066  000000
00101    07*    C                    X-COORDINATE.  WHEN SET TO 0, THE DATA      PRN00067  000000
00101    08*    C                    IS PLOTTED NORMALLY.  WHEN SET TO           PRN00068  000000
00101    09*    C                    NON-ZERO, THE MIRROR IMAGE OF THE           PRN00069  000000
00101    70*    C                    X-COORDINATE IS PLOTTED.                    PRN00070  000000
00101    71*    C                                                                PRN00071  000000
00101    72*    C          MIRY      IS AN INTEGER FLAG TO INDICATE WHETHER      PRN00072  000000
00101    73*    C                    OR NOT TO PLOT THE MIRROR IMAGE OF THE      PRN00073  000000
00101    74*    C                    Y-COORDINATE.  WHEN SET TO 0, THE DATA      PRN00074  000000
00101    75*    C                    IS PLOTTED NORMALLY.  WHEN SET TO           PRN00075  000000
00101    76*    C                    NON-ZERO, THE MIRROR IMAGE OF THE           PRN00076  000000
00101    77*    C                    Y-COORDINATE IS PLOTTED.                    PRN00077  000000
00101    78*    C                                                                PRN00078  000000
00101    79*    C          KLASS     IS A ONE-CHARACTER HOLLERITH CODE,          PRN00079  000000
00101    80*    C                    LEFT-ADJUSTED IN THE WORD, REPRESENTING     PRN00080  000000
00101    81*    C                    THE SECURITY CLASSIFICATION TO BE           PRN00081  000000
00101    82*    C                    PRINTED ON THE PLOT.                        PRN00082  000000
00101    83*    C                    = 'U', UNCLASSIFIED                         PRN00083  000000
00101    84*    C                    = 'C', CONFIDENTIAL                         PRN00084  000000
00101    85*    C                    = 'S', SECRET                               PRN00085  000000
00101    86*    C                    = 'T', TOP SECRET                           PRN00086  000000
00101    87*    C                    = BLANK, NO CLASSIFICATION WILL BE          PRN00087  000000
00101    88*    C                           PRINTED.                             PRN00088  000000
00101    89*    C                                                                PRN00089  000000
00101    90*    C                                                                PRN00090  000000
00101    91*    C                                                                PRN00091  000000
00101    92*    C                                                                PRN00092  000000
00101    93*    C                                                                PRN00093  000000
00101    94*    C     AFTER INITIALIZATION, THE DATA IS PLOTTED ON A POINT-BY-    PRN00094  000000
00101    95*    C     POINT BASIS THROUGH THE FOLLOWING CALL:                    PRN00095  000000
00101    96*    C                                                                PRN00096  000000
00101    97*    C                                                                PRN00097  000000
00101    98*    C     CALL PLOTPT (X, Y)                                         PRN00098  000000
00101    99*    C                                                                PRN00099  000000
00101   100*    C                                                                PRN00100  000000
00101   101*    C     WHERE X AND Y ARE THE X AND Y PARAMETERS, RESPECTIVELY.    PRN00101  000000
00101   102*    C                                                                PRN00102  000000
00101   103*    C                                                                PRN00103  000000
00101   104*    C                                                                PRN00104  000000
00101   105*    C                                                                PRN00105  000000
00101   106*    C                                                                PRN00106  000000
00101   107*    C     ONCE ALL DATA HAS BEEN PLOTTED, THE PLOT IS PRINTED OUT    PRN00107  000000
00101   108*    C     THROUGH THE FOLLOWING CALL:                                PRN00108  000000
00101   109*    C                                                                PRN00109  000000
00101   110*    C                                                                PRN00110  000000
00101   111*    C     CALL PRINTP                                                PRN00111  000000
00101   112*    C                                                                PRN00112  000000
00101   113*    C                                                                PRN00113  000000
00101   114*    C                                                                PRN00114  000000
00101   115*    C                                                                PRN00115  000000
00101   116*    C                                                                PRN00116  000000
00101   117*    C     THE ABOVE PROCEDURES MAY BE REPEATED AS MANY TIMES AS      PRN00117  000000
00101   118*    C     DESIRED FOR ADDITIONAL PRINTER PLOTS.                      PRN00118  000000
```

```
00101   119*     C                                                                                  PRN00119   000000
00101   120*     C                                                                                  PRN00120   000000
00101   121*     C                                                                                  PRN00121   000000
00101   122*     C                                                                                  PRN00122   000000
00101   123*     C                                                                                  PRN00123   000000
00101   124*     C                                                                                  PRN00124   000000
00101   125*     C                                                                                  PRN00125   000000
00101   126*     C                                                                                  PRN00126   000000
00101   127*     C                                                                                  PRN00127   000000
00103   128*           IMPLICIT LOGICAL (L)                                                         PRN00128   000000
00103   129*     C                                                                                  PRN00129   000000
00104   130*           DIMENSION  IWORD(51,17), IBLANK(6), IPLUS(6), IMINUS(6), ISTAR(6)            PRN00130   000000
00105   131*           DIMENSION  JCLASS(4,2), KCLASS(5,2), XANNOT(6), JYLABL(24)                   PRN00131   000000
00106   132*           DIMENSION  IXLABL(4), IYLABL(4), ITITLE(4)                                   PRN00132   000000
00106   133*     C                                                                                  PRN00133   000000
00106   134*     C                                                                                  PRN00134   000000
00106   135*     C                                                                                  PRN00135   000000
00107   136*           DATA  (IBLANK(I), I=1,6) /0007777777777, 0770077777777,                      PRN00136   000000
00107   137*          1                          0777700777777, 0777777007777,                      PRN00137   000000
00107   138*          2                          0777777770077, 0777777777700/                      PRN00138   000000
00107   139*     C                                                                                  PRN00139   000000
00111   140*           DATA  (IMINUS(I), I=1,6) /0410000000000, 0004100000000,                      PRN00140   000000
00111   141*          1                          0000041000000, 0000000410000,                      PRN00141   000000
00111   142*          2                          0000000004100, 0000000000041/                      PRN00142   000000
00111   143*     C                                                                                  PRN00143   000000
00113   144*           DATA  (IPLUS(I), I=1,6) /0420000000000, 0004200000000,                       PRN00144   000000
00113   145*          1                          0000042000000, 0000000420000,                      PRN00145   000000
00113   146*          2                          0000000004200, 0000000000042/                      PRN00146   000000
00113   147*     C                                                                                  PRN00147   000000
00115   148*           DATA  (ISTAR(I), I=1,6) /0500000000000, 0005000000000,                       PRN00148   000000
00115   149*          1                          0000050000000, 0000000500000,                      PRN00149   000000
00115   150*          2                          0000000005000, 0000000000050/                      PRN00150   000000
00115   151*     C                                                                                  PRN00151   000000
00117   152*           DATA  ISIXBK /6H      /,  IFIVEM /6H----- /,  ISIXM /6H------/             PRN00152   000000
00117   153*     C                                                                                  PRN00153   000000
00123   154*           DATA  ILIST /6/                                                              PRN00154   000000
00123   155*     C                                                                                  PRN00155   000000
00123   156*     C                                                                                  PRN00156   000000
00125   157*           DATA  (JCLASS(1,I), I=1,2) /6HUNCLAS, 6HSIFIED/,                             PRN00157   000000
00125   158*          1      (JCLASS(2,I), I=1,2) /6HCONFID, 6HENTIAL/,                            PRN00158   000000
00125   159*          2      (JCLASS(3,I), I=1,2) /6HSECRET, 6H      /,                            PRN00159   000000
00125   160*          3      (JCLASS(4,I), I=1,2) /6HTOP  S, 6HECRET /                            PRN00160   000000
00125   161*     C                                                                                  PRN00161   000000
00125   162*     C                                                                                  PRN00162   000000
00125   163*     C                                                                                  PRN00163   000000
00125   164*     C                                                                                  PRN00164   000000
00125   165*     C                                                                                  PRN00165   000000
00125   166*     C            THE STATEMENTS FROM HERE THROUGH STATEMENT 7400 ARE FOR               PRN00166   000000
00125   167*     C            INITIALIZATION OF THE PRINTER PLOT.                                   PRN00167   000000
00125   168*     C                                                                                  PRN00168   000000
00125   169*     C                                                                                  PRN00169   000000
00132   170*           ICLASS = 0                                                                   PRN00170   000000
00133   171*           IF (KLASS .NE. 'U')  GO TO 6120                                              PRN00171   000000
00135   172*           ICLASS = 1                                                                   PRN00172   000003
00136   173*           GO TO 6200                                                                   PRN00173   000005
00137   174*      6120 IF (KLASS .NE. 'C')  GO TO 6130                                              PRN00174   000007
00141   175*           ICLASS = 2                                                                   PRN00175   000011
```

```
00142   176*          GO TO 6200                                                    PRN00176   000013
00143   177*    6130  IF (KLASS .NE. 'S')  GO TO 6140                             PRN00177   000015
00145   178*          ICLASS = 3                                                    PRN00178   000017
00146   179*          GO TO 6200                                                    PRN00179   000021
00147   180*    6140  IF (KLASS .NE. 'T')  GO TO 6200                             PRN00180   000023
00151   181*          ICLASS = 4                                                    PRN00181   000025
00152   182*    6200  LMIRX = .FALSE.                                             PRN00182   000030
00153   183*          LMIRY = .FALSE.                                               PRN00183   000030
00154   184*          IF (MIRX .NE. 0)  LMIRX = .TRUE.                              PRN00184   000031
00156   185*          IF (MIRY .NE. 0)  LMIRY = .TRUE.                              PRN00185   000035
00160   186*          DO 6300  I=1,4                                                PRN00186   000047
00163   187*          ITW = 5 - I                                                   PRN00187   000047
00164   188*          IF (ITITLE(ITW) .NE. ' ')  GO TO 6350                         PRN00188   000052
00166   189*    6300  CONTINUE                                                      PRN00189   000061
00170   190*          ITW = 0                                                       PRN00190   000061
00171   191*    6350  DO 6400  I=1,4                                                PRN00191   000063
00174   192*          IXW = 5 - I                                                   PRN00192   000070
00175   193*          IF (IXLABL(IXW) .NE. ' ')  GO TO 6450                         PRN00193   000073
00177   194*    6400  CONTINUE                                                      PRN00194   000102
00201   195*          IXW = 0                                                       PRN00195   000102
00202   196*    6450  DO 6500  I=1,4                                                PRN00196   000104
00205   197*          IYW = 5 - I                                                   PRN00197   000111
00206   198*          IF (IYLABL(IYW) .NE. ' ')  GO TO 7000                         PRN00198   000114
00210   199*    6500  CONTINUE                                                      PRN00199   000123
00212   200*          IYW = 0                                                       PRN00200   000123
00212   201*    C               COMPUTE THE X AND Y AXIS PRINTER PLOT MULTIPLYING   PRN00201   000123
00212   202*    C               FACTORS.                                            PRN00202   000123
00213   203*     7000 XPRFAC = 100.0 / (XMAX - XMIN)                                PRN00203   000125
00214   204*          YPRFAC = 50.0 / (YMAX - YMIN)                                 PRN00204   000131
00214   205*    C               THE STATEMENTS FROM HERE TO STATEMENT 7100 TRANSFER THE PRN00205   000131
00214   206*    C               Y-AXIS PRINTER PLOT LABEL INTO THE ARRAY 'JYLABL'   PRN00206   000131
00214   207*    C               (DIMENSIONED BY 24), WHICH WILL BE PRINTED OUT      PRN00207   000131
00214   208*    C               VERTICALLY.  ONE CHARACTER OF THE LABEL IS PLACED IN THE PRN00208   000131
00214   209*    C               FIRST CHARACTER OF EACH WORD OF THE JYLABL ARRAY.  WHEN PRN00209   000131
00214   210*    C               THERE ARE FEWER THAN THE MAXIMUM OF 4 WORDS, THE LABEL PRN00210   000131
00214   211*    C               IS CENTERED IN THE ARRAY, WITH THE UNUSED WORDS ON EACH PRN00211   000131
00214   212*    C               END BLANKED OUT.                                    PRN00212   000131
00215   213*          IA = 1                                                        PRN00213   000136
00216   214*          IB = 3 * (4 - IYW)                                            PRN00214   000140
00217   215*          IF (IB .EQ. 0)  GO TO 7050                                    PRN00215   000144
00221   216*          IF (IYW .EQ. 0)  IB = 24                                      PRN00216   000146
00223   217*     7030 DO 7040  I=IA,IB                                              PRN00217   000153
00226   218*     7040 JYLABL(I) = ISIXBK                                            PRN00218   000160
00230   219*          IF (IB .EQ. 24)  GO TO 7100                                   PRN00219   000162
00232   220*     7050 DO 7080  I=1,IYW                                              PRN00220   000166
00235   221*          DO 7070  N=1,6                                                PRN00221   000174
00240   222*          IB = IB + 1                                                   PRN00222   000174
00241   223*          NA = (6-N) * 2                                                PRN00223   000177
00242   224*          ITEMP = IYLABL(I) / 8**NA                                     PRN00224   000203
00243   225*          JTEMP = ITEMP / 64                                            PRN00225   000214
00244   226*     7070 JYLABL(IB) = (ITEMP  -  (64 * JTEMP))  *  8**10               PRN00226   000216
00246   227*     7080 CONTINUE                                                      PRN00227   000226
00250   228*          IF (IB .GE. 24)  GO TO 7100                                   PRN00228   000226
00252   229*          IA = IB + 1                                                   PRN00229   000232
00253   230*          IB = 24                                                       PRN00230   000235
00254   231*          GO TO 7030                                                    PRN00231   000237
00254   232*    C                                                                   PRN00232   000237
```

218

```
00254    233*    C                                                                PRN00233  000237
00254    234*    C                                                                PRN00234  000237
00254    235*    C            THE STATEMENTS FROM STATEMENT 7100 THROUGH STATEMENT PRN00235  000237
00254    236*    C            7400 INITIALIZE THE PRINTER PLOT GRID.  THIS IS     PRN00236  000237
00254    237*    C            ACCOMPLISHED BY FIRST BLANKING OUT THE ENTIRE GRID, THEN PRN00237  000237
00254    238*    C            MASKING THE GRID PATTERN INTO THE ARRAY.            PRN00238  000237
00255    239*     7100 DO 7140  I=1,51                                            PRN00239  000244
00260    240*          DO 7120  J=1,17                                            PRN00240  000244
00263    241*     7120 IWORD(I,J) = ISIXBK                                        PRN00241  000244
00265    242*     7140 CONTINUE                                                   PRN00242  000257
00267    243*          DO 7180  I=1,51,10                                         PRN00243  000257
00272    244*          DO 7160  J=1,16                                            PRN00244  000257
00275    245*     7160 IWORD(I,J) = ISIXM                                         PRN00245  000257
00277    246*     7180 IWORD(I,17) = IFIVEN                                       PRN00246  000261
00301    247*          IAA = 5                                                    PRN00247  000267
00302    248*          IB = -4                                                    PRN00248  000271
00303    249*          IC = 106                                                   PRN00249  000273
00304    250*          DO 7400  IA=1,2                                            PRN00250  000300
00307    251*          IF (IA .EQ. 2)  IAA = 10                                   PRN00251  000306
00311    252*          IB = IB + 5                                                PRN00252  000312
00312    253*          IC = IC - 5                                                PRN00253  000315
00313    254*          DO 7390  K=IB,IC,10                                        PRN00254  000323
00316    255*          ITEMP = (K-1) / 6                                          PRN00255  000332
00317    256*          J = ITEMP + 1                                              PRN00256  000336
00320    257*          ITEMP = ITEMP * 6                                          PRN00257  000340
00321    258*          N = K - ITEMP                                              PRN00258  000343
00322    259*          MASK = IPLUS(N)                                            PRN00259  000345
00323    260*          MASKBK = IBLANK(N)                                         PRN00260  000350
00324    261*          DO 7300  I=1,51,IAA                                        PRN00261  000363
00327    262*     7300 IWORD(I,J) = OR( AND(IWORD(I,J), MASKBK), MASK )           PRN00262  000363
00331    263*          IF (IA .EQ. 2)  GO TO 7390                                 PRN00263  000370
00333    264*          MASK = IMINUS(N)                                           PRN00264  000372
00334    265*          DO 7380  JA=2,47,5                                         PRN00265  000377
00337    266*          JB = JA + 3                                                PRN00266  000402
00340    267*          DO 7360  I=JA,JB                                           PRN00267  000405
00343    268*     7360 IWORD(I,J) = OR( AND(IWORD(I,J), MASKBK), MASK )           PRN00268  000412
00345    269*     7380 CONTINUE                                                   PRN00269  000425
00347    270*     7390 CONTINUE                                                   PRN00270  000425
00351    271*     7400 CONTINUE                                                   PRN00271  000425
00353    272*          RETURN                                                     PRN00272  000425
00353    273*    C                                                                PRN00273  000425
00353    274*    C                                                                PRN00274  000425
00353    275*    C                                                                PRN00275  000425
00353    276*    C                                                                PRN00276  000425
00353    277*    C            THE STATEMENTS FROM HERE TO STATEMENT 7700          PRN00277  000425
00353    278*    C            PLOT THE POINT ON THE PRINTER PLOT.  THESE STATEMENTS PRN00278  000425
00353    279*    C            ARE EXECUTED ONCE FOR EACH PLOTTED POINT.           PRN00279  000425
00353    280*    C                                                                PRN00280  000425
00353    281*    C                                                                PRN00281  000425
00354    282*          ENTRY PLOTPT (X, Y)                                        PRN00282  000430
00354    283*    C                                                                PRN00283  000430
00354    284*    C                                                                PRN00284  000430
00354    285*    C            COMPUTE THE COORDINATES (TAKING THE MIRROR IMAGE IF PRN00285  000430
00354    286*    C            REQUIRED) OF THE PLOTTED POINT ON THE PRINTER PLOT.  THE PRN00286  000430
00354    287*    C            COORDINATES ARE (K,I), WHERE K IS THE NUMBER OF SPACES PRN00287  000430
00354    288*    C            FROM LEFT TO RIGHT, AND I IS THE LINE NUMBER FROM TOP TO PRN00288  000430
00354    289*    C            BOTTOM.                                             PRN00289  000430
```

219

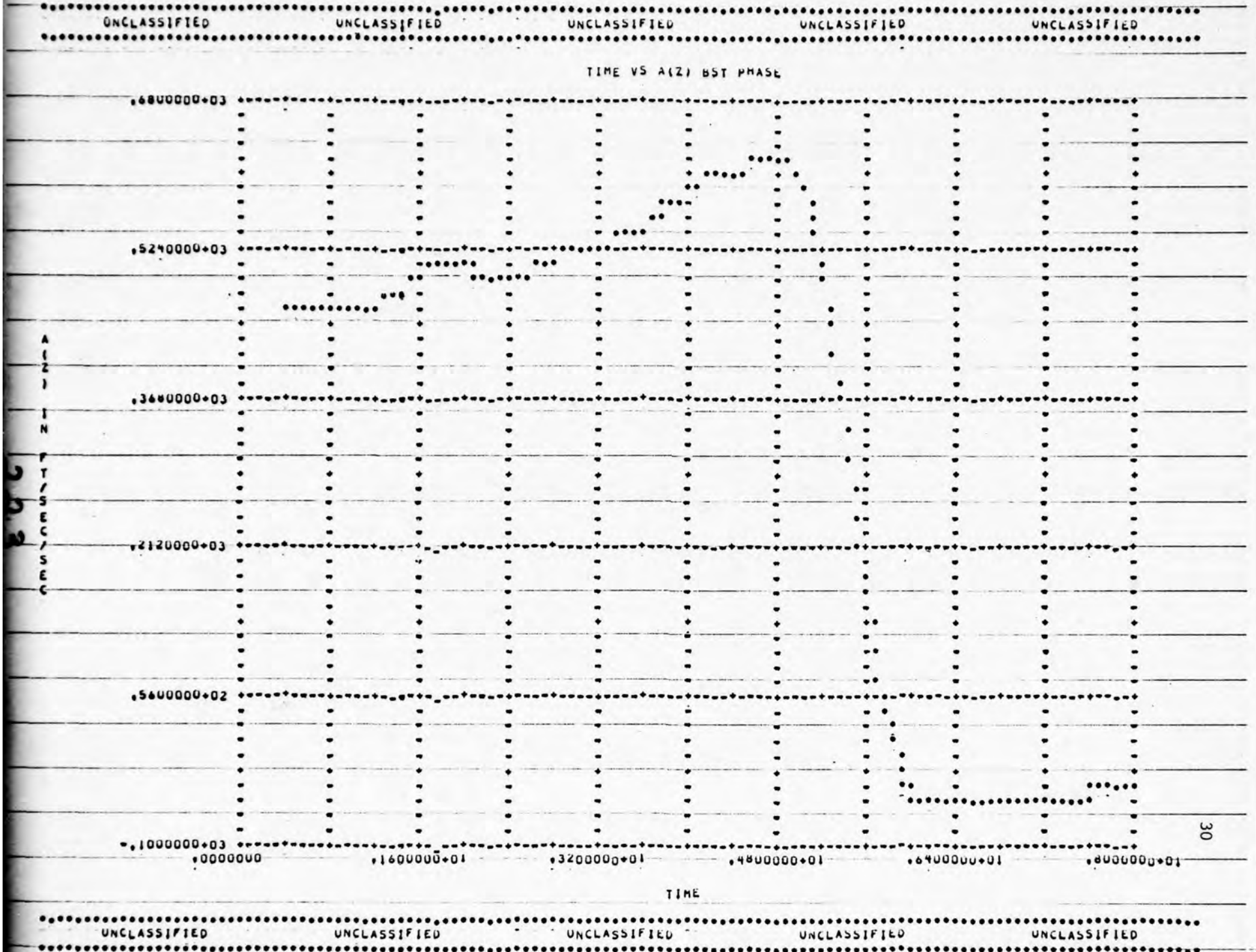
```
00356   290*          K = (X - XMIN) * XPRFAC  +  1.5                                  PRN00290   000430
00357   291*          IF ((K .LE. 0) .OR. (K .GT. 101))  RETURN                        PRN00291   000443
00361   292*          I = (Y - YMIN) * YPRFAC  +  1.5                                  PRN00292   000463
00362   293*          IF ((I .LE. 0) .OR. (I .GT. 51))  RETURN                         PRN00293   000476
00364   294*          IF (LMIRX)  K = 102 - K                                          PRN00294   000516
00366   295*          IF (.NOT. LMIRY)  I = 52 - I                                     PRN00295   000523
00366   296*    C             J IS THE NUMBER OF THE WORD (FROM LEFT TO RIGHT, 1 TO    PRN00296   000523
00366   297*    C             17) IN WHICH THE PLOTTED POINT LIES.                     PRN00297   000523
00366   298*    C             N IS THE ACTUAL CHARACTER (OR 6-BIT BYTE) WITHIN THE     PRN00298   000523
00366   299*    C             WORD, THAT THE PLOTTED POINT OCCUPIES.                   PRN00299   000523
00370   300*          ITEMP = (K-1) / 6                                                PRN00300   000530
00371   301*          J = ITEMP + 1                                                    PRN00301   000534
00372   302*          ITEMP = ITEMP * 6                                                PRN00302   000536
00373   303*          N = K - ITEMP                                                    PRN00303   000543
00373   304*    C             MASK IN AN ASTERISK FOR THE PLOTTED POINT.               PRN00304   000543
00374   305*          IWORD(I,J) = OR( AND(IWORD(I,J), IBLANK(N)), ISTAR(N))           PRN00305   000545
00375   306*          RETURN                                                           PRN00306   000553
00375   307*    C                                                                      PRN00307   000553
00375   308*    C                                                                      PRN00308   000553
00375   309*    C                                                                      PRN00309   000553
00375   310*    C                                                                      PRN00310   000553
00375   311*    C             THE STATEMENTS FROM HERE TO STATEMENT 8500 WRITE         PRN00311   000553
00375   312*    C             OUT THE PRINTER PLOT.                                    PRN00312   000553
00375   313*    C                                                                      PRN00313   000553
00375   314*    C                                                                      PRN00314   000553
00375   315*    C                                                                      PRN00315   000553
00376   316*          ENTRY PRINTP                                                     PRN00316   000556
00376   317*    C                                                                      PRN00317   000556
00376   318*    C                                                                      PRN00318   000556
00377   319*          WRITE (ILIST, 7701)                                              PRN00319   000556
00401   320*     7701 FORMAT (1H1)                                                     PRN00320   000563
00401   321*    C             TRANSFER TO STATEMENT 7750 IF NO CLASSIFICATION LABEL IS PRN00321   000563
00401   322*    C             TO BE WRITTEN.                                           PRN00322   000563
00402   323*          IF (ICLASS .EQ. 0)  GO TO 7750                                   PRN00323   000563
00402   324*    C             STORE THE CLASSIFICATION IN FIVE LOCATIONS AND WRITE IT  PRN00324   000563
00402   325*    C             OUT.                                                     PRN00325   000563
00404   326*          DO 7720  I=1,5                                                   PRN00326   000565
00407   327*          KCLASS(I,1) = JCLASS(ICLASS,1)                                   PRN00327   000571
00410   328*     7720 KCLASS(I,2) = JCLASS(ICLASS,2)                                   PRN00328   000572
00412   329*          WRITE (ILIST, 7731)  ((KCLASS(I,J), J=1,2), I=1,5)               PRN00329   000575
00423   330*     7731 FORMAT (1H , 130(1H*), /, 1H , 7X, 2A6, 4(14X, 2A6), /, 1H ,     PRN00330   000614
00423   331*         1        130(1H*) )                                              PRN00331   000614
00424   332*          WRITE (ILIST, 7741)                                              PRN00332   000614
00426   333*     7741 FORMAT (1H )                                                     PRN00333   000622
00426   334*    C             TRANSFER TO STATEMENT 7800 IF THERE IS NO TITLE.         PRN00334   000622
00427   335*     7750 IF (ITW .EQ. 0)  GO TO 7800                                      PRN00335   000622
00427   336*    C             TRANSFER TO THE APPROPRIATE WRITE STATEMENT TO WRITE OUT  PRN00336   000622
00427   337*    C             A CENTERED TITLE.                                        PRN00337   000622
00431   338*          GO TO (7760, 7770, 7780, 7790), ITW                              PRN00338   000623
00432   339*     7760 WRITE (ILIST, 7761)  ITITLE(1)                                   PRN00339   000635
00435   340*     7761 FORMAT (1H , 70X, A6)                                            PRN00340   000644
00436   341*          GO TO 7795                                                       PRN00341   000644
00437   342*     7770 WRITE (ILIST, 7771)  (ITITLE(I), I=1,2)                          PRN00342   000646
00445   343*     7771 FORMAT (1H , 67X, 2A6)                                           PRN00343   000660
00446   344*          GO TO 7795                                                       PRN00344   000660
00447   345*     7780 WRITE (ILIST, 7781)  (ITITLE(I), I=1,3)                          PRN00345   000662
00455   346*     7781 FORMAT (1H , 64X, 3A6)                                           PRN00346   000674
```

220

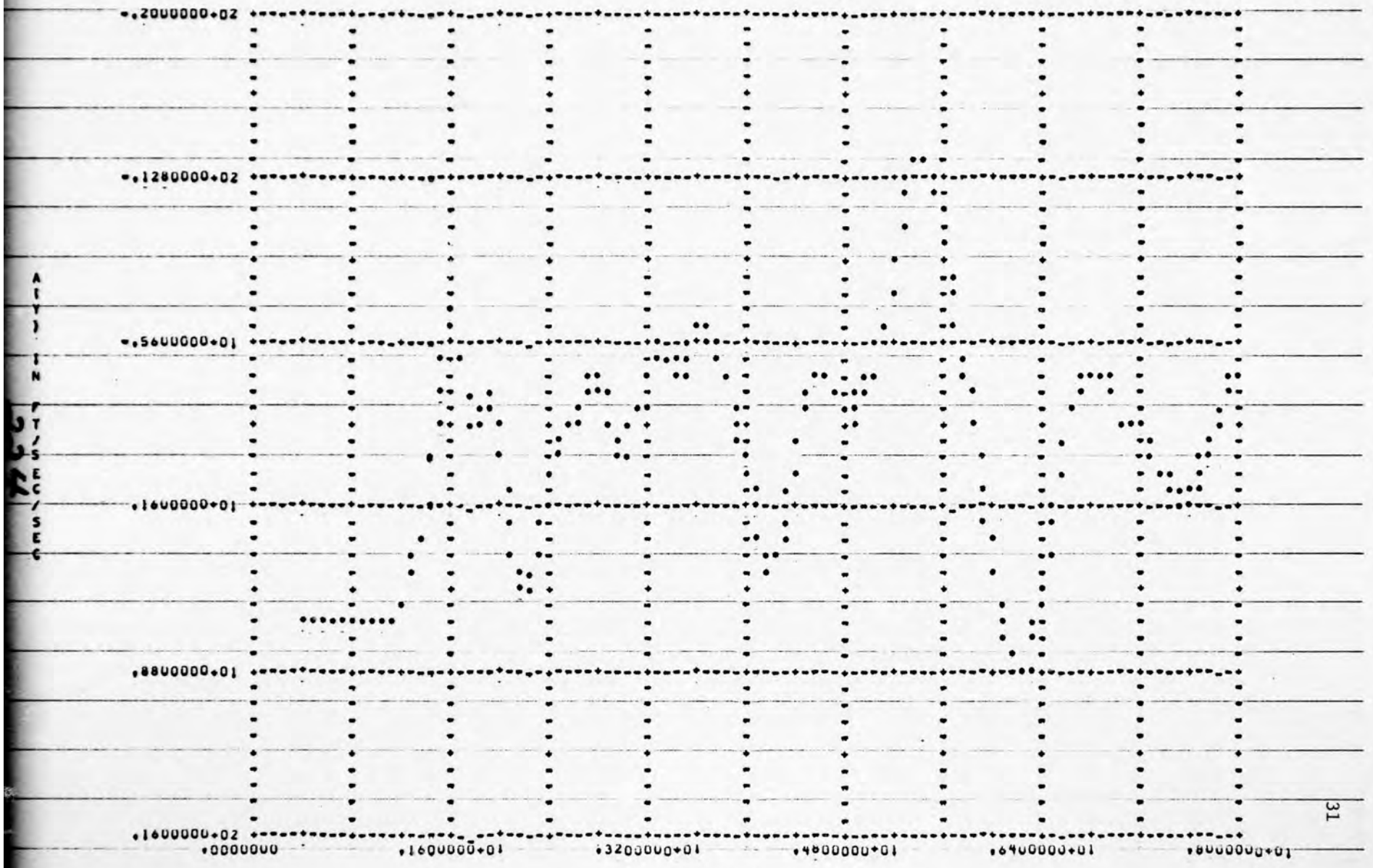
```
00456   347*          GO TO 7795                                                    PRN00347   000674
00457   348*    7790  WRITE (ILIST, 7791)  (ITITLE(I), I=1,4)                      PRN00348   000676
00465   349*    7791  FORMAT (1H , 61X, 4A6)                                       PRN00349   000711
00466   350*    7795  WRITE (ILIST, 7741)                                          PRN00350   000711
00466   351*          C             COMPUTE THE INCREMENT FOR THE PRINTER PLOT Y-AXIS  PRN00351   000711
00466   352*          C             NUMERICAL ANNOTATION.                         PRN00352   000711
00470   353*    7800  YPRINC = 0.2 * (YMAX - YMIN)                                PRN00353   000716
00471   354*          YINCT = ABS(0.001 * YPRINC)                                 PRN00354   000721
00472   355*          YINCMP = 6.0                                                PRN00355   000724
00473   356*          N = 9                                                       PRN00356   000726
00474   357*          IA = 1                                                      PRN00357   000730
00475   358*          IB = 13                                                     PRN00358   000732
00475   359*          C             THE STATEMENTS FROM HERE TO STATEMENT 7860 WRITE OUT THE  PRN00359   000732
00475   360*          C             FIRST 13 LINES AND THE LAST 14 LINES OF THE PRINTER  PRN00360   000732
00475   361*          C             PLOT.                                         PRN00361   000732
00476   362*    7820  DO 7860  I=IA,IB                                            PRN00362   000735
00501   363*          N = N + 1                                                   PRN00363   000744
00502   364*          IF (N .EQ. 10)  GO TO 7840                                  PRN00364   000747
00504   365*          WRITE (ILIST, 7831)  (IWORD(I,J), J=1,17)                   PRN00365   000751
00512   366*    7831  FORMAT (1H , 22X, 17A6)                                     PRN00366   000764
00513   367*          GO TO 7860                                                  PRN00367   000764
00514   368*    7840  N = 0                                                       PRN00368   000766
00515   369*          YINCMP = YINCMP - 1.0                                       PRN00369   000770
00516   370*          YANNOT = YMIN + (YINCMP * YPRINC)                           PRN00370   000773
00517   371*          IF (ABS(YANNOT) .LE. YINCT)  YANNOT = 0.0                   PRN00371   000777
00521   372*          WRITE (ILIST, 7851)  YANNOT, (IWORD(I,J), J=1,17)           PRN00372   001004
00530   373*    7851  FORMAT (1H , 5X, E16.6, 1X, 17A6)                           PRN00373   001022
00531   374*    7860  CONTINUE                                                    PRN00374   001022
00533   375*          IF (IB .EQ. 51)  GO TO 7920                                 PRN00375   001022
00533   376*          C             THE STATEMENTS FROM HERE THROUGH STATEMENT 7910 WRITE  PRN00376   001022
00533   377*          C             OUT THE 24 LINES IN THE CENTER OF THE PRINTER PLOT WHICH  PRN00377   001022
00533   378*          C             CAN CONTAIN PART OF THE Y-AXIS LABEL.         PRN00378   001022
00535   379*          DO 7910  I=14,37                                            PRN00379   001030
00540   380*          N = N + 1                                                   PRN00380   001033
00541   381*          IF (N .EQ. 10)  GO TO 7890                                  PRN00381   001036
00543   382*          WRITE (ILIST, 7881)  JYLABL(I-13), (IWORD(I,J), J=1,17)     PRN00382   001040
00552   383*    7881  FORMAT (1H , A1, 21X, 17A6)                                 PRN00383   001054
00553   384*          GO TO 7910                                                  PRN00384   001054
00554   385*    7890  N = 0                                                       PRN00385   001056
00555   386*          YINCMP = YINCMP - 1.0                                       PRN00386   001060
00556   387*          YANNOT = YMIN + (YINCMP * YPRINC)                           PRN00387   001063
00557   388*          IF (ABS(YANNOT) .LE. YINCT)  YANNOT = 0.0                   PRN00388   001067
00561   389*          WRITE (ILIST, 7901)  JYLABL(I-13), YANNOT, (IWORD(I,J), J=1,17)  PRN00389   001074
00571   390*    7901  FORMAT (1H , A1, 4X, E16.6, 1X, 17A6)                       PRN00390   001113
00572   391*    7910  CONTINUE                                                    PRN00391   001113
00574   392*          IA = 38                                                     PRN00392   001113
00575   393*          IB = 51                                                     PRN00393   001115
00576   394*          GO TO 7820                                                  PRN00394   001117
00576   395*          C             COMPUTE AND STORE THE SIX VALUES USED TO NUMERICALLY  PRN00395   001117
00576   396*          C             ANNOTATE THE X-AXIS, THEN WRITE OUT THE X-AXIS NUMERICAL  PRN00396   001117
00576   397*          C             ANNOTATION.                                   PRN00397   001117
00577   398*    7920  XPRINC = 0.2 * (XMAX - XMIN)                               PRN00398   001121
00600   399*          XINCT = ABS(0.0005 * XPRINC)                                PRN00399   001124
00601   400*          XINCMP = -1.0                                               PRN00400   001127
00602   401*          DO 7930  I=1,6                                              PRN00401   001134
00605   402*          XINCMP = XINCMP + 1.0                                       PRN00402   001134
00606   403*          XANNOT(I) = XMIN  +  (XINCMP * XPRINC)                      PRN00403   001136
```

```
00607   404*          IF (ABS(XANNOT(I)) .LE. XINCT)  XANNOT(I) = 0.0                  PRN00404   001141
00611   405*    7930  CONTINUE                                                         PRN00405   001150
00613   406*          WRITE (ILIST, 7935)  (XANNOT(I), I=1,6)                          PRN00406   001150
00616   407*    7935  FORMAT (1H , 8X, 6(4X, E16.6))                                  PRN00407   001160
00616   408*    C             TRANSFER TO STATEMENT 7960 IF THERE IS NO X-AXIS LABEL.  PRN00408   001160
00617   409*          IF (IXW .EQ. 0)  GO TO 7960                                      PRN00409   001160
00621   410*          WRITE (ILIST, 7741)                                              PRN00410   001162
00621   411*    C             TRANSFER TO THE APPROPRIATE WRITE STATEMENT TO WRITE OUT  PRN00411   001162
00621   412*    C             A CENTERED X-AXIS LABEL.                                 PRN00412   001162
00623   413*          GO TO (7940, 7945, 7950, 7955), IXW                              PRN00413   001167
00624   414*    7940  WRITE (ILIST, 7761)  IXLABL(1)                                   PRN00414   001201
00627   415*          GO TO 7960                                                       PRN00415   001210
00630   416*    7945  WRITE (ILIST, 7771)  (IXLABL(I), I=1,2)                         PRN00416   001212
00636   417*          GO TO 7960                                                       PRN00417   001224
00637   418*    7950  WRITE (ILIST, 7781)  (IXLABL(I), I=1,3)                         PRN00418   001226
00645   419*          GO TO 7960                                                       PRN00419   001240
00646   420*    7955  WRITE (ILIST, 7791)  (IXLABL(I), I=1,4)                         PRN00420   001242
00654   421*    7960  IF (ICLASS .EQ. 0)  RETURN                                      PRN00421   001255
00654   422*    C             WRITE OUT THE CLASSIFICATION LABEL.                      PRN00422   001255
00656   423*          WRITE (ILIST, 7741)                                              PRN00423   001261
00660   424*          WRITE (ILIST, 7731) ((KCLASS(I,J), J=1,2), I=1,5)                PRN00424   001266
00671   425*          RETURN                                                           PRN00425   001305
00671   426*    C                                                                      PRN00426   001305
00672   427*          END                                                              PRN00427   001404
```

222


TIME VS A(Z) BST PHASE
.6800000+03
.5240000+03
.3680000+03
.2120000+03
.5600000+02
-.1000000+03
A(Z) IN FT/SEC/SEC
.0000000
.1600000+01
.3200000+01
.4800000+01
.6400000+01
.8000000+01
TIME

TIME VS A(Y) BST PHASE



D. Accounting of Data Points Processed

Inevitably, some graphics jobs will not produce the output that the customer expected, and the resourceful graphics programmer will prepare in advance for this contingency if he is to avoid spending many of his future days troubleshooting other people's production problems. Since most production graphics problems are in fact caused by the customer's inattention to the input data, what we would like to do is give him information with every plot that will allow him to discover the cause of the malfunction before he resorts to contacting either the software development personnel or the operations personnel.

The single most valuable tool in this respect is the simple matter of accurately accounting for all data points processed by the graphics software. This usually consists of nothing more than a summary of the total number of points examined, along with a breakdown of their disposition as to whether they were actually plotted, were off scale, and so forth. If, for example, the user receives a blank plot, the summary can immediately tell him whether the expected data was even presented to the plot program in the first place, and if it was, the summary will indicate the probable cause of its failure to be plotted.

It is useful to classify the status of data points according to how they are handled by the plot program. A sampling of some commonly used classifications are:

Blank--A data point that has been replaced by a blank or transparent word that serves as filler where the particular parameter does not exist because it was not available or was not recorded.

Off-scale--A data point which has either or both coordinates lying outside the limits specified by the extremities of the axes.

Outside start or end times--Where the option exists to specify further restricting limits (usually based on some interval of time), a data point that is on-scale, but not within the start or end time limits.

Thinned-out--A data point that passes all of the above tests, but is thinned out to reduce the density of the plotted points.

An examination of the characteristics of these various conditions will show that they should be tested in the order listed, in order to provide the most meaningful information to the analyst. Some categories can be broken down even further, such as off-scale data points which could be off in X, Y, both X and Y, and so forth, but this is usually not necessary. Figure 2 shows a practical approach to the overall accounting process, and the following page shows an example of what the summary might look like on the user's computer listing.

Another useful feature which is related to the accounting of data points is to save and print out the coordinates of the first and last data points plotted, and if desired, the first and last points examined. This gives the analyst the ability to track down at least two specific data elements at the extremities of the plot when troubleshooting a malfunction. Core and processing requirements for this task are trivial.

Finally, the plots and grids themselves should be counted, and summarized at the end of the job. The grids are counted separately since several plots may be placed on a single grid, so the number of grids is not necessarily the same as the number of plots. A typical summary might state something like "126 PLOTS ON 84 GRIDS." This way, the analyst knows at a glance whether the job was completed in its entirety.

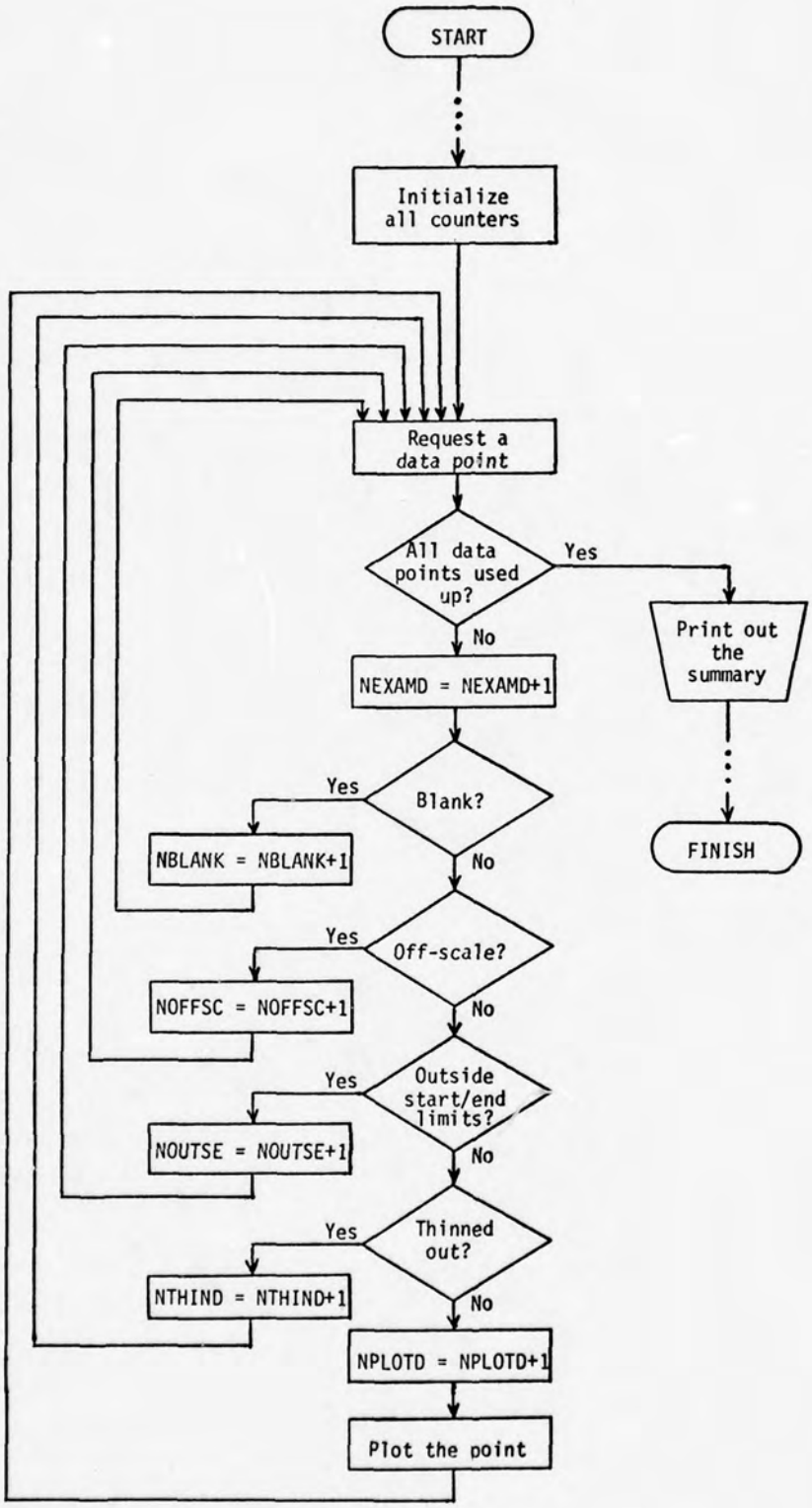

START
Initialize all counters
Request a data point
All data points used up?
Yes
No
Print out the summary
FINISH
NEXAMD = NEXAMD+1
Blank?
Yes
No
NBLANK = NBLANK+1
Off-scale?
Yes
No
NOFFSC = NOFFSC+1
Outside start/end limits?
Yes
No
NOUTSE = NOUTSE+1
Thinned out?
Yes
No
NTHIND = NTHIND+1
NPLOTD = NPLOTD+1
Plot the point

Figure 2

```
        1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  0108211    X VS Z  BOOST PHASE   X IN K FEET          Z IN K FEET

        1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  02082        3   149  349

        1         2         3         4         5         6         7         8
12345678901234567890123456789012345678901234567890123456789012345678901234567890
  03082        0   8    8         -5        11        -3        15    1         31

THE DATA TO BE PLOTTED IS FROM PASS OBJECT 11 OF MODULE SEQUENCE NUMBER  3.

X-COORDINATE IS FROM RELATIVE WORD   1, ABSOLUTE WORD  11.

Y-COORDINATE IS FROM RELATIVE WORD   3, ABSOLUTE WORD  13.

X-COORDINATE UNITS CONVERSION MULTIPLICATIVE AND ADDITIVE FACTORS ARE    .10000000-02 AND    .00000000   , RESPECTIVELY.

Y-COORDINATE UNITS CONVERSION MULTIPLICATIVE AND ADDITIVE FACTORS ARE    .10000000-02 AND    .00000000   , RESPECTIVELY.

POINT PLOT.

GRID TYPE =  31          X-AXIS  2/IN.          Y-AXIS  10/IN.

X-AXIS LENGTH =      8.000 INCHES.          Y-AXIS LENGTH =      6.000 INCHES.

XMIN =    -.50000000+01          XMAX =    .11000000+02                    YMIN =    -.30000000+01          YMAX =    .15000000+02

THIS IS PLOT NUMBER    2.

THIS PLOT STARTS AT SEARCH RECORD NUMBER    3

FIRST POINT = (    .22076223-01,    .27634051-01)          LAST POINT = (    .10975479+02,    .14264942+02)

  1779 POINTS BLANK,     716 POINTS OFF-SCALE,       0 POINTS OUTSIDE START/END VALUES,       0 POINTS THINNED OUT.

   229 POINTS PLOTTED, OUT OF   2724 POINTS EXAMINED.
```

V. The White Sands Production Graphics System

White Sands Missile Range is one of the world's busiest test ranges, where enormous quantities of data are collected by both optical and electronic means. A wide variety of data reduction programs reduce the raw data on a production basis in order to provide the range users with complete and timely data reports for use in analysis of the individual test missions. Most of these reports contain considerable graphics, which are generated by a production graphics system which has undergone a number of changes over the years.

Prior to 1971, several large flat-bed plotters were in use, producing a total of fewer than fifty plots per day. These machines were operated off-line, and the customer had to run the computer job, retrieve the plot tape, personally deliver the tape to the plotting room along with a work order specifying grid size, scale factors, plot mode (point or line), ink color, and so forth. The user specifications were entered into the plotters through switches, and the equipment was manually calibrated for each plot, based on the stated paper size and scaling factors. Once completed, the customer picked up the plots and tape, disposed of the tape as necessary, then turned in the plots to a reports section to be manually labled and annotated. The entire process was not only tedious, but susceptible to human error at every stage.

Due to an increasing workload, and a continuing obsolescence of the existing graphics equipment, a changeover was made that converted the graphics operation to a high-production system that eliminates all human intervention that affects the content of the plots. The operational procedure is shown schematically in Figure 3.

The core of the system is a CalComp Model 835 CRT plotter which draws the plots on a cathode ray tube and automatically records them on 35 mm film. The customer himself simply submits his job to the computer, where an off-line plot tape is generated. The computer listing, with full and complete printer plots, is returned directly to him, while the plot tape is automatically delivered directly to the plotter without any action on his part. Plot tapes are grouped at the plotter (classified tapes, of course, are handled separately), and at designated times of the day all of the tapes present are run. The operator has only to mount the tape, push the start button, and dismount the tape when finished.

Every plot is fully labeled and numerically annotated (inclduing security classification when needed) as it is drawn. The operating speed of the plotter is such that a single plot generally takes no more than five seconds to complete, even though the grid itself must also be drawn. The film, of course, advances automatically between plots. A tape containing thirty complete plots can be fully processed by the operator in three minutes or so. The graphics software generates a header and trailer frame at the beginning and end of each job for identification purposes, along with a unique serial number which is recorded in small numerals in the far corner of each plot.

Once all tapes have been recorded, the film reel is removed and processed in a small automatic developer which requires little more than that the operator mount the reel, push the loose end of the film in a hole, turn on the water, and watch it wind up on a reel on the other side. The processing time for a 200-foot reel (1600 plots) is about 15 minutes, though preparation and cleanup require a few minutes additional.

The film is then taken to a semi-automatic printer for final printing. The plots vary in size, according to their intended use, and the customer specifies the size at the time of the computer run. The operator at the printer does have to change paper sizes when instructed to do so by the customer's header frame. Maximum plot size is 11 by 17 inches. The printing process requires about fifteen seconds per plot.

Upon completion of the printing, the plots are returned directly to the user. Since all labeling and annotation is done by the plotter, there is no further clerical process, and the plots may be inserted directly into the data reports. Once the finished plots have been delivered, the tapes are automatically degaussed, and the film is handled as classified waste.

Naturally, there is a degree of flexibility built into the operation to handle special requirements such as the occasional expedite priority job. And not all graphics at White Sands are processed on this system. There are a number of other plotters available, and various organizations use them for special purposes.

As carefree as the operation sounds, it has not been without its problems. Perhaps the first to become apparent was due to the fact that the system prints in only one color--black, on white paper. It was very difficult to convince the old-timers, who were accustomed to multiple ink colors, that they could live just as well with plots using one color in multiple modes (line, point, symbol, etc.). Next was the matter of size. The system can produce plots up to 11 x 17 inches, which is really quite large, but the previous plotters could handle 30 x 30 inch paper. No one knew what he would do with a plot greater than 11 x 17, but many were still hard to convince.

Overall reliability did not turn out to be as good as expected, either. This was primarily due to the fact that there are three major links in the chain: the plotter, the developer, and the printer. If just one of them breaks down, the entire system grinds to a halt. The problem is controlled primarily by increased attention to preventive maintenance. The particular model of plotter, incidentally, has been discontinued by the manufacturer, and at some future date this may impact the maintenance program, but it has had no effect as yet.

With the increased capability of the system came increased demand. Whereas graphics requests had been running at a few hundred per month, peak output under the present system has soared to over 15,000 plots per month. The increased use of graphics, however, has cut the need for detailed data listings in some areas, and often facilitates the analyst's job. Some samples of typical output from the White Sands production graphics system are shown on the following pages.

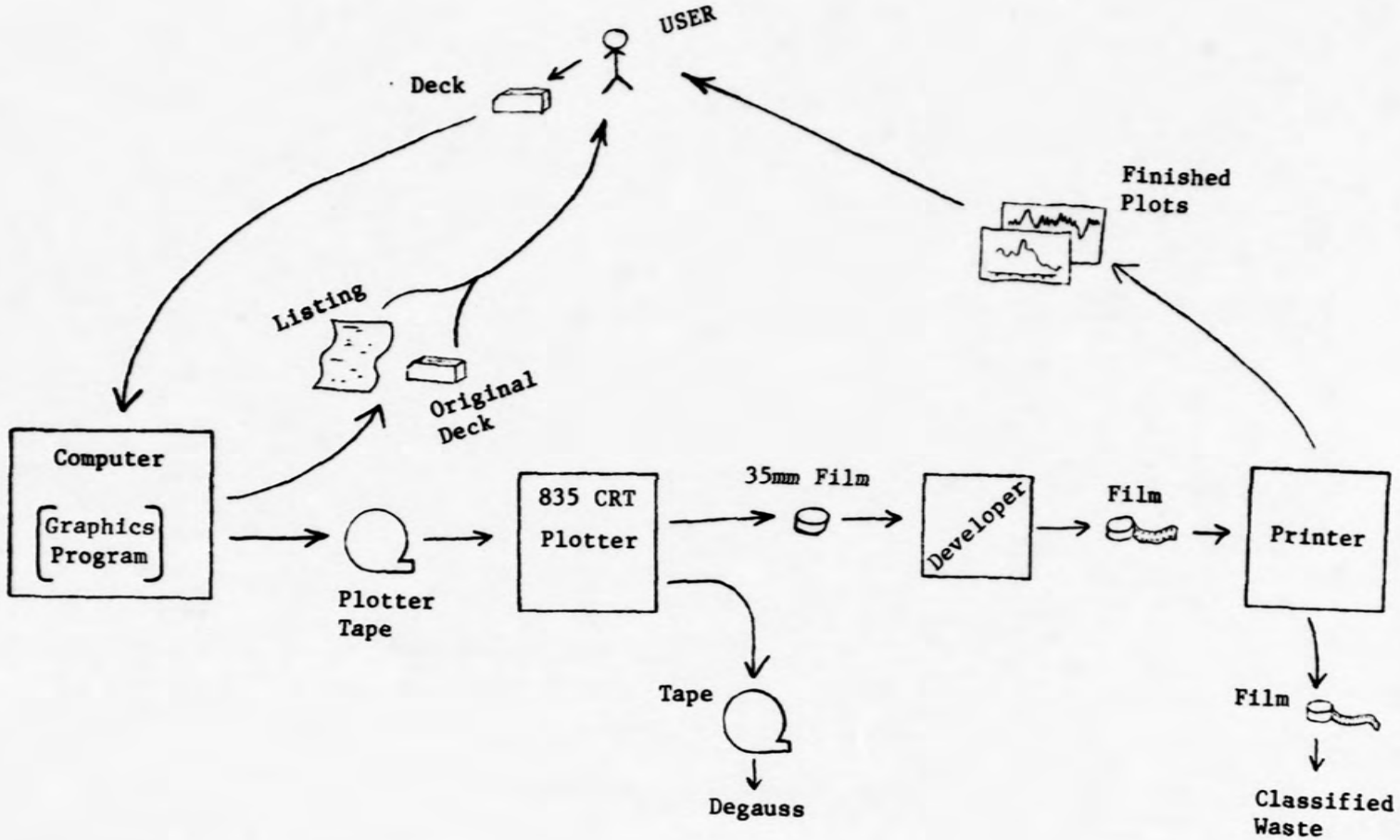
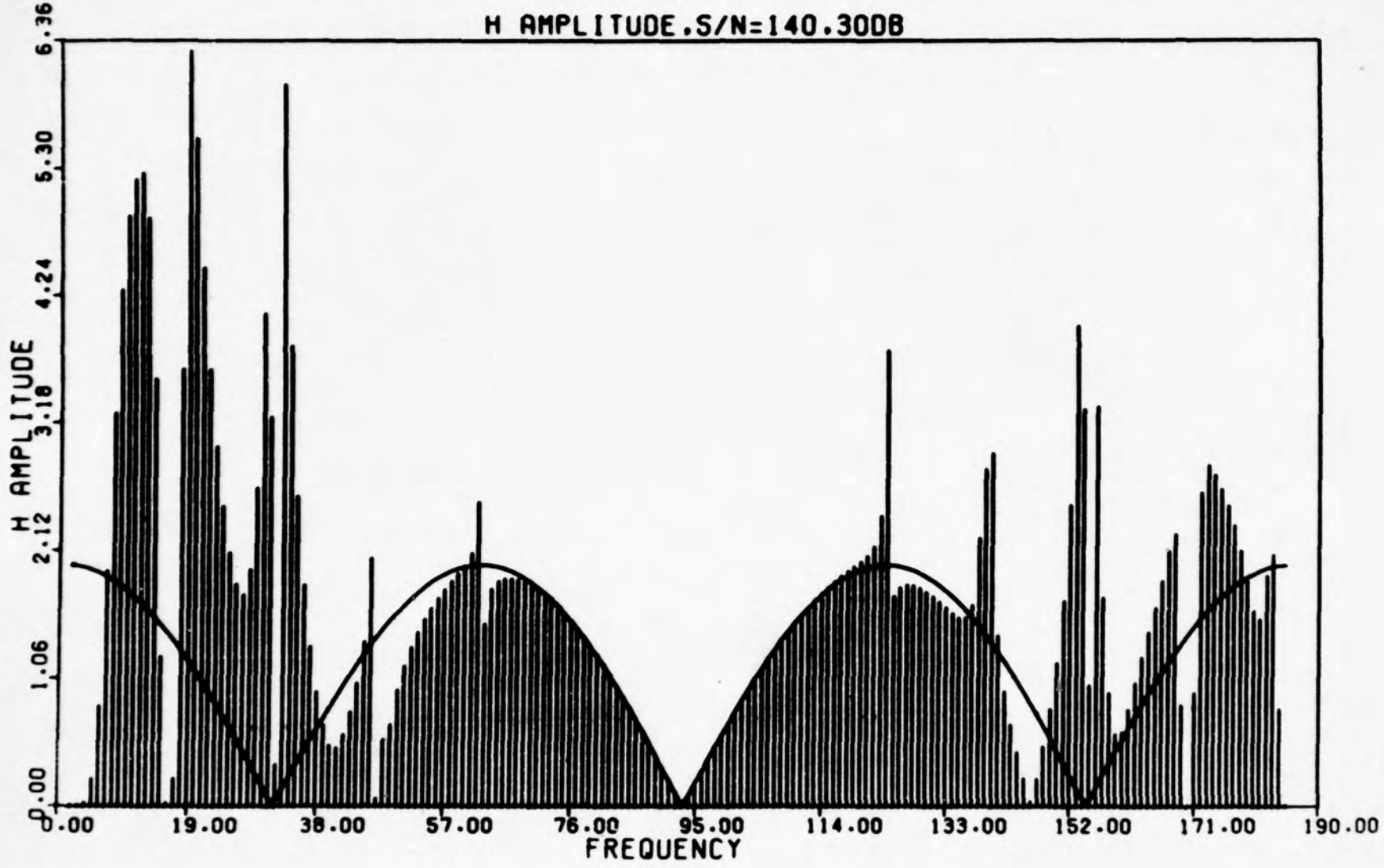
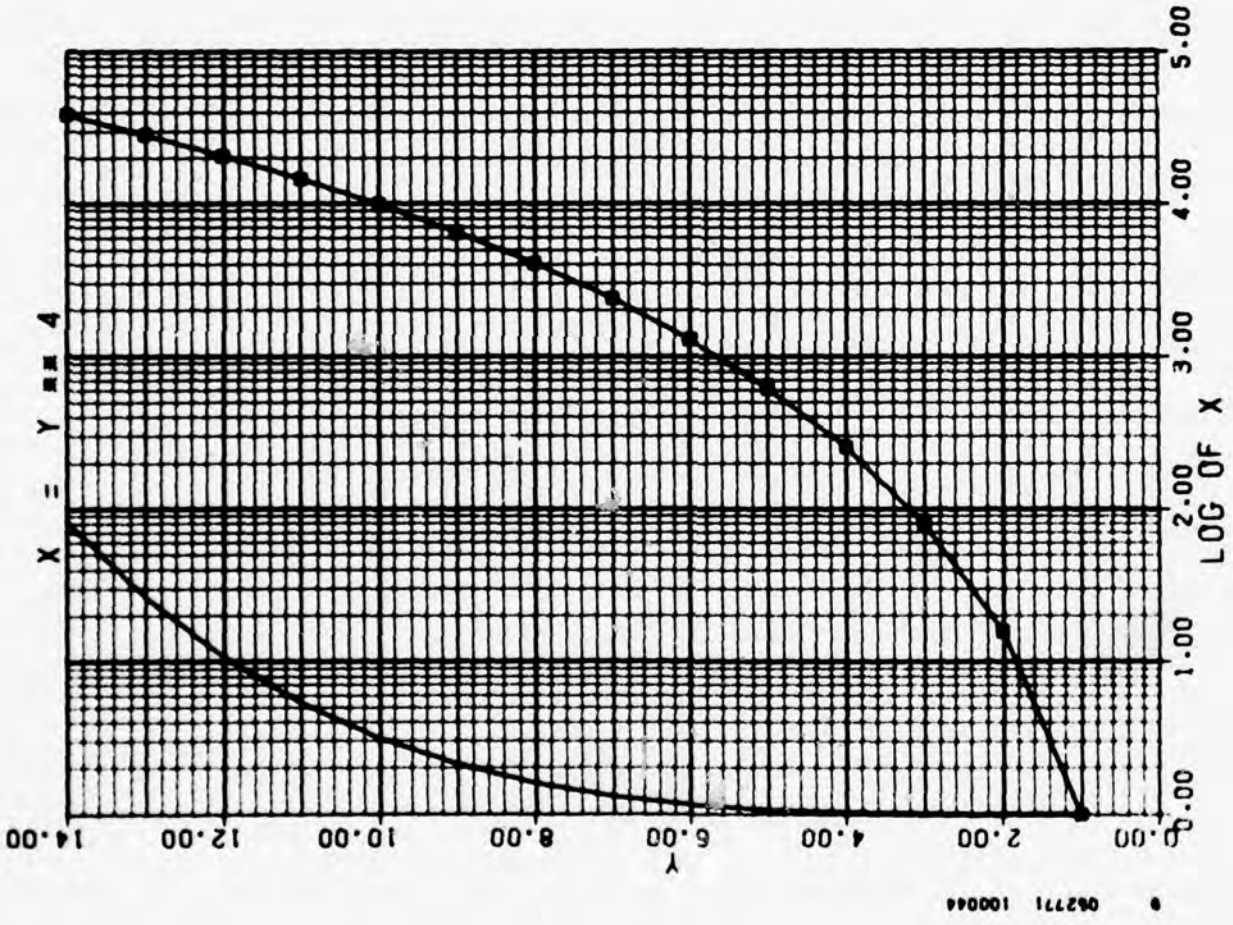


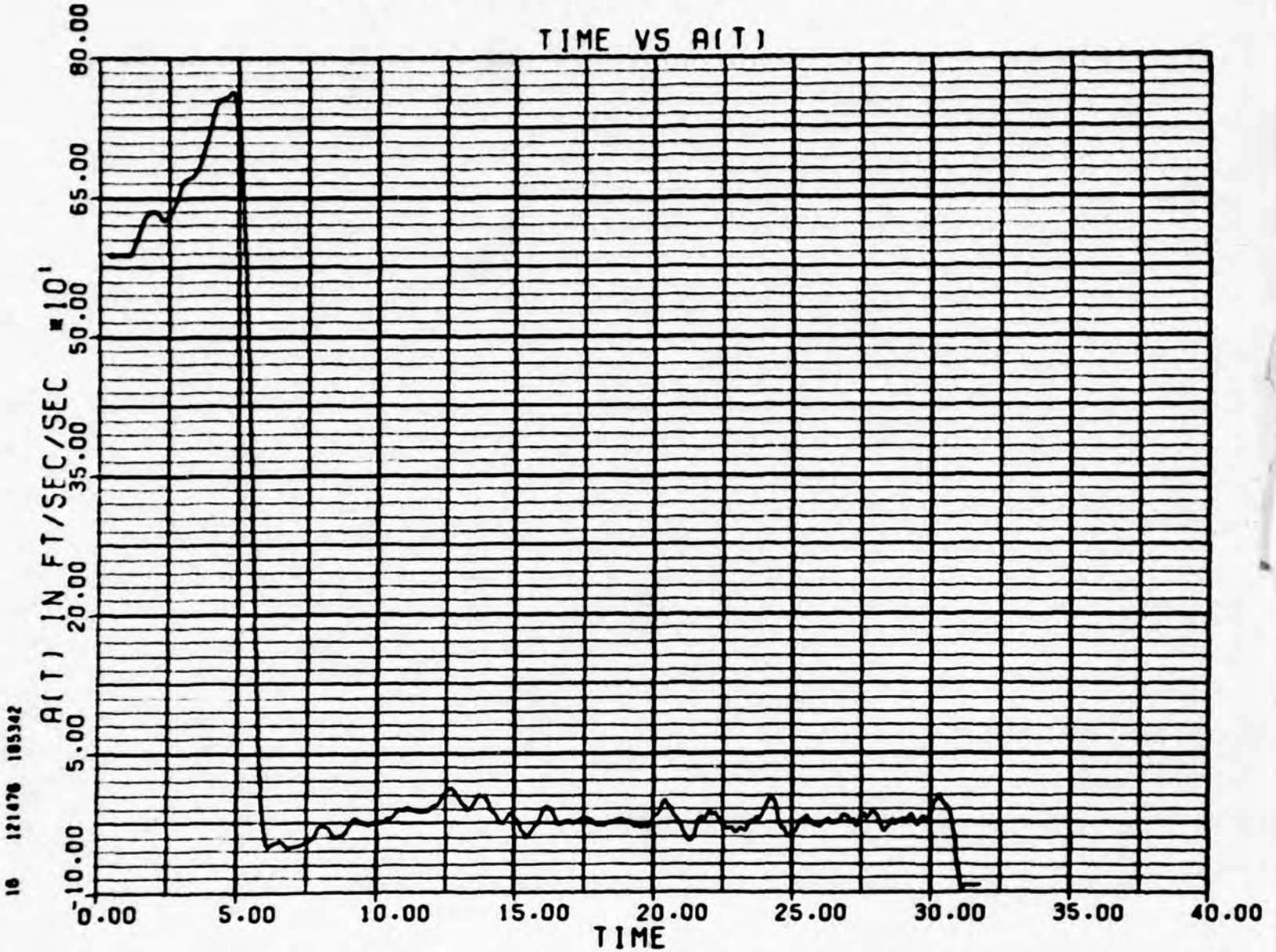
Figure 3

White Sands Production Graphics System


H AMPLITUDE.S/N=140.30DB
H AMPLITUDE
0.00
1.06
2.12
3.18
4.24
5.30
6.36
0.00
19.00
38.00
57.00
76.00
95.00
114.00
133.00
152.00
171.00
190.00
FREQUENCY

VDS4-.QDS4....AVS** TIME
TIME (SEC.)
130.00
125.00
120.00
115.00
110.00
105.00
100.00
95.00
90.00
85.00
80.00
0.00
14.29
28.57
42.86
57.14
71.43
85.71
100.00


X = Y ** 4
LOG OF X
Y
0.00
1.00
2.00
3.00
4.00
5.00
0.00
2.00
4.00
6.00
8.00
10.00
12.00
14.00


TIME VS A(T)
A(T) IN FT/SEC/SEC *10'
80.00
65.00
50.00
35.00
20.00
5.00
-10.00
0.00
5.00
10.00
15.00
20.00
25.00
30.00
35.00
40.00
TIME
18 121478 18342

MATHEMATICAL TRADE-OFFS FOR MANAGERIAL CONTROL

John L. Lazaruk
Systems and Economic Analysis Division
Office of the Comptroller
U.S. Army Communications Command
Ft. Huachuca, Az. 85613

ABSTRACT. Recent investigations in system theory have generated a mathematical definition of a trade-off. This concept arises naturally in the context of economic decisions, systems analysis, and systems engineering. This paper presents a mathematical structure of a trade-off between system or subsystem attributes which insures that the suboptimization process inherent in the design of large scale systems yields an optimization of the overall system design. In addition, a methodology is developed based on a hierarchy of functions associated with the life cycle of large scale systems using mathematical trade-offs for the managerial control of the engineering, implementation, and operation of large scale systems.

1. INTRODUCTION. Large scale system exhibit certain characteristics such as an overall purpose, the interaction of humans and machines, and a variety of subsystems which transcend the expertise of a single discipline. Generally, the user, operator, and builder are distinct. Examples of large scale systems are military weapon systems, communications systems, and social welfare systems.

The characteristics of a large scale system lead to a hierarchy of of functions corresponding to Figure 1.

HIERARCHY OF SYSTEM FUNCTIONS

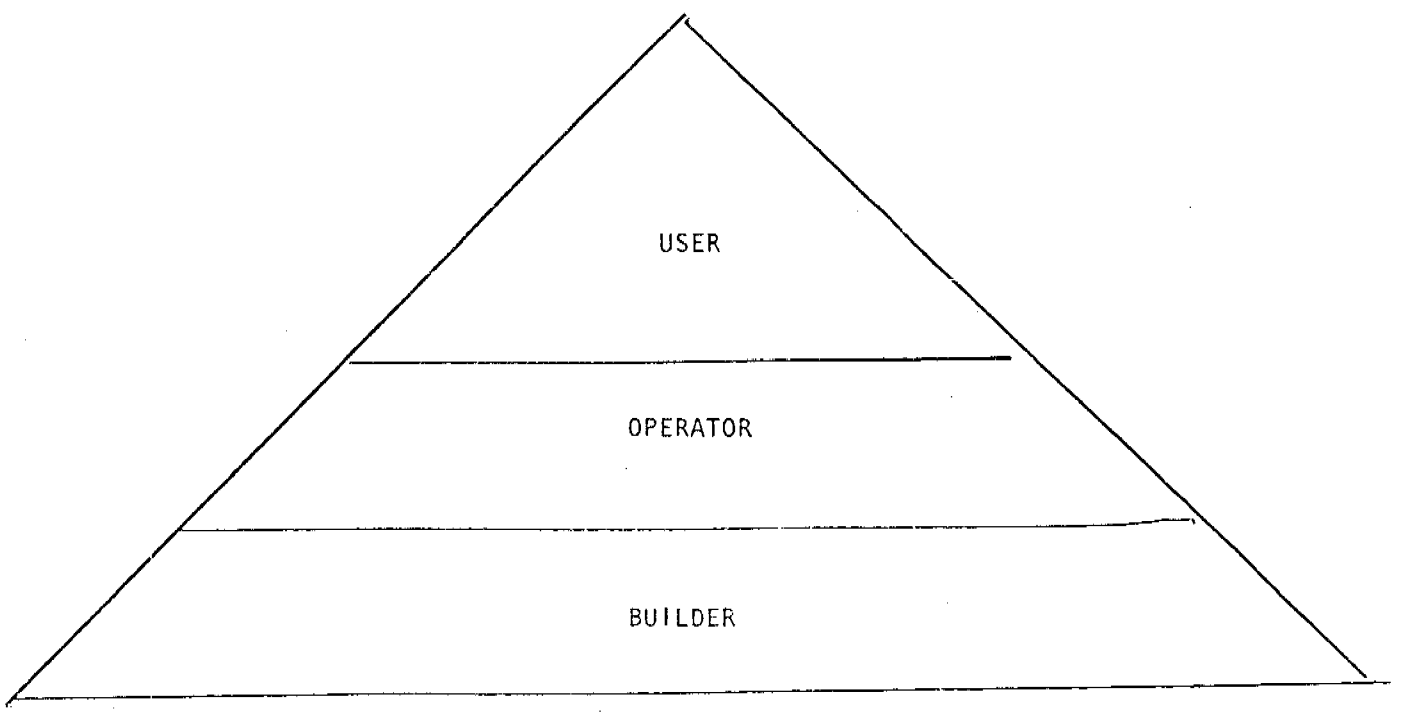


FIGURE 1

The hierarchy is based on the interrelationship of functional purposes. The overall purpose of the system is to satisfy the requirements of the user. The user requirements are satisfied through the operation of the system by an operator. Thus, the system is built for the operator to satisfy the requirements of the user.

As a result of this tri-level hierarchy of system functions, corresponding hierarchy of effectiveness can be constructed as in Figure 2.

HIERARCHY OF SYSTEM EFFECTIVENESS

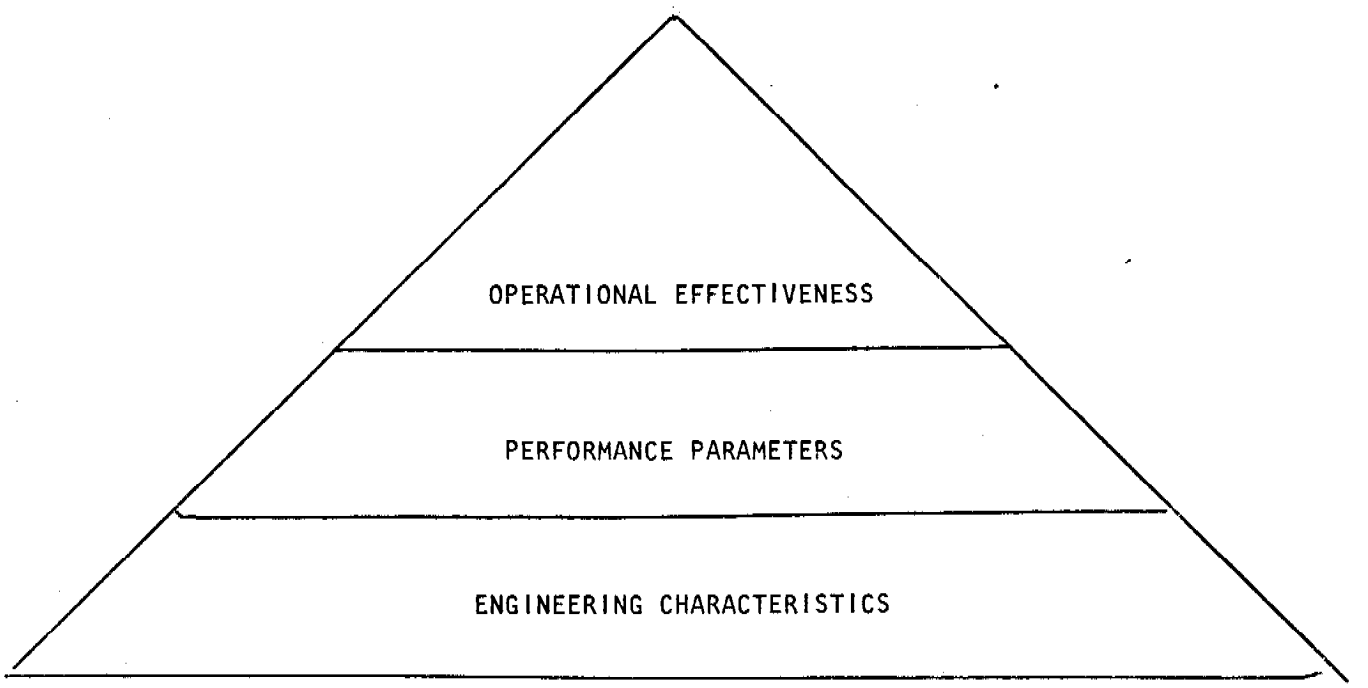


FIGURE 2

The operational effectiveness level of a system pertains to characteristics of conditions which are of concern or are experienced by the user. An example of a measure of operational effectiveness is the length of time the system takes to satisfy a user's requirement, speed of service. The performance parameters are the characteristics or conditions which are of concern or are experienced by the operator. An example of a performance parameter is the percentage of the time a system is capable of performing its intended functions, the system availability. The engineering characteristics are the last level in the hierarchy of system effectiveness discussed in this paper. The engineering characteristics are the conditions or characteristics the builder uses to produce a system, an example is the mean time between failures, MTBF of a system component.

These three aspects are also visible in the basic structure of a system. The black box model of a system exhibited in Figure 3 can be related to the tri-level hierarchial structure of system effectiveness.

BLACKBOX SYSTEM MODEL

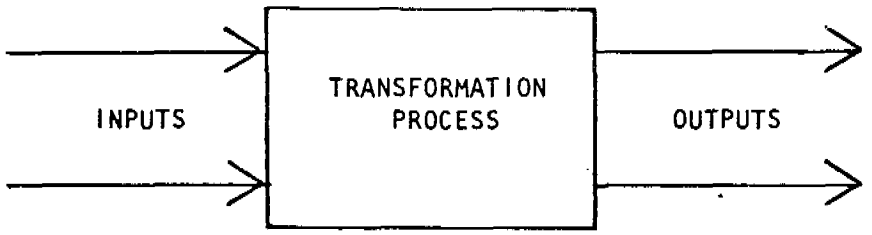


FIGURE 3

The base of the effectiveness hierarchy, engineering characteristics, correspond to the technical specifications of the transformation process. The performance parameters correspond to measurements of the way the transformation process performs. The operational effectiveness of a system corresponds to measurements of the way the system's input/outputs effect the user.

A major implication of this structure in the management of large scale systems is a delegation of responsibilities and a process of suboptimization. This process is a result of decomposing a system into components which are then optimized or improved independently of one another. What is proposed in this paper is a method of controlling the suboptimization process to insure actions taken with respect to the components are compatible with the overall objectives of the system and its optimization by using trade-off functions in the hierarchial structure described above.

2. MEASUREMENTS AND TRADE-OFF FUNCTIONS:

Coombs, Raiffa, and Thrall (reference 3) define a measurement as a function from a set objects to a mathematical system. This paper is concerned with two types of measurement: measures of effectiveness and preference measures. If S is a set of systems and R is a subset of real numbers, a measure of effectiveness is a function $\mu:S\rightarrow R$. If P is a partially ordered set, a function $\rho:S\rightarrow P$ is a preference measure if $\rho(x)\geq\rho(y)$ for $x,y\in S$ implies x is as good as y.

In the management of large scale systems, there are generally many attributes, characteristics, or conditions for which a manager is responsible. If $\{\rho_i:S\rightarrow P_i\}$ represents managerial preference with respect to n different attributes, a preference measure $\tau:S\rightarrow P$ which is compatible with the managerial preference of the individual attributes should have the following property:

(1) If $x,y\in S$ and $\rho_i(x)\geq\rho_i(y)$ for each i, then $\tau(x)\geq\tau(y)$.

This property is equivalent to one of two properties used by Wymore to define the concept of a trade-off between orderings in reference 11. The second property in Wymore's definition is equivalent to

(2) If $x,y\in S$ such that $\rho_i(x)\geq\rho_i(y)$ for each i and there exists j such that $\rho_j(x)>\rho_j(y)$, then $\tau(x)>\tau(y)$.

As is often the case, a measure of effectiveness may also be a preference measure where the partial order is the natural order of the real numbers. For a given attribute, more is generally better. However, it is possible that an increase in a particular attribute may not cause one system to be preferred to another system with a value for the attribute. In fact, the US Army Logistics Management Center has published an analysis which includes indifference curves which bound regions in which an attribute may be better for one alternative than another and yet the decision maker is not willing to say the one alternative is better than the other. This means that there exists situations in which the function representing the preference for the combined set of attributes does not satisfy property (2). A function, $\tau:S\rightarrow P$ will be called a trade-off function over $\{\rho_i:S\rightarrow P_i\}$ if property (1) is satisfied.

The next Theorem is fundamental to the methodology for the management of large scale systems described in this paper. It is the essential characteristic of trade-off functions for use with the hierarchial management structure for large scale systems which was described in Section I .

Theorem 1: Let A be a set of preference measures of a set of systems, S, and $\{A_k\}$ a partition of A. If $\sigma_k:S\rightarrow P_k$ is a trade-off over A_k and $\tau:S\rightarrow P$ is a trade-off over $\{\sigma_k:S\rightarrow P_k\}$, then τ is a trade-off over A.

Proof: If $x,y\in S$ such that $\rho(x)\geq\rho(y)$ for each $\rho\in A$, then $\sigma_k(x)\geq\sigma_k(y)$ for each k. Therefore, $\tau(x)\geq\tau(y)$. Thus, τ is a trade-off over A.

The next Theorem provides the basis for insuring that the suboptimization process described in Section I is compatible with the preferences for the system expressed by a trade-off function. Before proceeding with the theorem, a definition of completeness is needed.

A set of systems, S, is complete with respect to $\{\rho_i : S \to P_i\}$, a set of preference measures, if for any element, a, in the cartasian product of $\rho(s)$ there exists $y \in S$ such that $\rho_i(y) \geq a_i$ for each i.

Theorem 2: If S is complete with respect to $\{\rho_i : S \to P_i\}$ and $\tau : S \to P$ is a trade-off function which satisfies property (2) and each P_i and P are linearly ordered, then $\tau(x)$ is a maximum over $\tau(S)$ if, and only if, $\rho_i(x)$ is a maximum over $\rho_i(S)$ for each i.

Proof: If $\tau(x) \geq \tau(y)$ for each $y \in S$, suppose there is a j such that $\rho_j(y) > \rho_j(x)$. Since S is complete, there exists $z \in S$ such that $\rho_i(z) \geq \rho_i(x)$ for $i \neq j$ and $\rho_j(z) \geq \rho_j(y)$. Since τ is a trade-off which satisfies property (2) $\tau(z) > \tau(x)$. Thus $\tau(x)$ is not maximum over $\tau(S)$.

If $\rho_i(x)$ is a maximum over $\rho_i(S)$ for each i then, for $y \in S$ $\rho_i(x) \geq \rho_i(y)$ for each i. Hence, $\tau(x) \geq \tau(y)$ since τ is a trade-off.

3. MANAGERIAL CONTROL THROUGH TRADE-OFFS

The structure of the managerial problem in the design of large scale systems under consideration in this paper has four elements:

(1) A set of characteristics or conditions which require separate expertise and also define the system management situation in the sense that preferences in the performance of the system are based on the characteristics specified. This means that the characteristics and/or conditions specified are complete.

(2) A method of measuring each of the characteristics determined in the first element above.

(3) Standards or tolerances which are used to compare the characteristics measures in order to insure the compatibility of the attribute values with the overall objective of the system.

(4) The means to alter the characteristic whose difference from the standard exceeds the tolerance.

The model which forms the basis of the managerial control theory presented in this paper concentrates on the first three elements. It is assumed that the manager by definition has the last property. What we are concerned with is providing the responsible managers with the information required to perform the control function rationally and effectively at each hierarchial level.

The first step in the methodology of managerial control of systems through trade-offs is to establish the top level characteristics or conditions to be controlled. The top level attributes in the management of large scale systems should be relatable to the user functions in the functional hierarchy described in Section I . The characteristics or conditions for the operator's managers to be concerned with are traded off by upper level attributes. Proceeding in the same manner through all the levels of the functional hierarchy, the measures established at each level are trade-offs of lower level attributes and are traded off by the higher level attributes.

Theorem 1 insures that the higher level attributes are trade-offs of subsets at each lower level. Theorem 2 supplies conditions under which the suboptimization process necessary in large scale systems is consistent with the overall objectives of the system.

4. TRADE-OFFS IN A COMMUNICATION SYSTEM (An Example)

The purpose of a communication system is to transfer information between two or more points. Figure 3 is a model of a simple communication system providing circuits between two points.

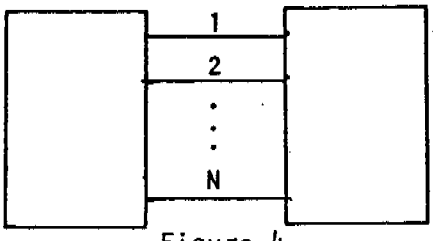


Figure 4

Suppose the measurable purpose or overall objective of the system is to provide a required number of circuit hours per month where a circuit hour is defined as one circuit operating for 1 hour. Availability is a measure of the time a circuit is useable. It is expressable in terms of two engineering criteria which measure the maintainability of the system, Mean Time Between Failure (MTBF), and Mean Time to Repair (MTTR).

The availability, Av, is defined to be $\frac{MTBF}{MTBF + MTTR}$

If λ is the failure rate in failures per hour and ρ is the repair rate in repairs per hour, then

$$Av = \frac{\rho}{\lambda + \rho}$$

It can be shown that the availability is a trade-off over $1/\lambda$ and ρ. Assume each circuit in the system has the same maintainability characteristics. The number of circuit hours per day for the given system is a

trade-off between the number of circuits and the availability of each circuit as represented by the following equation:

$$CHD = 24NAv$$

where CHD = Circuit Hours Per Day
N = Number of Circuits
Av = Circuit Availability

Figure 4 summarizes the functional hierarchy and the appropriate measures.

FUNCTIONAL LEVEL	MEASURE
USER	CHD
OPERATOR	N
	Av
BUILDER	N
	$1/\lambda$
	ρ

Figure 5

The standards or tolerances on N, $1/\lambda$ and ρ are established or linked to the user requirements through the trade-off relationship.

BIBLIOGRAPHY

1. Kenneth J. Arrow, Social Choice and Individual Values, 2nd Edition, John Wiley & Sons, 1963.

2. L.D. Attoway, Criteria and the Measurement of Effectiveness in Systems Analysis in Defense, Ed.by E.S. Quade & W.I. Boucher, American Elsevier Publishing Company, Inc., New York, 1968.

3. C.H. Coombs Howard Raiffa, and R.M. Thrall, Mathematical Models and Measurement Theory in Decision Processes Edited by R. M. Thrall, C.H. Coombs, and R.L. Davis, John Wiley & Sons 1954 pp 19-37.

4. Cost and Operational Effectiveness Handbook, TRADOC Pamphlet 11-8.

5. Decision Risk Analysis for the SUPER DAWG Missile System, US Army Logistics Management Center, Fort Lee Virginia, ALM-63-3792-H.

6. Harry H. Goode & Robert E. Machol, SYSTEM ENGINEERING An Introduction to The Design of Large-scale Systems, McGraw Hill, Inc., 1957.

7. Richard Johnson, Fremont Kast, and James Rosenzweig, The Theory and Management of Systems, McGraw Hill, 1973.

8. John L. Kelley, General Topology, D. Van Nostrand Company, 1955.

9. K.R. MacCrimmon, Decision Making among Multi-Attribute Alternatives: A Survey and Consolidated Approach, Rand Corporation Memorandum RM-4823-ARPA December, 1968.

10. Kenneth O. May, Intransitivity Utility and The Aggregation of Preference Pattersons, Econometrica, Vol.22, January 1954, Number pp 1-13.

11. A Wayne Wymore, Systems Engineering Methodology for Interdisciplinary Teams, John Wiley & Sons, Inc., 1976.

12. Wayne Wymore & Kingsley Forry, "Trades-off in Systems and Economics Analysis", To Be Published.

RADAR CROSS-SECTION DATA REDUCTION

Ernest J. Sanchez
National Range Operations Directorate
U.S. Army White Sands Missile Range

Abstract

The main theme of the paper is the exact solution for the normalized backscatter cross-section of a perfectly conducting sphere utilizing Mie Theory. The Mie equation was used along with spherical Bessel and Hankel functions, Recurrence Relations and Mie Theory with proper boundary conditions applied. The paper takes you through step by step procedure of the exact solution plus the tie-in of these results to the Radar Cross-Section Data Reduction System.

The problem of the scattering of electomagnetic waves from a sphere has received considerable attention due to a large extent to the fact that the rigorous solution has been known for a long time.

The rigorous solution, known as the Mie series, allows numerical results to be obtained to a high degree of accuracy.

The subroutine utilized for the Mie series calculation is ADECRS. The most critical part of subroutine is the evaluation of the required Bessel and Hankel functions. ADECRS was modified from a Fortran II program written by J. Rheinstein at M.I.T. for calculating the scattering of electromagnetic waves by a lossless, layered spherical Dielectric or Eaton lens. ADECRS contains built-in checking procedures, which are independent of the calculation algorithms to insure that sufficient accuracy in these functions as well as in other parts of the calculation is retained to give at least five significant figures in the result. The results from ADECRS have been compared with other published data for accuracy.

The exact solution for the backscatter cross-section for a perfectly conducting sphere is computed on the Univac 1108 computer utilizing double precision throughout the Fortran V subroutine ADECRS.

Following the method of Mie, as outlined in Stratton, the backscatter cross-section of a perfectly conducting sphere is represented by

$$\sigma_s = \frac{\pi r^2}{\rho^2} \left[\sum_{n=1}^{\infty} (-1)^n (2n + 1) (a_n^s - b_n^s) \right]^2$$

The paper contains an introduction, a mathematical analysis and derivation of Radar Cross-Section equations; a bibliography, plus 17 illustrations and a glossary section.

I. INTRODUCTION

A. Program Objective

The RCS165 Module is a collection of subroutines with a main program monitor developed for the A-Scope, AGC, and AGC-VCO radar cross-section (RCS) data reduction. RCS165's main function is to convert the digitized raw data to useful radar cross-section data on a daily production basis.

A new version of the Radar Cross-Section Data Reduction System was deemed necessary due primarily to the many additions and changes that have been made to the subroutines for greater clarity, completeness, efficiency and new program requirements.

The author would like to express his thanks to Liz Duran our secretary, for her tireless effort in typing this new document version.

B. Changes from the Previous Version

This is the first version of the program under MIPS.

C. Operating Environment

RCS165 utilizes Fortran V, and operates on the Univac 1108 with the EXEC 8 operating system as a Stand Alone Program. This module is of an auxiliary nature.

Core storage area is conserved by the utilization of overlaying technique (segmentation) in the subsystem.

A total of ten major output formats are generated by this system to cover all present WSMR requirements. Tape Record format and listing information was typed so as to allow for zeroxing of a specified format as requested by user without having to include other formats.

MATHEMATICAL ANALYSIS AND LOGIC

A. MATHEMATICAL ANALYSIS

1. GLOSSARY

The following glossary is provided so that readers may find most of the definitions and description symbols utilized in this section in one place. Many of the terms not defined here will be defined in the text in order to provide greater clarity and continuity. A term will normally be underlined to denote a definition.

Sections 2 through 12 contain the derivation of the major equations utilized in the radar cross-section data reduction subsystem module, RCS165.

1. GLOSSARY (CONTINUED)

A-SCOPE

Type A or A-Scope is a type of data presentation where the echo strength or target amplitude is indicated by vertical displacement (Y-axis) of the luminous spot and range by horizontal (X-axis) displacement on a cathode-ray tube. There is no angle information displayed on this type of data presentation.

Optical cameras are setup and calibration procedures are followed for A-Scope recording on film. The data is recorded on 35 mm film in three ways; Type I film is video pulse streak camera recording (strip film), Type II film is framing camera recording, and Type III film is intensity modulation streak camera recording. The timing on film is IRIG timing. This technical report covers the data reduction of strip and frame film data. A-Scope type data is primarily utilized when target discrimination, individual RCS for entire pulse, peak RCS, or other signature analysis is required.

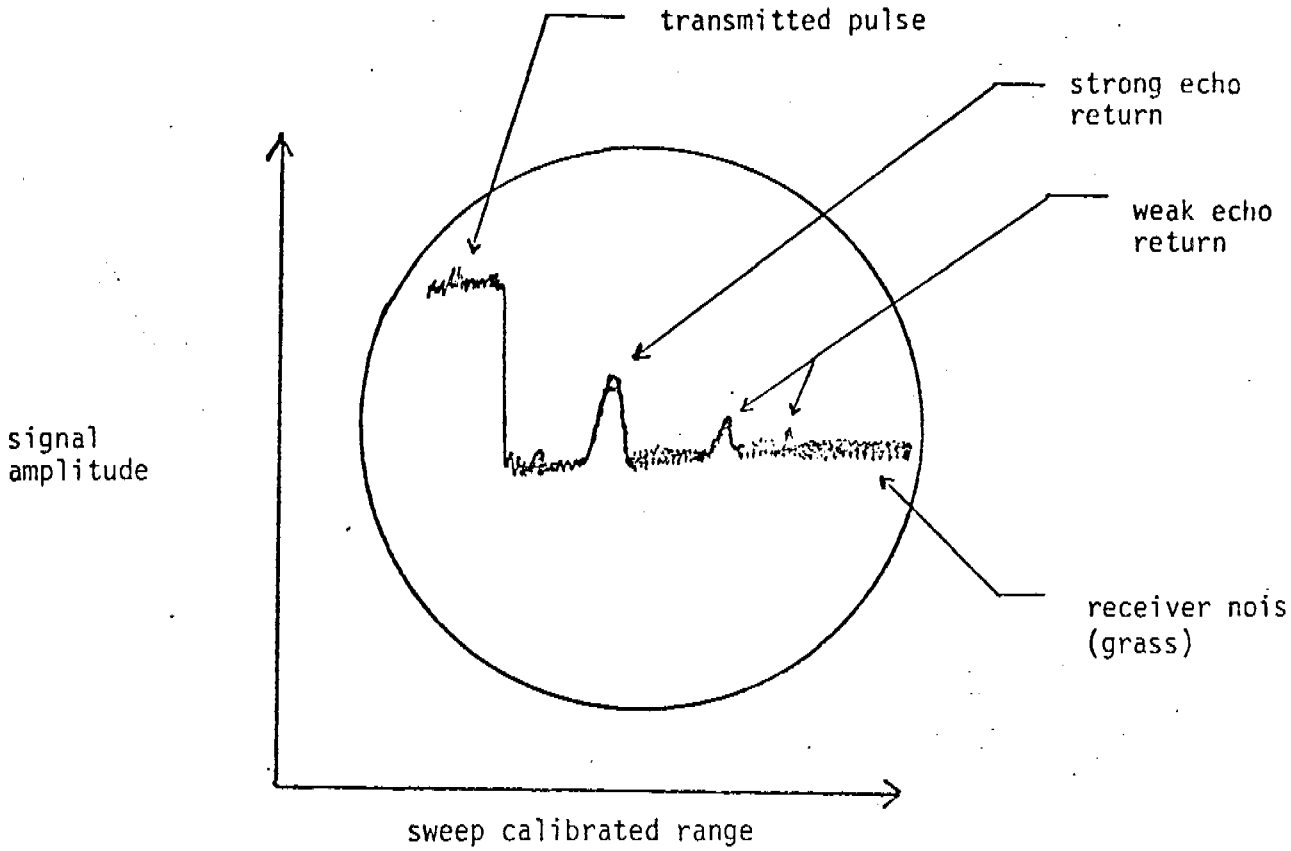
Another type of presentation is Type P or the plan position indicator (PPI) scope where a radial displacement of the spot from the center of the screen indicates range and the direction of the radial displacement indicates azimuth angle.

A third type is Type B in which vertical displacement indicates range and horizontal displacement indicates azimuth angle. There are many other display systems in current use, but only the A-Scope type is currently being reduced at WSMR. (See figure 2.1)

AUTOMATIC GAIN CONTROL (AGC)

The function of the AGC is to maintain the d-c level of the receiver output constant and to smooth or eliminate as much of the noiselike amplitude fluctuations as possible without disturbing the extraction of the desired error signal at the conical-scan frequency. One other purpose of the AGC in any receiver is to prevent saturation by large signals.

The digitized AGC data recorded on radar field tape is pre-calibrated on the boresight tower in 5 db steps from 0 db-noise level. This data is then stored in the Milgo ARCADE Computer memory and during mission the target video is referenced to stored AGC precal. ARCADE is an abbreviation for Automatic Radar Control and Data Equipment. The reference AGC level is then recorded on radar field tape. The digitized raw data is produced as a positive displacement above the noise level. Each radar launches their own spheres and conducts the sphere calibrations immediately after mission completion. This type of data is normally converted to peak RCS.



A-SCOPE DISPLAY SYSTEM
(Figure 2.1)

1. GLOSSARY (CONTINUED)

AUTOMATIC GAIN CONTROL-VOLTAGE CONTROL OSCILLATOR (AGC-VCO)

AGC-VCO describes a type of raw data available from the FPS16's. The detected video is recorded on analog tape by a telemetry ~~van~~ station utilizing FM methods. The data is later digitized to machine counts and from there to engineering units of db through the use of pre or post step calibrations. This type of data is normally converted to peak RCS.

BEAMWIDTH (θ_b)

The symbol theta sub b denotes beamwidth. The width of the beam is usually specified by the beam angle between half-power points. This is the angle between lines on opposite sides of the beam axis along which the power density is half as great as it is on the axis. The beam angle serves as a measure of the angular accuracy and angular resolution of a radar set.

DECIBEL (db)

A unit for expressing the magnitude of change in electrical power level. The bel is the fundamental division of a logarithmic scale expressing the ratio of two amounts of power, the number of bels denoting such a ratio being the logarithm to the base 10 of this ratio. The decibel is one-tenth of a bel. For example, with P_1 and P_2 designating two amounts of power in watts and n the number of decibels denoting their ratio,

$$n(db) = 10 \log_{10} \frac{P_1}{P_2}$$

DECIBEL REFERENCED TO ONE SQUARE METER (dbsm)

If X is in square meters and X_{ref} is equal to 1 square meter then

$$X(dbsm) = 10 \log_{10}(X/X_{ref}).$$

In this technical report, radar cross-section is computed in square meters first, and then the above equation is utilized to compute dbsm.

1. GLOSSARY (CONTINUED)

ELECTROMAGNETIC WAVE

An electromagnetic wave consists of coupled electric (E) and magnetic (B) field oscillations.

FOOT

The definition of a foot is the one used by the United States Coast and Geodetic, rather than the international definition of one inch equals 2.54 centimeters. This was selected since all the site surveys are based on this definition. Exactly 3937 feet are equal to 1200 meters exactly. These values were obtained from IRIG Document 104-64, "A Glossary of Range Terminology."

FREQUENCY (f)

The number of waves (cycles) that go by a point in a unit of time (seconds). It is normally seen as megacycles (Mc) for radar frequency; such as, 5490 Mc per second for an FPS16.

PEAK TRANSMITTED POWER (P_o)

The parameter P_o is actually the rms power during the pulse. It is usually expressed in megawatts or 10^9 milliwatts.

POLARIZATION

For a plane electromagnetic wave, the electric field vector must always point in a direction perpendicular to the direction of propagation. If this direction is constant, the electric field lies entirely in one plane and the electromagnetic wave is said to be linearly polarized. If the direction rotates with time at a constant rate the wave is said to be elliptically polarized. A circular polarized wave is a special case of elliptical polarization. (See vertical and circular polarization).

1. GLOSSARY (CONTINUED)

POLARIZATION (CIRCULAR)

If the electric field vector appears to be rotating clockwise to an observer looking in the direction of propagation, the polarization is said to be right circular or right elliptical. Counterclockwise rotation looking in the direction of propagation is, of course, designated left circular or left elliptical. This is the IEEE standard definition. The reader should be warned that many authors use the opposite definition. The FPS16's did have the capability for circular polarization at one time, but this has been removed.

POLARIZATION (VERTICAL)

Linear polarization can be divided into two orthogonal polarizations which are called horizontal and vertical polarization with the planes containing the electric field lying parallel and perpendicular to the earth's surface. The FPS16's transmit and receive a perpendicular (vertical) field. Associated with each electric field there is a magnetic field whose direction is perpendicular to both the electric field and the direction of propagation, but polarization is defined in terms of the electric field vector.

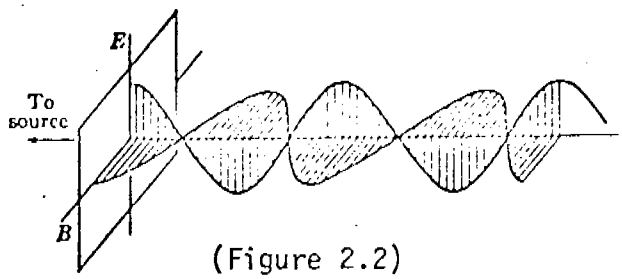
POWER RETURN OR POWER RECEIVED (P_r)

The radar receiver return power is not measured directly but rather is recorded as a power level (S) so many db above (or below) the receiver noise power (N). The receiver noise power is considered to be constant for all practical purposes but may contain a slight drift. The raw data from A-Scope, AGC, AGC-VCO, and the sanborns (strip charts) is therefore a record of the ratio (S/N) expressed in decibels (db). All four types of raw data can be expressed as a positive displacement above noise or in direct proportion to the step calibrations.

1. GLOSSARY (CONTINUED)

POYNTING VECTOR

The poynting vector describes the flow of energy in an electromagnetic wave. Its direction is that of the wave and its magnitude is equal to the rate at which energy is being transported by the wave per unit cross-sectional area (watts/m^2).



(Figure 2.2)

The variations in an electromagnetic wave occur simultaneously in both fields so that maxima and minima occur at the same times and places. The direction of the electric and magnetic fields are perpendicular to each other and to the direction the wave is moving. Waves are therefore transverse. The speed of the waves depends only upon the electric and magnetic properties of the medium they travel in and not upon the amplification of the field variations.

PROPAGATION VELOCITY

The propagation velocity is the velocity of light in a vacuum. The value obtained from IRIG Document 104-64 is 299.7925 meters per microsecond.

PULSE DURATION (τ)

The symbol tau denotes pulse duration which is the length of time a transmitter emits energy. Pulse duration determines range resolution. A pulse duration of one microsecond would result in a range resolution cell of approximately 492 feet or 1 microsecond = 149.89625 meters.

1. GLOSSARY (CONTINUED)

PULSE REPETITION FREQUENCY (PRF)

PRF is the rate at which pulses are transmitted for a periodic pulse train; for example, 160 PRF or 160 pulses every 10^6 μ seconds. One of the uses of a radar's PRF is to determine maximum range from which echoes can be returned without ambiguity. The propagation velocity (c) is divided by 2PRF. An example would be R(max) = 299.7925 m per μ sec divided by 2(160 pulse per 10^6 μ sec) which is approximately equal to 936,852 meters or about 580 statute miles.

RADAR CROSS-SECTION (σ) or (RCS)

The symbol sigma or RCS denotes radar cross-section or alternatively back-scattering cross-section. It's defined as the area intercepting that amount of power which, when scattered isotropically, produces an echo equal to that observed from the target. An alternate definition is the cross-sectional area of a perfectly conducting sphere that would return the same power to the radar as does the actual target. RCS is normally expressed in M^2 or dbsm. A third definition is the area which would intercept sufficient power out of the transmitted field to produce the given echo by isotropic re-radiation. All three definitions are essentially equivalent to each other.

The FPS16's are monostatic or back-scatter radars. This means that the receiving and transmitting antennas are one and the same. Back-scatter cross-section of a sphere (σ_s) is the value in M^2 or dbsm computed in sub-routine ADECRS by the utilization of Mie theory for exact solution of a perfectly conducting sphere.

RANGE (R)

Range is defined as the propagation velocity multiplied by delta time divided by 2. Delta time is the time required electromagnetic energy wave to travel out to object being tracked and back.

SIGNAL STRENGTH (S)

A measure of the power output of a radar at a particular location, normally the radar receiver. To measure signal strength (S) at radar receiver, a sphere calibration should be utilized.

SIGNAL-TO-NOISE RATIO (S/N)

In radar, the ratio of the value of the signal to that of the noise. The ratio is expressed in decibels and is expressed as the ratio of root-mean-square signal voltage during the pulse to root-mean-square noise voltage, as measured at any point in the i-f portion of the receiver following the introduction of substantial gain and restriction of the noise bandwidth.

WAVELENGTH (λ)

The symbol lambda denotes the wavelength of the energy transmitted by radar in meters per cycle. It is calculated by dividing the propagation velocity (c) by the frequency of radar.

WHITE SANDS CARTESIAN SYSTEM (WSCS) LEFT HAND SYSTEM

A Cartesian coordinate system with origin at the intersection of latitude (ϕ_0) 33° 05'0.000" North and longitude (λ_0) 106° 20'0.000" West. At this origin, the XY plane is tangent to the Clark Spheroid of 1866 with the semi-major axis (a) equal to 6,378,206.4 meters and the semi-minor axis (b) equal to 6,356,583.8 meters. The eccentricity squared major (e^2) = $1-b^2/a^2$ is equal to 0.00676 86579 97291. The eccentricity squared minor ($(e^1)^2$) is equal to 0.00681 47849 45915. The origin has a value of 500,000.00 feet East (E_0), 500,000.00 feet North (N_0), and 0.00 feet for (Z_0). The Y-axis is an east-west line in the tangent plane passing through the origin and increasing positively eastward. The X-axis is a north-south line in the tangent plane passing through the origin and increasing positively northward. The Z-axis is perpendicular to the tangent plane and increasing positively upward. There are numerous systems available for WSMR-provided data which are fully documented in Data Reduction Division Technical Report titled, "Coordinate Systems Related to WSMR," July 1964 where above values were obtained.

2. RADAR RANGE EQUATION

In this derivation, no propagation effects are considered. The radar set and the missile are assumed to be isolated bodies in space. The antenna is positioned so that maximum radiation is directed toward the object.

The term "target" will refer to the object whose radar cross section is desired.

Let R = the slant range from the radar to the target

P_o = the peak power of the rectangular pulse transmitted from the antenna.

This peak power, P_o, is spread over the spherical wavefront traveling outward from the antenna. When the wavefront reaches the target, the area of its spherical surface is $4\pi R^2$, and if the radiation were uniform in all directions the power density at the target would be $P_o/4\pi R^2$. Because the antenna is directional, the distribution of power is not uniform. The power density is greater than $P_o/4\pi R^2$ at the target and less at other positions on the wavefront. Thus the power density at the target is

$$(2.1) \qquad (P_o/4\pi R^2)G_t$$

where G_t is a factor greater than 1 to take account of the concentration of energy in the direction of the missile. A fictitious antenna that radiates uniformly in all directions is called an isotropic antenna, and G_t is called the power gain of the transmitting antenna relative to an isotropic antenna. The gain G_t is the ratio of the actual power density at the target to the power density that would be produced by an isotropic antenna. Part of the transmitted wave is reflected from the target. Wavefronts of the reflected wave are expanding spheres centered at the target. Suppose, temporarily, the target is a sphere having a perfectly conducting surface and cross sectional area. Such a sphere collects the power in an area σ of the incident wave and reradiates this power uniformly in all directions, provided the diameter of the sphere is large relative to the length of the radio waves. The reflected wavefront has an area of $4\pi R^2$ when it reaches the radar antenna, and therefore the power density of the reflected wave at the radar antenna is

$$(2.2) \qquad (P_o/4\pi R^2)(G_t)(\sigma/4\pi R^2)$$

Actual targets may extract more or less energy from the transmitted wave than that contained in their cross sectional area. Furthermore, they absorb some of this energy and reradiate only the remaining part, and the radiation is usually far from uniform in direction. Nevertheless, the reflecting ability

of any target can be described by an area σ, called the radar cross section of the object, which is the cross sectional area of a perfectly conducting sphere that would return the same power to the radar as does the actual target. Equation 2.2 thus applies to all targets if σ is considered to stand for radar cross section.

The power absorbed from the reflected wave by the radar antenna is proportional to the power density of the wave when it reaches the antenna. Therefore, the received power, P_r, is:

(2.3) $$P_r = (P_o/4\pi R^2)(G_t)(\sigma/4\pi R^2)A$$

(The Radar Equation)

$$= (P_o G_t \sigma A) / (4\pi)^2 R^4$$

where A is the constant of proportionality. This constant has units of area and is called the effective area of the receiving antenna. Equation 2.3 is known as the radar range equation.[7]

In the FPS16's the antenna disk is a circular paraboloid which is the surface generated by rotating a parabolic curve about its axis. The properties of the parabola which makes it particularly useful for focusing radiant energy into a directional beam are characterized by two ray considerations: First, any ray from the focus is reflected in a direction parallel to the axis of the parabola; and second, the distance traveled by any ray from the focus to the parabola and by reflection to a plane perpendicular to the parabola axis is independent of its path, and therefore such a plane represents a wave front of uniform phase.[25]

Calculations of the antenna efficiency based on the aperture distribution set up by the primary pattern as well as the spillover indicate theoretical efficiencies of about 80 per cent for paraboloidal antennas when compared across the aperture, poor polarization characteristics, and antenna mismatch reduce the efficiency to the order of 55 to 65 per cent for ordinary parabolic reflector. It should be noted that the antenna aperture of a parabolic reflector is the area projected on a plane perpendicular to its axis and is not the surface area.[4]

The aperture efficiency factor, Neta sub a (η_a) is included for variations in illumination over the aperture. A circular aperture with edge illumination reduced to suppress rate lobes will have η_a equal to 60 per cent.[26]

2. RADAR RANGE EQUATION (CONTINUED)

The theoretical gain for circular aperture may then be represented by

$$(2.4) \qquad G_r = \frac{4\pi A n_a}{\lambda^2} = \frac{4\pi(\pi r^2) n_a}{\lambda^2}$$

if $\frac{r}{\lambda} >> 1$ where r is the radius of the reflector and lambda (the wavelength) is the propagation velocity divided by the frequency of radar.[25] For the FPS16, r/λ is approximately equal to 32.

The half angle subtended by paraboloid at focus is about 35 degrees for a gain efficiency of aperture of about 60 per cent.

The beamwidth in degrees may be approximated utilizing the following equation for circular aperture.

$$(2.5) \qquad \theta_b \cong \frac{1.2\lambda}{D} \left(\frac{360}{2\pi}\right)$$

where D is the diameter of reflector.

Thus equation 2.3 may be rewritten in the form:

$$(2.6) \qquad P_r = \frac{P_o G_t G_r \sigma \lambda^2}{(4\pi)^3 R^4}$$

If the same antenna is used for transmitting and receiving we have $G_t = G_r = G$ and we may write 2.3 as

$$(2.7) \qquad P_r = \frac{P_o G^2 \lambda^2 \sigma}{(4\pi)^3 R^4}$$

This latter form will be used in the rest of the section to derive the necessary formula.

3. BACKSCATTER CROSS-SECTION FOR A PERFECTLY CONDUCTING SPHERE (σ_s)

The problem of the scattering of electromagnetic waves from a sphere has received considerable attention due to a large extent to the fact that the rigorous solution has been known for a long time.

The rigorous solution, known as the Mie series, allows numerical results to be obtained to a high degree of accuracy.

The subroutine utilized for the Mie series calculation is RCSCRS. The most critical part of subroutine is the evaluation of the required Bessel and Hankel functions. RCSCRS was modified from a Fortran II program written by J. Rheinstein at M.I.T. for calculating the scattering of electromagnetic waves by a lossless, layered spherical Dielectric or Eaton lens, see references. RCSCRS contains built in checking procedures, which are independent of the calculation algorithms to insure that sufficient accuracy in these functions as well as in other parts of the calculation is retained to give at least five significant figures in the result. The results from RCSCRS have been compared with other published data for accuracy.

The exact solution for the backscatter cross-section for a perfectly conducting sphere is computed on the Univac 1108 computer utilizing double precision throughout the Fortran V subroutine RCSCRS.

Following the method of Mie, as outlined in Stratton,[22] the backscatter cross-section of a perfectly conducting sphere is represented by

$$\sigma_s = \frac{\pi r^2}{\rho^2} \left[\sum_{n=1}^{\infty} (-1)^n (2n+1) (a_n^s - b_n^s) \right]^2 \tag{2.8}$$

Since the values of the real and imaginary coefficients Re $\{a_n\}$, Im $\{a_n\}$, Re $\{b_n\}$, Im $\{b_n\}$, decrease rapidly for $n > \rho$, it is necessary to carry summation only to $n = 2\rho + 5$ or a maximum value for n of about 245 for $r/\lambda = 19.00$. The following notation is utilized.

$$\rho = \frac{2\pi r}{\lambda} \tag{2.9}$$

where r is the sphere radius, and λ = velocity of light/frequency.

$$(2.10) \qquad a_n^s = -j_n(\rho) \,/\, h_n^{(2)}(\rho)$$

where $j_n(\rho)$ and $h_n^{(2)}$ are spherical Bessel and Hankel functions and primes denote differntiation with respect to argument.

$$(2.11) \qquad b_n^s = -[\rho\, j_n(\rho)]' \,/\, [\rho\, h_n^{(2)}(\rho)]'$$

The above equation was utilized along with spherical Bessel and Hankel functions which are evaluated right in the program RCSCRS by the use of recurrence relations.

Most of the problems and difficulties that are encountered are related to the generation of the necessary spherical Bessel and Neumann functions and to the evaluation of the determinants.

The spherical Bessel and Neumann functions are calculated by the substitution of $\Lambda_n(\rho)$ and $\Lambda_{-n}(\rho)$ for $j_n(\rho)$ and $n_n(\rho)$, respectively. These functions are related by

$$(2.11\text{-}1) \qquad j_n(\rho) = \rho^n \Lambda_n(\rho) \frac{2^n\, n!}{(2n+1)!}$$

and

$$(2.11\text{-}2) \qquad n_n(\rho) = \rho^{-(n+1)} \Lambda_{-n}(\rho) \frac{(-1)^n (2n)!}{2^n\, n!}$$

The function $\Lambda_{-n}(\rho)$ can be evaluated very simply since

$$(2.11\text{-}3) \qquad \Lambda_{-o}(\rho) = \cos x$$

(2.11-4) $$\Lambda_{-1}(\rho) = \rho \sin \rho + \cos \rho$$

and the recursion relation

(2.11-5) $$\Lambda_{-n-1}(\rho) = \Lambda_{-n}(\rho) - \frac{\rho^2}{4(-n - 0.5)(0.5 - n)} \Lambda_{-n+1}(\rho)$$

can be used for calculating higher orders.

The function $\Lambda_n(\rho)$ is more difficult to evaluate.

Assume that the argument is ρ and the largest desired order is m. Choose k equal to approximately the maximum of m + 20, ρ + 30 or 35. Then set

(2.11-6) $$H_k(\rho) = 0.0$$

(2.11-7) $$H_{k-1}(\rho) = 10^{-6}$$

then, employing the recursion relation,

(2.11-8) $$H_{j-1}(\rho) = H_j(\rho) - \frac{\rho^2 H_{j+1}(\rho)}{4(j + 0.5)(j + 1.5)}$$

determine all $H_j(\rho)$, $j = k - 2, k - 3, \ldots, 1$

Evaluate

$$\Lambda_1(\rho) = \left(\frac{\sin \rho}{\rho} - \cos \rho\right) \frac{3}{\rho^2} \tag{2.11-9}$$

and determine the ratio

$$Y = \Lambda_1(\rho) \,/\, H_1(\rho) \tag{2.11-10}$$

It will then be found that

$$\Lambda_n(\rho) = Y H_n(\rho) \tag{2.11-11}$$

where $n = 2,3,4, \ldots, m$

and that the required functions are evaluated.

The functions obtained in this manner may be checked by employing the relation

$$\Lambda_n(\rho)\, \Lambda_{-n-1}(\rho) - \frac{\rho^2}{4(n+0.5)(n+1.5)}\, \Lambda_{n+1}(\rho)\, \Lambda_{-n}(\rho) = 1 \tag{2.11-12}$$

The determinants, as employed, have a form which may be evaluated in a relatively simple manner. If c(i,j) denotes the element in row i and column j, the nonzero terms are c(i,i-1), c(i,i), c(i,i+1), c(i,i+2), for i odd, and c(i,i-2), c(i,i-1), c(i,i), c(i,i+1), for i even. However, c(1,0), c(2,0), c(2N-1,2N+1), and c(2N,2N+1) do not exist.

After some manipulation of the determinants, the Mie coefficients may be put into a form such that

$$\mathrm{Re}\,\{a_n\} = \frac{-N_n^2}{N_n^2 + D_n^2} \tag{2.11-13}$$

and

$$\text{(2.11-14)} \qquad \mathrm{Im}\{a_n\} = \frac{N_n D_n}{N_n^2 + D_n^2}$$

The determinants N_n and D_n are identical except for the terms c(1,1) and c(2,1). Similar forms are obtained for the coefficients b_n by simply multiplying each element of rows 2i + 1, i = 0,1,2, . . . by the factor $\varepsilon(i + 1)$. where $\varepsilon = 1$ for a dielectric in free space. All c(i,j) are real.

If the above notation is employed, the determinant may be evaluated by the following scheme.

Evaluate

$$\text{(2.11-15)} \qquad B(2,2) = c(1,1)\, c(2,3) - c(1,3)\, c(2,1)$$

$$\text{(2.11-16)} \qquad B(2,1) = c(1,1)\, c(2,4) - c(1,4)\, c(2,1)$$

$$I = 3$$

$$\text{(2.11-17)} \qquad B(1,1) = c(I,I-1)\, B(2,2) - C(I,I-2)\, B(2,1)$$

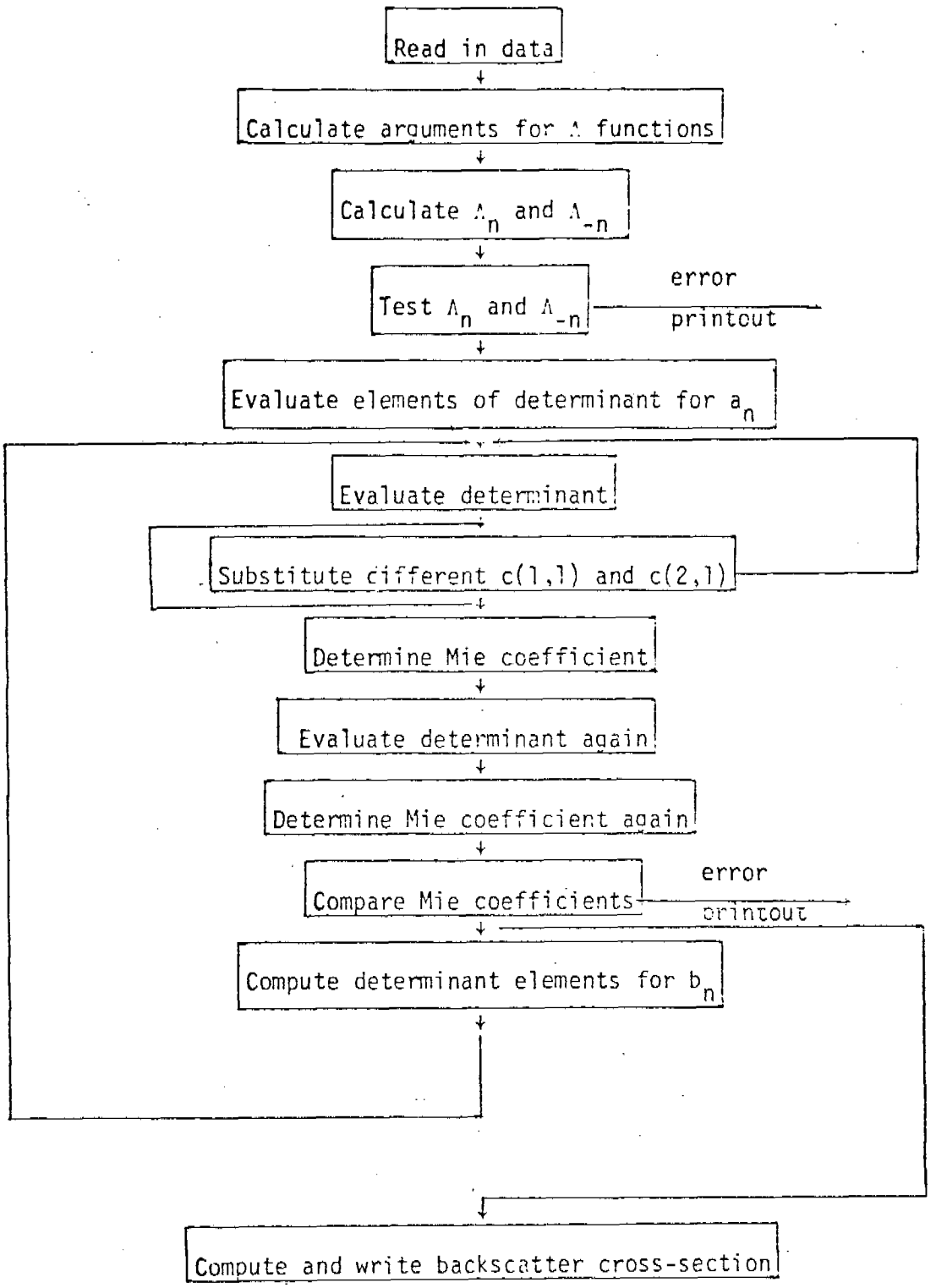
$$\text{(2.11-18)} \qquad B(1,2) = c(I + 1, I - 1)\, B(2,2) - c(I + 1, I-2)\, B(2,1)$$

$$\text{(2.11-19)} \qquad B(2,2) = c(I + 1, I)\, B(1,1) - c(I,I)\, B(1,2)$$

$$\text{(2.11-20)} \qquad B(2,1) = c(I + 1, I + 1)\, B(1,1) - c(I, I + 1)\, B(1,2)$$

Then set I = I + 2 and recompute B(1,1), B(1,2), etc. The value of the determinant is the value of B(2,2) when I has reached the value 2N - 1. This scheme for evaluating the determinants was found by employing Laplace's expansion by minors. As a check, the determinants are evaluated a second time by employing a similar scheme found by a different form of Laplace's expansion.

A simplified flow chart follows to aid readers through subroutine RCSCRS. (see figure 2.2-1)


Read in data
Calculate arguments for A functions
Calculate A_n and A_{-n}
Test A_n and A_{-n}
error printout
Evaluate elements of determinant for a_n
Evaluate determinant
Substitute different c(1,1) and c(2,1)
Determine Mie coefficient
Evaluate determinant again
Determine Mie coefficient again
Compare Mie coefficients
error printout
Compute determinant elements for b_n
Compute and write backscatter cross-section

(Figure 2.2-1)

3. BACKSCATTER CROSS-SECTION FOR A PERFECTLY CONDUCTING SPHERE (σ_s) (CONTINUED)

The data in table and plots has been normalized so as to be dimensionless. The following table is included to provide the position of the minima and maxima of the normalized backscatter from a conducting sphere.

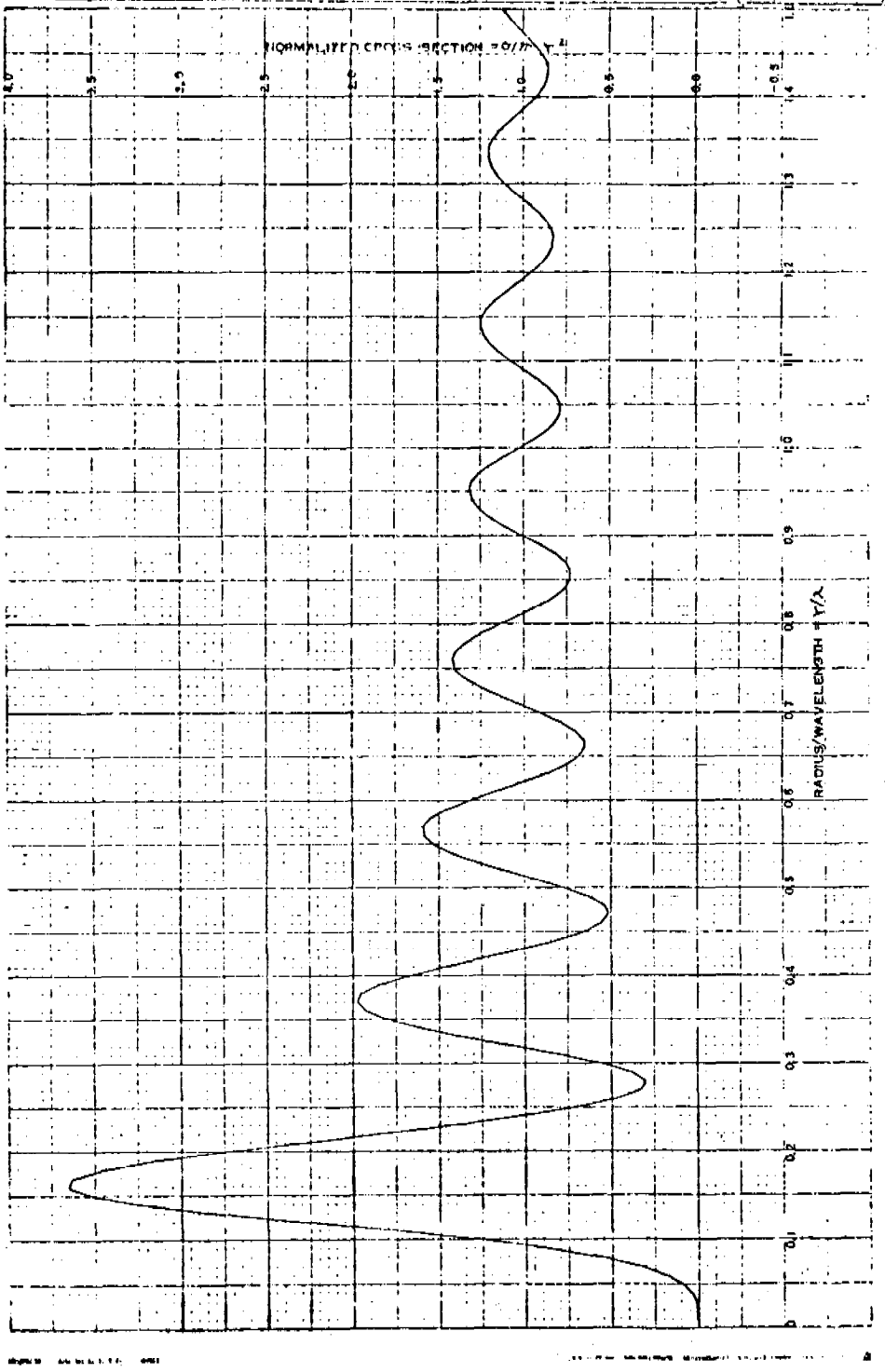
POSITION OF MAXIMA	CROSS-SECTION	POSITION OF MINIMA	CROSS-SECTION
r/λ	$\sigma/\pi r^2$	r/λ	$\sigma/\pi r^2$
0.1636	3.65495	0.2775	0.285041
0.3717	1.96958	0.4707	0.505600
0.5659	1.58864	0.6626	0.634592
0.7578	1.41048	0.8543	0.716232
0.9495	1.30698	1.0458	0.772456
1.1413	1.24023	1.2374	0.813423
1.3331	1.19398	1.4292	0.844286
1.5249	1.16016	1.6210	0.868137
1.7169	1.13451	1.8130	0.887011
1.9089	1.11453	2.0051	0.902239
2.1011	1.09864	2.1973	0.914714
2.2933	1.08576	2.3896	0.925057
2.4857	1.07515		

(Figure 2.3)

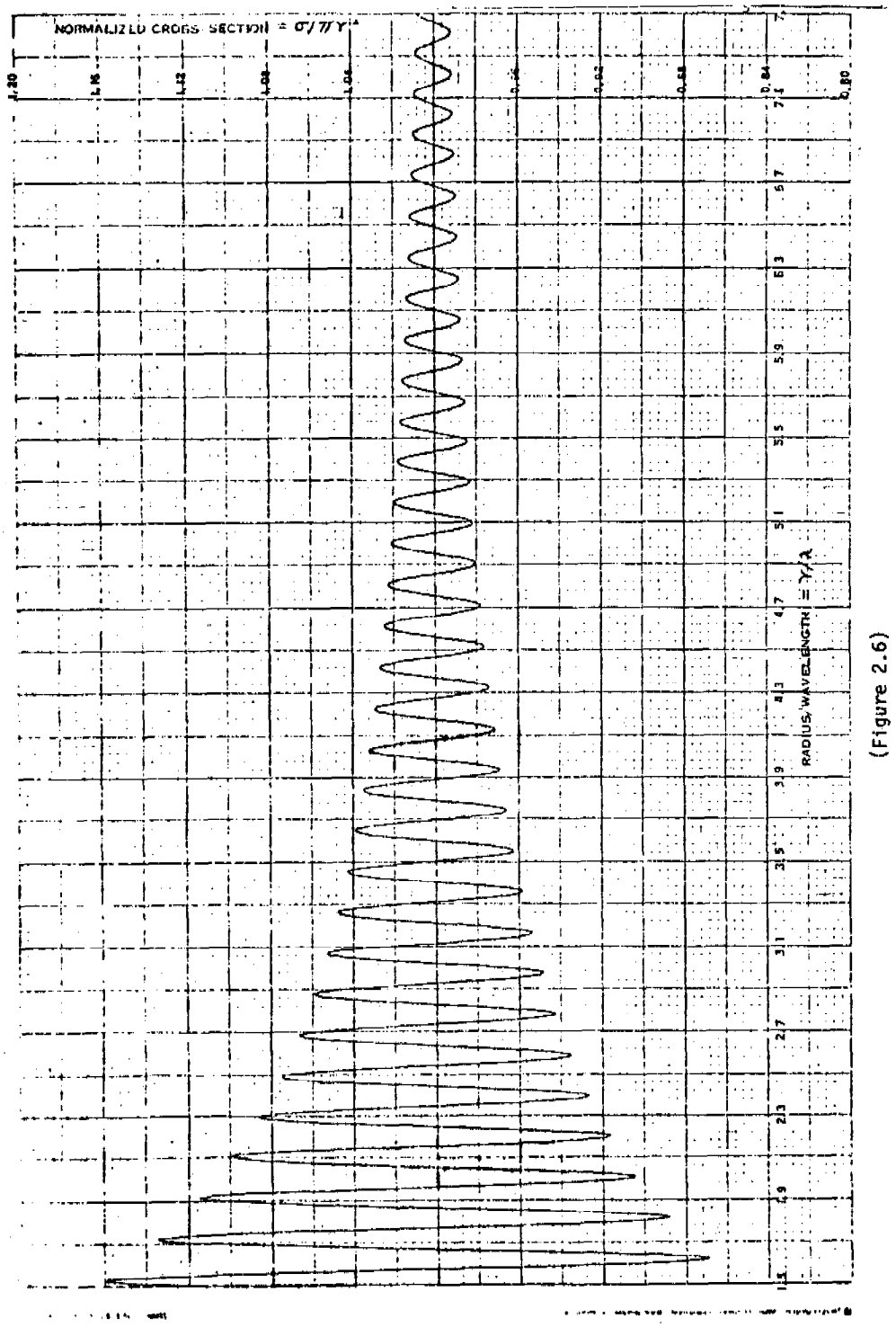
The following plots were generated from RCSCRS data and were plotted to four significant figures utilizing the EAI 3440 digital plotter.

(Figure 2.4)

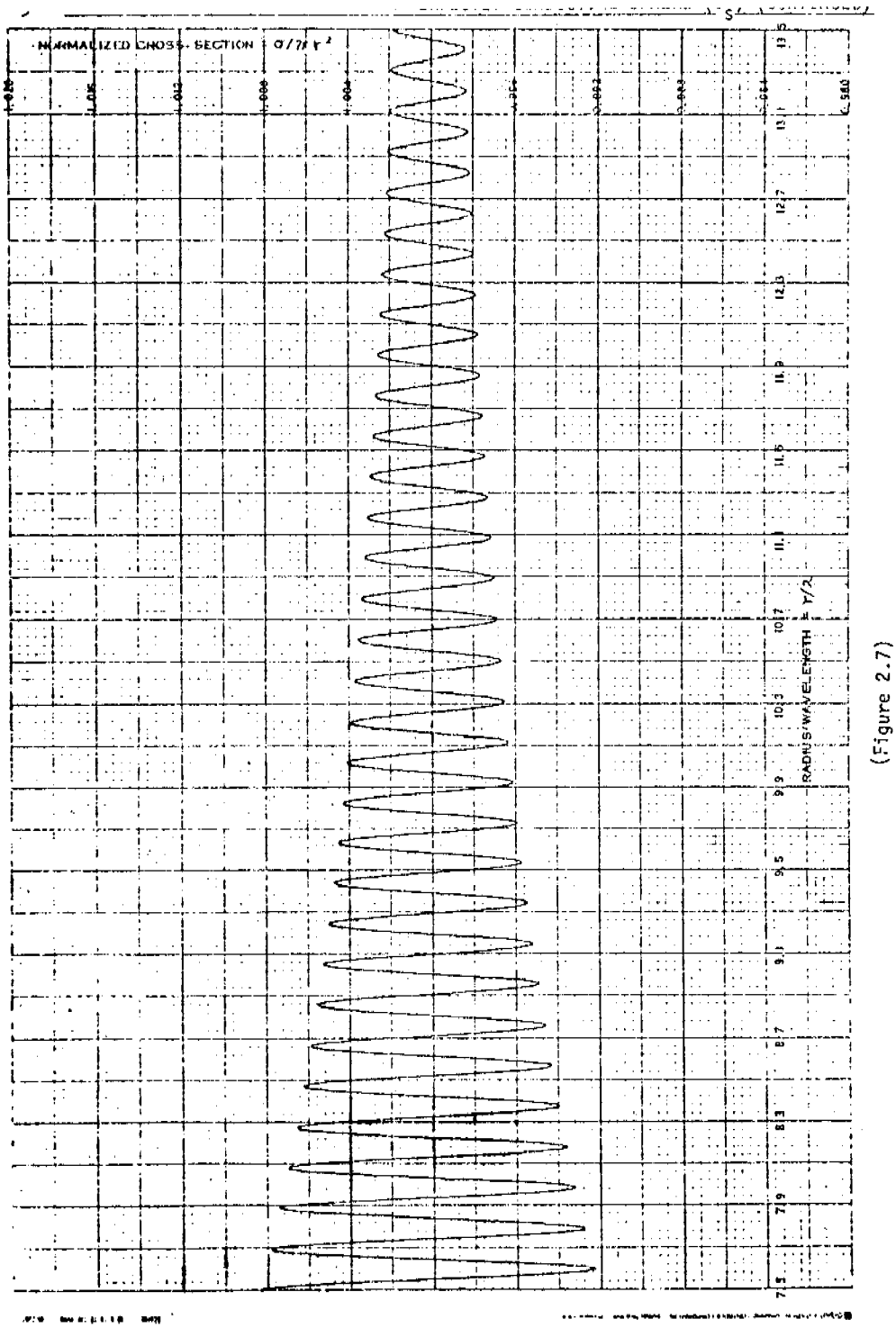
3. BACKSCATTER CROSS-SECTION FOR A PERFECTLY CONDUCTING SPHERE (σ_c) (CONTINUED)



(Figure 2.5)

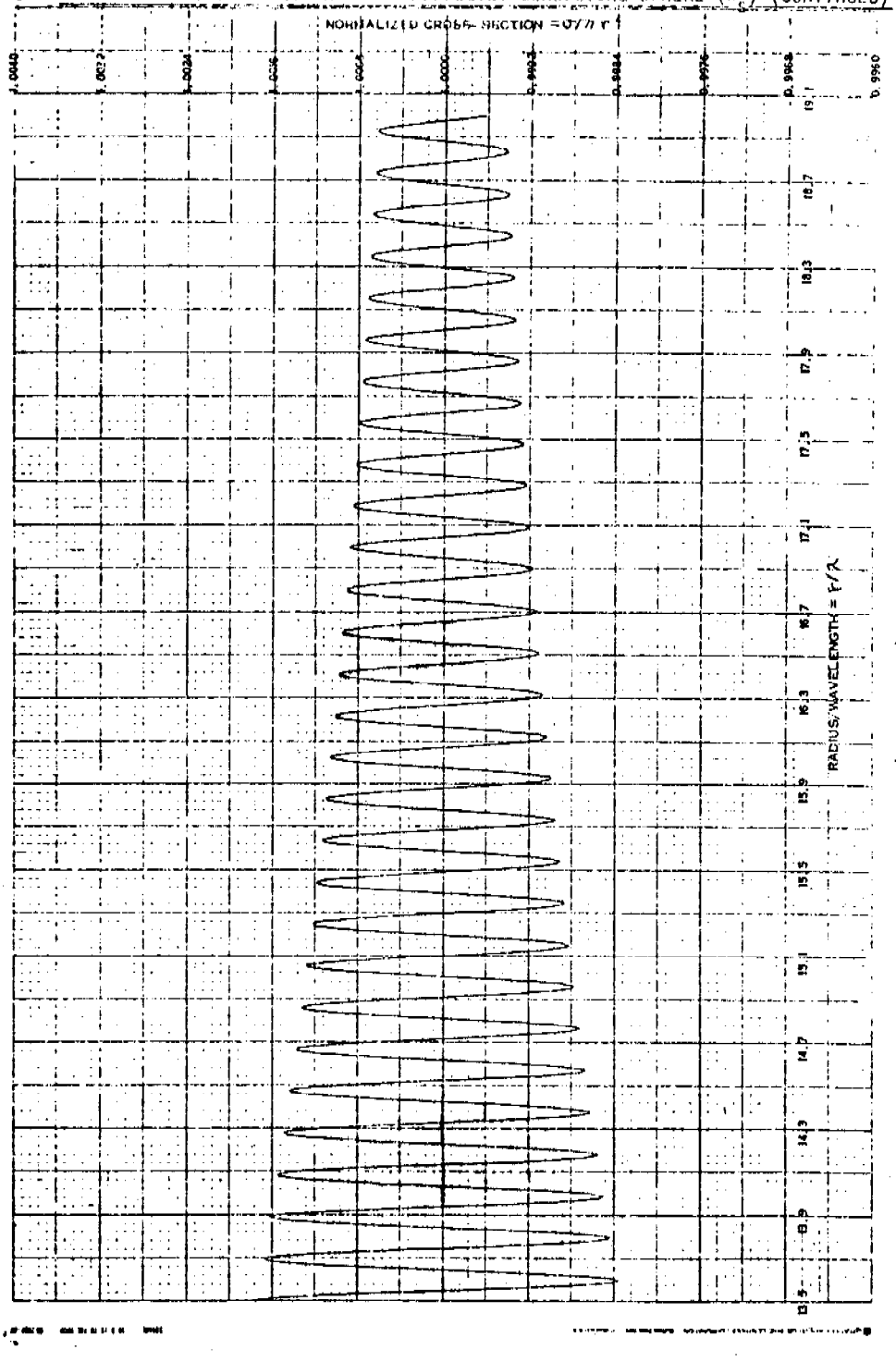


(Figure 2.6)



(Figure 2.7)

3. BACKSCATTER CROSS-SECTION FOR A PERFECTLY CONDUCTING SPHERE (σ_c) (CONTINUED)



(Figure 2.8)

4. SIGNAL STRENGTH

Signal strength (S) is the measurement of the power output of a radar at a particular location. To generate (S), a sphere calibration should be employed. The signal strength in milliwatts may be computed utilizing

$$S = \frac{P_o G^2 \lambda^2 \sigma_s}{(4\pi)^3 R_s^4} \tag{2.12}$$

$$S(\text{dbm}) = 10 \log (S/1 \text{ milliwatt}) \tag{2.13}$$

5. SIGNAL-TO-NOISE RATIO (S/N)

The radar equation as derived by D. Barton is as follows for i-f signal-to-noise ratio utilized to measure radar effectiveness.

$$(2.14) \qquad P_r' = \frac{S}{N} = \frac{P_o G^2 \lambda^2 \sigma}{(4\pi)^3 k T_o B \overline{NF}_o L R^4}$$

This i-f or "single pulse" S/N ratio does not measure the over-all effectiveness of the radar, but serves as an intermediate step in a number of further calculations.

Departure from free-space propagation conditions must be accounted for by appropriate components of the total system loss factor L or by modification of the antenna gain G.

The transmitted power is assumed to lie within the bandwidth B of the receiver, after reflection from the target and propagation through the medium.

In the above equation NF_o is the operating receiving system noise factor, S is the signal strength, k is Boltzmann's constant, (1.38×10^{-23} watt per cps per degree Kelvin), $T_o = 290°$ K, B is the equivalent noise bandwidth of the receiver in cps. A more rigorous discussion of the above may be found in Radar System Analysis by David K. Barton where this section has been taken from.

6. SPHERE EQUATION

The sphere equation is basically the radar range equation in which the units of P_r have been converted from watts to db. We will assume in this subsection that the values P_o, G, λ, and σ in equation 2.14 are constant. The assumption that σ and aspect angles be constant at all times means, for all practical purposes, that a sphere must be the object tracked.

With the above assumptions equation 2.14 becomes:

$$(2.15) \qquad P_r = \frac{K}{R^4} \quad \text{where } K = (P_o G^2 \lambda^2 \sigma_s) \,/\, (4\pi)^3 k T_o B \overline{NF}_o L$$

By taking the log of both sides of 2.15 we obtain

$$(2.16) \qquad \log P_r = \log K - 4 \log R$$

Multiply both sides of 2.16 by 10

$$(2.17) \qquad 10 \log P_r = 10 \log K - 40 \log R$$

Since $10 \log P_r = P_r\,(db) + 10 \log P_{ref}$ for P_r in watts and P_{ref} the reference power, we may substitue in 2.17 and get

$$(2.18) \qquad P_r(db) + 10 \log P_{ref} = 10 \log K - 40 \log R$$

Or

$$P_r(db) = - 40 \log R + 10 \log K - 10 \log P_{ref}$$

But K and P_{ref} are constant and thus we can make the substitution

$$b = 10 \log K - 10 \log P_{ref}$$

And obtain

$$(2.19) \qquad P_r(db) = - 40 \log R + b \quad \text{(The sphere equation)}$$

Thus the received power in db varies linearly as the log of the range when the target being tracked is a sphere or has constant cross section.

6. SPHERE EQUATION (CONTINUED)

A BBAR is then computed to be utilized in cross-section equation later. The equation is written:

$$\bar{b}(db) = \frac{1}{n} \sum_{i=1}^{n} [P_{ri}(db) + 40 \log R_i] \tag{2.20}$$

Or

$$\bar{b}(db) = \frac{1}{n} \sum_{i=1}^{n} b_i \tag{2.21}$$

7. CROSS-SECTION EQUATION

The problem with using the range equation directly to solve for radar cross section is that of determining P_o, G, and λ. While P_o and λ are not difficult to calculate, G is extremely hard to find in a practical situation. The reason is that receiver gain must also be considered. Henceforth G will be used to mean "power gain" in the system. Also, it may be possible on some radars to vary P_o and λ. We may assume however, that for a short time interval, say one or two hours, that P_o, and G, and λ remain constant. This assumption means that the radar operators must not change these values during this time period.

In this subsection the notation will be as previously described but an additional subscript t will mean that term refers to the target, an s will mean the term refers to the sphere. e.g., P_r has been used for received power, thus P_{rt} means power received from the target and P_{rs} means power received from the sphere.

To derive the cross section equation we will divide the range equation for the target by the range equation for the sphere, then solve the resultant for σ_t. It is assumed that throughout the target and sphere tracks that P_o, G, and λ remain the same. Thus we have:

$$P_{rt} = \frac{P_o G^2 \lambda^2 \sigma_t}{(4\pi)^3 R_t^4} \tag{2.22}$$

$$P_{rs} = \frac{P_o G^2 \lambda^2 \sigma_s}{(4\pi)^3 R_s^4} \tag{2.23}$$

Therefore

$$\frac{P_{rt}}{P_{rs}} = \frac{\sigma_t}{\sigma_s} \frac{R_s^4}{R_t^4} \tag{2.24}$$

Hence we get

$$\sigma_t = (P_{rt}/P_{rs})(R_t/R_s)^4 \sigma_s \tag{2.25}$$

7. CROSS-SECTION EQUATION (CONTINUED)

Since the power term in 2.25 refers to power received, the subscript r is usually dropped and the equation written:

$$\sigma_t = (P_t/P_s)(R_t/R_s)^4\sigma_s \tag{2.26}$$

The cross-section equation 2.26 is further changed by letting:

$$RFAC = \frac{\sigma_s}{P_s R_s^4} \tag{2.27}$$

Taking the log of both sides, then multiplying both sides by 10 and using definition $\log \frac{X}{YZ} = \log X - \log Y - \log Z$ on equation 2.27 we get:

$$10 \log RFAC = 10 \log \sigma_s - (10 \log P_s + 40 \log R_s) \tag{2.28}$$

Since $X(db) = 10 \log (\frac{X}{X_{ref}})$, we may substitute in 2.28 and we have

$$RFAC(db) = \log \sigma_s - [P_s(db) + 40 \log R_s]$$

Then utilizing 2.20 we get

$$RFAC(db) = 10 \log \sigma_s - \overline{b} \tag{2.29}$$

Taking the antilog of 2.29 we obtain in watts,

$$RFAC = 10^{0.1[RFAC\ (db)]} \tag{2.30}$$

7 · CROSS-SECTION EQUATION (CONTINUED)

Taking the antilog of power received from target, we have in watts,

$$P_t = 10^{0.1\,[P_t(db)]} \tag{2.31}$$

All parameters are now in proper units and are used conjunctionally to form new cross-section equation as follows:

$$\sigma_t(m^2) = P_t R_t^4 RFAC \tag{2.32}$$ (The Cross-section equation)

To obtain σ_t in db per square meter (DBSM), we use the definition

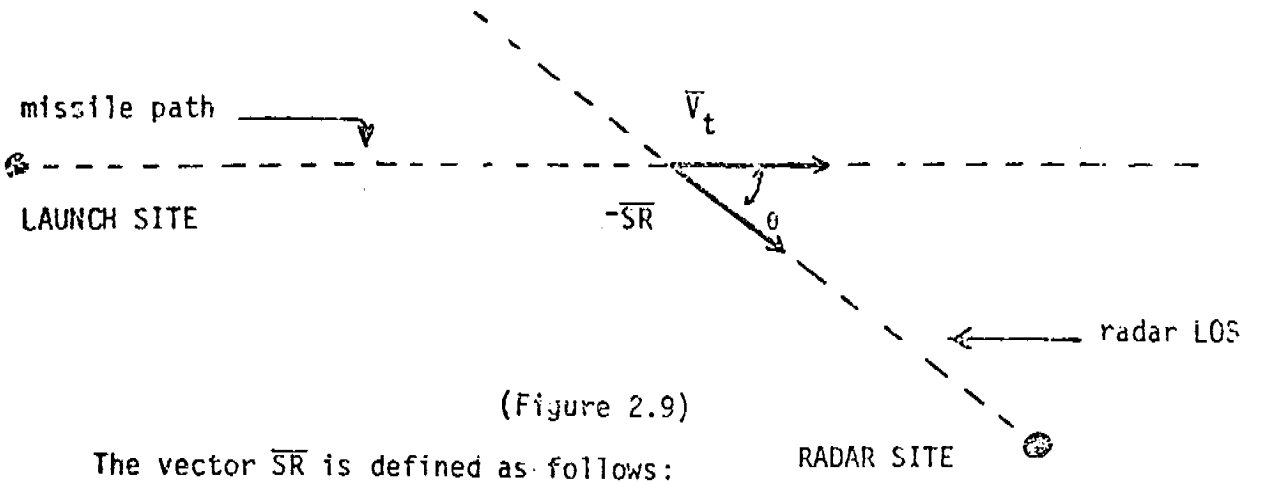
$$X(db) = 10 \log \left(\frac{X}{X_{ref}}\right) \text{ where } X_{ref} = 1\ m^2$$

and we get

$$\sigma_t(dbsm) = 10 \log \sigma_t\ (m^2) \tag{2.33}$$

8. ASPECT OR VIEW ANGLE (θ)

The aspect or view angle denoted by the symbol theta is the angle between the total velocity vector and the slant range vector. This can also be defined as the angle between unit vector along longitudinal axis of target and a unit vector along the radar line of sight.



(Figure 2.9)

The vector $\overline{SR}$ is defined as follows:

$$\overline{SR} = \bar{D}_x + \bar{D}_y + \bar{D}_z \tag{2.34}$$

From the distance formula we obtain the magnitude of the slant range vector as follows:

$$SR = |\overline{SR}| \tag{2.35}$$

$$SR = \sqrt{D_x^2 + D_y^2 + D_z^2} \tag{2.36}$$

where

$$D_x = (X_{tgt} - X_{rad}) \tag{2.37}$$

$$D_y = (Y_{tgt} - Y_{rad})$$

$$D_z = (Z_{tgt} - Z_{rad})$$

8. ASPECT OR VIEW ANGLE (α) (CONTINUED)

The vector $\overline{V}_t$ is defined as follows:

$$\overline{V}_t = \overline{V}_x + \overline{V}_y + \overline{V}_z \tag{2.38}$$

The magnitude of the velocity is as follows:

$$V_t = \left| \overline{V}_t \right| \tag{2.39}$$

$$V_t = \sqrt{V_x^2 + V_y^2 + V_z^2} \tag{2.40}$$

$$\overline{A} \cdot \overline{B} = A\, B \cos \alpha \tag{2.41}$$

$$\cos \alpha = \frac{\overline{A} \cdot \overline{B}}{AB} \quad \text{where} \quad o \leq \alpha \leq \pi$$

$$\text{if } \overline{A} = A_1 i + A_2 j + A_3 k \tag{2.42}$$

$$\text{and } \overline{B} = B_1 i + B_2 j + B_3 k$$

$$\text{then } \overline{A} \cdot \overline{B} = A_1 B_1 + A_2 B_2 + A_3 B_3 \tag{2.43}$$

By utilizing the dot product definition 2.43 we get

$$\alpha = \arccos \frac{-(V_x D_x + V_y D_y + V_z D_z)}{V_t SR} \tag{2.44}$$

8. ASPECT OR VIEW ANGLE (α) (CONTINUED)

Now α is defined as follows:

If α is negative radians, set $\theta = \alpha + \pi$

If α is positive radians, $\theta = \alpha$

where $0 \leq \theta \leq \pi$

9. ROOT MEAN SQUARE (RMS) EQUATION

The root-mean-square (RMS) of the deviations from the mean is derived as follows:

$$\text{Let } \bar{y} = \frac{1}{n} \Sigma_{i=1}^{n} y_i \quad \text{(Arithmetic mean)}$$

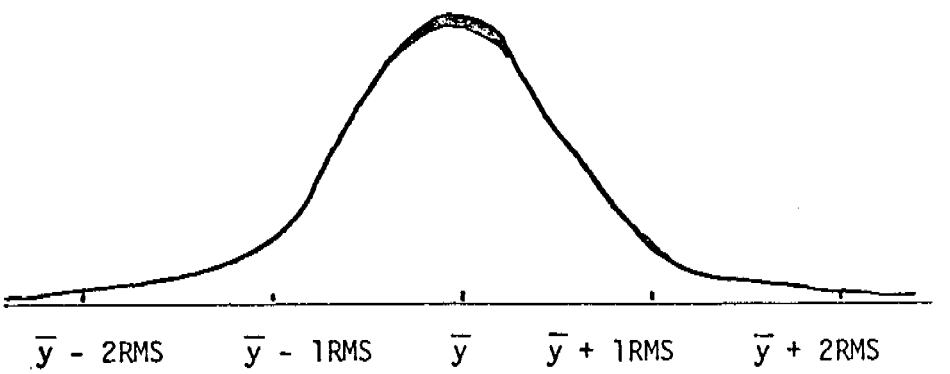
For large values of n (certainly n>30), the following equation is used for RMS.

$$\text{RMS} = \sqrt{\frac{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}{n}} \tag{2.45}$$

For a small population (n≤30), the following equation is used for RMS.

$$\text{RMS} = \sqrt{\frac{\Sigma_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}} \tag{2.46}$$

For normal distribution, we will utilize the following,



(Figure 2.10)

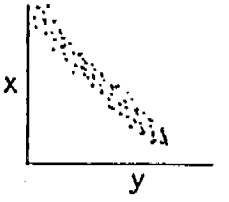
where one standard deviation on either side of mean covers 68.27% of the area underneath the curve, ±2 deviations is 95.45% and ±3 deviations is 99.73%.

Root-mean-square deviation is utilized in subroutines RCSLSQ, RCSCAL, and RCSXSC.

10. CURVE FITTING AND THE METHOD OF LEAST SQUARES

In this derivation, the notation and wording has been changed to conform with other sections of this appendix.

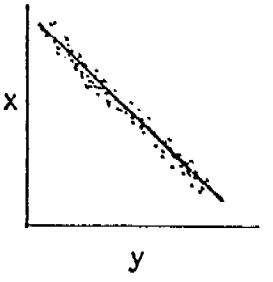
Let us start with a scatter diagram in figure below



(Figure 2.11)

where x is a dependent variable and y is independent.

From the scatter diagram we can pass a smooth curve which "fits" the given set of data as in figure 2.12.



(Figure 2.12)

We will call R_i the difference between x_i and the smooth curve. This difference R_i is called the deviation. It is also sometimes called the error or residual and may be positive, negative, or zero.

DEFINITION: Of all smooth curves approximating a given set of data points, the curve having the property that

$$R_1^2 + R_2^2 + \cdots + R_n^2 \text{ is a minimum is called a best fitting curve.}$$

The above may be rewritten as

$$\Sigma_{i=1}^{n} R_i^2 = S \text{ where S is at a minimum} \tag{2.47}$$

A curve having this property is said to fit the data in the least square sense and can be linear, quadratic (parabolic), cubic, or a fourth degree curve, or larger according to method used. We will derive the linear curve below.

10. CURVE FITTING AND THE METHOD OF LEAST SQUARES (CONTINUED)

The least square line is derived as follows:

Let $y = Ax + B$

$y = f(x)$

It follows that

$R_i = B + Ax_i - y_i$

$R_i = f(x_i) - y_i$

Let $S = \Sigma_{i=1}^{n} R_i^2$

(2.48) $S = \Sigma_{i=1}^{n} (B + Ax_i - y_i)^2$

Taking the derivative and setting the results equal to zero we obtain the following:

(2.49) $$\frac{\delta S}{\delta B} = 2 \Sigma_{i=1}^{n} (B + Ax_i - y_i) = 0$$

(2.50) $$\frac{\delta S}{\delta A} = 2 \Sigma_{i=1}^{n} x_i(B + Ax_i - y_i) = 0$$

Let us substitute

$\Sigma_{i=1}^{n} x_i = \Sigma x$, and $\Sigma_{i=1}^{n} y_i = \Sigma y$ from here on in.

Dividing both sides of equations (2.49) and (2.50) by 2 and clarifying we get the following set of normal equations

(2.51) $Bn + A\Sigma x - \Sigma y = 0$

(2.52) $B\Sigma x + A\Sigma x^2 - \Sigma y_i x_i = 0$

Solving simultaneously and letting $A = \frac{D}{C}$ we get

$$A = \frac{D}{C} = \frac{\Sigma xy - \frac{1}{n}\Sigma x\ \Sigma y}{\Sigma x^2 - \frac{1}{n}\Sigma x\ \Sigma x} \tag{2.53}$$

the coefficient A in $y = Ax + B$, and

$$B = \frac{\frac{1}{n}(\Sigma y\ \Sigma\ x^2 - \Sigma x\ \Sigma xy)}{\Sigma x^2 - \frac{1}{n}\ \Sigma x\ \Sigma x} \tag{2.54}$$

Now by substitution we may change equation 2.54 to obtain

$$B = \frac{\frac{1}{n}\Sigma y\ \Sigma x^2 - \frac{1}{n}\Sigma x\ \Sigma xy}{C}$$

The above equation is further manipulated to

$$B = \frac{\Sigma y\ \Sigma x^2}{nC} - \left(\frac{\Sigma x\ \Sigma x\ \Sigma y}{n^2C} - \frac{\Sigma x\ \Sigma x\ \Sigma y}{n^2C}\right) - \frac{\Sigma x\ \Sigma xy}{nC}$$

which can be factored out to

$$B = \frac{\Sigma y}{nC}\left(\Sigma x^2 - \frac{\Sigma x\ \Sigma x}{n}\right) - \frac{\Sigma x}{nC}\left(\Sigma xy - \frac{\Sigma x\ \Sigma y}{n}\right)$$

Now by substitution we have

$$B = \frac{C\Sigma y}{nC} - \frac{D\Sigma x}{nC}$$

which by further clarifying and remembering that $A = \frac{D}{C}$ we may alter to

$$B = \frac{\Sigma y}{n} - A\frac{\Sigma x}{n} \tag{2.55}$$

the coefficient B in $y = Ax + B$. Hence equation 2.53 and 2.55 are utilized in subroutines to compute coefficients A and B. Curve fitting using the method of least squares is utilized in RCSLSQ for "unskewing" the pulse on film due to photographic process.

11. STANDARD ERROR OF ESTIMATE

If we let y_{est} represent the value of y for given values of x as estimated from $y = Ax + B$, a measure of the scatter about the regression line of y on x is given by

$$s = \sqrt{\frac{\Sigma (y - y_{est})^2}{n}} \tag{2.56}$$

which is called the standard error of estimate of y on x.

Squaring equation 2.56 we have

$$s^2 = \frac{\Sigma (y - y_{est})^2}{n}$$

Now we substitute $y_{est} = Ax + B$ and obtain

$$s^2 = \frac{\Sigma (y - Ax - B)^2}{n}$$

Expanding above we get

$$s^2 = \frac{\Sigma[y(y - Ax - B) - Ax(y - Ax - B) - B(y - Ax - B)]}{n}$$

Now we clarify the above expression by first applying the definition on summations

$$\Sigma(ax + by - cz) = a\,\Sigma x + b\,\Sigma y - c\Sigma z$$

11. STANDARD ERROR OF ESTIMATE (CONTINUED)

and then utilize the normal equations 2.51 and 2.52 to obtain

$$s^2 = \frac{\Sigma y^2 - A\Sigma xy - B\Sigma y}{n}$$

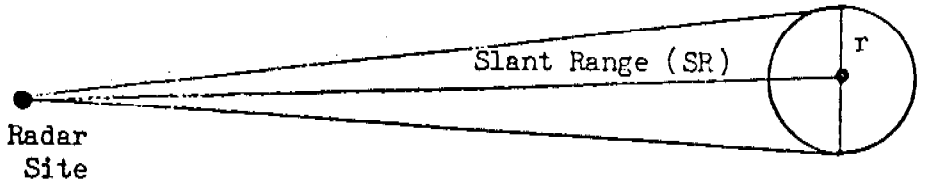
Taking the square root we finally have

(2.57) $$s = \sqrt{\frac{\Sigma y^2 - A\Sigma xy - B\Sigma y}{n}}$$ (The standard error of estimate of y on x)

The coefficients A and B used in equation 2.57 are obtained from equation 2.53 and 2.55. The standard error of estimate is utilized in RCSLSQ and RCSCL1 for computing threshold value.

12. DENSITY OF A CHAFF CLOUD

The Density of a Chaff Cloud is defined here as the radar cross-section per volume. The radar cross-section $\sigma_t(m^2)$ is as shown in equation (2.32).



(Figure 2.13)

The beamwidth Θ_b is approximated utilizing equation (2.5) for circular aperture where the frequency of the radar and the diameter of the reflector are taken into consideration.

The constant for depth penetration is the transmitted pulse duration T in microseconds converted to kilometers as described in the glossary.

The area of a circle is πr^2. The radius r of the circle was found by using the following equation where SR is the Slant Range in kilometers.

(2.58) $$r = SR \sin\left(\frac{\Theta_b}{2}\right)$$

The volume V is defined as the area of the circle multiplied by the pulse duration, tau, and the units are kilometers cubed.

(2.59) $$V = \pi r^2 T$$

The density D is then

(2.60) $$D = \frac{\sigma_t(m^2)}{\pi r^2 T}$$

where the units of measurement are RCS(meters squared)/Volume(kilometers cubed).

This computation is incorporated into subroutine RCSXSC only.

II. MATHEMATICAL ANALYSIS AND LOGIC

B. Logical Procedures

All deck setups have been standardized between each subroutine for conformity within itself and with other data reduction subsystems.

The module may be utilized to process one subroutine by itself or to process a multi-call "stacked job" in a computer pass or run.

Flexibility is built into the subsystem for future expansion or growth. This can be accomplished by following the technique presently used in this system for addition of new subroutines to meet additional WSMR data reduction requirements.

The author wishes to acknowledge all contributions to the "state of the art" of radar cross-section data software programs and technical reports by various people throughout the years, especially Darold W. Comstock, Richard H. Dale, Eugene H. Dirk, Jr., and Leonard D. Erickson, in NR-A, White Sands Missile Range, and Graham Hall, Harlan F. Lerum, J.V. Migliorato and Neil E. Feichtner at the Air Force Ballistic RE-entry System/Ballistic Missile Research System (ABRES/BMRS), Data Center at Holloman Air Force Base. The list is not necessarily complete, and the author apologizes for any omissions.

Due to the dynamic aspects and continual evolutionary process involved in data reduction, the "state of the art" of radar cross-section data is continually being improved upon because of advances in computers and computer technology, scientific equipment, and the discovery of better data reduction and programming techniques; hence suggestions, ideas, contributions, or valuable criticism are welcome, and will be incorporated in future versions.

A simple definition of radar cross-section data is the size of an object as it appears to a radar regardless of its actual size. There is no simple relationship between the actual size and the radar cross-section. One of the methods used to find this relationship is by the use of scale models in an indoor test range called radar anechoic chambers. The models are place on turn tables so that various aspect (view) angles may be obtained. Another method is by use of sphere drops or sphere raises by balloon as a method of comparison. The sphere technique is the method used in this report. Radar cross-section can then be defined as the cross-section area of a perfectly conducting sphere at the same range as the target which would return the same power as the target. Normally, a tethered sphere is lofted on a balloon or a sphere is dropped from an aircraft after a mission has been completed. The power return bounced back and the slant range is measured and recorded for sphere and this becomes the BBAR value in the reduction of mission data. Radar cross-section is then computed from BBAR, power return from mission, slant range from radar trajectory data, the radar theoretical back-scattering cross-section for a perfectly conducting sphere of known size and frequency, and the radar equation. The RCS165 module utilizes the fact that radar cross-section depends on radar frequency, the angle at which beam strikes target, the polarization of the signal, plus many other radar constants and variables unique to each radar such as RAM, RAMPART, FPS16, TTR, HAPDAR, RTMS, DR, and MAR. The flexibility to handle above mentioned radars is built into the subroutines comprising this monitor.

B. Logical Procedure (CONTINUED)

There are three distinct types of data that are handled by RCS165 For better clarity, each one will be discussed separately.

The first is A-Scope. A-Scope is a type of data presentation where the echo strength or target amplitude is indicated by vertical displacement (Y-axis) of the luminous spot and range by horizontal (X-axis) displacement on a cathode-ray tube. There is no angle information displayed on this type of data presentation. The radar signal or burst of energy is bounced off the target and optical cameras are setup and calibration procedures are followed for A-Scope recording film. The data is recorded on 35 mm film in three ways; Type I film is video pulse streak camera recording (strip film), Type II is framing camera recording, and Type III film is intensity modulation streak camera recording. The timing on film is IRIG timing. This technical report covers the data reduction of strip and frame film data. The film is assessed and then read on our PFR-3 A-Scope Film Reader System. The A-Scope Film Reader System (AFRS) was designed by Lincoln Laboratory, Massachusetts Institute of Technology as a part of an Advanced Research Project Agency (ARPA) project and built by Information International Incorporated, Cambridge, Massachusetts to read film on a semi-automatic basis. The AFR System utilizes the Programmed Data Processor (PDP-1) computer to operate and control input-output devices. A-Scope type data is primarily utilized when target discrimination, individual RCS for entire pulse, peak RCS, or other signature analysis is required.

The second type is the Automatic Gain Control (AGC) type. The digitized AGC data recorded on radar field tape is pre-calibrated on the boresight tower in 5 db steps from 0 db to noise level. This data is then stored in the Milgo (ARCADE) Computer memory and during mission the target video is referenced to stored AGC precal. ARCADE is an abbreviation for Automatic Radar Control and Data Equipment. The reference AGC level is then recorded on radar field tape. The digitized raw data is produced as a positive displacement above the noise level. Each radar launches their own spheres and conducts the sphere calibrations immediately after mission completion. This type of data is normally converted to peak RCS.

The third type is the Automatic Gain Control - Voltage Control Oscillator (AGC-VCO). The detected video is recorded on analog tape by a telemetry van station utilizing FM methods. The data is later digitized to machine counts and from there to engineering units of db through the use of pre or post step calibrations. This type of data is normally converted to peak RCS. All three types utilize sphere calibrations. The third type utilizing telemetry is used very little for obtaining peak RCS as type two is faster. Type three is utilized more for radar system analysis work.

B. Logical Procedures (CONTINUED)

Radar signature data and radar cross-section data are two terms that are synonymous and are used interchangeably in this report.

Radar Cross-section is often measured in DBSM (db per meters squared) which is plotted as shown in figure 1.1. Analysis of this pattern would indicate shape of vehicle, the number of protrusions, whether target is spinning and if so at what rate, plus other information according to needs of customer. RCS is often plotted against time, altitude, aspect angle, and slant range.

Figure 1.2 shows the characteristics of instrumentation Radar AN/FPS16 and Figure 1.3 is a map of the "on range" radar sensor instrumentation sites for FPS16, chain, surveillance, and project radars.

Target discrimination and penetration aid techniques are but two of the many areas dependent on the output of a program like RCS165. One of the fields where analysis of RCS data is required is in re-entry body studies. As a vehicle re-enters the atmosphere, the shock waves formed cause a layer of electrons on the body of the vehicle. This plasma sheath causes a drastic change in cross-section as compared to free space cross-section. The ionized field or wake which follows the vehicle also causes a reduction or enlargement of cross-section. The study of this phenomena in signature data analysis can be utilized in the development of antiradar signature decoys or in the determination of enemy warheads in a mass of re-entry vehicles by cross-section data predictions. Its measurements (the recognition of shapes such as conical, cylindrical, spherical, or any geometrical combination) and numerous other techniques are now being refined by Signature Data Analyst contractors throughout the nation. The Signature Analyst can then be expected to come up with a reasonable approximation of the unknown body by knowing the characteristic returns of certain bodies and the recognition and cataloging of different shapes.[30]

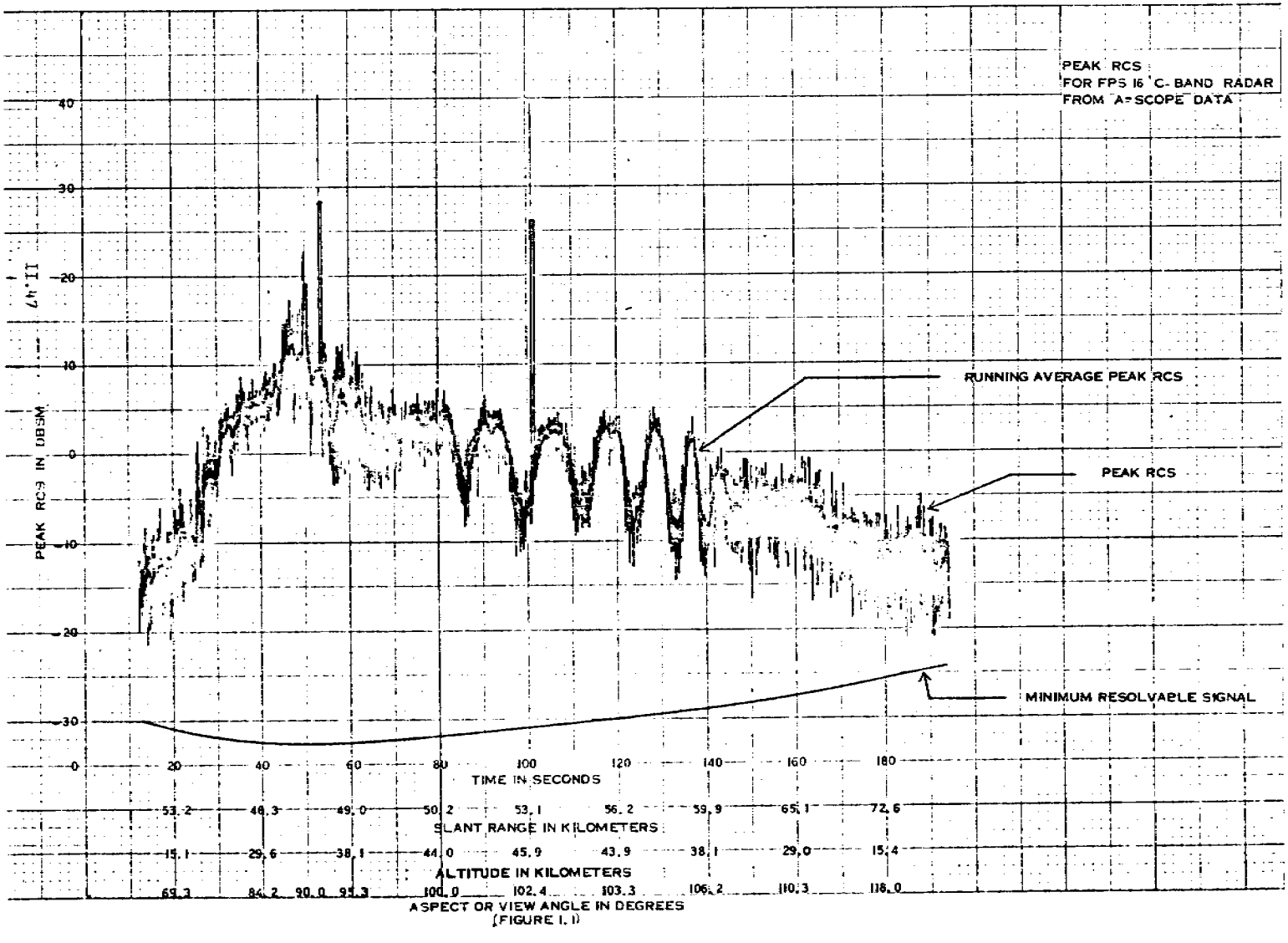
Since warhead must be identified in time for defensive action, a computer is necessary. A computer, however, tends to take things too literally: if a return differs slightly from the description which is given to it, the computer will not recognize the object. But with new computers which are capable of learning and with improved optical technique for pattern recognition, perhaps this problem is closer to a solution.

Another big area where RCS Data Analysis plays an important role is in Electronic Countermeasures (ECM) studies. ECM embraces such techniques as jamming transmitters, set-on receivers, false target repeater, simulation of different aircraft radar signatures by the use of decoys, chaff clouds cut to an appropriate size to resonate at the frequency of the radar to be jammed, and other techniques by which the enemy is denied the use of the electromagnetic spectrum for radar, communications, navigation and guidance.[31]

B. Logical Procedures (CONTINUED)

The future of RCS is very bright indeed. The resolution capability of radar is constantly being upgraded. With the use of synthetic-spectrum radar, chirped radar, phased-array systems, etc., the resolution should improve to the point where the image of a distant target will correspond more to its physical than its electrical features.

The behavior of wake and plasma phenomena is becoming better understood as more advances are made. Coupled with new approaches to computer printout, this should provide displays of greater validity. The entire field of RCS in re-entry physics, radar profiling, passive radar detection, and large-scale air traffic control has barely been opened.[29]


PEAK RCS
FOR FPS 16 C-BAND RADAR
FROM A-SCOPE DATA
RUNNING AVERAGE PEAK RCS
PEAK RCS
MINIMUM RESOLVABLE SIGNAL
PEAK RCS IN DBSM
40
30
20
10
0
-10
-20
-30
0
20
40
60
80
100
120
140
160
180
TIME IN SECONDS
53.2
46.3
49.0
50.2
53.1
56.2
59.9
65.1
72.6
SLANT RANGE IN KILOMETERS
15.1
29.6
38.1
44.0
45.9
43.9
38.1
29.0
15.4
ALTITUDE IN KILOMETERS
69.3
84.2
90.0
95.3
100.0
102.4
103.3
106.2
110.3
118.0
ASPECT OR VIEW ANGLE IN DEGREES

(FIGURE I. 1)

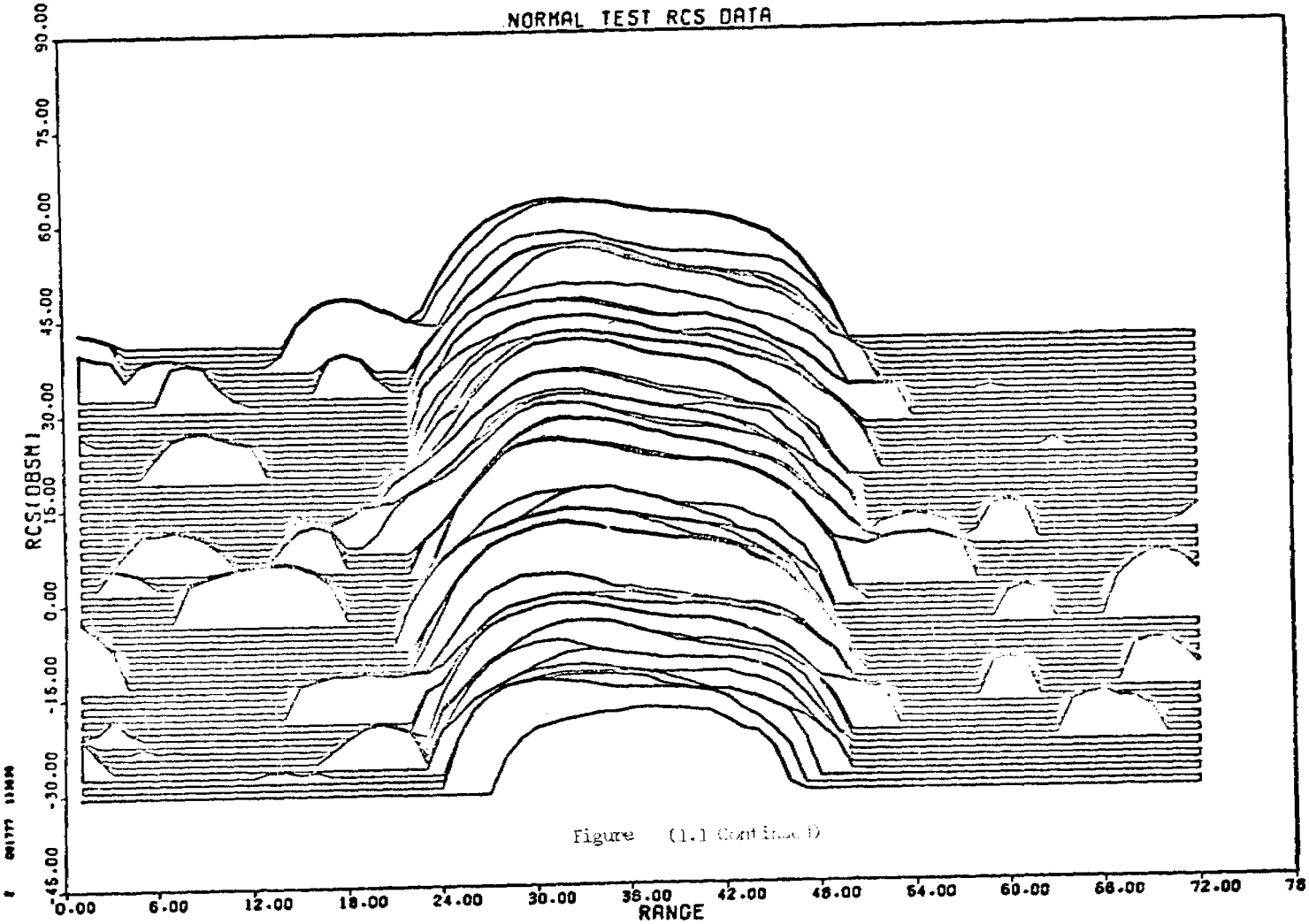


Figure (1.1 Continued)

TABLE OF CHARACTERISTICS OF INSTRUMENTATION RADAR AN/FPS-16 [26,27,28]

RECEIVER SYSTEM

System Noise Figure	- < 11 db
I.F. Center Frequency	- 30 MC
Bandwidths	- Wide 8.0 MC, Narrow 1.6 MC
Local Oscillators	- Two - Skin and Beacon
AFC	- Skin or Beacon
Non-Tracking I.F.	- Non-gated manually gain controlled receiver, video added to tracking video for display only

RANGE TRACKING SYSTEM

Range	- 1,000,000 yards
Gate Widths	- Pulse widths 0.25 μs, 0.5 μs, 1.0 μs Acquisition 1.0 μs, 1.25 μs, 1.75 μs Tracking 0.5 μs, 0.75 μs, 1.25 μs
Aided Tracking	- Yes
Maximum Slew Rate	- 40,000 yards per second
Maximum Tracking Rate,	- 12,000 yards per second
Automatic Lock-On	- Search ±1000 yards, and auto lock
Servo Bandwidth	- Continuously adjustable manually or automatically between 0.5 cps (K_v = 2000) and 6.0 cps (K_v = 3000)
Range Accuracy	- ±3.5 to ±15 yards

ANGLE TRACKING SYSTEM

Aided Tracking	- Yes
Maximum Slew Rate	- Azimuth 48° per second, Elevation 37° per second
Maximum Tracking Rate	- Azimuth 42° per second, Elevation 22.5° per second
Servo Bandwidth	- Continuously adjustable manually or automatically between 0.5 cps (K_v = 150) and 4.0 cps (K_v = 300)
Scan	- Circle scan of adjustable radius and rate. Also sector scan in azimuth and elevation
Angle Accuracy	- ±0.05 mil to ±0.3 mils

DATA OUTPUTS

Potentiometer	
Sycro	- Army or Navy speeds
Digital	- Serial straight binary Range - 20 bits (0.5 yard quanta) Angle - 17 bits (< .05 mils quanta)

(Figure 1.2)

TABLE OF CHARACTERISTICS OF INSTRUMENTATION RADAR AN/FPS-16 [26,27,28]

ANTENNA PEDESTAL

Axes	- Azimuth and Elevation
Weight	- 12,000 pounds
Plunging Capability	- Yes
Azimuth Coverage	- Continuous 360°
Elevation Coverage	- 10° to 190°

ANTENNA SYSTEM

Size	- 12 foot diameter parabola
Feed	- 4 - Horn Monopulse
Gain	-44.5 db
Beamwidth @3 db pts	- 1.2 degrees
Polarization	- Vertical
Rotary Joints and Waveguide	- 3 Megawatt capability

TRANSMITTER

TYPE - MAGNETRON

	Peak Power	Frequency	Max. Duty Cycle
R-113, R-123, R-127	3MW	5450-5825 MC	0.001
Eight Other FPS16's	1MW Fixed Tuned	5480 ± 35MC	0.0010
	250 KW Tunable	5450 - 5825MC	0.0016

PRF - Internal: 3MW-12 steps between 142 to 1707 PPS
Internal: 1MW-18 steps between 71 to 1707 PPS
External: Any PRF between 160 to 1707 PPS

Pulse Width - 0.25 μs, 0.5μs and 1.0μs

Coding - Up to 5 pulses within Duty Cycle Limitations

(Figure 1.2)
CONTINUED

TABLE OF CHARACTERISTICS OF INSTRUMENTATION RADAR AN/FPS-16 [26,27,28]

DISPLAYS

Range	- Dual A-Scope Dials Digital-Octal numeral
Angle	- Dials Digital-Octal numeral

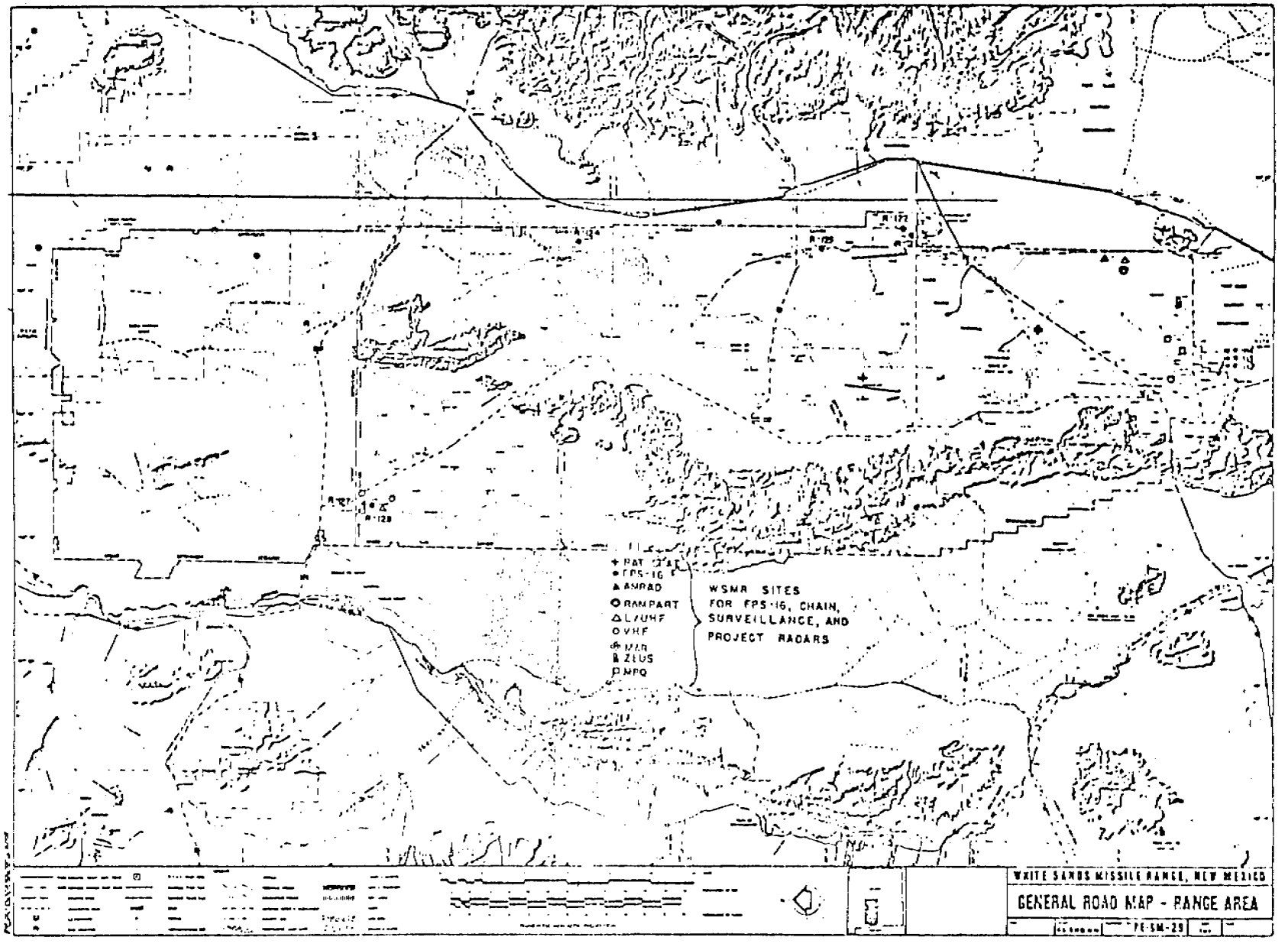
MONITORING

Noise Figure Measurement
Transmitter Power
Range and Angle Servo Performance
Significant Waveforms
Significant Voltages and Currents
Strip Chart Recorder and Signal Patch Panel

USE

Provides real-time information to the Missile Flight Safety Officer, trajectory, A-Scope, AGC data to the project, acquisition data to other range data gathering systems, and tracking data for drones and target aircraft. transmit digital data in real time.

(Figure 1.2)
CONTINUED


WSMR SITES
FOR FPS-16, CHAIN,
SURVEILLANCE, AND
PROJECT RADARS
WHITE SANDS MISSILE RANGE, NEW MEXICO
GENERAL ROAD MAP - RANGE AREA

(Figure 1.3)

XIV. BIBLIOGRAPHY

A. REFERENCES

1. Richard H. Dale, Analysis Branch Special Report on Radar Cross Sections, DRD-A, 14 June 63 White Sands Missile Range, New Mexico.

2. Irvin McClintock, The Cross-Section Measurement Capabilities of the AN/FPS-16, Instrumentation Development Directorate, January 1965, White Sands Missile Range, New Mexico.

3. E. H. Dirk Jr., Program 3-Scan Generalized Ascope Scan Program (GASP), October 1966, Analysis and Computation Directorate, National Range Operations, White Sands Missile Range, New Mexico.

4. Merril I. Skolnik, Introduction to Radar Systems, McGraw-Hill Book Company, 1962.

5. H. Weil, M. L. Barach, and T. A. Kaplan, Scattering of Electromagnetic Waves by Spheres, University of Michigan Engineering Research Institute Report 2255-20-T, July 1956.

6. R. W. King, T. T. Wu, The Scattering and Diffraction of Waves, Harvard University Press, Cambridge, Massachusetts, 1959.

7. Reintjes and Coate, Principles of Radar, McGraw-Hill Book Company, New York, 3rd edition.

8. M. Spiegel, Statistics, Schaum Publishing Co., New York, 1961.

9. Kaiser S. Kunz, Numerical Analysis, McGraw-Hill Book Company, New York 1961.

10. A. E. Knowlton, Standard Handbook for Electrical Engineers, McGraw-Hill Book Company, 9th edition.

11. H. H. Koelle, Handbook of Astronautical Engineering, McGraw-Hill Book Company, 1st edition.

XIV. BIBLIOGRAPHY

A. REFERENCES

12. Murray H. Protter, Charles B. Morrey, Jr., Modern Mathematical Analysis, Addison-Wesley Publishing Co., Inc., 1966.

13. Eugene H. Dirk Jr., ASCOPE DCS PROGRAM, 1967, Analysis and Computation Directorate, National Range Operations, White Sands Missile Range, New Mexico.

14. Ernest J. Sanchez, ADEADR-Signature Data Monitor - A-Scope, AGC, and AGC-VCO Cross Section System, AD 884 532 L, Technical Report Number 21, version 00, 3 May 71, Analysis and Computation Division, National Range Operations Directorate, White Sands Missile Range, New Mexico. W-037396 at Tech Lib WSMR.

15. Ernest J. Sanchez, The Exact Solution for the Normalized Back-scatter Cross-Section of a Perfectly Conducting Sphere Utilizing Mie Theory, Technical Report Number 31, version 00, 8 February 72, Analysis and Computation Division, National Range Operations Directorate, White Sands Missile Range, New Mexico.

16. Ernest J. Sanchez, ELDASY11 (ADERCS), AGC, and AGC-VCO - RADAR Cross Section - DATA REDUCTION SYSTEM, Technical Report Number 21, version 01, 8 September 72, Analysis and Computation Division, National Range Operations Directorate, White Sands Missile Range, New Mexico.

17. J. Rheinstein, Scattering of Electromagnetic Waves by a Eaton Lens, M.I.T. Lincoln Laboratory, TR-273 (June 1962).

18. J. Rheinstein, Tables of the Amplitude and Phase of the Backscatter from a Conducting Sphere, M.I.T. Lincoln Laboratory, TR 22-G-16 (June 1963).

XIV. BIBLIOGRAPHY

A. REFERENCES

19. J. Rheinstein, Backscatter from Spheres: A Short Pulse View, M.I.T. Lincoln Laboratory, TR 414 (April 1966).

20. Arthur Beiser, The Mainstream of Physics, Addison-Wesley Publishing Company, Inc., 1961.

21. MIPS Development Group, Modular Integrated Processing System (MIPS), Report #4, 2 Dec 71.

22. J. A. Stratton, Electromagnetic Theory, McGraw-Hill, New York, (1941), pp 563-567.

23. H. C. Van de Hulst, Light Scattering by Small Particles, John Wiley and Sons, New York, (1957) p. 286.

24. George T. Ruck, Donald E. Barrich, William D. Stuart, Clarance K. Krichbaum, Radar Cross Section Handbook Volume I and II, Plenum Press New York, London (1970).

25. Cutler C. C., Parabolic-antenna Design for Microwave, Proc IRE volume 35 pp 1284-1294, Nov. 1947.

26. Barton, David Knox, Radar System Analysis, Prentice-Hall, Inc., Electrical Engineering Series, New Jersey, 1964.

27. Barton, David Knox, Accuracy of a Monopulse Radar, Proc. Third National Military Electronics Convention IRE-PGMIL, June 30, 1959 pp 179-186.

28. White Sands Missile Range Instrumentation Capability Inventory, 10 April 64.

XIV. BIBLIOGRAPHY (CONTINUED)

A. REFERENCES

29. A. E. Judd, "Are Modern Military Radars Infallible", Popular Electronics, September 1971.

30. Edward A. Lacy, "Radar Signature Analysis", Electronic World, February 1967.

31. Aviation Week and Space Technology, "Electronic Countermeasures", February 1972.

32. Comstock, D.W.; Wright, H.H; Tipton, V.B.; Handbook of Data Reduction Methods, Technical Report, Data Reduction Division, White Sands Missile Range, New Mexico, 13 August 1964, Second Printing-24 Sep 64, Third Printing-1 Mar 69.

33. Ernest J. Sanchez; ELDASY II(ADERCS) A-SCOPE, AGC, AND AGC-VCO RADAR CROSS-SECTION DATA REDUCTION SYSTEM Technical Report number 21, Version 01, 8 Sep 72, AD904445L, W-037400.

34. Modular Integrated Processing Systems (MIPS) Description Document dated 29 Sep 72, prepared by MIPS Development Group, TSTR, New Mexico.

A NONLINEAR SINGULARLY PERTURBED VOLTERRA INTEGRODIFFERENTIAL EQUATION OCCURRING IN POLYMER RHEOLOGY

A.S. Lodge[1), 2)], J.B. McLeod[1)], and J.A. Nohel[1), 3), 4)]
Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

ABSTRACT

We study the initial value problem for the nonlinear Volterra integrodifferential equation

$$(+)\qquad \begin{cases} -\mu y'(t) = \displaystyle\int_{-\infty}^{t} a(t-s)\, F\left(y(t), y(s)\right) ds & (t > 0) \\ y(t) = g(t) & (-\infty < t \le 0), \end{cases}$$

where $\mu > 0$ is a small parameter, a is a given real kernel, and F, g are given real functions; $(+)$ models the elongation ratio of a homogeneous filament of a certain polyethylene which is stretched on the time interval $(-\infty, 0]$, then released and allowed to undergo elastic recovery for $t > 0$. Under assumptions which include physically interesting cases of the given functions a, F, g, we discuss qualitative properties of the solution of $(+)$ and of the corresponding reduced problem when $\mu = 0$, and the relation between them as $\mu \to 0^+$, both for t near zero (where a boundary layer occurs) and for large t. In particular, we show that in general the filament does not recover its original length, and that the Newtonian term $-\mu y'$ in $(+)$ has little effect on the ultimate recovery but significant effect during the early part of the recovery.

Sponsored by:

[1)] The United States Army under Contract Number DAAG29-75-C-0024;
[2)] National Science Foundation under Grant Number ENG 75-18397;
[3)] The United States Army under Grant Number DAHC04-74-G-0012;
[4)] National Science Foundation under Grant Number MCS 75-21868.

1. Introduction.

We study the nonlinear Volterra integrodifferential equation

$$(1.1) \qquad -\mu\, y'(t) = \int_{-\infty}^{t} a(t-s)\, F\big(y(t), y(s)\big)\, ds \qquad (t > 0; \quad ' = d/dt)$$

subject to the initial condition

$$(1.2) \qquad y(t) = g(t) \qquad (-\infty < t \leq 0)\,.$$

This initial value problem arises as a mathematical model for a process in polymer rheology which is described in Appendix A. In the specific problem discussed there μ is a positive parameter related to the viscosity, and the given real functions a, F, g take the forms

$$(1.3) \qquad a(t) = \sum_{k=1}^{m} a_k \exp(-t/\tau_k)$$

where a_k and τ_k are positive constants,

$$(1.4) \qquad F(y, z) = y^3/z^2 - z\,,$$

and

$$(1.5) \qquad g(t) = \begin{cases} 1 & \text{if } -\infty < t \leq -t_0 \quad (t_0 > 0) \\ e^{\kappa(t+t_0)} & \text{if } -t_0 < t \leq 0\,, \end{cases}$$

where κ is a positive constant (see equations (A5), (A 21), (A 24), (A25) in Appendix A). The unknown function y measures the ratio of the extended length to the original length of a homogeneous filament which

is stretched in accordance with (1.2) on the time interval $(-\infty, 0]$ and then released. The equation (1.1) then describes the process of elastic recovery. Also of interest, both mathematically and physically, is the reduced equation

$$0 = \int_{-\infty}^{t} a(t-s)\, F\big(y(t), y(s)\big)\, ds \qquad (t > 0), \tag{1.6}$$

with $y(t) = g(t)$ on $(-\infty, 0)$, and in particular the relation between the solutions of (1.1), (1.2) and of (1.6) as $\mu \to 0^+$; for small $\mu > 0$, (1.1), (1.2) may be regarded as a singular perturbation of (1.6).

The purpose of this paper is to discuss the qualitative behaviour of solutions of (1.1), (1.2) on the one hand and of (1.6) on the other, and the relation between them as $\mu \to 0^+$. Our analysis is not confined to the specific forms of a, F, g in (1.3) - (1.5), but rather we abstract the essential properties of these functions. Two results of particular interest are : (i) in general, the filament does not return to its original length, as confirmed by experiments (see Appendix A and Theorems 4 and 5); (ii) similarly to behaviour in singular perturbation problems for ordinary differential equations, the solution of (1.1), (1.2) for small $\mu > 0$ decreases rapidly as t increases near $t = 0$ to the solution of (1.6), becoming close to it in a "boundary layer" time interval of order $\mu|\log \mu|$, and thereafter remains close (Corollary 3.1 and Theorems 7 and 8). The introduction of the Newtonian term $-\mu y'$ in (1.1) has little effect on the ultimate behaviour of the solution.

From the mathematical point of view the theory of integrodifferential

equations with decreasing convex kernels and with nonlinearities consisting of functions of one variable is well known, see e.g. [7], [18] where references to other literature are given. A part of the novelty of the present analysis is that F in (1.1) and (1.6) is a function of two variables, $y(t)$ and $y(s)$.

We acknowledge with pleasure the various helpful discussions with colleagues during the preparation of this paper, in particular with M.G. Crandall, R.W. Dickey, F. Howes, J.J. Levin, S.O. Londen, S.V. Parter, A. Pazy, D.F. Shea, O.J. Staffans, L. Tartar, and W. Wasow.

2. Statement of Results.

Let $\mathbb{R}$ denote the real numbers, $\mathbb{R}^+$ the positive real numbers, and C^k the set of k times continuously differentiable functions.

We make the following assumptions on the functions a, F, g throughout:

$$(H_a)\quad \begin{cases} a \in C^1[0,\infty);\ a(t) > 0,\ a'(t) < 0 \quad (0 \le t < \infty); \\ a \in L^1(0,\infty);\ \log a(t) \text{ is convex} \quad (0 \le t < \infty), \text{ i.e.} \\ a'(t)/a(t) \text{ is nondecreasing;} \end{cases}$$

$$(H_F)\quad \begin{cases} F : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R};\ F(x,x) = 0 \text{ for every } x > 0; \\ F \in C^1(\mathbb{R}^+ \times \mathbb{R}^+) \text{ and } F_1(y,z) > 0,\ F_2(y,z) < 0 \\ (y, z \in \mathbb{R}^+), \text{ where the subscripts denote partial differentiation;} \end{cases}$$

$$(H_g)\quad \begin{cases} g : (-\infty, 0] \to \mathbb{R}^+;\ g(-\infty) = 1,\ g(0) > 1; \\ g \in C(-\infty, 0] \text{ and } g \text{ is nondecreasing.} \end{cases}$$

It is readily verified that the specific functions defined in (1.3) - (1.5) satisfy (H_a), (H_F) and (H_g) respectively. While for the majority of results these are the only conditions required, some additional assumptions are needed in Theorems 4 and 5, Corollaries 2.1 and 3.1, and Theorems 6 a

The first result concerns the global existence and uniqueness of the solution of (1.1), (1.2) for a fixed $\mu > 0$ and gives some useful properties of the solution.

Theorem 1. Let (H_a), (H_F), (H_g) be satisfied. Then for each $\mu > 0$, the initial value problem (1.1), (1.2) has a unique solution $\phi(t,\mu)$ on $[0,\infty)$ satisfying the following properties:

(2.1) $\quad \phi'(t,\mu) < 0 \quad$ and $\quad 1 < \phi(t,\mu) \leq g(0) \quad (0 \leq t < \infty)$;

(2.2) if g_1, g_2 satisfy (H_g) and if $g_1(t) \geq g_2(t)$ $(-\infty < t \leq 0)$, then the corresponding solutions $\phi_1(t,\mu)$, $\phi_2(t,\mu)$ of (1.1), (1.2) satisfy $\phi_1(t,\mu) \geq \phi_2(t,\mu)$ $(0 \leq t < \infty)$;

(2.3) if $\mu_1 > \mu_2$, then the corresponding solutions of (1.1), (1.2) satisfy $\phi(t,\mu_1) > \phi(t,\mu_2)$ $(0 < t < \infty)$.

An immediate consequence of (2.1), (2.3) is

Corollary 1.1. Let (H_a), (H_F), (H_g) be satisfied. Then for each fixed $\mu > 0$ one has

(2.4) $\quad \alpha(\mu) = \lim_{t \to \infty} \phi(t,\mu)$ exists and $\alpha(\mu) \geq 1$;

moreover, if $\mu_1 > \mu_2$, then $\alpha(\mu_1) \geq \alpha(\mu_2) \geq 1$.

Theorem 1 is proved in Section 3.

Remark 1.1. In the special case $a(t) \equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$ which satisfies (H_a) (with $\frac{a'(t)}{a(t)} \equiv -\lambda$), conclusions (2.1) and (2.4) can be strengthened respectively to the following:

$$(2.1') \qquad \phi'(t,\mu) < 0 \quad \text{and} \quad 1 < y_0 < \phi(t,\mu) \leq g(0) \quad (0 \leq t < \infty),$$

$$(2.4') \qquad \alpha(\mu) = \lim_{t \to \infty} \phi(t,\mu) \text{ exists} \quad \text{and} \quad \alpha(\mu) \geq y_0 > 1,$$

where (see the proof of (2.8) in Theorem 3) y_0 is uniquely defined by the equation

$$\int_{-\infty}^{0} e^{\lambda s} F(y_0, g(s))\, ds = 0 .$$

The proof of Remark 1.1 is carried out in the course of proving (2.1) (Case (ii)).

Remark 1.2. If one is interested only in global existence, rather than further properties of solutions of (1.1), (1.2), then the following existence result can readily be established.

Proposition 1. Let $a \in L^1(0,\infty)$, $a(t) > 0$ $(0 \leq t < \infty)$; let g satisfy (H_g); let $F \in C(\mathbb{R}^+ \times \mathbb{R}^+)$, $F(x,x) = 0$ for every $x \in \mathbb{R}^+$ and $F(y,z) < 0$ for $y < z$ and $F(y,z) > 0$ for $y > z$ $(0 < y, z < \infty)$. Then for each fixed $\mu > 0$ the initial value problem (1.1), (1.2) has at least one solution on $0 < t < \infty$.

The proof of this result follows from the following observations: (i) under the present hypotheses (1.1), (1.2) has a local solution on $0 \le t < t_0$ (see e.g. [17, Lemma 1.1], also [16]); (ii) if $y(t)$ is any solution of (1.1), (1.2) on $0 \le t < \infty$, then a slight modification of the early part of the proof of Theorem 1 shows that $1 < y(t) \le g(0)$ $(0 \le t < \infty)$; (iii) in view of (i), (ii), every such local solution can be extended (but not necessarily uniquely) to the interval $[0, \infty)$ (see e.g. [17, Lemma 1.2], also [16]).

We next obtain some estimates of the solution $\phi(t,\mu)$ of (1.1), (1.2) which are useful for the study of the asymptotic behaviour as $t \to \infty$, and which will be used to deduce the existence of a solution of the reduced problem (1.6).

Theorem 2. Let (H_a), (H_F), (H_g) be satisfied. Then there exists a constant $K_1 > 0$ (independent of μ) and constants $\mu_0 > 0$ and $\tilde{K} = \tilde{K}(\mu_0) > 0$ such that the solution $\phi(t,\mu)$ of (1.1), (1.2) satisfies the estimate

$$(2.5) \qquad 0 < -\phi'(t,\mu) \le \frac{\tilde{K}}{\mu} \exp(-K_1 t/\mu) + \tilde{K} \int_t^\infty a(s)ds \quad (0 \le t < \infty;\ 0 < \mu \le \mu_0).$$

As a consequence of Theorem 2 and the logarithmic convexity of a, one obtains

Corollary 2.1. If, in addition,

$$(2.6) \qquad \int_0^\infty t\,a(t)\,dt < \infty$$

is satisfied, then there exist constants $\mu_0 > 0$ and $K = K(\mu_0) > 0$ such that

$$(2.7) \qquad 0 < \phi(t,\mu) - \alpha(\mu) \le K \int_t^\infty (s-t)\, a(s)\, ds \quad (0 \le t < \infty;\ 0 < \mu \le \mu_0),$$

where $\alpha(\mu) = \lim_{t \to \infty} \phi(t,\mu)$.

Theorem 2 and Corollary 2.1 are proved in Section 4.

Concerning the reduced problem (1.6) we prove the following global results.

Theorem 3. Let (H_a), (H_F), (H_g) be satisfied. Then (1.6) has a unique (continuous) solution ϕ_0 on $[0,\infty)$ satisfying the following properties:

$$(2.8) \qquad \text{if } y_0 = \phi_0(0), \text{ then } 1 < y_0 < g(0);$$
$$\text{if } a(t) \not\equiv Ae^{-\lambda t},\ A > 0,\ \lambda > 0, \text{ then}$$

$$(2.9) \qquad \phi_0 \in C^1[0,\infty),\ \phi_0'(t) < 0 \text{ and } 1 < \phi_0(t) \le y_0 \quad (0 \le t < \infty);$$
$$\text{if } a(t) \equiv Ae^{-\lambda t},\ A > 0,\ \lambda > 0, \text{ then } \phi_0(t) \equiv y_0 \quad (0 \le t < \infty);$$

$$(2.10) \quad \begin{cases} \text{if } g_1, g_2 \text{ satisfy } (H_g) \text{ and if } g_1(t) \ge g_2(t) \quad (-\infty < t \le 0), \\ \text{then the corresponding solutions } \phi_0^{(1)}, \phi_0^{(2)} \text{ of (1.6) satisfy} \\ \phi_0^{(1)}(t) \ge \phi_0^{(2)}(t) \quad (0 \le t < \infty); \end{cases}$$

$$(2.11) \quad \begin{cases} \text{if } \phi(t,\mu) \text{ is the solution of (1.1), (1.2) for a fixed } \mu > 0, \\ \text{and if } \phi_0(t) \text{ is the solution of (1.6), then } \phi_0(t) < \phi(t,\mu) \\ (0 \le t < \infty). \end{cases}$$

As a consequence of (2.4), (2.9), (2.11) we have the first statement in

Corollary 3.1. Let (H_a), (H_F), (H_g) be satisfied. Then

(2.12) $$\alpha_0 = \lim_{t \to \infty} \phi_0(t) \text{ exists and } 1 \leq \alpha_0 \leq \alpha(\mu);$$

if $a(t) \equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$, then, in fact, $\alpha_0 = y_0 > 1$. If also (2.6) holds, then

(2.13) $$\lim_{\mu \to 0^+} \alpha(\mu) = \alpha_0$$

and

(2.14) $$0 < \phi(t,\mu) - \phi_0(t) \leq \alpha(\mu) - \alpha_0 + K \int_t^\infty (s-t)\, a(s)\, ds \quad (0 \leq t < \infty;\ 0 < \mu \leq \mu_0).$$

Theorem 3 and (2.13), (2.14) are proved in Section 5. In Theorem 8 below we establish a more precise result than (2.13), (2.14) under some additional assumptions.

Remark 3.1. An estimate similar to (2.5) holds for the solution ϕ_0 of the reduced problem (1.6), (but, of course, without the term $\frac{K}{\mu} \exp(-K_1 t/\mu)$). This can be proved as a special case of the proof of Theorem 2 by obtaining an estimate of the form (4.6) in Section 4, where f is now independent of μ, and using it and an estimate of the form (4.4) (also now independent of μ) in (5.12) of Section 5. Consequently, (2.7) also holds with $\phi(t,\mu)$ replaced by $\phi_0(t)$ and $\alpha(\mu)$ by α_0, and with $0 \leq t < \infty$.

The next task is to establish the physically important fact that the limiting value $\alpha(\mu)$ of the solution $\phi(t,\mu)$ of (1.1), (1.2) as $t \to \infty$ satisfies $\alpha(\mu) > 1$ $(\mu > 0)$, rather than the weak form $\alpha(\mu) \geq 1$ in (2.4). By properties (2.11), (2.12) it suffices to prove $\alpha_0 > 1$ $(\alpha_0 = \lim_{t \to \infty} \phi_0(t))$.

Theorem 4. Let (H_a), (H_F), (H_g) be satisfied. If, in addition, (2.6) is satisfied, then

$$\alpha_0 > 1 . \tag{2.15}$$

We remark that in the special case $a(t) \equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$, there is nothing to prove since $\phi_0(t) \equiv y_0 > 1$. Theorem 4 is proved in Section 6.

Theorem 4 is best possible in the following sense.

Theorem 5. Let (H_a), (H_g) be satisfied and let

$$F(y,z) = y - z . \tag{2.16}$$

If

$$\int_0^\infty sa(s)\,ds = \infty , \tag{2.17}$$

then

$$\lim_{t \to \infty} \phi_0(t) = 1 , \tag{2.18}$$

when ϕ_0 is the solution of (1.6) with F satisfying (2.16).

Remark 5.1. $F(y,z) = F_0(y-z)$, where $F_0 > 0$ is a constant, is, being linear, the simplest form of F consistent with assumptions (H_F).

Remark 5.2. A similar proof applied to (1.1), (1.2) shows that if the hypotheses of Theorem 5 hold, then $\lim_{t\to\infty} \phi(t,\mu) = 1$, where $\phi(t,\mu)$ is the solution of (1.1), (1.2) with F satisfying (2.16).

Theorem 5 is proved in Section 7.

In connection with the last statement in Remark 3.1 it is of interest to note that such an estimate for ϕ_0, (2.7), can be proved independently of Theorem 2, and we state this fact as a separate result in Theorem 6. However, it should also be noted that the estimate (2.5) (for ϕ_0') described in Remark 3.1 cannot be obtained from Theorem 6.

Theorem 6. Let (H_a), (H_F), (H_g) be satisfied, and let (2.6) hold. Then there exists a constant $K > 0$ such that, if $a(t) \not\equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$,

$$0 < \phi_0(t) - \alpha_0 \le K \int_t^\infty (s-t)a(s)ds \qquad (0 \le t < \infty). \tag{2.19}$$

Theorem 6 is proved in Section 8.

Remark 6.1. One can also prove the estimate (2.7) for the solution $\phi(t,\mu)$ of (1.1), (1.2) in the manner of Theorem 6, without using (2.5).

Our next task is to establish the existence of a boundary layer in a neighbourhood of $t = 0$ as $\mu \to 0^+$. For this purpose we consider the following approximation of the problem (1.1), (1.2) for small $t \ge 0$:

$$-\mu v'(t) = \int_{-\infty}^0 a(-s)F(v(t),g(s))ds \qquad (t > 0;\ v(0) = g(0)). \tag{2.20}$$

It will be observed that (2.20) is not a Volterra equation, but acts rather like an ordinary differential equation. Performing the stretching transformation

(2.21) $$t = \mu\tau$$

and setting $w(\tau) = v(t)$ transforms (2.20) to

(2.22) $$-\frac{dw}{d\tau} = \int_{-\infty}^{0} a(-s)F(w(\tau), g(s))ds \quad (\tau > 0;\ w(0) = g(0)) .$$

Theorem 7. Let (H_a), (H_F), (H_g) be satisfied. Then the initial value problem (2.22) has a unique solution $w = \xi(\tau)$ existing on $0 \le \tau < \infty$ and satisfying the following properties:

(2.23) $$\lim_{\tau \to \infty} \xi(\tau) = y_0 = \phi_0(0);\ 0 < \xi(\tau) - y_0 \le (g(0) - y_0)e^{-K\tau} \quad (0 \le \tau < \infty) ,$$

where ϕ_0 is the solution of (1.6) (see Theorem 3) and K is some positive constant.

Moreover, if $\phi(t,\mu)$ is the unique solution of (1.1), (1.2) in Theorem 1 and if $\xi(t/\mu)$ is the unique solution of (2.20) for $\mu > 0$, then for any $t_0 > 0$ there exists a constant $\bar{K} > 0$ (independent of μ) such that

(2.24) $$|\phi(t,\mu) - \xi(t/\mu)| \le \bar{K}t \qquad (0 \le t \le t_0;\ \mu > 0)$$

The estimate (2.24) establishes the existence of a boundary layer in a positive neighbourhood of $t = 0$. Theorem 7 is proved in Section 9.

In Corollary 3.1 we showed that $\alpha(\mu) \to \alpha_0$ as $\mu \to 0^+$, so that the solutions $\phi(t,\mu)$ of (1.1), (1.2) and $\phi_0(t)$ of (1.6) do not differ by much for small $\mu > 0$ and for large t. Our final result makes this

precise, under the additional assumptions that $ta(t) \in L^1(0,\infty)$ and $F \in C^2(\mathbb{R}^+ \times \mathbb{R}^+)$.

Theorem 8. Let (H_a), (H_F), (H_g) be satisfied. In addition, assume that $F \in C^2(\mathbb{R}^+ \times \mathbb{R}^+)$ and that (2.6) holds. Then there exist constants $K > 0$, $\mu_0 > 0$ and a function $\gamma \in C^1[0,\infty)$, γ positive, bounded and non-decreasing, such that

$$(2.25)\quad \phi_0(t) < \phi(t,\mu) < \phi_0(t) + (g(0) - \phi_0(0))\exp(-Kt/\mu) + \gamma(t)\mu|\log\mu| \qquad (0 \le t < \infty;\ 0 < \mu \le \mu_0).$$

In particular, as an immediate consequence of (2.25), there exists a constant $\tilde{K} > 0$ such that

$$(2.26)\quad 0 < \phi(t,\mu) - \phi_0(t) = O(\mu|\log\mu|), (\mu \to 0^+;\ \tilde{K}\mu|\log\mu| \le t < \infty).$$

Theorem 8 is proved in Section 10.

The method of proof of Theorem 8 uses the notions of upper and lower solutions of (1.1). The necessary preliminary material, which follows the lines of well known results (see e.g. [6], [19]) is collected in Appendix B. The inequality (2.25) is established by showing that the solution ϕ_0 of the reduced problem (1.6) is a lower solution of (1.1) on $0 \le t < \infty$ and by showing that

$$w(t,\mu) = \phi_0(t) + (g(0) - \phi_0(0))\exp(-Kt/\mu) + \gamma(t)\mu|\log\mu|$$

is an upper solution for suitably chosen K and γ (i.e. that

$$-\mu w'(t) < \int_{-\infty}^{t} a(t-s)F(w(t),w(s))ds \quad \text{for} \quad 0<t<\infty,$$

where $w(t) = g(t)$ on $-\infty < t < 0$). Inequality (2.25) then follows from Proposition 2B, Appendix B.

The question arises whether the order $O(\mu|\log \mu|)$ in (2.26) is best possible. In the linear case it is not; for if $\Gamma(y,z) = y - z$, one can establish the inequality

$$\phi_0(t) < \phi(t,\mu) < \phi_0(t) + (g(0) - \phi_0(0))\exp(-Kt/\mu) + \gamma(t)\mu$$

for $0 \leq t < \infty$ and $0 < \mu \leq \mu_0$, by the method of proof of Theorem 8. In addition, one can compute $\alpha(\mu)$ and α_0 in the linear case by the method of Laplace transforms and show that $\alpha(\mu) - \alpha_0$ is precisely of the order μ. In the general case, however, we have been unable to improve the estimate (2.25).

3. Proof of Theorem 1. The classical Picard successive approximations (or the Banach fixed point theorem) applied to the integrated form of (1.1), (1.2) show that for each fixed $\mu > 0$ there is a unique local solution $\phi(t,\mu)$ existing and in C^1 on some interval $[0,T]$, $T > 0$. To show that this solution can be continued (necessarily uniquely in view of the assumptions) to the interval $[0,\infty)$, it suffices to establish the inequalities (2.1) on any interval on which the solution $\phi(t,\mu)$ exists. For then the solution satisfies a priori upper and lower bounds, independent of T, and hence can be continued to the interval $[0,\infty)$ by a standard result [17].

To establish (2.1) on any interval on which $\phi(t,\mu)$ exists we have from (1.1), (1.2)

$$(3.1)\qquad -\mu\phi'(0,\mu) = \int_{-\infty}^{0} a(-s)F(g(0),g(s))ds\ .$$

The integral clearly exists since $a \in L^1(0,\infty)$ and $F(g(0),g(s))$ is bounded on $(-\infty,0]$ by $(H_F),(H_g)$. From (H_a), $a(-s) > 0$ $(-\infty < s \leq 0)$ and from (H_F), (H_g), $F(g(0),g(s)) \geq 0$ $(-\infty < s \leq 0)$, with the strict inequality holding for large negative s; therefore, $\phi'(0,\mu) < 0$. Since $\phi \in C^1$, one has by continuity that $\phi'(t,\mu) < 0$ $(0 \leq t < \alpha)$, for some $\alpha > 0$. We claim first that

$$(3.2)\qquad \phi(t,\mu) > 1 \qquad (0 \leq t \leq \alpha)\ .$$

Indeed, $\phi(0,\mu) = g(0) > 1$, and by continuity (3.2) holds at least on some interval to the right of $t = 0$. Suppose $0 < t_1 \leq \alpha$ is the first point at which $\phi(t_1,\mu) = 1$, and $1 < \phi(t,\mu) < g(0)$ $(0 < t < t_1)$. From (1.1) we have

$$(3.3)\qquad -\mu\phi\ (t_1,\mu) = \int_{-\infty}^{0} a(t_1 - s)F(1,g(s))ds + \int_{0}^{t_1} a(t_1 - s)F(1,\phi(s,\mu))ds\ .$$

By (H_a), $a(t_1 - s) > 0$ $(-\infty < s \leq t_1)$, and by (H_F), (H_g), $F(1,g(s)) \leq 0$ on $(-\infty,0]$, with the strict inequality holding near zero. Since ϕ is strictly decreasing on $[0,t_1]$, we also have $F(1,\phi(s,\mu)) < 0$ $(0 \leq s < t_1)$. Therefore each integral in (3.3) is negative and $\phi'(t_1,\mu) > 0$, which, in view of $\phi'(t,\mu) < 0$ $(0 \leq t < \alpha)$, is impossible; this proves (3.2).

We next claim:

(3.4) $$\phi'(t,\mu) < 0 ,$$

for as long as the solution exists. Indeed, suppose for contradiction that $\alpha > 0$ is the first point at which

(3.5) $$\phi'(\alpha,\mu) = 0 \quad \text{and} \quad \phi'(t,\mu) < 0 \qquad (0 \le t < \alpha) .$$

By the argument of the preceding paragraph we have

(3.6) $$1 < \phi(\alpha,\mu) < g(0) .$$

To prove (3.4) we compute $\phi''(\alpha,\mu)$ from (1.1) and we shall obtain an obvious contradiction of (3.5) by showing that

(3.7) $$\phi''(\alpha,\mu) < 0 ;$$

this implies that no such $\alpha > 0$ satisfying (3.5) exists and proves (3.4). Indeed, differentiating (1.1) (justified by (H_a), (H_F), (H_g) - note that by (H_a), $a' \in L^1(0,\infty)$) one has

(3.8) $$\begin{cases} -\mu\phi''(t,\mu) = \displaystyle\int_{-\infty}^{t} a'(t-s)F(\phi(t,\mu),\phi(s,\mu))ds \\ \qquad\qquad + \phi'(t,\mu)\displaystyle\int_{-\infty}^{t} a(t-s)F_1(\phi(t,\mu),\phi(s,\mu))ds . \end{cases}$$

Putting $t = \alpha$ and using (3.5) gives

(3.9) $$-\mu\phi''(\alpha,\mu) = \int_{-\infty}^{\alpha} a'(\alpha-s)F(\phi(\alpha,\mu),\phi(s,\mu))ds .$$

Thus to prove (3.7) we wish to show that

(3.10) $$I(\alpha) = \int_{-\infty}^{\alpha} a'(\alpha-s)F(\phi(\alpha,\mu),\phi(s,\mu))ds > 0 .$$

We shall need to consider two cases:

In case (i) $a(t)$ satisfies (H_a) with $a(t) \not\equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$ (i.e. $a'(t)/a(t) \not\equiv -\lambda$);

In case (ii) $a(t)$ satisfies (H_a) with $a(t) \equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$.

Case (i). Define a number $-\beta$, $\beta > 0$, by the relation $\phi(\alpha,\mu) = g(-\beta)$; $-\beta$ exists in view of (3.6) and (H_g). Since g may take the constant value $\phi(\alpha,\mu)$ on some interval $J \subset (-\infty,0]$, we define $-\beta$ uniquely by taking it to be the right-hand end point in such a case. We then have

$$(3.11)\quad I(\alpha) = \int_{-\infty}^{-\beta} a'(\alpha - s)F(g(-\beta),g(s))ds + \int_{-\beta}^{0} a'(\alpha - s)F(g(-\beta),g(s))ds$$

$$+ \int_{0}^{\alpha} a'(\alpha - s)F(\phi(\alpha,\mu),\phi(s,\mu))ds .$$

Since $\phi'(\alpha,\mu) = 0$ we also have from (1.1)

$$(3.12)\qquad 0 = \int_{-\infty}^{-\beta} a(\alpha - s)F(g(-\beta),g(s))ds + \int_{-\beta}^{0} a(\alpha - s)F(g(-\beta),g(s))ds$$

$$+ \int_{0}^{\alpha} a(\alpha - s)F(\phi(\alpha,\mu),\phi(s,\mu))ds .$$

We next define the function σ by the relation

$$\sigma(s) = \frac{a'(\alpha - s)}{a(\alpha - s)} \quad (-\infty < s \leq \alpha) ,$$

and we observe that the log convexity of a implies that $\sigma(s)$ is negative and nonincreasing; moreover, since $a'(t)/a(t) \not\equiv -\lambda$, $\lambda > 0$, $\sigma(s)$ is strictly decreasing, at least on some interval contained in $(-\infty,\alpha]$.

We rewrite (3.11) in the equivalent form

$$I(\alpha) = \int_{-\infty}^{-\beta} \sigma(s)h(s)ds + \int_{-\beta}^{\alpha} \sigma(s)h(s)ds \tag{3.13}$$

where

$$h(s) = \begin{cases} a(\alpha - s)F(g(-\beta), g(s)) & (-\infty < s \leq 0) \\ a(\alpha - s)F(\phi(\alpha,\mu), \phi(s,\mu)) & (0 < s \leq \alpha) \ ; \end{cases}$$

we also write (3.12) in the equivalent form

$$0 = \int_{-\infty}^{-\beta} h(s)ds + \int_{-\beta}^{\alpha} h(s)ds \ . \tag{3.14}$$

From the definition of $-\beta$ and (3.6) one has $1 < g(-\beta) < g(0)$. Therefore, (H_a), (H_F), (H_g) imply that

$$h(s) \geq 0 \qquad (-\infty < s \leq -\beta) \tag{3.15}$$

with strict inequality for large negative s, and

$$h(s) < 0 \qquad (-\beta < s < \alpha) \ . \tag{3.16}$$

Combining (3.13), (3.14) yields

$$I(\alpha) = \int_{-\infty}^{-\beta} (\sigma(s) - \sigma(-\beta))h(s)ds + \int_{-\beta}^{\alpha} (\sigma(s) - \sigma(-\beta))h(s)ds \ . \tag{3.17}$$

But

$$\sigma(s) \geq \sigma(-\beta) \quad (-\infty < s \leq -\beta) \ , \tag{3.18}$$

$$\sigma(s) \leq \sigma(-\beta) \quad (-\beta < s \leq \alpha) \ , \tag{3.19}$$

with strict inequalities holding either for s negative and large or for s near α. Using (3.15), (3.16) and (3.18), (3.19) in (3.17) shows that each integral in (3.17) is nonnegative and at least one of them is positive. This proves (3.10) and hence also (3.7) and (3.4). The proof of the global existence and uniqueness of the solution $\phi(t,\mu)$ of (1.1), (1.2) and of property (2.1) in case (i) is then completed by a straightforward continuation argument.

Case (ii). The above balancing argument establishing (3.7), and hence (3.4), cannot be used when $a(t) \equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$, since for this case it is readily established from (1.1) and (3.9) that $\phi''(\alpha,\mu) = 0$ whenever $\phi'(\alpha,\mu) = 0$. Instead we proceed as follows.

Define the number y_0, $1 < y_0 < g(0)$, by the equation $\int_{-\infty}^{0} e^{\lambda s} F(y_0, g(s))ds = 0$. (For the proof of the existence of a unique y_0 with this property in the general case of $a(t)$, which also applies here, see the proof of (2.8) in Section 5.) Since as in case (i) $\phi'(0,\mu) < 0$, a repetition of the argument (3.1) - (3.3) above with $a(t) = Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$, and "1" replaced by y_0 in (3.2) and the paragraph following (3.2), shows that $\phi(t,\mu) > y_0$ for $t \geq 0$, so long as $\phi'(t,\mu) < 0$. To show that $\phi'(t,\mu) < 0$ for all $t \geq 0$, differentiation of (1.1) with $a(t) = Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$, yields the equation

$$\mu\phi''(t,\mu) + \phi'(t,\mu)[\lambda\mu + \int_{-\infty}^{t} e^{-\lambda(t-s)} F_1(\phi(t,\mu), \phi(s,\mu))ds] = 0 ,$$

with

$$\phi(0,\mu) = g(0), \quad \phi'(0,\mu) = -\frac{1}{\mu}\int_{-\infty}^{0} e^{\lambda s} F(g(0), g(s))ds < 0 .$$

Since $F_1 > 0$, $\lambda > 0$, $\mu > 0$, an elementary argument shows that $\phi'(t,\mu) < 0$ so long as the solution $\phi(t,\mu)$ exists. This completes the proof of (2.1') in case (ii).

To prove (2.2) subtract the equations (1.1) for ϕ_1 and ϕ_2 and apply the mean value theorem obtaining

$$-\mu(\phi_1'(t,\mu) - \phi_2'(t,\mu)) = (\phi_1(t,\mu) - \phi_2(t,\mu))\int_{-\infty}^{t} a(t-s)F_1(\sigma(t), \phi_1(s,\mu))ds$$

$$+ \int_{-\infty}^{t} a(t-s)F_2(\phi_2(t,\mu), \tau(s))(\phi_1(s,\mu) - \phi_2(s,\mu))ds ,$$

where $\sigma(t)$ is between $\phi_1(t,\mu)$ and $\phi_2(t,\mu)$ and $\tau(s)$ is between $\phi_1(s,\mu)$ and $\phi_2(s,\mu)$. Let $t = t_0$ be the last point for which $\phi_1(t,\mu) \geq \phi_2(t,\mu)$: i.e. $\phi_1(t_0,\mu) = \phi_2(t_0,\mu)$, while $\phi_1(t,\mu) \geq \phi_2(t,\mu)$ for $-\infty < t < t_0$. Since $g_1(t) \geq g_2(t)$, $-\infty < t \leq 0$, it is clear that $t_0 \geq 0$. Both integrals exist since $a \in L^1(0,\infty)$, $F \in C^1$, and ϕ_1, ϕ_2 satisfy (2.1) (or (2.1')). From the definition of t_0 and $F_2 < 0$ one has $-\mu(\phi_1'(t_0,\mu) - \phi_2'(t_0,\mu)) \leq 0$, with the equality sign holding if and only if $g_1 \equiv g_2$ on $(-\infty, 0]$. Since $\mu > 0$ this implies $\phi_1'(t_0,\mu) - \phi_2'(t_0) > 0$ for $g_1 \not\equiv g_2$ and this is impossible.

To prove (2.3) put $z(t) = \phi(t,\mu_1) - \phi(t,\mu_2)$. Then from (1.1), (1.2) one has $z(0) = 0$ and

$$z'(0) = (-\frac{1}{\mu_1} + \frac{1}{\mu_2})\int_{-\infty}^{0} a(-s)F(g(0), g(s))ds > 0 .$$

By continuity suppose that there exists $T > 0$ such that $z(t) > 0$ $(0 < t < T)$ and $z(T) = 0$. Then

$$z'(T) = -\frac{1}{\mu_1}\int_{-\infty}^{T} a(T - s)F(\phi(T,\mu_1),\phi(s,\mu_1))ds$$

$$+\frac{1}{\mu_2}\int_{-\infty}^{T} a(T - s)F(\phi(T,\mu_2),\phi(s,\mu_2))ds .$$

But by (2.1) $\int_{-\infty}^{T} a(T - s)F(\phi(T,\mu_i),\phi(s,\mu_i))ds > 0$ $(i = 1, 2)$. Therefore $\mu_1 > \mu_2$ implies that

$$z'(T) > \frac{1}{\mu_1}\int_{-\infty}^{T} a(T - s)\{F(\phi(T,\mu_2),\phi(s,\mu_2)) - F(\phi(T,\mu_1),\phi(s,\mu_1))\}ds ,$$

which by the definition of T and $\phi(t,\mu_1) = \phi(t,\mu_2) = g(t)$ on $(-\infty, 0]$ yields

$$(3.20)\quad z'(T) > \frac{1}{\mu_1}\int_{0}^{T} a(T - s)\{F(\phi(T,\mu_2),\phi(s,\mu_2)) - F(\phi(T,\mu_2),\phi(s,\mu_1))\}ds .$$

Applying the mean value theorem in (3.20) gives

$$(3.21)\qquad z'(T) > \frac{1}{\mu_1}\int_{0}^{T} a(T - s)F_2(\phi(T,\mu_2),\zeta(s))(\phi(s,\mu_2) - \phi(s,\mu_1))ds ,$$

where $\zeta(s)$ is between $\phi(s,\mu_2)$ and $\phi(s,\mu_1)$. Since $a(T - s) > 0$ and $F_2 < 0$ by assumption and since $\phi(s,\mu_2) - \phi(s,\mu_1) < 0$ on $0 < s < T$ by the definition of T, (3.21) implies that $z'(T) > 0$. Therefore there cannot exist a $T > 0$ such that $z(T) = 0$; this proves (2.3), and completes the proof of Theorem 1.

4. Proof of Theorem 2 and Corollary 2.1. Let ϕ be the solution of (1.1), (1.2). We return to equation (3.8) obtained by differentiating (1.1) and we write (3.8) in the form

$$-\mu\phi''(t,\mu) = G(t,\mu)\phi'(t,\mu) + f(t,\mu) \quad (0 \le t < \infty) , \tag{4.1}$$

where

$$G(t,\mu) = \int_{-\infty}^{t} a(t-s)F_1(\phi(t,\mu),\phi(s,\mu))ds , \tag{4.2}$$

$$f(t,\mu) = \int_{-\infty}^{t} a'(t-s)F(\phi(t,\mu),\phi(s,\mu))ds . \tag{4.3}$$

Since $F_1 > 0$, $a \in L^1(0,\infty)$, $a(t) > 0$ $(0 \le t < \infty)$, and since ϕ satisfies (2.1), we have by (H_g),

$$0 < \gamma A \le G(t,\mu) \le \Gamma A \quad (0 \le t < \infty,\ \mu > 0) , \tag{4.4}$$

where

$$\begin{cases} \gamma = \inf\limits_{S} F_1(y,z),\ \Gamma = \sup\limits_{S} F_1(y,z),\ S = [1,g(0)] \times [1,g(0)] , \\ A = \int_0^{\infty} a(s)ds . \end{cases} \tag{4.5}$$

We next show that there exists a constant $K > 0$, independent of μ, such that

$$|f(t,\mu)| \le K \int_t^{\infty} a(s)ds \quad (0 \le t < \infty,\ \mu > 0) . \tag{4.6}$$

Write $f(t,\mu) = I_1 + I_2$, where

$$I_1 = \int_{-\infty}^{0} a'(t - s)F(\phi(t,\mu), g(s))ds ,$$

$$I_2 = \int_{0}^{t} a'(t - s)F(\phi(t,\mu), \phi(s,\mu))ds .$$

Consider I_2 first and recall, from the proof of Theorem 1, that $F(\phi(t,\mu), \phi(s,\mu)) < 0$ $(0 \leq s < t)$. Since $a' < 0$, one has, on letting $\sigma(t - s) = \frac{a'(t - s)}{a(t - s)}$ and on using the log convexity of a, that

$$0 < I_2 \leq (-\sigma(0)) \left| \int_{0}^{t} a(t - s)F(\phi(t,\mu), \phi(s,\mu))ds \right| . \tag{4.7}$$

But now (1.1) and conclusion (2.1) of Theorem 1, together with a simple consideration of signs of the terms on the right-hand side of (1.1), shows that

$$\left| \int_{0}^{t} a(t - s)F(\phi(t,\mu), \phi(s,\mu))ds \right| \leq \int_{-\infty}^{0} a(t - s)F(\phi(t,\mu), g(s))ds . \tag{4.8}$$

Moreover, the boundedness of ϕ and g and the continuity of F imply that

$$\int_{-\infty}^{0} a(t - s)F(\phi(t,\mu), g(s))ds \leq \sup_{\substack{0 \leq t < \infty \\ -\infty < s \leq 0}} |F(\phi(t,\mu), g(s))| \int_{t}^{\infty} a(s)ds . \tag{4.9}$$

Combining (4.7), (4.8), (4.9) shows the existence of a constant K, independent of μ, such that I_2 satisfies the estimate (4.6). In a similar way one finds on using the log convexity of a that

$$|I_1| \leq (-\sigma(0)) \sup_{\substack{0 \leq t < \infty \\ -\infty < s \leq 0}} |F(\phi(t,\mu), g(s))| \int_{t}^{\infty} a(s)ds .$$

Combining the estimates for I_1 and I_2 establishes (4.6).

Returning to (4.1) we have

$$(4.10)\qquad \frac{d}{dt}\left(-\phi'(t,\mu)\exp\left(\frac{1}{\mu}\int_0^t G(s,\mu)ds\right)\right) = \frac{f(t,\mu)}{\mu}\exp\left(\frac{1}{\mu}\int_0^t G(s,\mu)ds\right).$$

By (4.6) one has, for $0 \le t < \infty$ and $\mu > 0$,

$$(4.11)\qquad \frac{|f(t,\mu)|}{\mu}\exp\left(\frac{1}{\mu}\int_0^t G(s,\mu)ds\right) \le \frac{K}{\mu}\exp\left(\frac{1}{\mu}\int_0^t G(s,\mu)ds + \log\int_t^\infty a(s)ds\right).$$

Before integrating (4.10) observe that by $a,\ a' \in L^1(0,\infty)$ and by the log convexity of a one has

$$\frac{d}{dt}\left(\frac{1}{\mu}\int_0^t G(s,\mu)ds + \log\int_t^\infty a(s)ds\right) = \frac{G(t,\mu)}{\mu} - \frac{a(t)}{\int_t^\infty a(s)ds}$$

$$= \frac{G(t,\mu)}{\mu} + \frac{\int_t^\infty a'(s)ds}{\int_t^\infty a(s)ds} = \frac{G(t,\mu)}{\mu} + O(1)\quad (0 \le t < \infty,\ \mu > 0).$$

A simple but tedious calculation then shows that by use of (4.4), (4.6) in (4.11) there exist constants $\mu_0 > 0$, $\tilde{K} = \tilde{K}(\mu_0) > 0$ such that

$$(4.12)\qquad \left|\int_0^t \frac{f(\xi,\mu)}{\mu}\exp\left(\frac{1}{\mu}\int_0^\xi G(s,\mu)ds\right)d\xi\right|$$

$$\le \tilde{K}\exp\left(\frac{1}{\mu}\int_0^t G(s,\mu)ds + \log\left(\int_t^\infty a(s)ds\right)\right)\ (0 \le t < \infty;\ 0 < \mu \le \mu_0).$$

Then integrating (4.10) and using (4.12) and

$$-\phi'(0,\mu) = \frac{1}{\mu}\int_{-\infty}^0 a(-s)F(g(0),g(s))ds,$$

as well as conclusion (2.1) of Theorem 1, yields the estimate (2.5), where $K_1 = \gamma\Lambda$. This completes the proof of Theorem 2.

Proof of Corollary 2.1. Integrating (2.5) from t to infinity and using conclusions (2.1) and (2.4) of Theorem 1 yields

$$0 < \phi(t,\mu) - \alpha(\mu) \leq \tilde{K}e^{-K_1 t/\mu} + \tilde{K}\int_t^\infty \left(\int_\xi^\infty a(s)ds\right)d\xi \quad (0 < \mu \leq \mu_0;\ 0 \leq t < \infty) ;$$

an integration by parts gives

$$(4.13)\quad 0 < \phi(t,\mu) - \alpha(\mu) \leq \tilde{K}e^{-K_1 t/\mu} + \tilde{K}\int_t^\infty (s - t)a(s)ds \quad (0 < \mu \leq \mu_0;\ 0 \leq t < \infty) .$$

Since $a(t)$ is log convex, $\log a(t)$ is bounded below by an affine function; thus there exist constants $\alpha > 0$, $\beta > 0$ such that $a(t) \geq \alpha e^{-\beta t}$ $(0 \leq t < \infty)$. Hence given any $t_0 > 0$ there exists a constant which we again call $\mu_0 > 0$ such that the integral term in (4.13) dominates the first term for $t_0 \leq t < \infty$, $0 < \mu \leq \mu_0$; this, together with (4.13), establishes (2. 7) for $0 < t_0 \leq t < \infty$, $0 < \mu \leq \mu_0$, and one obtains the estimate (2.7) with $t_0 = 0$ by an appropriate modification of K. This completes the proof of Corollary 2.1.

5. Proof of Theorem 3 and Corollary 3.1. We observe first that if ϕ_0 is a (continuous) solution of (1.6) for $t \geq 0$, then $\phi_0(0) = \phi_0(0^+)$ must satisfy

$$(5.1)\qquad 0 = \int_{-\infty}^0 a(-s)F(\phi_0(0),\ g(s))\,ds .$$

This integral clearly exists for any number $\phi_0(0)$ since $a \in L^1(0,\infty)$ and (H_F), (H_g) hold; moreover by (H_F), the integral in (5.1) is a continuous, strictly increasing function of (the parameter) $\phi_0(0)$. The integral is negative if one chooses $\phi_0(0) = 1$; it is positive if one chooses $\phi_0(0) = g(0)$. Therefore, there exists a unique number $\phi_0(0) = y_0$ for which the integral in (5.1) vanishes and clearly $1 < y_0 < g(0)$. This proves (2.8).

We postpone to the end of this section the proof of the existence and uniqueness of the solution ϕ_0 of (1.6) and of the fact that $\phi_0 \in C^1[0,\infty)$ (with $\phi_0'(0) = \phi_0'(0^+)$), and we proceed to establish the remaining conclusions of Theorem 3. If $\phi_0 \in C^1[0,\infty)$, then differentiation of (1.6), justified by (H_a), (H_F), (H_g), yields

$$(5.2)\quad 0 = \int_{-\infty}^{t} a'(t-s)\, F(\phi_0(t), \phi_0(s))\, ds + \phi_0'(t) \int_{-\infty}^{t} a(t-s) F_1(\phi_0(t), \phi_0(s))\, ds \qquad (0 \le t < \infty).$$

We shall establish conclusions (2.9) - (2.11) by using (5.2). Note that the coefficient of ϕ_0' in (5.2) is bounded away from zero by (H_F).

A simple calculation shows that in the special case $a(t) \equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$, we have $\phi_0'(t) \equiv 0$ for $t > 0$ (from (5.2)), so that $\phi_0(t) \equiv y_0$ for $t \ge 0$, and this proves the remark below (2.9).

If $a(t) \not\equiv Ae^{-\lambda t}$, $A > 0$, $\lambda > 0$, we wish to show that (2.9) is satisfied. From (5.2) we first have

$$(5.3)\qquad -\phi_0'(0) = \frac{\int_{-\infty}^{0} a'(-s)I'(y_0, g(s))ds}{\int_{-\infty}^{0} a(-s)F_1(y_0, g(s))ds} .$$

By (H_a), (H_F), (H_g) the denominator in (5.3) is positive. To determine the sign of the numerator we note first that by the definition of y_0

$$(5.4)\qquad \int_{-\infty}^{0} a(-s)F(y_0, g(s))ds = 0 ,$$

and by the log convexity of a, repeating the argument in Theorem 1 which shows that the integral $I(\alpha) > 0$ in (3.10) (here we take $\alpha = 0$, $\phi(\alpha,\mu) = y_0$, $\phi(s,\mu) = g(s)$, and define $-\beta$, $\beta > 0$, by the relation $g(-\beta) = y_0$), we find

$$(5.5)\qquad \int_{-\infty}^{0} a'(-s)F(y_0, g(s))ds > 0 .$$

Therefore, by (5.3), (5.4), (5.5), one has $\phi_0'(0) < 0$.

Since $\phi_0 \in C^1[0,\infty)$, by continuity one has $\phi_0'(t) < 0$, $0 \le t < \alpha$, for some $\alpha > 0$. We claim, as in the analogous part of the proof of Theorem 1 (see the proof of (3.2)), that

$$(5.6)\qquad \phi_0(t) > 1 \qquad (0 \le t \le \alpha) .$$

Since $\phi_0(0) = y_0 > 1$ one has by continuity that (5.6) holds on some

interval to the right of $t = 0$. Letting $0 < t_1 \leq \alpha$ be the first point at which $\phi_0(t_1) = 1$, and using $y_0 > \phi_0(t) > 1$ for $0 < t < t_1$, we substitute in (1.6) with $t = t_1$ and see that the right-hand side is strictly negative. This contradiction proves (5.6).

We next show that

$$\phi_0'(t) < 0 \qquad (0 \leq t < \infty) . \tag{5.7}$$

Here the procedure differs somewhat from the analogous part of the proof of Theorem 1 in that we do not need to compute ϕ_0''. Suppose $t = \alpha$ is the first point at which

$$\phi_0'(\alpha) = 0 \quad \text{and} \quad \phi_0'(t) < 0 \qquad (0 \leq t < \alpha) . \tag{5.8}$$

Moreover, by (5.2) we have

$$\phi_0'(\alpha) = - \frac{\int_{-\infty}^{0} a'(\alpha - s)F(\phi_0(\alpha), g(s))ds + \int_{0}^{\alpha} a'(\alpha - s)F(\phi_0(\alpha), \phi_0(s))ds}{\int_{-\infty}^{0} a(\alpha - s)F_1(\phi_0(\alpha), g(s))ds + \int_{0}^{\alpha} a(\alpha - s)F_1(\phi_0(\alpha), \phi_0(s))ds} . \tag{5.9}$$

The denominator in (5.9) is clearly positive by (H_a), (H_F), (H_g). The balancing argument, involving (1.6) with $t = \alpha$ and the log convexity of a which has been previously employed in the proof of (3.10) and (5.5), shows that

$$\int_{-\infty}^{0} a'(\alpha - s)F(\phi_0(\alpha), g(s))ds + \int_{0}^{\alpha} a'(\alpha - s)F(\phi_0(\alpha), \phi_0(s))ds > 0 .$$

Therefore $\phi_0'(\alpha) < 0$, contradicting (5.8) and proving (5.7). This completes the proof of (2.9).

The proof of property (2.10) is similar to that of (2.2) in Theorem 1 and is omitted.

To prove property (2.11) we observe first that since $\phi(0,\mu) = g(0)$ and since $\phi_0(0) = y_0$, (2.11) is true at $t = 0$ by (2.8), and therefore, by continuity, (2.11) is true in some interval of $t \geq 0$. Suppose that $\phi(t,\mu) - \phi_0(t)$ is zero for the first time at $t = t_0 > 0$ and $\phi(t,\mu) - \phi_0(t) > 0$ $(0 \leq t < t_0)$. Then from (1.1), (1.6) at $t = t_0$, one obtains by subtraction

$$(5.10)\quad -\mu\phi'(t_0,\mu) = \int_0^{t_0} a(t_0 - s)[F(\phi(t_0,\mu),\phi(s,\mu)) - F(\phi_0(t_0),\phi_0(s))]\,ds\ .$$

Applying the mean value theorem to the difference under the integral (note that $\phi(t_0,\mu) = \phi_0(t_0)$ by the definition of t_0) and using $F_2(y,z) < 0$ and $\phi(s,\mu) - \phi_0(s) > 0$ $(0 \leq s < t_0)$, it is evident that the right-hand side of (5.10) is negative. This implies that $\phi'(t_0,\mu) > 0$ which contradicts (2.1) and proves (2.11).

It remains to prove the existence, uniqueness and differentiability of the solution ϕ_0 of (1.6) for $0 \leq t < \infty$. Let $0 < \varepsilon \leq t \leq T < \infty$ and let $\{\mu_n\}_{n=0}^{\infty}$ be an arbitrary sequence with $\mu_{n+1} < \mu_n$, $\lim_{n\to\infty} \mu_n = 0$. Consider the sequence $\{\phi(t,\mu_n)\}_{n=0}^{\infty}$ of solutions of the initial value problem (1.1), (1.2) on $\varepsilon \leq t \leq T$. By Theorem 1, $1 < \phi(t,\mu_n) < g(0)$, and $\phi(t,\mu_{n+1}) < \phi(t,\mu_n)$, $\varepsilon \leq t \leq T$, $n = 0, 1, \ldots$; therefore, $z(t) = \lim_{n\to\infty} \phi(t,\mu_n)$, $\varepsilon \leq t \leq T$, exists. By the estimate (2.5) of Theorem 2 the sequence $\{\phi(t,\mu_n)\}_{n=0}^{\infty}$ is equicontinuous on the interval $\varepsilon \leq t \leq T$ for

$0 < \mu_n \le \mu_0$, and so, by the Ascoli-Arzelà theorem and the fact that the sequence is already known to converge pointwise to $z(t)$, we see that the convergence to $z(t)$ is uniform for $\varepsilon \le t \le T$, $\varepsilon > 0$ being arbitrary, and so $z(t)$ is continuous for $t > 0$. Passing to the limit as $n \to \infty$ in the equation (1.1) for $\phi(t, \mu_n)$, we see, again using the estimate (2.5) and Lebesgue's convergence theorem, that $z(t)$ satisfies (1.6) for $t > 0$.

Since

$$\phi(t_1, \mu_n) > \phi(t_2, \mu_n) \quad \text{if} \quad 0 < t_1 < t_2 , \tag{5.11}$$

we see in the limit as $n \to \infty$ that $z(t)$ is a nonincreasing function of t and so $z(0+)$ exists. If we define $z(0) = z(0+)$, then clearly $z(t)$ satisfies (1.6) for $t = 0$ as well as for $t > 0$.

Identifying z with the required solution ϕ_0, we can show ϕ_0 to be continuously differentiable by forming difference quotients in (1.6), applying the mean value theorem and showing that we can pass to the limit to prove both the existence of ϕ_0' and the equation (5.2).

To prove uniqueness let $u, v \in C[0,T]$, $T > 0$ arbitrary, $u(0) = v(0) = y_0$, $u(t) = v(t) = g(t)$ on $-\infty < t < 0$, be two continuous solutions of (1.6) on $[0,T]$ with $u(t), v(t) > 1$ on $[0,T]$. Then by (H_F) and the mean value theorem we have

$$\begin{aligned} 0 = (u(t) - v(t))\Big[& \int_{-\infty}^{0} a(t-s)\, F_1\big(\xi_1(t,s), g(s)\big)\, ds \\ & + \int_0^t a(t-s)\, F_1\big(\xi_2(t,s), \eta_1(t,s)\big)\, ds \Big] \\ & + \int_0^t a(t-s) F_2\big(\xi_3(t,s), \eta_2(t,s)\big)\big(u(s) - v(s)\big)\, ds , \quad 0 \le \end{aligned} \tag{5.12}$$

for some $\xi_1(t,s)$, $\xi_2(t,s)$, $\xi_3(t,s)$ between $u(t)$ and $v(t)$, and for some $\eta_1(t,s)$, $\eta_2(t,s)$ between $u(s)$ and $v(s)$. But then $a \in L^1(0,\infty)$, the continuity of F_1, F_2, $F_1 > 0$, and the continuity of u, v on $[0,T]$, together with Gronwall's inequality applied to (5.12), imply that $u(t) \equiv v(t)$, $0 \le t \le T$, where $T > 0$ is arbitrary, establishing uniqueness of a continuous solution ϕ_0 of (1.6) on $[0,\infty)$ with $\phi_0(t) > 1$, $0 \le t < \infty$. This completes the proof of Theorem 3.

Proof of Corollary 3.1. To establish (2.13) note first that by Corollary 2.1, $\phi(T,\mu) \to \alpha(\mu)$ as $T \to \infty$ uniformly in μ for $0 < \mu \le \mu_0$. Let $\eta > 0$ be given. Using (2.4), (2.11), (2.12) choose $T > 0$ so large that $|\alpha(\mu) - \phi(t,\mu)| < \frac{\eta}{3}$ and $0 < \phi_0(t) - \alpha_0 < \frac{\eta}{3}$ for $t \ge T$, $0 < \mu \le \mu_0$. By the convergence of $\phi(t,\mu)$ to $\phi_0(t)$ as $\mu \to 0^+$ on $0 < t < \infty$ choose $0 < \bar{\mu}_0 \le \mu_0$ sufficiently small that $0 < \phi(T,\mu) - \phi_0(T) < \frac{\eta}{3}$, $0 < \mu \le \bar{\mu}_0$. Then

$$0 \le \alpha(\mu) - \alpha_0 \le |\alpha(\mu) - \phi(T,\mu)| + \left(\phi(T,\mu) - \phi_0(T)\right) + \left(\phi_0(T) - \alpha_0\right) < \eta,$$

for $0 < \mu \le \bar{\mu}_0$, proving (2.13).

The first inequality in (2.14) follows from (2.11). To prove the second inequality in (2.14) note that

$$\begin{aligned} \phi(t,\mu) - \phi_0(t) &= \phi(t,\mu) - \alpha(\mu) + \alpha(\mu) - \phi_0(t) \\ &\le \phi(t,\mu) - \alpha(\mu) + \alpha(\mu) - \alpha_0 \quad , \end{aligned}$$

where the last step follows by Theorem 3; (2.14) now follows by estimating $\phi(t,\mu) - \alpha(\mu)$ by (2.7). This completes the proof of Corollary 3.1.

6. Proof of Theorem 4. In view of the monotonicity property (2.10) of solutions of (1.6) with respect to the function g, it suffices to prove the result for the function g given by

$$g(t) = \begin{cases} 1+\delta & \text{if} \quad -\eta \le t \le 0 \\ 1 & \text{if} \quad t < -\eta \end{cases}, \quad \delta > 0, \ \eta > 0 . \tag{6.1}$$

(Strictly speaking, this function g does not satisfy the hypothesis (H_g), being discontinuous, but it is readily verified that the proof below would be essentially unaltered if the given g were replaced by a continuous g approximating sufficiently closely to it.) Since $1 \le g(t) \le 1+\delta$, $-\infty < t \le 0$, property (2.9) implies that $1 \le \phi_0(t) \le 1+\delta$, $0 \le t < \infty$. This means that the arguments of F in (1.6) are close to 1, if $\delta > 0$ is sufficiently small which will be the case in what follows. For this reason and in order to simplify the calculations we assume, consistent with (H_F) and without loss of generality, that

$$F_1(1,1) = 1, \quad F_2(1,1) = -1 ; \tag{6.2}$$

note that $F(x,x) = 0$ $(x > 0)$ implies that $F_1(x,x) = -F_2(x,x)$.

Substituting (6.1) into (1.6) yields (note $a \in L^1(0,\infty)$)

$$F(\phi_0(t),1) \int_{-\infty}^{-\eta} a(t-s)ds + F(\phi_0(t),1+\delta) \int_{-\eta}^{0} a(t-s)ds + \int_0^t a(t-s)F(\phi_0(t),\phi_0(s))ds = 0 . \tag{6.3}$$

Using (6.2), the mean value theorem, and $F \in C^1(\mathbb{R}\times\mathbb{R})$ yields

$$(6.4)\quad \begin{cases} (\phi_0(t)-1)\int_{-\infty}^{-\eta} a(t-s)ds + (\phi_0(t)-1-\delta)\int_{-\eta}^{0} a(t-s)ds \\ \qquad + o(\delta)\int_{-\infty}^{0} a(t-s)ds + \int_0^t a(t-s)(\phi_0(t)-\phi_0(s))ds \\ \qquad + \int_0^t a(t-s)[\,o(\phi_0(t)-\phi_0(s))]\,ds = 0\ , \end{cases}$$

where (by $a \in L^1(0,\infty)$, $a(t) > 0$) the terminology $w(t) = o(\delta)\int_{-\infty}^{0} a(t-s)ds$ means that for every $\varepsilon > 0$ one has $|w(t)| \le \varepsilon\delta\int_t^{\infty} a(\xi)d\xi$ for $t \ge 0$ and for $\delta > 0$ sufficiently small. An equivalent form of (6.4) is

$$(6.5)\quad (\phi_0(t)-1)\int_0^{\infty} a(\xi)d\xi - \int_0^t a(t-s)(\phi_0(s)-1)ds = \delta\int_t^{t+\eta} a(\xi)d\xi + o(\delta)\int_t^{\infty} a(\xi)d\xi + \int_0^t a(t-s)[\,o(\phi_0(t)-\phi_0(s))]\,ds\ .$$

Since $1 \le \phi_0(t) \le 1+\delta$ on $0 \le t < \infty$, (H_F) implies that the first term in (6.3) is positive, while the second and third terms are negative. Thus the third term in (6.3) must be in modulus less than the first term. This in turn implies that the last term in (6.5) (or the last integral in (6.4)) must be $o(\delta)\int_t^{\infty} a(\xi)d\xi$. Therefore, putting $z(t) = \phi_0(t) - 1$ in (6.5) yields the equivalent equation

$$(6.6)\quad z(t)A - a * z(t) = \delta\int_t^{t+\eta} a(s)ds + o(\delta\int_t^{\infty} a(s)ds)\ \ (0 \le t < \infty)\ ,$$

where $*$ denotes the convolution and $A = \int_0^{\infty} a(s)ds$.

We shall now apply a simple Tauberian theorem for (real) Laplace transforms [20; Theorem 4.3, p. 192] to solutions of (6.6). Put

$$\psi(t) = \int_t^{t+\eta} a(s)ds, \quad \omega(t) = \int_t^{\infty} a(s)ds , \tag{6.7}$$

$$\hat{z}(p) = \int_0^{\infty} e^{-pt} z(t)dt \qquad (p > 0) . \tag{6.8}$$

Since z is bounded and continuous on $[0,\infty)$, and since $a \in L^1(0,\infty)$, $\hat{z}(p)$ as well as $\hat{\psi}(p)$ and $\hat{\omega}(p)$ exist for $p > 0$. Noting that $A = \hat{a}(0)$ and that multiplication of (6.6) by e^{-pt} and integration preserves the o-relation when p is real, we obtain on solving for $\hat{z}(p)$

$$\hat{z}(p) = \frac{\delta\hat{\psi}(p) + o(\delta)\hat{\omega}(p)}{\hat{a}(0) - \hat{a}(p)} \quad (p > 0) . \tag{6.9}$$

By assumptions (H_a), (2.6), and Lebesgue's dominated convergence theorem, and integration by parts we also have

$$\hat{\omega}(0) = \int_0^{\infty}\int_t^{\infty} a(s)dsdt = \int_0^{\infty} sa(s)ds , \tag{6.10}$$

$$\hat{\psi}(0) = \int_0^{\infty}\int_t^{t+\eta} a(s)dsdt \le \int_0^{\infty}\int_t^{\infty} a(s)dsdt = \int_0^{\infty} sa(s)ds . \tag{6.11}$$

Moreover, by Fubini's theorem and (2.6)

$$\frac{d}{dp}\hat{a}(p) = -\int_0^{\infty} te^{-pt}a(t)dt \qquad (p \ge 0) , \tag{6.12}$$

so that by the mean value theorem and (2.6),

$$\hat{a}(0) - \hat{a}(p) \sim p\int_0^{\infty} ta(t)dt \qquad (p \to 0^+) . \tag{6.13}$$

Hence using (6.10), (6.11), (6.13), letting $p \to 0^+$ in (6.9), and taking $\delta > 0$ sufficiently small, there exists a constant $K(\delta) > 0$ such that

$$\hat{z}(p) \sim \frac{K(\delta)}{p} \qquad (p \to 0^+) . \tag{6.14}$$

The above-mentioned Tauberian theorem, together with the fact that $\lim_{t\to\infty} z(t)$ exist, then implies

$$z(t) \sim K(\delta) > 0 \qquad (t \to +\infty) , \tag{6.15}$$

and by the definition of z (6.15) yields

$$\phi_0(t) \sim 1 + K(\delta) > 1 \qquad (t \to +\infty) ,$$

for $\delta > 0$ sufficiently small. This completes the proof of Theorem 4.

7. Proof of Theorem 5. Using (2.16), equation (1.6) may be written in the form

$$Ay(t) - a * y(t) = \int_{-\infty}^{0} a(t - s)g(s)ds \qquad (0 \le t < \infty) , \tag{7.1}$$

where $A = \int_0^\infty a(s)ds$ and $a * y(t) = \int_0^t a(t - s)y(s)ds$. Letting, as in the proof of Theorem 4, $y(t) = z(t) - 1$,

(7.1) becomes the linear Volterra equation

$$Az(t) - a * z(t) = G(t) \qquad (0 \le t < \infty) , \tag{7.2}$$

where

$$G(t) = \int_{-\infty}^{0} a(t - s)(g(s) - 1)ds \qquad (0 \le t < \infty) . \tag{7.3}$$

We proceed as in the proof of Theorem 4, but taking any g satisfying (H_g), rather than the special g of (6.1). Taking the Laplace transform and noting that $A = \hat{a}(0)$ we obtain

$$\hat{z}(p) = \frac{\hat{G}(p)}{\hat{a}(0) - \hat{a}(p)} . \tag{7.4}$$

By Theorem 3, $\lim_{t \to \infty} z(t) = z_\infty$ exists and $z_\infty \geq 0$. If $z_\infty > 0$, a standard Abelian theorem for Laplace transforms [20, Cor. 1b, p. 182] states that $\hat{z}(p) \sim \frac{z_\infty}{p}$ $(p \to 0^+)$. We shall apply this result to (7.4) for p real. Since by (H_g) $g(t) \geq 1$ $(-\infty < t \leq 0)$, and since g is nondecreasing with $g(t) > 1$ near $t = 0$, we have by Fubini's theorem

$$0 < \hat{G}(p) = \int_0^\infty e^{-pt} \int_{-\infty}^0 a(t-s)(g(s) - 1)dsdt \tag{7.5}$$

$$= \int_{-\infty}^0 (g(s) - 1) \int_0^\infty e^{-pt} a(t - s)dtds \quad (p > 0) .$$

Let $\varepsilon > 0$ be given; choose a number $N = N(\varepsilon) > 0$, and using (H_g) divide the interval $(-\infty, 0]$ into two parts, such that

$$g(s) - 1 \leq \varepsilon \qquad (s \leq -N) . \tag{7.6}$$

Then (7.6) used in (7.5) yields

$$0 < \hat{G}(p) \leq \varepsilon \int_{-\infty}^0 e^{-ps} \int_{-s}^\infty e^{-p\theta} a(0)d\theta ds + O(1) \qquad (p \to 0^+) ,$$

or equivalently

$$0 < \hat{G}(p) \le \varepsilon \int_0^\infty e^{p\sigma} \int_\sigma^\infty e^{-p\theta} a(\theta)\,d\theta\,d\sigma + O(1) \qquad (p \to 0^+) . \tag{7.7}$$

By an integration by parts and $a \in L^1(0,\infty)$ we also have

$$\hat{a}(0) - \hat{a}(p) = \int_0^\infty (1 - e^{-pt}) a(t)\,dt = \int_0^\infty (1 - e^{-pt}) e^{pt} e^{-pt} a(t)\,dt \tag{7.8}$$

$$= p \int_0^\infty e^{pt} \int_t^\infty e^{-p\theta} a(0)\,d\theta\,dt > 0 \qquad (p > 0) .$$

Therefore, (7.7), (7.8) substituted into (7.4) yields

$$\hat{z}(p) \le \frac{\varepsilon \int_0^\infty e^{pt} \int_t^\infty e^{-p\theta} a(0)\,d\theta\,dt + O(1)}{p \int_0^\infty e^{pt} \int_t^\infty e^{-p\theta} a(0)\,d\theta\,dt} \qquad (p \to 0^+) . \tag{7.9}$$

But by (2.17)

$$\lim_{p \to 0^+} \int_0^\infty e^{pt} \int_t^\infty e^{-p\theta} a(0)\,d\theta\,dt = +\infty ,$$

and this fact used in (7.9) shows that for any $\varepsilon > 0$

$$\hat{z}(p) \le \frac{\varepsilon}{p} + o\left(\frac{1}{p}\right) \qquad (p \to 0^+) .$$

Since $\varepsilon > 0$ is arbitrary, $z_\infty > 0$ is impossible. This completes the proof of Theorem 5.

8. Proof of Theorem 6. Let ϕ_0 be the unique solution of (1.6) (see Theorem 3). Thus ϕ_0 satisfies

$$(8.1)\quad \int_{-\infty}^{0} a(t-s)F(\phi_0(t), g(s))ds + \int_{0}^{t} a(t-s)F(\phi_0(t), \phi_0(s))ds = 0 \quad (0 \le t < \infty).$$

By $a \in L^1(0,\infty)$, (H_F), (H_g), and the boundedness of ϕ_0 there exists a constant $K > 0$ such that

$$(8.2)\qquad 0 < \int_{-\infty}^{0} a(t-s)F(\phi_0(t), g(s))ds \le K \int_{t}^{\infty} a(s)ds \quad (0 \le t < \infty).$$

(The first inequality in (8.2) follows by simple consideration of signs in (8.1).) Applying (H_F) and the mean value theorem we have

$$F(\phi_0(t), \phi_0(s)) = F_1(\phi_0(s), \phi_0(s))(\phi_0(t) - \phi_0(s)) + o(\phi_0(t) - \phi_0(s)),$$

which by the boundedness of ϕ_0, $F_1 > 0$, and the continuity of F_1 implies that there exist constants M_1, $M_2 > 0$ such that

$$(8.3)\quad M_1(\phi_0(s) - \phi_0(t)) \le -F(\phi_0(t), \phi_0(s)) \le M_2(\phi_0(s) - \phi_0(t)) \quad (0 \le s \le t < \infty).$$

Using (8.2), (8.3) in (8.1) shows that

$$(8.4)\qquad \int_{0}^{t} a(t-s)(\phi_0(t) - \phi_0(s))ds = \psi(t) \qquad (0 \le t < \infty)$$

where

$$(8.5)\qquad \psi(t) = O\left(\int_{t}^{\infty} a(s)ds\right).$$

Let

$$z(t) = \phi_0(t) - \alpha_0 \quad (\alpha_0 = \lim_{t \to \infty} \phi_0(t)) . \tag{8.6}$$

Then, from (8.4), z satisfies the linear Volterra equation

$$z(t) \int_0^t a(s)ds - \int_0^t a(t - s)z(s)ds = \psi(t) \quad (0 \leq t < \infty) . \tag{8.7}$$

Writing $\int_0^t a(s)ds = \int_0^\infty a(s)ds - \int_t^\infty a(s)ds$ and noting that $z(t)$ is bounded shows that (8.7) may be written in the form

$$\begin{cases} z(t)A - a * z(t) = w(t) \quad (0 \leq t < \infty; A = \int_0^\infty a(s)ds > 0) \\ w(t) = \psi(t) + z(t) \int_t^\infty a(s)ds = O(\int_t^\infty a(s)ds) \quad (0 \leq t < \infty). \end{cases} \tag{8.8}$$

We now solve (8.8) by Laplace transforms. By the argument of Theorem 4 and using $\int_0^\infty ta(t)dt < \infty$, we find that if

$$\hat{w}(0) = \int_0^\infty w(t)dt \neq 0$$

(note that this integral exists in view of the estimate satisfied by w and $ta(t) \in L^1(0, \infty)$), then $\lim_{t \to \infty} z(t) = z_\infty \neq 0$. But, by Theorem 3, $z_\infty = 0$. Therefore,

$$\int_0^\infty w(t)dt = 0 . \tag{8.9}$$

Now integrate (8.8) over $[0,T]$, $T > 0$, obtaining

$$A\int_0^T z(s)ds - \int_0^T a * z(t)dt = \int_0^T w(t)dt .$$

But, by (8.9), the last equation can be written as

$$A\int_0^T z(s)ds - \int_0^T a * z(t)dt = -\int_T^\infty w(t)dt . \tag{8.10}$$

Interchanging the order of integration in the double integral on the left-hand side of (8.10) and using the estimate for w in (8.8), as well as $ta(t) \in L^1(0,\infty)$, yields

$$A\int_0^T z(s)ds - \int_0^T z(s)\int_0^{T-s} a(\sigma)d\sigma ds = O\Big(\int_T^\infty (t-T)a(t)dt\Big) \quad (0 \le T < \infty) . \tag{8.11}$$

Using the definition of A and combining the two integrals on the left-hand side of (8.11) yields

$$\int_0^T z(s)\int_{T-s}^\infty a(\sigma)d\sigma ds = O\Big(\int_T^\infty (t-T)a(t)dt\Big) . \tag{8.12}$$

Finally, using the fact that z is decreasing on $0 \le T < \infty$, we obtain from (8.12)

$$z(T)\int_0^T\int_{T-s}^T a(\sigma)d\sigma ds = O\Big(\int_T^\infty (t-T)a(t)dt\Big),$$

which on interchanging the order of integration yields

$$z(T)\int_0^T \sigma a(\sigma)d\sigma = O\Big(\int_T^\infty (t-T)a(t)dt\Big). \tag{8.13}$$

This implies (2.19) and completes the proof of Theorem 6.

9. Proof of Theorem 7. By standard results the initial value problem (2.22) has a unique local solution $w = \xi(t)$ having

$$-\xi'(0) = \int_{-\infty}^{0} a(-s)F(g(0), g(s))ds > 0 . \tag{9.1}$$

Thus ξ decreases initially. To continue the local solution we proceed similarly to the proof of Theorem 1. First, we show that for as long as $\xi'(\tau) < 0$ one has

$$\xi(\tau) > y_0 , \tag{9.2}$$

where y_0 is (see Theorem 3) the unique value for which

$$\int_{-\infty}^{0} a(-s)F(y_0, g(s))ds = 0 . \tag{9.3}$$

We observe that equation (2.22) may be regarded as an autonomous ordinary differential equation having the point y_0 as its only critical point. Recall from Theorem 3 that $(\xi(0) =)g(0) > y_0$. If ξ assumes the value y_0 at $\tau = \tau_1$, then by (9.3)

$$-\xi'(\tau_1) = \int_{-\infty}^{0} a(-s)F(y_0, g(s))ds = 0 ,$$

which is impossible if $\xi'(\tau_1) < 0$. This proves (9.2). On the other hand, $\xi'(0) < 0$ implies by continuity that $\xi'(\tau) < 0$ for $0 \leq \tau < \alpha$, $\alpha > 0$. If $\alpha > 0$ is the first point at which $\xi'(\alpha) = 0$, then

$$0 = -\xi'(\alpha) = \int_{-\infty}^{0} a(-s)F(\xi(\alpha), g(s))ds .$$

This by uniqueness of y_0 implies that $\xi(\alpha) = y_0$, and therefore $w = \xi(\tau)$ and $w \equiv y_0$ would be two solutions of (2.22) through the point (α, y_0), contradicting uniqueness. Therefore,

$$\xi'(\tau) < 0 , \tag{9.4}$$

for as long as the solution exists. Now (9.2), (9.4) and a standard continuation argument yield the global existence and uniqueness of the solution $w = \xi(\tau)$ of (2.22) such that

$$\xi'(\tau) < 0, \; y_0 < \xi(\tau) \leq g(0) \quad (0 \leq \tau < \infty) , \tag{9.5}$$

which implies that $\lim_{\tau \to \infty} \xi(\tau) = \xi_\infty$ exists and $\xi_\infty \geq y_0$.

To prove (2.23) we combine (2.22) and (9.3) obtaining

$$-\frac{d\xi}{d\tau} = \int_{-\infty}^{0} a(-s)(F(\xi(\tau), g(s)) - F(y_0, g(s))ds \quad (0 \leq \tau < \infty) . \tag{9.6}$$

Applying the mean value theorem and (H_F) yields the existence of a $\theta(\tau,s)$, $0 < \theta(\tau,s) < 1$, such that

$$-\frac{d\xi}{d\tau} = \int_{-\infty}^{0} a(-s)F_1(y_0 + \theta(\tau,s)(\xi(\tau) - y_0), g(s))ds[\xi(\tau) - y_0] . \tag{9.7}$$

Let $S = [y_0, g(0)] \times [1, g(0)]$, $A = \int_0^\infty a(s)ds > 0$. By (H_F)

$$\gamma = \inf_{(y,z) \in S} F_1(y,z) > 0 .$$

Therefore using this and (9.5) in (9.7) gives the differential inequality

$$-\frac{d\xi}{d\tau} \geq \gamma A(\xi(\tau) - y_0) \quad (0 < \tau < \infty) . \tag{9.8}$$

Integrating (9.8) and using the initial condition $\xi(0) = g(0)$ yields the second statement in (2.23), which also implies the first statement in (2.23).

Applying the transformation (2.21) it is clear that $y(t) = \xi(t/\mu)$ is the unique solution of the initial value problem (2.20) for $\mu > 0$, $0 \le t < \infty$. Thus

$$(9.9) \qquad \begin{cases} -\mu \frac{d}{dt} \xi(\frac{t}{\mu}) = \int_{-\infty}^{0} a(-s) F(\xi(\frac{t}{\mu}), g(s)) ds \quad (\mu > 0,\ 0 < t < \infty) \\ \xi(0) = g(0) . \end{cases}$$

Let $\phi(t,\mu)$ be the unique solution of (1.1), (1.2) on $0 \le t < \infty$, $\mu > 0$ (see Theorem 1), which can be written in the form

$$(9.10) \qquad -\mu \frac{d}{dt} \phi(t,\mu) = \int_{-\infty}^{0} a(-s) F(\phi(t,\mu), g(s))\, ds$$

$$+ t \int_{-\infty}^{0} a'(-s + \theta(t,s)t) F(\phi(t,\mu), g(s)) ds$$

$$+ \int_{0}^{t} a(t - s) F(\phi(t,\mu), \phi(s,\mu)) ds \quad (0 \le t < \infty,\ \mu > 0) ,$$

where $0 < \theta(t,s) < 1$ comes from the application of the mean value theorem to $a(t - s) - a(-s)$. Note that since $F(\phi(t,\mu), g(s))$ is bounded and $a' \in L^1(0,\infty)$, the second integral in (9.10) clearly exists. Subtracting (9.9) from (9.10), using the mean value theorem once more, and observing that the last two terms in (9.10) are $O(t)$ as $t \to 0^+$, uniformly in $\mu > 0$, yields

(9.11) $$-\mu \frac{d}{dt}[\phi(t,\mu) - \xi(\frac{t}{\mu})] = \mathfrak{A}(t,\mu)[\phi(t,\mu) - \xi(\frac{t}{\mu})] + O(t)$$

$$(t \to 0^+, \quad \text{uniformly in} \quad \mu > 0),$$

where

(9.12) $$\mathfrak{A}(t,\mu) = \int_{-\infty}^{0} a(-s)F_1(\xi(\frac{t}{\mu}) + \theta(t,s,\mu)(\phi(t,\mu) - \xi(\frac{t}{\mu})), g(s))ds,$$

$0 < \theta(t,s,\mu) < 1$. Letting $S = [1, g(0)] \times [1, g(0)]$, $A = \int_0^\infty a(s)ds$, and using (H_a), (H_F), (H_g) implies

(9.13) $$\inf_{\substack{0 \le t \le t_0 \\ \mu > 0}} \mathfrak{A}(t,\mu) \ge \gamma A > 0,$$

where $\gamma = \inf_S F_1(y,z)$ is independent of μ. Using (9.13) in (9.11), integrating (9.11), and applying the initial condition $\phi(0,\mu) = \xi(0) = g(0)$ yields the existence of a constant $K = K(t_0)$, independent of μ, such that

$$|\phi(t,\mu) - \xi(\frac{t}{\mu})| \le K \int_0^t \frac{s}{\mu} e^{-\frac{\gamma A}{\mu}(t-s)} ds \qquad (0 \le t \le t_0 : \mu > 0)$$

$$\le \frac{Kt}{\mu} \int_0^t e^{-\frac{\gamma A}{\mu}(t-s)} ds \le \frac{Kt}{\gamma A},$$

which is (2.24) with $\overline{K} = K/\gamma A$. This completes the proof of Theorem 7.

10. Proof of Theorem 8. We show that

$$w(t,\mu) = \phi_0(t) + (g(0) - \phi_0(0))\exp(-Kt/\mu) + \gamma(t)\mu|\log\mu|, \quad 0 \le t < \infty, \tag{10.1}$$

where ϕ_0 is the solution of (1.6), is an upper solution of (1.1) for $0 < t < \infty$ for suitable choices of the constant $K > 0$ and the function γ. By Appendix B it suffices to show that w defined by (10.1) for $0 \le t < \infty$ and $w(t) = g(t)$ $(-\infty < t < 0)$ satisfies the integrodifferential inequality

$$-\mu w'(t,\mu) < \int_{-\infty}^{t} a(t-s)F(w(t,\mu), w(s,\mu))ds \quad (0 < t < \infty). \tag{10.2}$$

We begin by defining γ and K in (10.1). For reasons which will become apparent below let

$$\gamma(t) = \gamma_0 \exp(K_1 \int_0^t (\int_\sigma^\infty a(\tau)d\tau)d\sigma) \quad (0 \le t < \infty) \tag{10.3}$$

where $\gamma_0 > 0$, $K_1 > 0$ are constants specified below. In view of (2.6), $\gamma(\infty)$ exists with $\gamma(\infty) = \gamma_0 e^{BK_1}$, where

$$B = \int_0^\infty \tau a(\tau)\, d\tau .$$

Thus the function w defined by (10.1) satisfies the inequality

$$1 \le w(t,\mu) \le g(0) + \gamma(\infty)\mu|\log\mu| \qquad (0 \le t < \infty). \tag{10.4}$$

Let J denote the closed interval $[1, 2g(0)]$ and let R denote the rectangle $J \times J$. Let $M > 0$ be a constant such that

$$\sup_{(y,z)\in R} \{|F(y,z)|, |F_1(y,z)|, |F_2(y,z)|, |F_{11}(y,z)|, |F_{12}(y,z)|, |F_{22}(y,z)|\} \le M \tag{10.5}$$

and let $m > 0$ be a constant such that (see (H_F))

$$\inf_{(y,z)\in R} F_1(y,z) \geq m . \tag{10.6}$$

Then for any choice of γ_0, K_1, independent of μ, one has for μ sufficiently small that the points

$$(w(t,\mu), w(s,\mu)),\ (w(t,\mu), \phi_0(s)),\ (\phi_0(t), w(s,\mu)),\ (\phi_0(t), \phi_0(s)) \in R$$

for $0 \leq s \leq t < \infty$; this statement is also true for $t < 0$ since $w(t,\mu) = \phi_0(t) = g(t)$ for $t < 0$. Therefore the values of F and its first and second partial derivatives at these points are by (10.5) bounded by M.

We define K in (10.1) by $K = \frac{1}{2} mA$, where $A = \int_0^\infty a(s)ds$.

With these definitions of K and γ it remains to verify (10.2). We begin by doing this on the interval $0 < t \leq \frac{2}{K} \mu |\log \mu|$. To simplify the exposition let RHS denote the integral on the right-hand side of (10.2) for t on any interval under consideration. We shall also suppress the parameter μ in $w(t,\mu)$ when no confusion arises. By the mean value theorem we have for $0 < t \leq \frac{2}{K} \mu |\log \mu|$:

$$\text{RHS} = \int_{-\infty}^{t} a(t-s) F\big((w-\phi_0)(t) + \phi_0(t), (w-\phi_0)(s) + \phi_0(s)\big) ds \tag{10.7}$$

$$= (w-\phi_0)(t) \int_{-\infty}^{t} a(t-s) F_1\big(\xi(t,s), \eta(t,s)\big) ds + O(\mu |\log \mu|),$$

where we have used the facts that ϕ_0 satisfies (1.6) and that

$(w - \phi_0)(s) = 0$ for $s < 0$. In (10.7) the O-term does not exceed

$$\frac{2Ma(0)}{K}\,\mu|\log\mu|\{(g(0) - \phi_0(0)) + \gamma(t)\mu|\log\mu|\}$$

and the point $(\xi(t, s), \eta(t, s))$ lies in R. Thus (10.1), (10.6) used in (10.7) yields

(10.8) $$\text{RHS} \geq mA\{(g(0) - \phi_0(0))e^{-Kt/\mu} + \gamma(t)\mu|\log\mu|\} + O(\mu|\log\mu|).$$

We can clearly choose γ_0 in (10.3) sufficiently large that the term $mA\gamma(t)\mu|\log\mu|$ in (10.8) is at least twice the O-term if μ is sufficiently small, and we leave K_1 arbitrary at the moment. Then (10.8) certainly yields

(10.9) $$\text{RHS} \geq mA\{(g(0) - \phi_0(0))e^{-Kt/\mu} + \frac{1}{2}\gamma_0\mu|\log\mu|\}.$$

On the other hand, it follows from (10.1) and (10.3) using $\gamma'(t) > 0$ $(0 < t < \infty)$ that

(10.10) $$-\mu w'(t) \leq -\mu\phi_0'(t) + K(g(0) - \phi_0(0))e^{-Kt/\mu} \quad (t > 0),$$

and so from the choice of K that for μ sufficiently small the desired inequality (10.2) holds on the interval $0 < t \leq \frac{2}{K}\mu|\log\mu|$.

We next verify (10.2) on the interval $\frac{2}{K}\mu|\log\mu| \leq t \leq K_2$ where K_2 is a positive constant, independent of μ, to be determined. Observe that for $t \geq \frac{2}{K}\mu|\log\mu|$,

(10.11) $$(w - \phi_0)(t) \leq \gamma(t)\mu|\log\mu| + O(\mu^2).$$

Thus by the mean value theorem and the fact that $w(t) = \phi_0(t) = g(t)$ $(-\infty < t <$ one has

(10.12) $$\text{RHS} = (w - \phi_0)(t) \int_{-\infty}^{t} a(t - s) F_1\big(\xi(t, s),\ \eta(t, s)\big)\, ds$$

$$+ \int_{0}^{2\mu |\log \mu|/K} a(t - s)(w - \phi_0)(s) F_2\big(\xi(t, s),\ \eta(t, s)\big)\, ds$$

$$+ \int_{2\mu |\log \mu|/K}^{t} a(t - s)(w - \phi_0)(s) F_2\big(\xi(t, s),\ \eta(t, s)\big)\, ds \quad \left(t \geq \frac{2}{K}\mu |\log \mu|\right),$$

where the point $(\xi(t,s), \eta(t,s))$ lies in R. Let I_1, I_2, I_3 denote the first, second, and third integrals in (10.12). Then by (10.1), (10.5), and γ nondecreasing one has

(10.13) $$|I_2| \leq \frac{2Ma(0)}{K} \gamma(t)\mu^2 |\log \mu|^2 + \frac{Ma(0)}{K} \mu(g(0) - \phi_0(0)),$$

and similarly

(10.14) $$|I_3| \leq M[\gamma(t)\mu |\log \mu| + O(\mu^2)] \int_{2\mu|\log \mu|/K}^{t} a(t - s)ds .$$

Note that I_2, I_3 are each negative while I_1 is positive, and for $\mu > 0$ sufficiently small

(10.15) $$I_1 \geq \frac{3}{4} mA\ \gamma(t)\mu |\log \mu| .$$

Hence if K_2 is chosen independent of μ and sufficiently small, it follows from (10.12) - (10.15) that for μ sufficiently small

(10.16) $$\text{RHS} \geq \frac{1}{2} mA\gamma(t)\mu |\log \mu| \qquad \left(\frac{2\mu |\log \mu|}{K} \leq t \leq K_2\right) .$$

From (10.1), (10.3) one has again (10.10) valid on the interval being considered. Comparing (10.10) and (10.16) one finds that (10.2) is satisfied on the interval $\frac{2}{K}\mu|\log\mu| \le t \le K_2$, with K_2 chosen as above, if $\mu > 0$ is sufficiently small. Note that in this and the previous part of the proof we have not used the special form of $Y(t)$ in any significant way.

Finally, we establish (10.2) for any $t > K_2$. For any such t (10.12) is valid and the estimate for I_2 given by (10.13) holds. In fact, (10.13) can be strengthened to give

$$|I_2| \le \frac{2M}{K} a\left(t - \frac{2\mu|\log\mu|}{K}\right) Y\left(\frac{2\mu|\log\mu|}{K}\right) \mu^2 |\log\mu|^2$$

$$+ \frac{M}{K} a\left(t - \frac{2\mu|\log\mu|}{K}\right) \mu \left(g(0) - \phi_0(0)\right),$$

and since a'/a is bounded, we have

$$a\left(t - \frac{2\mu|\log\mu|}{K}\right) = a(t) - \frac{2\mu|\log\mu|}{K} a'(\tau),$$

where

$$t - \frac{2\mu|\log\mu|}{K} < \tau < t,$$

and so

$$a\left(t - \frac{2\mu|\log\mu|}{K}\right) = a(t)\left(1 + O(\mu|\log\mu|)\right).$$

Hence, for μ sufficiently small,

$$|I_2| \leq \frac{4M}{K}\, a(t)\, \gamma\left(\frac{2\mu|\log\mu|}{K}\right) \mu^2 |\log\mu|^2$$

(10.17)
$$+ \frac{M}{K}\, a(t)\, \mu\left(g(0) - \phi_0(0)\right) .$$

Further, since a'/a is bounded, it follows on integration that

(10.18)
$$a(t) \Big/ \int_t^\infty a(\tau)\, d\tau$$

is bounded.

We can write

$$I_1 = (w - \phi_0)(t) \left\{ \int_{-\infty}^{2\mu|\log\mu|/K} + \int_{2\mu|\log\mu|/K}^{t} \right\} a(t-s)\, F_1(\xi(t,s), \eta(t,s))\, ds$$

$$= I_{11} + I_{12} , \quad \text{say} .$$

Certainly,

$$I_1 > m\gamma(t)\, \mu |\log\mu| \int_{-\infty}^{0} a(t-s)\, ds$$

$$= m\gamma(t)\, \mu |\log\mu| \int_t^\infty a(\tau)\, d\tau ,$$

so that, if μ is sufficiently small, we can use (10.17) and (10.18) to obtain

(10.19)
$$I_1 + I_2 \geq \frac{1}{2}\, m\gamma(t)\, \mu |\log\mu| \int_t^\infty a(\tau)\, d\tau .$$

But, from (10.10) and Theorem 2 (or more precisely Remark 3.1), the left-hand side of (10.2) does not exceed

$$\mu \tilde{K} \int_t^\infty a(\tau)d\tau + K\big(g(0) - \phi_0(0)\big) e^{-Kt/\mu} \quad . \tag{10.20}$$

We have already remarked in the proof of Corollary 2.1 that $a(t) \geq \alpha e^{-\beta t}$ for positive constants $\alpha, \beta,$ and that consequently the second term in (10.20) is negligible compared with the first if μ is sufficiently small, $t \geq K_2$. Comparing (10.19) and (10.20), we see that (10.2) is certainly satisfied for μ sufficiently small and $t \geq K_2$ if

$$I_{12} + I_3 \geq 0 \quad .$$

Now

$$I_{12} + I_3 = \int_{2\mu|\log\mu|/K}^{t} a(t-s)\,[\,(w-\phi_0)(t)\,F_1\big(\xi(t,s), \eta(t,s)\big) + (w-\phi_0)(s)\,F_2\big(\xi(t,s), \eta(t,s)\big)\,]\,ds \quad . \tag{10.21}$$

The hypothesis $F(x,x) = 0$ implies that $F_1(x,x) = -F_2(x,x)$, and the bounds on the first and second partial derivatives of F imply that the first derivatives of F_2/F_1 are bounded by $2M^2/m^2$. Hence the expression in $[\ldots]$ in (10.21) may be written

$$(w-\phi_0)(s)\,F_1\big(\xi(t,s), \eta(t,s)\big)\left\{\frac{(w-\phi_0)(t)}{(w-\phi_0)(s)} + \frac{F_2\big(\xi(t,s), \eta(t,s)\big)}{F_1\big(\xi(t,s), \eta(t,s)\big)}\right\} \tag{10.22}$$

$$\geq (w-\phi_0)(s)\,F_1\big(\xi(t,s), \eta(t,s)\big)\left\{\frac{(w-\phi_0)(t)}{(w-\phi_0)(s)} - 1 - \frac{2M^2}{m^2}\,|\,\xi(t,s) - \eta(t,s)|\right\}.$$

From the application of the mean value theorem in which they arise, we see that

$$\xi(t, s) = \phi_0(t) + \theta(t, s)\,(w - \phi_0)(t),$$

$$\eta(t, s) = \phi_0(s) + \theta(t, s)\,(w - \phi_0)(s),$$

where $0 < \theta < 1$.

Thus

$$|\xi(t, s) - \eta(t, s)| \leq |\phi_0(t) - \phi_0(s)| + |(w - \phi_0)(t) - (w - \phi_0)(s)| ,$$

and

$$\frac{(w - \phi_0)(t)}{(w - \phi_0)(s)} - 1 - \frac{2M^2}{m^2} |\xi(t, s) - \eta(t, s)|$$

$$\geq \left\{\frac{(w - \phi_0)(t)}{(w - \phi_0)(s)} - 1\right\} \left(1 - \frac{2M^2}{m^2} (w - \phi_0)(s)\right) \tag{10.23}$$

$$- \frac{2M^2}{m^2} |\phi_0(t) - \phi_0(s)| .$$

But since, in the range of integration with which we are now concerned,

$$2\mu |\log \mu| / K \leq s \leq t ,$$

we have

$$\frac{(w - \phi_0)(t)}{(w - \phi_0)(s)} - 1 = \frac{(\gamma(t) - \gamma(s))\,\mu |\log \mu| + (g(0) - \phi_0(0))(e^{-Kt/\mu} - e^{-Ks/\mu})}{\gamma(s)\,\mu |\log \mu| + O(\mu^2)} .$$

If μ is sufficiently small, the denominator does not exceed $\frac{3}{2}\gamma(s)\mu|\log\mu|$, and using (10.3) we have

$$\frac{(w-\phi_0)(t)}{(w-\phi_0)(s)} - 1 \geq \frac{2}{3}\left\{ \exp\left(K_1 \int_s^t \left(\int_\sigma^\infty a(\tau)\,d\tau \right) d\sigma \right) - 1 + O\left(\frac{1}{\mu^2|\log\mu|} \int_s^t e^{-K\sigma/\mu}\,d\sigma \right) \right\}$$

(10.24)

$$\geq \frac{2}{3} K_1 \int_s^t \left(\int_\sigma^\infty a(\tau)\,d\tau \right) d\sigma + O\left(\frac{1}{\mu^2|\log\mu|} \int_s^t e^{-K\sigma/\mu}\,d\sigma \right)$$

But we have already remarked that a is bounded below by a negative exponential, so that

$$a(\tau) \geq \alpha e^{-\beta\tau} .$$

Further, for μ sufficiently small,

$$\frac{1}{\mu^2|\log\mu|} e^{-K\sigma/\mu} \leq e^{-\beta\sigma} \quad \text{if} \quad \sigma \geq 2\mu|\log\mu|/K .$$

For this is equivalent to saying that

$$\left(\frac{K}{\mu} - \beta\right)\sigma \geq -\log(\mu^2|\log\mu|) = 2|\log\mu| - \log|\log\mu| ,$$

and this is certainly true even if $\sigma = 2\mu|\log\mu|/K$ and μ is

sufficiently small. Hence, if μ is sufficiently small and K_1 (still at our choice) is fixed large enough, the O-term in (10.24) does not exceed half the first term, and substituting back in (10.23) and noting that $(w-\phi_0)(s)$ is small for s in the relevant range $s \geq 2\mu\,|\log \mu|/K$ and μ sufficiently small, we see that the right-hand side of (10.23) is not less that

$$(10.25) \qquad \frac{1}{6} K_1 \int_s^t \left(\int_\sigma^\infty a(\tau)\,d\tau \right) d\sigma - \frac{2M^2}{m^2} \,|\, \phi_0(t) - \phi_0(s) \,| \quad .$$

Finally, from Theorem 2 (or more precisely Remark 3.1),

$$\begin{aligned} |\phi_0(t) - \phi_0(s)| &= \left| \int_s^t \phi_0'(\sigma)\,d\sigma \right| \\ &\leq \tilde{K} \int_s^t \left(\int_\sigma^\infty a(\tau)\,d\tau \right) d\sigma \quad . \end{aligned}$$

and so, if K_1 is fixed sufficiently large, (10.25) is not less than

$$\frac{1}{12} K_1 \int_s^t \left(\int_\sigma^\infty a(\tau)\,d\tau \right) d\sigma \quad .$$

Substituting this into (10.22) and then into (10.21), we conclude that $I_{12} + I_3 > 0$, and the theorem is proved.

Appendix A : Statement of physical problem and formulation of mathematical model

Molten plastics commonly exhibit large elastic recovery; for example, a filament of a certain polyethylene* ('Melt 1' at 150°C), when elongated at a rate of 1 cm/sec/cm from an initial length of, say, 1 cm to a length of 55 cm and then released, will reach a final length of about 5 cm [15]. Such elastic and other rheological properties are of interest in the processing of plastics and rubber and also as examples of materials with 'memory'

The correct equations for describing the isothermal behaviour of a given molten plastic are not yet known. One set which has been used with some success [8] for 'Melt 1' can be expressed as follows in body tensors:

$$\text{(A1)} \qquad \underline{\pi}(P,t) + p\underline{\gamma}^{-1}(P,t) = -\eta \frac{\partial \underline{\gamma}^{-1}(P,t)}{\partial t} + \int_{-\infty}^{t} a(t-s)\underline{\gamma}^{-1}(P,s)ds \ ;$$

$$\text{(A2)} \qquad \frac{\partial}{\partial s} \det \underline{\gamma}(P,s) = 0 \qquad (-\infty < s \leq t) \ ;$$

$$\text{(A3)} \qquad \underline{\nabla}_t \cdot \underline{\pi}(P,t) = \rho(\underline{\alpha} - \underline{\Xi}) \ ;$$

$$\text{(A4)} \qquad \underline{R}\{\underline{\gamma}(P,s)\} = 0 \qquad (-\infty < s \leq t) \ .$$

(A1) is the so-called 'rubberlike liquid' constitutive equation relating the symmetric contravariant stress tensor $\underline{\pi}(P,t)$ at particle P and time t to the values of the reciprocal $\underline{\gamma}^{-1}$ of the symmetric covariant

*The U. S. production (8×10^9 lb) of polyethylene in 1973 exceeded that of any other one polymer.

metric tensor $\underline{Y}(P,s)$ at times in the interval $-\infty < s \leq t$ (see (A12) below). The material properties are determined by the nonnegative constant η and the nonnegative constants $a_1,\ldots,a_m, \tau_1,\ldots,\tau_m$ in the 'memory function'

$$a(t) = \sum_{r=1}^{m} a_r \exp(-t/\tau_r) \qquad (a_r \geq 0,\ \tau_r > 0)\ . \tag{A5}$$

(See (A 26) and the subsequent explanation; for other possible constitutive equations see [1] and [5].)

If the integral term were omitted, (A1), with the constant volume condition (A2), would describe an incompressible Newtonian liquid of viscosity η and could be used, with the stress equation of motion (A3), to derive the Navier-Stokes equations; there would be no elastic recovery possible. $\underline{\alpha}$ and $\underline{\Xi}$ are contravariant vectors describing acceleration and body force per unit mass, respectively; ρ denotes the density. p is a scalar function of P and t, of the nature of a hydrostatic pressure, introduced in conjunction with the incompressibility condition (A2). $\underline{\nabla}_t$ in (A3) denotes the covariant derivative operator formed with $\underline{Y}(P,t)$, and the dot denotes contraction [8, p. 193]. (A4) expresses the vanishing of the fourth rank Riemann-Christoffel curvature tensor $\underline{R}$ constructed with $\underline{Y}(P,s)$: this expresses the fact that the body manifold is Euclidean at each instant s and so admits a body coordinate system that is instantaneously rectangular Cartesian at s [8, p. 202]. Equations (A1 - 4) are sufficient in number to determine, in principle, the unknown variables $\underline{\pi}$, $\underline{Y}$, and p when the remaining quantities (together with suitable boundary and initial conditions) are given.

When referred to an arbitrary body coordinate system $B : \{P\} \to \mathbb{R}^3$, (A1) and (A2) yield equations

$$\pi^{ij}(\xi,t) + p(\xi,t)\gamma^{ij}(\xi,t) = -\eta \frac{\partial \gamma^{ij}(\xi,t)}{\partial t} + \int_{-\infty}^{t} a(t-s)\gamma^{ij}(\xi,s)ds \,, \tag{A6}$$

$$\frac{\partial}{\partial s}\det \gamma_{ij}(\xi,s) = 0 \qquad (-\infty < s \leq t) \,, \tag{A7}$$

where ξ in this and in similar contexts is short for (ξ^1,ξ^2,ξ^3), the coordinates of P in B; π^{ij}, γ^{ij}, and γ_{ij} denote components of $\underline{\pi}$, $\underline{\gamma}^{-1}$, and $\underline{\gamma}$, respectively, $i,j = 1,2,3$. For an arbitrary B, (A3) yields a complicated set of equations. One can, however, always choose a B that is Cartesian at the instant t at which (A3) applies (i.e., B is such that $\gamma_{ij}(\xi,t)$ is independent of ξ for $i,j = 1,2,3$), and a space coordinate system $C : \{Q\} \to R^3$ that is rectangular Cartesian; (A3) then yields the following three partial differential equations:

$$\sum_{j,k=1}^{3} \frac{\partial \pi^{kj}(\xi,t)}{\partial \xi^k} \frac{\partial f^i(\xi,t)}{\partial \xi^j} = \rho\left\{\frac{\partial^2 f^i(\xi,t)}{\partial t^2} - X^i\right\} \qquad (i = 1,2,3) \,. \tag{A8}$$

The motion of the body is now described by the three equations

$$x^i = f^i(\xi,s) \qquad (i = 1,2,3) \,, \tag{A9}$$

where x^i denote the coordinates in C of the place Q occupied by particle P at time s. X^i denotes the components in C of the external body force per unit mass.

(A4) yields a very complicated set of nonlinear, second order partial differential equations in $\gamma_{ij}(\xi,s)$ whose solution is of the form

(A10) $$\gamma_{ij}(\xi, s) = \sum_{k=1}^{3} \frac{\partial f^k(\xi, s)}{\partial \xi^i} \frac{\partial f^k(\xi, s)}{\partial \xi^j} \qquad (i, j = 1, 2, 3)$$

for arbitrary B and rectangular Cartesian C [3,11]. We may thus use the three functions $f^i(\xi, s)$ in place of the six functions $\gamma_{ij}(\xi, s)$ $(= \gamma_{ji}(\xi, s))$ as unknowns. On using (A10) and the matrix equation

(A11) $$[\gamma^{ij}(\xi, s)] = [\gamma_{ij}(\xi, s)]^{-1},$$

we may express (A6) and (A7) in terms of π^{ij}, f^i, and p: on substituting the resulting expressions for π^{ij} into (A8), we finally obtain three nonlinear, partial-integro-differential equations, which, with the single equation resulting from (A7), yields a set of four equations for the four unknown functions f^i, p; the independent variables are ξ^i, s.

The final equations are nonlinear in f^i, p although the rubberlike liquid constitutive equation (A1) is linear in the tensors $\underline{\pi}$, $\underline{\gamma}^{-1}$. The nonlinearity comes from the constant volume condition (A2) and from the zero-curvature condition (A4) whose solution (A10) is quadratic in the unknown functions f^i. The nonlinearity arising from the products in the left-hand side of (A8) can be removed trivially by choosing B to coincide with C at time t, so that $f^i(\xi, t) = \xi^i$ and $\partial f^i/\partial \xi^j = \delta_{ij}$ at time t.

A very considerable simplification of the above equations is obtained for flow histories which are homogeneous (or uniform) under conditions in which the inertial and body force terms on the right-hand side of (A3) can be neglected. Such histories are of little or no interest in classical

hydrodynamics (where the constitutive equations are given by (A1) with $a(t) \equiv 0$) but are of fundamental importance in polymer rheology where highly viscous molten plastics can be subjected to uniform elongation in filament form or to two-way stretching in sheet form; results of carefully controlled experiments of this type can be used to test the applicability of constitutive equations such as (A1). A flow history is homogeneous if, for any two instants s, t, we have $\underset{t}{\nabla} \underline{\gamma}(P,s) = 0$ [8, p. 247], from which it can be shown that $\underset{t}{\nabla} \underline{\gamma}^{-1}(P,s) = 0$; since $\underset{t}{\nabla}$ commutes with the operators $\partial/\partial t$ and $\int a(t-s)\ldots ds$, it follows from (A1) (taking p to be independent of P, as a trial solution) that $\underset{t}{\nabla} \underline{\pi}(P,t) = 0$ (showing that the stress is homogeneous) and hence also that (A3) (with the right-hand side zero) is satisfied. It also follows from the above definition of a homogeneous flow history that a body coordinate system $B : P \to \xi$ exists that is Cartesian in every state, i.e., which is such that $\gamma_{ij}(\xi,s)$ is independent of ξ for all s [8, p. 247], and hence (A4) is satisfied. The behaviour in homogeneous flow histories with inertial and body forces neglected is thus governed by (A6) and (A7) with ξ absent, and there is no longer a need to introduce a space coordinate system or the functions f^i: one can instead use γ_{ij} or γ^{ij} as the unknowns, for example, in the case of problems involving the calculation of free elastic recovery: in such problems, the flow history (and hence $\gamma_{ij}(s)$) would be specified throughout some interval $-\infty < s < t_1$, say:

for $t_1 < s < t$, the stress would be zero, and the elastic recovery would be determined by solving the set (A6) (with $\pi^{ij} = 0$) and (A7) for $p(t)$ and $\gamma_{ij}(s)$ $(t_1 < s < t)$. These equations thus form a simultaneous system of nonlinear Volterra integrodifferential equations, in which the nonlinearity arises from the incompressibility condition.

In this paper, we consider the particular case of the above in which the specified flow history is one of simple elongation (at constant volume). The variable p can be eliminated, and the recovery behaviour is then governed by a single nonlinear Volterra equation, which we now derive.

In any B, the separation P_0P at time t between any two neighbouring particles P_0, P is given by the equation

$$(P_0P)_t^2 = \sum_i \sum_j \gamma_{ij}(\xi, t)\,\delta\xi^i\,\delta\xi^j \tag{A12}$$

where ξ^i and $\xi^i + \delta\xi^i$ are the coordinates in B of P_0 and P. For any two times s, t, there are three material lines through any given P_0 that are mutually orthogonal at s and at t; in the strain $s \to t$, infinitesimal material line elements tangential to these three material lines at P_0 change in length by factors $\lambda_i(P_0, s, t)$ which are given by the positive roots in λ of the equation

$$\det\{\gamma_{ij}(\xi, t) - \lambda^2 \gamma_{ij}(\xi, s)\} = 0\,. \tag{A13}$$

The factors λ_i are called 'principal elongation ratios'. A flow is 'shear free' if there exists a body coordinate system B that is always

orthogonal, i.e., such that $\gamma_{ij} = 0$ when $i \neq j$ [8, p. 81]. For a shear-free flow, the principal elongation ratios are given by the length changes of the coordinate lines, and the roots of (A13) are given by

$$\lambda_i^2 \gamma_{ii}(\xi, s) = \gamma_{ii}(\xi, t) \qquad (i = 1, 2, 3) . \tag{A14}$$

For a shear-free flow that is homogeneous, B is always Cartesian, i.e., the γ_{ii} are independent of ξ*. A shear-free flow is a 'simple elongation' if two principal elongation ratios, λ_2 and λ_3 say, are always equal; the ξ^1-coordinate lines are then called directions of elongation. The constant volume condition (A7) reduces to the equation $\lambda_1\lambda_2\lambda_3 = 1$ and hence, for simple elongation at constant volume, we have

$$\lambda_2 = \lambda_3 = \lambda_1^{-\frac{1}{2}} \tag{A15}$$

We now consider the following problem:

$-\infty < s \leq -t_0$ Zero stress; no flow; B rectangular Cartesian; hence

$$\pi^{ij} = 0, \quad \gamma_{ij} = \gamma^{ij} = \delta_{ij} \tag{A16}$$

$-t_0 < s \leq 0$ Homogeneous simple elongation at constant volume and constant rate κ, i.e.,

$$\frac{d\lambda_1(-t_0, s)}{ds} = \kappa\lambda_1(-t_0, s), \quad \lambda_2 = \lambda_3 = \lambda_1^{-\frac{1}{2}} . \tag{A17}$$

$0 < s \leq t$ Zero stress; free elastic recovery:

$$\pi^{ij} = 0 . \tag{A18}$$

---

*And hence λ_i are independent of P_0.

We wish to calculate $\gamma_{ij}(\xi, s)$ for $0 < s \leq t$. As a trial solution, it is reasonable to suppose that the elastic recovery will involve a homogeneous simple elongation at constant volume with the ξ^1-coordinate lines again as directions of elongation (or contraction). For convenience, we write

$$y(s) = \lambda_1(-t_0, s) . \tag{A19}$$

Since the entire flow history is a homogeneous simple elongation at constant volume, we have, from (A11), (A14), (A15), (A16), and (A19),

$$\left.\begin{aligned} \gamma^{11}(\xi, s) = y^{-2}(s),\ \gamma^{22}(\xi, s) = \gamma^{33}(\xi, s) = y(s), \\ \gamma^{ij} = 0 \quad (i \neq j) . \end{aligned}\right\} \quad (-\infty < s \leq t) \tag{A20}$$

From (A16) and (A17), we have

$$y(s) = \begin{cases} 1 & (-\infty < s \leq -t_0) , \\ \exp\{\kappa(s + t_0)\} & (-t_0 < s \leq 0) . \end{cases} \tag{A21}$$

$y(s)$ is to be calculated for $0 < s \leq t$ so as to satisfy (A6) and (A7).

(A7) is satisfied by (A20). Using (A20) and (A18), the six equations (A6) for $t > 0$ reduce to the following two:

$$(i = j = 1) \quad p(t)y^{-2}(t) = -\eta \frac{dy^{-2}(t)}{dt} + \int_{-\infty}^{t} a(t - s)y^{-2}(s)ds , \tag{A22}$$

$$(i = j = 2, 3) \quad p(t)y(t) = -\eta \frac{dy(t)}{dt} + \int_{-\infty}^{t} a(t - s)y(s)ds . \tag{A23}$$

The unknown function $p(t)$ may be eliminated by multiplying (A22) by $y^3(t)$ and then subtracting (A23). The resulting equation may be written in the form

$$-\mu \frac{dy(t)}{dt} = \int_{-\infty}^{t} a(t-s)F(y(t),y(s))ds \qquad (t > 0)\,, \tag{A24}$$

where

$$\mu = 3\eta,\ F(y,z) = (y^3/z^2) - z\,. \tag{A25}$$

These are the equations used in the text above. Finally, we add some brief remarks about the physical basis and applicability of the rubberlike liquid constitutive equation (A1) which has been discussed elsewhere [8, pp. 143, 223-236].

(A1) has been derived from two different molecular theories: the 'bead-spring' theory of Rouse and Zimm for very dilute solutions of deformable long molecules in an incompressible Newtonian solvent of viscosity η, and the network theory of Green and Tobolsky, Yamamoto, and Lodge which is developed for concentrated polymer solutions and undiluted or molten polymers. It is curious that two different molecular theories should yield constitutive equations of the same form, but the reason for this is known: the differences between the two sets of equations at the molecular level do not survive the averaging process used to go from the molecular quantities to the macroscopic quantities $\underline{\pi}$ and $\underline{Y}$ which appear in the constitutive equation (A1) [9]. The memory function constants a_r, τ_r are specified in terms of three unknown constants by the bead-spring theory but are not specified by the network theory, which also leaves η unspecified.

According to the network theory, the integral term in (A1) arises from the thermal motion of a network composed of long, deformable polymer molecules temporarily linked together at a few points called entanglements or temporary junctions which are assumed to be created and lost at constant rates which are unaffected by the flow history. The concentration $N(t)dt$ of network strands which were created in the interval $(0, dt)$ and are still in the network at time t (a 'strand' being that part of a polymer molecule lying between two consecutive junctions) is given by an equation of the form

$$N(t) = \sum_{r=1}^{m} C_r \exp(-t/\tau_r) \tag{A26}$$

where, for simplicity, it has been assumed that the set of all strands can be sorted into m subsets, labelled $1, 2, \ldots, m$, such that, in the r^{th} subset, all strands were created at the same rate C_r (per millilitre) and have the same probability $1/\tau_r$ per second of leaving the network [10]. The memory function in (A1) is given by the equation $a(t) = kTN(t)$, where k is Boltzmann's constant and T is the absolute temperature. Thus $a_r = kTC_r > 0$, and (A5) is proved.

According to the network theory, then, it follows that $a(t) > 0$, because there is always a nonzero concentration of strands of age t, and that $a'(t) < 0$ because strands of age $t(> 0)$ can only be lost and not created: strands are created with age 0 only. It also follows from (A5) that

$$(A27)\qquad \left.\begin{array}{l} (-1)^k a^{(k)}(t) > 0 \\ \text{and } a^{(k+1)}(t)/a^{(k)}(t) \text{ is nondecreasing} \end{array}\right\} (k = 0,1,2,\ldots) .$$

$a^{(k)}(t)$ denotes the k^{th} derivative of $a(t)$. (A27) represents the properties of $a(t)$ some of which are used in the present analysis.

The constitutive equation obtained by putting $\eta = 0$ in (A1) leads to the 'reduced' equation (1.6) and has been tested for 'Melt 1' by comparing the predictions with results of a series of experiments performed by Meissner [13,15]. The constants in (A5), with $m = 5$, were chosen to fit stress growth data in simple elongation at low rates. (A1) (with $\eta = 0$) then gave good agreement with stress growth data in simple elongation at higher rates, with elastic recovery data following elongation and following shear, and with stress growth data in shear flow [8, pp. 225-231; 2; 4], provided that the total strain from rest was limited to moderate values; at higher strains, there was serious disagreement between theory and data: the predicted stresses and the predicted recoveries were greater than the observed. The present analysis of the elongational recovery problem shows that inclusion of the term in η leads to a reduction in the predicted recovery, which is in the right direction to give better agreement with experiment. The term in η has been added to represent the possible effect of the presence of a viscous solvent (in the case of a concentrated polymer solution) or of polymer of low molecular weight (in the case of a molten polymer). The possible effects of such a term are also of some

interest in connection with certain 'fast-strain' tests of the Gaussian network hypothesis which have recently been proposed as a possible method of testing certain of the network theory assumptions when separated from the others [8, pp. 231-236; 12]. It is recognized, however, that other modifications to (A1) are required if better agreement between all the predictions and data referred to above is to be obtained [14]. It should, perhaps, be added that the homogeneous elongation with neglect of inertial and body forces treated in the present analysis represents a reasonable idealization of the conditions of Meissner's elongation experiments on 'Melt 1': a long filament of high viscosity (5×10^5 poise) was floated on a bath of an inert oil, and the homogeneity of elongation was always checked by weighing samples into which the filament was cut after elongation: the variation of elongation ratio along the filament was about 3% or less [15].

Appendix B : Upper and lower solutions

Our purpose is to collect the results on integrodifferential inequalities needed in the proof of Theorem 8. These are rather similar to classical results of this type for ordinary differential equations and Volterra equations developed e.g. in [6] and [19]. However, our situation is sufficiently different that it is simpler to give an independent short exposition of what is used, rather than to apply more general known results.

In what follows let $D_- u(t)$ denote the lower left-hand Dini derivative, $D^- u(t)$ the upper left-hand Dini derivative, $D_+ u(t)$ the lower right-hand Dini derivative, and $D^+ u(t)$ the upper right-hand Dini derivative of a continuous function u. When $D^+ u(t) = D_+ u(t)$, we denote this common value by $u'_+(t)$, the right-hand derivative of u, and similarly $u'_-(t)$ denotes the left-hand derivative of u.

It will be convenient to write the initial value problem (1.1), (1.2) in the form

$$\text{(B1)} \qquad \begin{cases} -\mu y'(t) = \int_0^t F(y(t), y(s))ds + f_a(y)(t) \\ y(0) = g(0) \qquad (\mu > 0;\ 0 < t < \infty), \end{cases}$$

where

$$\text{(B2)} \qquad f_a(y)(t) = \int_{-\infty}^0 a(t - s)F(y(t), g(s))ds .$$

Throughout this appendix we shall assume

(B3) $a(t) \geq 0$, $a \in L^1(0, \infty)$, F satisfies (H_F), g satisfies (H_g).

The basic result needed is (compare Theorem 1.2.1 of [6]):

Proposition 1B. Let the assumptions (B3) be satisfied. Let $v, w \in C([0,\gamma); \mathbb{R})$, $\gamma > 0$, be given functions satisfying the following properties:

(B4) $$v(0) < w(0) ,$$

(B5) $$\begin{cases} -\mu D_- v(t) \geq \int_0^t a(t-s)F(v(t), v(s))ds + f_a(v)(t) \quad (0 < t < \gamma) , \\ -\mu D_- w(t) < \int_0^t a(t-s)F(w(t), w(s))ds + f_a(w)(t) \quad (0 < t < \gamma) . \end{cases}$$

Then $v(t) < w(t) \quad (0 \leq t < \gamma)$.

Proof. Define the set $Z = \{ t \in [0,\gamma) : w(t) \leq v(t)\}$. If Proposition 1B is false the set $Z \neq \phi$: let $t_1 = \inf Z$. By (B4), $t_1 > 0$ and

(B6) $$v(t_1) = w(t_1), \; v(t) < w(t) \quad (0 \leq t < t_1) .$$

Taking $h < 0$, $|h|$ small, one has $v(t_1 + h) < w(t_1 + h)$ and

$$\frac{v(t_1 + h) - v(t_1)}{h} > \frac{w(t_1 + h) - w(t_1)}{h} .$$

Taking the limit inferior as $h \to 0^-$ this implies

$$D_- v(t_1) \geq D_- w(t_1) ,$$

and therefore,

$$-\mu D_- v(t_1) \leq -\mu D_- w(t_1) \qquad (\mu > 0) .$$

Applying this and $v(t_1) = w(t_1)$ in inequalities (B5) yields the inequality

(B7) $$\int_0^{t_1} a(t_1 - s)F(v(t_1), v(s)ds < \int_0^{t_1} a(t_1 - s)F(w(t_1), w(s))ds .$$

On the other hand, the definition of t_1 and (B6), together with the assumption $\Gamma_2 < 0$ in (H_Γ), implies that

$$\Gamma(v(t_1), v(s)) > \Gamma(w(t_1), w(s)) \qquad (0 \le s < t_1) ,$$

so that, since $a(t) \ge 0$,

$$\text{(B8)} \qquad \int_0^{t_1} a(t_1 - s)\Gamma(v(t_1), v(s))ds > \int_0^{t_1} a(t_1 - s)\Gamma(w(t_1), w(s))ds .$$

Thus (B8) contradicts (B7) and the set $Z = \phi$. This proves Proposition 1B.

Definition. We shall say that w is an upper solution of the initial value problem (B1) on $0 < t < \gamma$ if and only if $w \in C([0, \gamma); R)$, $w(0) > g(0)$, $w'_+(t)$ exists on $(0, \gamma)$ and

$$-\mu w'_+(t) < \int_0^t a(t - s)\Gamma(w(t), w(s))ds + f_a(w)(t) \qquad (0 < t < \gamma) .$$

A similar definition holds for a lower solution with the inequalities reversed. We remark that as a consequence of Theorem 3 the solution ϕ_0 of the reduced equation (1.6) (recall that it was proved in Theorem 3 that $\phi'_0(t) < 0$ $(0 < t < \infty)$) is a lower solution of (1.1) (or (B1)) on $0 < t < \infty$.

The main result for the application in Theorem 8 is:

Proposition 2B. Let the assumptions (B3) be satisfied. Let ϕ be the solution, let w be an upper solution, and let v be a lower solution of (B1) on $0 < t < \gamma$. Then

$$\text{(B9)} \qquad v(t) < \phi(t) < w(t) \qquad (0 \le t < \gamma) .$$

In Theorem 8 one takes $v(t) = \phi_0(t)$, where ϕ_0 is the solution of (1.6), and one shows that w given by (10.1) is an upper solution.

Proof of Proposition 2B. We shall prove the second inequality in (B9); the first is proved in a similar way. Since $w(0) > g(0)$ the result follows directly from Proposition 1B with v replaced by ϕ, $D_- v$ by ϕ', and $D_- w$ by $w'_+(t)$ (and one uses Lemma 1.2.2 of [6]).

REFERENCES

1. B. Bernstein, E. A. Kearsley, and L. J. Zapas, A Study of Stress Relaxation with Finite Strain. Trans. Soc. Rheol. 7 (1963), 391-410.
2. Hui Chang, Elongational Flow and Spinnability of Viscoelastic Fluids. Ph.D. Dissertation, Dept. of Engineering Mechanics, Univ. of Wisconsin-Madison (1973).
3. L. P. Eisenhart, Riemannian Geometry. Princeton Univ. Press, 1926.
4. T. A. Huang, Time-dependent First Normal Stress Difference and Shear Stress Generated by Polymer Melts in Steady Shear Flow. Ph.D. Dissertation, Dept. of Engineering Mechanics, Univ. of Wisconsin-Madison (1976).
5. A. Kaye, Non-Newtonian Flow in Incompressible Fluids. CoA Note No. 134, The College of Aeronautics, Cranfield, Bletchley, England (1962).
6. V. Lakshmikantham and S. Leela, Differential and Integral Inequalities. Vol. I. Academic Press, 1969.
7. J. J. Levin and J. A. Nohel, Perturbations of a Nonlinear Volterra Equation. Mich. Math. J. 12 (1965), 431-447.
8. A. S. Lodge, Body Tensor Fields in Continuum Mechanics. Academic Press, 1974.
9. A. S. Lodge, Concentrated Polymer Solutions. Proc. Fifth Int. Congr. Rheol. 4, 169-178 (University of Tokyo Press, 1970).
10. A. S. Lodge, Constitutive Equations from Molecular Network Theories for Polymer Solutions. Rheol. Acta 7 (1968), 379-392.

11 A. S. Lodge, The Compatibility Conditions for Large Strains. Quart. J. Mech. Appl. Math. 4 (1951), 85-93.

12. A. S. Lodge and J. Meissner, On the Use of Instantaneous Strains, Superposed on Shear and Elongational Flows of Polymeric Liquids, to Test the Gaussian Network Hypothesis and to Estimate the Segment Concentration and its Variation during Flow. Rheol. Acta 11 (1972), 351-352.

13. A. S. Lodge and J. Meissner, Comparison of Network Theory Predictions with Stress/Time Data in Shear and Elongation for a Low-density Polyethylene Melt. Rheol. Acta 12 (1973), 41-47.

14. G. Marrucci and D. Acierno, Non-affine Deformation in Impermanent Networks of Polymer Chains. Proc. Seventh Int. Congr. Rheol. 538-539 (Swedish Soc. Rheol., Chalmers Univ. of Techn., Gothenburg, Sweden, 1976)

15. J. Meissner, Dehnungsverhalten von Polyäthylen-Schmelzen. Rheol. Acta 10 (1971), 230-242.

16. R. K. Miller, Nonlinear Volterra Integral Equations. W. A. Benjamin, Inc., 1971.

17. J. A. Nohel, Some Problems in Nonlinear Volterra Integral Equations. Bull. Amer. Math. Soc. 68 (1962), 323-329.

18. J. A. Nohel and D. F. Shea, Frequency Domain Methods for Volterra Equations. Advances in Math. 22(1976), 278-304.

19. W. Walter, Differential and Integral Inequalities. Springer-Verlag, 1970.

20. D. V. Widder, The Laplace Transform . Princeton Univ. Press, 1941.

VIBRATIONS OF A HELICOPTER ROTOR BLADE USING FINITE ELEMENT-UNCONSTRAINED VARIATIONAL FORMULATIONS

J. J. Wu and C. N. Shen
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. In the past several years, a numerical method has been developed which is a generalized Rayleigh-Ritz - finite element discretization using the combined concept of Lagrange multipliers and adjoint variables. This approach enables one to deal with problems associated with nonconservative forces, coupling effects and all types of boundary conditions in a routine fashion; and it appears promising in solving the vibration and dynamic stability problems associated with the complicated equations of a helicopter rotor blade. This paper presents the first application of the general method to the vibration problem of such a rotor blade.

The basic differential equations in this paper are taken from the linear, but fully coupled set developed by Houbolt and Brooks in 1956. These equations are further reduced to a simplest possible case and yet still containing the coupling of flap and root torsion modes. An unconstrained, adjoint variational statement has been established which is both the necessary and sufficient condition for the coupled differential equations and some general, but physical meaningful boundary conditions. The finite element matrix equations are then derived from this variational statement illustrating the way that coupling terms could be handled in general.

The numerical results from some demonstrative examples show that instability of flutter can occur in the range of operational rotor speed due to the coupled motion of flapping and root torsion without any aerodynamic force, if the torsional spring (or the pitch control link) is not sufficiently stiff. This instability does not appear to have been reported previously.

1. INTRODUCTION. An analytical investigation on vibrations and dynamic stability of helicopter rotor blades usually consists of two phases (1) the derivation of the governing differential equations to include parameters and variables considered physically important, and (2) the formulation of solutions for the equations derived and data interpretations. This paper deals with the second phase of such an investigation.

Due to the slenderness of a helicopter rotor blades, its aerodynamic cross-section and the requirements on the craft's maneuverability, there are a large number of interacting parameters and the resulting differential equations are, as a rule, nonlinear, coupled in terms of field variables. In addition, the aerodynamic forces, coriolis forces due to rotation of the blade are nonconservative in nature and the effects due to structural damping and various boundary conditions must be evaluated.

Considerable attention has been given recently to the derivation of a consistent set of nonlinear differential equations together with aerodynamic forces. This fact is amply demonstrated by the work of Friedman and Tong [1] and that of Hodges and Ormiston [2]. In both references [1] and [2], brief reviews on earlier work on helicopter blade equations can be found. As for obtaining solutions to these equations there does not appear to exist a general and efficient method to deal with the difficulties associated with nonlinearities, nonconservative forces, coupling terms, various boundary conditions, damping effects and periodic excitations. For example, Miller and Ellis [3], Ham [4] and Friedman and Tong [1] have obtained approximate solution by including in their solution formulation only the lowest modes of vibrations, Hodges and Ormiston [2,5] employed Galerkin technique in their numerical examples. One of the disadvantages of this approach is its inability to handle general boundary conditions.

Using the combined concept of adjoint variable and Lagrange multipliers, variational statements can be established for a wide range of linear problems with nonconservative forces and very versatile boundary conditions [6]. Thus a generalized Rayleigh-Ritz approximation scheme can be established for the obtaining of solutions of these otherwise difficult-to-solve problems. In conjunction with finite element discretization, this approach has been amply demonstrated in such applications as nonconservative stability, damping effects and very general boundary conditions [7,8,9]. In the present study, the solution formulation is limited to the blade vibration considering only the coupling of flapping and root torsion. Although the present method can conceivably be extended for solutions of nonlinear problems, it is desirable first to have a thorough understanding to solutions of linear problems. For this purpose, one can go back two decades and use the equations consistently derived by Houbolt and Brooks in 1956 [10]. The original set of equations was derived for elastic distributed torsion. It can be adapted to model root torsion if proper boundary conditions are introduced. This is shown in Sections 1 and 2. The physical justification for emphasizing root torsion over the distributed torsion was due to the pitch control link at the inboard end of the blade and was used by Miller and Ellis [3] and again by Ham [4]. It should be clear that, in the present formulation, to include distributed torsion is simply a matter of increasing the number of degrees-of-freedom of the discrete system. The basis of the solution formulation and the technique of handling the coupling terms are given in Sections 3 and 4. Finally, the numerical results obtained indicate that "flutter instability" can occur simply due to the coupling effect considered without any aerodynamic loads.

2. STATEMENT OF THE PROBLEM - DIFFERENTIAL EQUATIONS. As a first step to demonstrate the application of the unconstrained variational - finite element formulation to helicopter rotors, the vibration of a rotor blade considering the coupling of flap and root torsion modes of motion is analyzed (Figure 1). For this purpose, the linear set of equations, derived by Houbolt and Brooks, including the coupling flap and distributed torsion is rewritten here [10]:

$$(EIw'' - Te_A\phi)'' - (Tw')' - (\Omega^2 mxe\phi)' + m(\ddot{w} + e\ddot{\phi}) = L_z, \qquad (1)$$

$$- [(GJ + Tk_A^2)\phi']' - Te_A w'' + \Omega^2 mxew'$$

$$+ \Omega^2 m(k_{m2}^2 - k_{m1}^2 + ee_o)\phi + mk_m^2 \ddot{\phi} + me\ddot{w} = M \qquad (2)$$

where $w = w(x,t)$ and $\phi = \phi(x,t)$ denote the flapping deflection and distributed torsion of the rotor blade respectively. A prime (') denotes differentiation with respect to x, the coordinate along the blade elastic axis and a dot ($\cdot$), differentiation with respect to the time t. Other symbols are defined in the following and are consistent with the notation in reference [10].

EI = flexural rigidity of the cross section.

GJ = torsional rigidity of the cross section.

ℓ = length of the blade.

e = chordwise distance between elastic axis (E.A.) and centre of gravity (C.G.) of a cross-section, positive if C.G. is ahead of E.A.

e_A = chordwise distance between E.A. and centre of tensile area (C.T.) of a cross section, positive if C.T. is ahead of E.A.

e_o = chordwise distance at the root between E.A. and the axis about which the blade is rotating, positive if E.A. is ahead.

k_A = polar radius of a gyration of the tensile area of a cross-section w.r.t. E.A. in a cross-section.

k_{m1} = radius of gyration of the total area of a cross-section about the major neutral axis (axis 1-1 in Figure 2).

k_{m2} = radius of gyration of the total area of a cross-section about an axis perpendicular to the major neutral axis and through E.A. (axis m2-m2 in Figure 2).

k_m = polar radius of gyration of the total area of cross-section about E.A. ($k_m^2 = k_{m1}^2 + k_{m2}^2$).

Ω = blade angular velocity.

$T = T(x) = \int_x^{\ell} m\Omega^2 x dx$, which is the tensile force at location x.

L_z = aerodynamic lift per unit length of the blade.

M = aerodynamic torque per unit length of the blade.

In reference [10], a more general set of linear equations with fully coupled flap, lag and distributed torsion modes of motion were derived. The coupling terms have been shown to be due to the noncoincidence of the elastic centre, centre of gravity and tension centre, the centrifugal force and the built-in angle of twist. For a rotor blade without the built-in twist angle, the lag mode of motion is uncoupled from the general set of equations and the remaining coupled equations are Eqs. (1) and (2) considered here.

Since only the "free vibration" of the rotor blade with a coupling between flexural and root torsion is considered in this paper, the terms due to aerodynamic forces are set to zero and the torsional displacement is only a function of time and not a function of x. Thus

$$M = L_z = 0, \quad \phi = \phi(t) \tag{3}$$

and Eqs. (1) and (2) reduce to the following:

$$EIw'''' - (Tw')' + m\ddot{w} - \Omega^2 me\phi + me\ddot{\phi} = 0 \tag{4}$$

$$\Omega^2 m(k_{m2}^2 - k_{m1}^2 + ee_o)\phi + mk_m^2 \ddot{\phi} - Te_A w'' + m\Omega^2 exw' + me\ddot{w} = 0 \tag{5}$$

In Eqs. (4) and (5), it is also assumed that the blade has constant E, I and m throughout its length. To simplfy solution formulations as much as possible, Eqs. (4) and (5) will be transformed into dimensionless forms and appropriate dimensionless parameters will be introduced. This process will also facilitate parametric studies.

Let

$$\bar{w} = \frac{w}{\ell}, \quad \bar{\phi} = \phi$$

$$\bar{x} = \frac{x}{\ell}, \quad \bar{t} = \frac{t}{c} \tag{6}$$

where the constant c has a real time dimension and will be defined later in Eqs. (12). Eqs. (4) and (5) become

$$\frac{EI\ell}{\ell^4}\frac{\partial^4 \bar{w}}{\partial \bar{x}^4} - \frac{\ell}{\ell^4}\frac{\partial}{\partial \bar{x}}\left(T\frac{\partial \bar{w}}{\partial \bar{x}}\right) + \frac{m\ell}{c^2}\frac{\partial^2 \bar{w}}{\partial \bar{t}^2} - m\Omega^2 e\bar{\phi} + \frac{me}{c^2}\frac{\partial^2 \bar{\phi}}{\partial \bar{t}^2} = 0 \tag{7}$$

and

$$m\Omega^2 k^2 \bar{\phi} + \frac{mk_m^2}{c^2}\frac{\partial^2 \bar{\phi}}{\partial \bar{t}^2} - \frac{Te_A\ell}{\ell^2}\frac{\partial^2 \bar{w}}{\partial \bar{x}^2} + \frac{m\Omega^2 e\ell^2}{\ell} x \frac{\partial \bar{w}}{\partial \bar{x}} + \frac{me\ell}{c^2}\frac{\partial^2 \bar{w}}{\partial \bar{t}^2} = 0 \tag{8}$$

where

$$k^2 = k_{m2}^2 - k_{m1}^2 + ee_o \tag{9}$$

Multiplying Eq. (7) by $\frac{\ell^3}{EI}$, one has

$$\frac{\partial^4 \bar{w}}{\partial \bar{x}^4} - \frac{\partial}{\partial \bar{x}} \left(\frac{T\ell^2}{EI} \frac{\partial \bar{w}}{\partial \bar{x}}\right) + \frac{m\ell^4}{EI} \frac{1}{c^2} \frac{\partial^2 \bar{w}}{\partial \bar{t}^2} - \frac{m\ell^4}{EI} \frac{e}{\ell} \frac{1}{c^2} c^2 \Omega^2 \bar{\phi}$$

$$+ \frac{m\ell^4}{EI} \frac{1}{c^2} \frac{e}{\ell} \frac{\partial^2 \bar{\phi}}{\partial \bar{t}^2} = 0 \tag{10}$$

Multiplying Eq. (8) by $\frac{c^2}{m\ell^2}$, one has

$$c^2\Omega^2 \frac{k^2}{\ell^2} \bar{\phi} + \frac{k^2 m}{\ell^2} \frac{\partial^2 \bar{\phi}}{\partial \bar{t}^2} - \frac{c^2 T}{m\ell^2} \frac{e_A}{\ell} \frac{\partial^2 \bar{w}}{\partial \bar{x}^2} + c^2\Omega^2 \frac{e}{\ell} \bar{x} \frac{\partial \bar{w}}{\partial \bar{x}} + \frac{e}{\ell} \frac{\partial^2 \bar{w}}{\partial \bar{t}^2} = 0 \tag{11}$$

Now let

$$\left.\begin{aligned} c^2 &= \frac{m\ell^4}{EI} \\ \bar{e} = \frac{e}{\ell}\ ,\ \bar{e}_A &= \frac{e_A}{\ell}\ ,\ \bar{e}_o = \frac{e_o}{\ell} \\ \bar{k} = \frac{k}{\ell}\ ,\ \bar{k}_m &= \frac{k_m}{\ell} \\ \bar{\Omega} &= c\Omega \\ \bar{T}(\bar{x}) = \frac{T(x)\ell^2}{EI} &= \frac{1}{2} \bar{\Omega}^2 (1-\bar{x}^2) \end{aligned}\right\} \tag{12}$$

Eqs. (10) and (11) then become

$$\frac{\partial^4 \bar{w}}{\partial \bar{x}^4} - \frac{\partial}{\partial \bar{x}} \left(\bar{T}(\bar{x}) \frac{\partial \bar{w}}{\partial \bar{x}}\right) + \frac{\partial^2 \bar{w}}{\partial \bar{t}^2} - \bar{\Omega}^2 \bar{e} \bar{\phi} + \bar{e} \frac{\partial^2 \bar{\phi}}{\partial \bar{t}^2} = 0 \tag{13}$$

and

$$\bar{\Omega}^2\bar{k}^2\bar{\phi} + k_m^2 \frac{\partial^2\bar{\phi}}{\partial\bar{t}^2} - \bar{T}(\bar{x})\,\bar{e}_A \frac{\partial^2\bar{w}}{\partial\bar{x}^2} + \bar{\Omega}^2\bar{e}\bar{x} \frac{\partial\bar{w}}{\partial\bar{x}} + \bar{e} \frac{\partial^2\bar{w}}{\partial^2\bar{t}^2} = 0 \tag{14}$$

With all quantities in Eqs. (13) and (14) in dimensionless forms, one can omit the bars altogether and write:

$$w'''' - (Tw')' + \ddot{w} - \Omega^2 e\phi + e\ddot{\phi} = 0 \tag{15}$$

and

$$\Omega^2 k^2\phi + k_m^2 \ddot{\phi} - Te_A w'' + \Omega^2 exw' + e\ddot{w} = 0 \tag{16}$$

Furthermore, it is assumed that

$$w(x,t) = w(x)e^{\lambda t} \tag{17}$$

$$\phi(t) = \phi e^{\lambda t}$$

Thus the final set of equations upon which the present solution formulations are based, is the following:

$$w'''' - (Tw')' + \lambda^2 w - \Omega^2 e\phi + \lambda^2 e\phi = 0 \tag{18}$$

and

$$\Omega^2 k^2\phi + \lambda^2 k_m^2 \phi - Te_A w'' + \Omega^2 exw' + \lambda^2 ew = 0 \tag{19}$$

3. AN UNCONSTRAINED VARIATIONAL STATEMENT AND BOUNDARY CONDITIONS. Some of the unique features of the unconventional variational formulation are that all the boundary conditions are natural boundary conditions and that the set of the differential equations, together with all the boundary conditions, is the direct consequence of a variational statement and vise versa. The construction of such a variational statement is simply a process of integration-by-parts from a bilinear functional of the original differential equations multiplied by the variation of the adjoint field variable, into some other bilinear functional with lowest possible derivations of both the original field variable and the variations of the adjoint variable. With a proper choice of the generalized Lagrange multipliers, any physical meaningful boundary conditions consistent with the physical meaning of the given differential equations themselves can be resulted from the variational statement as natural boundary conditions. This general process was treated elsewhere [7] and will not be repeated here. Presently, an unconstrained variational statement will be given which leads to the original differential equations and a set of a very general boundary conditions.

Let us consider the variational statement

$$\delta I = 0 \tag{20a}$$

with

$$I = I_1 + I_2 \tag{20b}$$

$$I_1 = \int_0^1 [w''w^{*''} + Tw'w^{*'} + \lambda^2 ww^* + e(\lambda^2 - \Omega^2)\phi w^*]dx$$

$$+ k_1 w(0)w^*(0) + k_2 w'(0)w^{*'}(0) \tag{20c}$$

and

$$I_2 = \int_0^1 [(\Omega^2 k^2 + \lambda^2 k_m^2)\phi\phi^* + e_A T'w'\phi^* + e(\Omega^2 xw'\phi^* + \lambda^2 w\phi^*)]dx + k_3\phi\phi^* \tag{20d}$$

where a star(*) denotes the adjoint variables and the variations are totally unconstrained. The Lagrange multipliers k_1, k_2 and k_3 are the spring constants, in dimensionless form, for deflection, bending and torsion at the hub ($x = 0$) respectively. To show that the unconstrained variational statement of Eqs. (20) leads to the original differential equations and the necessary boundary conditions, one considers

$$(\delta I)_{w,\phi} = 0 \tag{21a}$$

where $(\delta I)_{w,\phi}$ means taking the variation of I with w,ϕ not varied. Thus

$$(\delta I)_{w,\phi} = (\delta I_1)_{w,\phi} + (\delta I_2)_{w,\phi} = 0 \tag{21b}$$

with

$$(\delta I_1)_{w,\phi} = \int_0^1 [w''\delta w^{*''} + Tw'\delta w^{*'} + \lambda^2 w\delta w^* + e(\lambda^2 - \Omega^2)\phi\delta w^*]dx$$

$$+ k_1 w(0)\delta w^*(0) + k_2 w'(0)\delta w^{*'}(0) \tag{21c}$$

and

$$(\delta I_2)_{w,\phi} = \int_0^1 [(\Omega^2 k^2 + \lambda^2 k_m^2)\phi\delta\phi^* + e_A T'w'\delta\phi^*$$

$$+ e(\Omega^2 xw' + \lambda^2 w)\delta\phi^*]dx + k_3\phi\delta\phi^* \tag{21d}$$

Performing integration-by-parts, one arrives at

$$(\delta I_1)_{w,\phi} = \int_0^1 [w'''' - (Tw')' + \lambda^2 w + e(\lambda^2 - \Omega^2)\phi]\delta w^* dx$$

$$+ w''(1)\delta w^{*\prime}(1) - [w'''(1) - T(1)w'(1)]\delta w^*(1)$$

$$- [w''(0) - k_2 w'(0)]\delta w^{*\prime}(0)$$

$$+ [w'''(0) - T(0)w'(0) + k_1 w(0)]\delta w^*(0) \qquad (22a)$$

and

$$(\delta I_2)_{w,\phi} = \int_0^1 [(\Omega^2 k^2 + \lambda^2 k_m^2)\phi - e_A T w'' + e(\Omega^2 x w' + \lambda^2 w)]\delta\phi^* dx$$

$$+ [e_A T(1)w'(1) - e_A T(0)w'(0) + k_3\phi]\delta\phi^* \qquad (22b)$$

Thus Eqs. (21) is the necessary and sufficient condition for the following differential equations and boundary conditions:

D.E.:

$$w'''' - (Tw')' + \lambda^2 w + e(\lambda^2 - \Omega^2)\phi = 0 \qquad (23a)$$

$$(\Omega^2 k^2 + \lambda^2 k_m^2)\phi - e_A T w'' + e(\Omega^2 x w' + \lambda^2 w) = 0 \qquad (23b)$$

B.C.:

$$w''(1) = 0, \quad w'''(1) = 0 \qquad (24a,24b)$$

$$w''(0) - k_2 w'(0) = 0 \qquad (24c)$$

$$w'''(0) - \frac{\Omega^2}{2} w'(0) + k_1 w(0) = 0 \qquad (24d)$$

and

$$-\frac{e_A}{2}\Omega^2 w'(0) + k_3\phi = 0 \qquad (24e)$$

Note that in Eqs. (24) the fact that

$$T(1) = 0$$

and

$$T(0) = \frac{1}{2}\Omega^2 \qquad (25)$$

has been used.

It is observed that the boundary conditions of Eqs. (24c) and (24d) represent very general support conditions at the hub. The special case $k_1 = \infty$ and $k_2 = 0$ corresponds to a hinged blade, while the case $k_1 = \infty$, $k_2 = \infty$ corresponds to a hingeless blade. These two special cases are the only ones considered in the literature. The boundary condition (24e) indicates a coupling between the flexural motion w and root torsion at the hub due to the centrifugal forces and the noncoincidence of the elastic axis and the tension axis.

4. MATRIX EQUATIONS FROM FINITE ELEMENT DISCRETIZATION. In this section, we shall briefly describe the formulation of the matrix equation of the approximate solution from the variational statement given in the previous section. From Eqs. (21), one can write

$$(\delta I)_{w,\phi} = 0 ;$$

or,

$$(\delta I)_{w,\phi} = \int_0^1 (w''\delta w^{*''} + Tw'\delta w^{*'} - e\Omega^2\phi\delta w^*)dx$$

$$+ k_1 w(0)\delta w^*(0) + k_2 w'(0)\delta w^{*'}(0) + \lambda^2 \int_0^1 (w\delta w^* + e\phi\delta w^*)dx$$

$$+ \int_0^1 [\Omega^2 k^2 \phi\delta\phi^* + e_A T' w'\delta\phi^* + e\Omega^2 x w'\delta\phi^*]dx + k_3\phi\delta\phi^*$$

$$+ \lambda^2 \int_0^1 (k_m^2\phi\delta\phi^* + ew\delta\phi^*)dx \qquad (26)$$

It should be noted that Eq. (26) is a quite general equation. Various types of approximate solutions can be obtained depending on the choice of the coordinate functions. The motivation of using the finite element discretizations, which corresponds to the choice of a set of piecewise analytic functions, is twofold: for the ease of extending this formulation to problems of irregular geometry and for the adaptations to general finite element computer systems.

In the present formulation, however, only blades of uniform cross-sections will be considered and the elements are assumed to be of the same length. Thus, one introduces a local (element) coordinate ξ which relates to the global (entire blade) coordinate x such that

$$\xi = \xi^{(i)} = L(x - \frac{i-1}{L}) \qquad (27)$$

where L is the number of elements and i denotes the i-th element. Using the notation

$$w^{(i)}(\xi) = w[x(\xi^{(i)})], \text{ etc.} \qquad (28)$$

for simplicity and noting that

$$d\xi = Ldx$$

$$w'(x) = Lw^{(i)'}(\xi), \text{ etc.} \qquad (29)$$

one then has, from Eq. (26):

$$(\delta I)_{w,\phi} = 0$$

$$= \sum_{i=1}^{L} \{\int_0^1 L^3 w^{(i)\prime\prime}(\xi)\delta w^{*(i)\prime\prime}(\xi)d\xi$$

$$+ \int_0^1 L\, T^{(i)}(\xi)w^{(i)\prime}(\xi)\delta w^{*,(i)}(\xi)d\xi - \frac{e\Omega^2\phi}{L}\int_0^1 \delta w^{*(i)}d\xi\}$$

$$+ k_1 w^{(1)}(0)\delta w^{*(1)}(0) + k_2 L^2 w^{(i)\prime}(0)\delta w^{*(i)\prime}(0)$$

$$+ \lambda^2 \sum_{i=1}^{L} \{\frac{1}{L}\int_0^1 w^{(i)}(\xi)\delta w^{*(i)}(\xi)d\xi + \frac{e\phi}{L}\int_0^1 \delta w^{*(i)}(\xi)d\xi\}$$

$$+ \sum_{i=1}^{L} \{\Omega^2 k^2 \phi\delta\phi^* \frac{1}{L}\int_0^1 d\xi^{(i)} + e_A L \int_0^1 T^{(i)\prime}(\xi)w^{(i)\prime}(\xi)d\xi\delta\phi^*$$

$$+ \frac{e\Omega^2}{L}\int_0^1 [\xi + (i-1)]w^{(i)\prime}(\xi)d\xi\,\delta\phi^*\} + k_3\phi\delta\phi^*$$

$$+ \lambda^2 \{\frac{k_m^2}{L}\phi\delta\phi^* \int_0^1 d\xi^{(i)} + \sum_{i=1}^{L}\frac{e}{L}\int_0^1 w^{(i)}(\xi)d\xi\}\delta\phi^* \qquad (30)$$

where

$$T^{(i)}(\xi) = -b[\xi^2 + 2(i-1)\xi + (i-1)^2 - L^2] \qquad (31)$$

and

$$b = \frac{\Omega^2}{2L^2} \qquad (32)$$

At this point, it is appropriate to introduce the shape function and generalized coordinates. Let

$$w^{(i)}(\xi) = \underset{\sim}{a}^T(\xi)\underset{\sim}{W}^{(i)}$$

$$w^{*(i)}(\xi) = \underset{\sim}{a}^T(\xi)\underset{\sim}{W}^{*(i)} \qquad (33)$$

where the superscript T denotes the transpose of a matrix,

$$\underset{\sim}{W}^{(i)T} = \{ W_1^{(i)} \quad W_2^{(i)} \quad W_3^{(i)} \quad W_4^{(i)} \}$$

$$\underset{\sim}{W}^{*(i)T} = \{ W_1^{*(i)} \quad W_2^{*(i)} \quad W_3^{*(i)} \quad W_4^{*(i)} \} \tag{34}$$

and

$$\underset{\sim}{a}^T(\xi) = \{1 - 3\xi^2 + 2\xi^3, \quad \xi - 2\xi^2 + \xi^3, \quad 3\xi^2 - 2\xi^3, \quad -\xi^2 + \xi^3\} \tag{35}$$

Using Eqs. (31), (32) and in terms of $W^{(i)}$, $W^{*(i)}$, Eq. (30) becomes

$$(\delta I)_{w,\phi} = 0$$

$$= \sum_{i-1}^{L} \delta W^{*(i)T} \{ L^3 \int_0^1 a'' a''^T d\xi - bL [\int_0^1 \xi^2 a' a'^T d\xi + 2(i-1) \int_0^1 \xi a' a'^T d\xi$$

$$+ \{ L^2 - (i-1)^2 \} \int_0^1 a' a'^T d\xi] \underset{\sim}{W}^{(i)} \}$$

$$+ \sum_{i=1}^{L} \delta \underset{\sim}{W}^{*(1)T} (-2ebL) \int_0^1 \underset{\sim}{a} d\xi \; \phi$$

$$+ \delta \underset{\sim}{W}^{*(1)T} k_1 \underset{\sim}{a}(0) \underset{\sim}{a}^T(0) \underset{\sim}{W}^{(1)} + \delta \underset{\sim}{W}^{*(1)} k_2 L^2 \underset{\sim}{a}'(0) \underset{\sim}{a}'^T(0) \underset{\sim}{W}^{(1)}$$

$$+ \lambda^2 \{ \sum_{i=1}^{L} \delta \underset{\sim}{W}^{*(i)T} \frac{1}{L} \int_0^1 \underset{\sim\sim}{a a}^T d\xi \; W^{(i)} + \frac{e}{L} \int_0^1 \underset{\sim}{a} d\xi \; \phi \}$$

$$+ \delta \phi^* \Omega^2 k^2 \phi + \delta \phi^* k_3 \phi$$

$$+ \delta \phi^* \sum_{i=1}^{L} 2(e - e_A) bL \int_0^1 \xi a'^T(\xi) d\xi + (i-1) \int_0^1 \underset{\sim}{a}'^T d\xi \; \underset{\sim}{W}^{(i)}$$

$$+ \lambda^2 \delta \phi^* [k_m^2 \phi + \sum_{i=1}^{L} \frac{e}{L} \int_0^1 \underset{\sim}{a}^T d\xi \; \underset{\sim}{W}^{(i)}] \tag{36}$$

Or,

$$(\delta I)_{w,\phi} = 0$$

$$= \sum_{i=1}^{L} \delta \underset{\sim}{W}^{*}(i)^{T} \{L^{3}\underset{\sim}{C} - bL[\underset{\sim}{E} + 2(i-1)\underset{\sim}{D} - (L^{2} - (i-1)^{2})\underset{\sim}{B}]\} \underset{\sim}{W}^{(i)}$$

$$+ \sum_{i=1}^{L} \delta \underset{\sim}{W}^{*}(i)^{T} (-2\, ebL)\underset{\sim}{p}\,\phi + \delta \underset{\sim}{W}^{*}(i)^{T} [k_{1}\underset{\sim}{F} + k_{2}L\,\underset{\sim}{G}]\underset{\sim}{W}^{(1)}$$

$$+ \lambda^{2}[\sum_{i=1}^{L} \delta \underset{\sim}{W}^{*}(i)^{T} [\frac{1}{L}\underset{\sim}{A}\,\underset{\sim}{W}^{(i)} + \frac{e}{L}\underset{\sim}{p}\,\phi]$$

$$+ \delta\phi^{*} \sum_{i=1}^{L} 2(e - e_{A})bL\,(\underset{\sim}{s}^{T} + (i-1)\underset{\sim}{r}^{T})\,\underset{\sim}{W}^{(i)}$$

$$+ \delta\phi^{*}\,(\Omega^{2}k^{2} + k_{3})\phi + \lambda^{2}\,\delta\phi^{*}\,(\sum_{i=1}^{L} \frac{e}{L}\,p^{T}\,\underset{\sim}{W}^{(i)} + k_{m}^{2}\phi) \qquad (37)$$

where the element matrices are defined as follows,

$$\underset{\sim}{A} = \int_{0}^{1} \underset{\sim}{a}\underset{\sim}{a}^{T} d\xi, \qquad \underset{\sim}{B} = \int_{0}^{1} \underset{\sim}{a}'\underset{\sim}{a}'^{T} d\xi, \qquad \underset{\sim}{C} = \int_{0}^{1} \underset{\sim}{a}''\underset{\sim}{a}''^{T} d\xi$$

$$\underset{\sim}{D} = \int_{0}^{1} \xi \underset{\sim}{a}'\underset{\sim}{a}'^{T} d\xi, \qquad \underset{\sim}{E} = \int_{0}^{1} \xi^{2} \underset{\sim}{a}'\underset{\sim}{a}'^{T} d\xi$$

$$\underset{\sim}{F} = \underset{\sim}{a}(0)\,\underset{\sim}{a}^{T}(0), \qquad \underset{\sim}{G} = \underset{\sim}{a}'(0)\underset{\sim}{a}'^{T}(0)$$

$$\underset{\sim}{p} = \int_{0}^{1} \underset{\sim}{a}\, d\xi, \qquad \underset{\sim}{r} = \int_{0}^{1} \underset{\sim}{a}'\, d\xi, \qquad \underset{\sim}{s} = \int_{0}^{1} \xi\, \underset{\sim}{a}'\, d\xi \qquad (38)$$

The numerical values of these matrices are given in the Appendix.

In terms of the global coordinates defined in the following

$$\underset{\sim}{W} = \begin{Bmatrix} \underset{\sim}{W} \\ \phi \end{Bmatrix}, \qquad \underset{\sim}{\bar{W}}^{*} = \begin{Bmatrix} \underset{\sim}{W}^{*} \\ \phi \end{Bmatrix} \qquad (39)$$

and

$$\underset{\sim}{W}^{T} = \{W_{1}^{(1)}\ W_{2}^{(1)}\ W_{3}^{(1)}\ W_{4}^{(1)}\ W_{3}^{(2)}\ W_{4}^{(2)} \ldots W_{3}^{(L)}\ W_{4}^{(L)}\}$$

$$\underset{\sim}{W}^{*T} = \{W_{1}^{*(1)}\ W_{2}^{*(1)}\ W_{3}^{*(1)}\ W_{4}^{*(1)}\ W_{3}^{*(2)}\ W_{4}^{*(2)} \ldots W_{3}^{*(L)}\ W_{4}^{*(L)}\} \qquad (40)$$

One has from Eq. (37):

$$(\delta I)_{w,\phi} = 0$$

$$= \delta\bar{\underset{\sim}{W}}^{*T}(\bar{\underset{\sim}{K}} + \lambda^2\bar{\underset{\sim}{M}})\bar{\underset{\sim}{W}}$$

$$= \{\delta\underset{\sim}{W}^{*T}\delta\phi^{*}\} \begin{bmatrix} \bar{K}_{11} + \lambda\bar{M}_{11} & \bar{K}_{12} + \lambda\bar{M}_{12} \\ \bar{K}_{21} + \lambda\bar{M}_{21} & \bar{K}_{22} + \lambda\bar{M}_{22} \end{bmatrix} \begin{Bmatrix} \underset{\sim}{W} \\ \phi \end{Bmatrix} \tag{41}$$

The global matrices $\bar{K}_{ij}$, $\bar{M}_{ij}$, i,j = 1,2 are assembled from the element matrices defined in Eqs. (38) in the following manner.

$$\delta\underset{\sim}{W}^{*T}\bar{K}_{11}\underset{\sim}{W} = \sum_{i=1}^{L} \delta\underset{\sim}{W}^{*(i)T}\{L^3(\underset{\sim}{C} + b\underset{\sim}{B}) + bL(2\underset{\sim}{D} - \underset{\sim}{B} - \underset{\sim}{E})$$

$$+ 2ibL(\underset{\sim}{B} - \underset{\sim}{D}) - i^2bL\underset{\sim}{B}\}\underset{\sim}{W}^{(i)} + \delta\underset{\sim}{W}^{*(1)T}\{k_1\underset{\sim}{F} + k_2L^2\underset{\sim}{G}\}\underset{\sim}{W}^{(1)} \tag{42a}$$

$$\delta W^{*T}\bar{K}_{12}\phi = \sum_{i=1}^{L} \delta\underset{\sim}{W}^{*(i)T}(-2ebL\,\underset{\sim}{p})\phi \tag{42b}$$

$$\delta\phi^{*}\bar{K}_{21}\underset{\sim}{W} = \delta\phi^{*}\sum_{i=1}^{L} 2(e - e_A)bL[\underset{\sim}{s}^T - \underset{\sim}{r}^T + i\underset{\sim}{r}^T]\underset{\sim}{W}^{(i)} \tag{42c}$$

$$\delta\phi^{*}\bar{K}_{22}\phi = \delta\phi^{*}(\Omega^2k^2 + k_3)\phi \tag{42d}$$

$$\delta\underset{\sim}{W}^{*}\bar{M}_{11}\underset{\sim}{W} = \sum_{i=1}^{L} \delta\underset{\sim}{W}^{*(i)T}\frac{1}{L}\underset{\sim}{A}\,\underset{\sim}{W}^{(i)} \tag{42e}$$

$$\delta\underset{\sim}{W}^{*}\bar{M}_{12}\phi = \sum_{i=1}^{L} \delta\underset{\sim}{W}^{*(i)T}\frac{e}{L}\,\underset{\sim}{p}\,\phi \tag{42f}$$

$$\delta\phi^{*}\bar{M}_{21}\underset{\sim}{W} = \delta\phi^{*}\sum_{i=1}^{L}\frac{e}{L}\,\underset{\sim}{p}^T\,\underset{\sim}{W}^{(i)} \tag{42g}$$

$$\delta\phi^{*}\bar{M}_{22}\phi = \delta\phi^{*}\,k_m^2\,\phi \tag{42h}$$

Since $\delta W^{*T} = \{\delta W^{*T}\ \delta\phi^{*}\}$ in Eq. (41) is unconstrained, Eq. (41) leads directly to

$$(\bar{K} + \lambda^2 \bar{M})\bar{W} = 0 \tag{43}$$

which is the final matrix eigenvalue equation to be solved.

5. RESULTS AND DISCUSSION. Prior to the presentation of the demonstrative numerical results, it will be worthwhile to make some observations on the nondimensionalized differential equations (15) and (16) and the boundary conditions (24).

i. For e = 0, the flapping motion w(x,t) and the root torsion $\phi(t)$ are essentially uncoupled. The eigenvalue solutions of w reduce to that of a rotating beam and agree well with the available data of Boyce, DiPrima and Handelman [11]. The torsional vibration frequency has only one eigenvalue solution and, if $e_A = 0$ also, it varies linearly with the rotor speed Ω as expected.

ii. For $e \neq 0$ and $e_A = 0$, the motions are generally coupled. It is observed that if e and ϕ both change sign, the equation remains unaltered. Thus, as far as eigenvalues are concerned, they depend only on the absolute values of e. The solution of ϕ for a negative e, however, is 180° out of phase compared with the one for a positive e of the same magnitude.

iii. For $e \neq 0$ and $e_A \neq 0$ the motions are generally coupled. It is observed that if e, e_A and ϕ all three change signs, the governing equations remain unaltered.

Some demonstrative calculations will now be given*. The eigenvalue λ is generally a complex number.

$$\lambda = \lambda_R + i\lambda_I \tag{44}$$

From equations (17) and (44), it is clear that the system is unstable of divergence when λ_R is nonzero positive and $\lambda_I = 0$. When $\lambda_R = 0$, on the other hand, λ_I (LAMBDA)** represents the nondimensional frequency of vibration and it can then be plotted against the nondimensional blade rotating speed Ω (OMEGA), as shown in Figures 3 through 7. When λ is complex, one of the square roots of λ^2 must have nonzero positive real part. The system is then unstable since λ appears in the equations only as λ^2. The value of λ generally become complex as the two branches of the frequency curve coelesce. The "critical" speeds can be located in these figures by noting these points of coelescence.

*An extensive parametric study of a rotor blade instability in vacuum due to the coupled flap-(root)torsion motion will be presented in a separate paper forthcoming.

**The symbols appeared in the parentheses are those used in Figures 3 through 7.

For these sample calculations here, the parameters used are typical for a medium size rotor blade of 20 ft. in length, for example. It is further assumed that (the nondimensionalized quantities)

$$e(E) = 0.003, \qquad e_o(E_o) = 0$$

$$k_{m1}(KM1) = 0.0025, \qquad k_{m2}(KM2) = 0.0100$$

The boundary conditions at the hub are those of a hingeless blade. The values of k_1(K-1) and k_2(K-2) are set to 10^8 as approximations to infinity. The torsional spring constant k_3(K-3) has been set to zero.

In Figure 3, $e_A(E_A)$ is taken to be 0.003. Thus

$$\gamma \text{ (GAMMA)} = \frac{e_A}{e} = \frac{-e_A}{-e} = 1.00 .$$

Here the lowest branch of λ_I is essentially for the torsional motion. The second and third lowest branches are essentially the first and the second for flapping motion. For the range of rotor speed shown*, $0 \leq \Omega \leq 25$, the coupling is not sufficient to have instability. In the subsequent figures as e_A = 0.0015, 0, -0.0015 and -0.003 (and γ = 0.5, 0, -0.5 and -1.0), the effect of coupling becomes more and more severe. In Figure 4, the two lowest branches of eigenvalues appear to draw closer compared with those in Figure 3. They actually coelesce in Figure 5 at a critical speed about Ω = 16.5. As e_A continue to decrease (increase) in algebraic sense while holding e a positive (negative) constant the critical speed of flutter instability tends to decrease and thus the structure becomes more critical. This is observed in Figures 6 and 7.

ACKNOWLEDGEMENT. The authors wish to express their appreciation to Mr. Richard F. Haggerty for his assistance in obtaining the numerical results. They also wish to thank Miss Ellen Fogarty for her patience and skill in typing the manuscript.

REFERENCES.

1. P. Friedman and P. Tong, "Nonlinear Equations for Bending of Rotating Beams with Application to Linear Flap-lag Stability of Hingeless Rotors," NASA CR-114485, May 1972.

2. D. H. Hodges and R. A. Ormiston, "Stability of Elastic Bending and Torsion of Uniform Cantilevered Rotor Blades in Hover," 14th Structures, Structural Dynamics and Materials Conference, AIAA Paper No. 73-405, 1973.

3. R. H. Miller and C. W. Ellis, "Helicopter Blade Vibration and Flutter," Journal of American Helicopter Society, Vol. 1, No. 3, pp. 19-38, 1956.

*At Ω (OMEGA) = 25, which is nondimensional, the real rotor speed is roughly 1000 rpm for a typical value of $c = \sqrt{(\rho A \ell^4/EI)}$ = 0.0040min.

4. N. D. Ham, "Helicopter Blade Flutter," AGARD Report No. 607, 1973.

5. D. H. Hodges and R. A. Ormiston, "Nonlinear Equations for Bending of Rotating Beams with Application to Linear Flap-lag Stability of Hingeless Rotors," NASA TM X-2770, 1973.

6. J. J. Wu, "On Adjoint Operators Associated with Boundary Value Problems," Journal of Sound and Vibration, Vol. 39, No. 2, pp. 195-206, 1975.

7. J. J. Wu, "A Unified Finite Element Approach to Column Stability Problems," Developments in Mechanics, Vol. 8, pp. 279-294, 1975.

8. J. J. Wu, "Effects of Support Flexibility on the Stability of a Beam Under a Followed Thrust and Inertia," Developments in Theoretical and Applied Mechanics, Vol. 8, pp. 391-402, 1976.

9. J. J. Wu, "Column Instability Under Nonconservative Forces, With Internal and External Damping - Finite Element Using Adjoint Variational Principles," Developments in Mechanics, Vol. 7, pp. 501-514, 1973.

10. J. C. Houbolt and G. W. Brooks, "Differential Equations of Motion for Combined Flapwise Bending, Chordwise Bending and Torsion of Twisted Rotor Blades," NACA Report No. 1346, 1956.

11. W. E. Boyce, R. C. DiPrima and G. H. Handelman, "Vibrations of Rotating Beams of Constant Section," Proceedings of the Second U.S. National Congress of Applied Mechanics, Ann Arbor, Mich., pp. 165-173 (1954).

APPENDIX

The numerical value of some of the matrices used in Section 4 are given here.

$$\underset{\sim}{A} = \int_0^1 \underset{\sim}{a}\underset{\sim}{a}^T d\xi = \begin{bmatrix} \frac{13}{25} & \frac{11}{210} & \frac{9}{70} & -\frac{13}{420} \\ \frac{11}{210} & \frac{1}{105} & \frac{13}{420} & -\frac{1}{140} \\ \frac{9}{70} & \frac{13}{420} & \frac{13}{35} & -\frac{11}{210} \\ -\frac{13}{420} & -\frac{1}{140} & -\frac{11}{210} & \frac{1}{105} \end{bmatrix} \tag{A-1}$$

$$\underset{\sim}{B} = \int_0^1 \underset{\sim}{a}'\underset{\sim}{a}'^T d\xi = \begin{bmatrix} \frac{6}{5} & \frac{1}{10} & -\frac{6}{5} & \frac{1}{10} \\ \frac{1}{10} & \frac{2}{15} & -\frac{1}{10} & -\frac{1}{30} \\ -\frac{6}{5} & -\frac{1}{10} & \frac{6}{5} & -\frac{1}{10} \\ \frac{1}{10} & -\frac{1}{30} & -\frac{1}{10} & \frac{2}{15} \end{bmatrix} \tag{A-2}$$

$$\underset{\sim}{C} = \int_0^1 \underset{\sim}{a}''\underset{\sim}{a}''^T d\xi = \begin{bmatrix} 12 & 6 & -12 & 6 \\ 6 & 4 & -6 & 2 \\ -12 & -6 & 12 & -6 \\ 6 & 2 & -6 & 4 \end{bmatrix} \tag{A-3}$$

$$\underset{\sim}{D} = \int_0^1 \xi \underset{\sim}{a}' \underset{\sim}{a}'^T d\xi = \begin{bmatrix} \frac{3}{5} & \frac{1}{10} & -\frac{3}{5} & 0 \\ \frac{1}{10} & \frac{1}{30} & \frac{1}{10} & \frac{1}{60} \\ -\frac{3}{5} & -\frac{1}{10} & \frac{3}{5} & 0 \\ 0 & -\frac{1}{60} & 0 & \frac{1}{10} \end{bmatrix} \tag{A-4}$$

$$\underset{\sim}{E} = \int_0^1 \xi^2 \underset{\sim}{a}' \underset{\sim}{a}'^T d\xi = \begin{bmatrix} \frac{12}{35} & \frac{1}{14} & -\frac{12}{35} & -\frac{1}{35} \\ \frac{1}{14} & \frac{1}{105} & -\frac{1}{14} & -\frac{1}{70} \\ -\frac{12}{35} & -\frac{1}{14} & \frac{12}{35} & \frac{1}{35} \\ -\frac{1}{35} & -\frac{1}{70} & \frac{1}{35} & \frac{3}{35} \end{bmatrix} \tag{A-5}$$

$$\underset{\sim}{F} = \underset{\sim}{a}(0)\underset{\sim}{a}^T(0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{A-6}$$

$$\underset{\sim}{G} = \underset{\sim}{a}'(0)\underset{\sim}{a}'^T(0) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{A-7}$$

$$\underset{\sim}{p} = \int_0^1 \underset{\sim}{a} d\xi = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{12} \\ \frac{1}{2} \\ -\frac{1}{12} \end{bmatrix} \tag{A-8}$$

$$\underset{\sim}{r} = \int_0^1 \underset{\sim}{a}' d\xi = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \tag{A-9}$$

$$\underset{\sim}{s} = \int_0^1 \xi \, \underset{\sim}{a}' d\xi = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{12} \\ \frac{1}{2} \\ \frac{1}{12} \end{bmatrix} \tag{A-10}$$

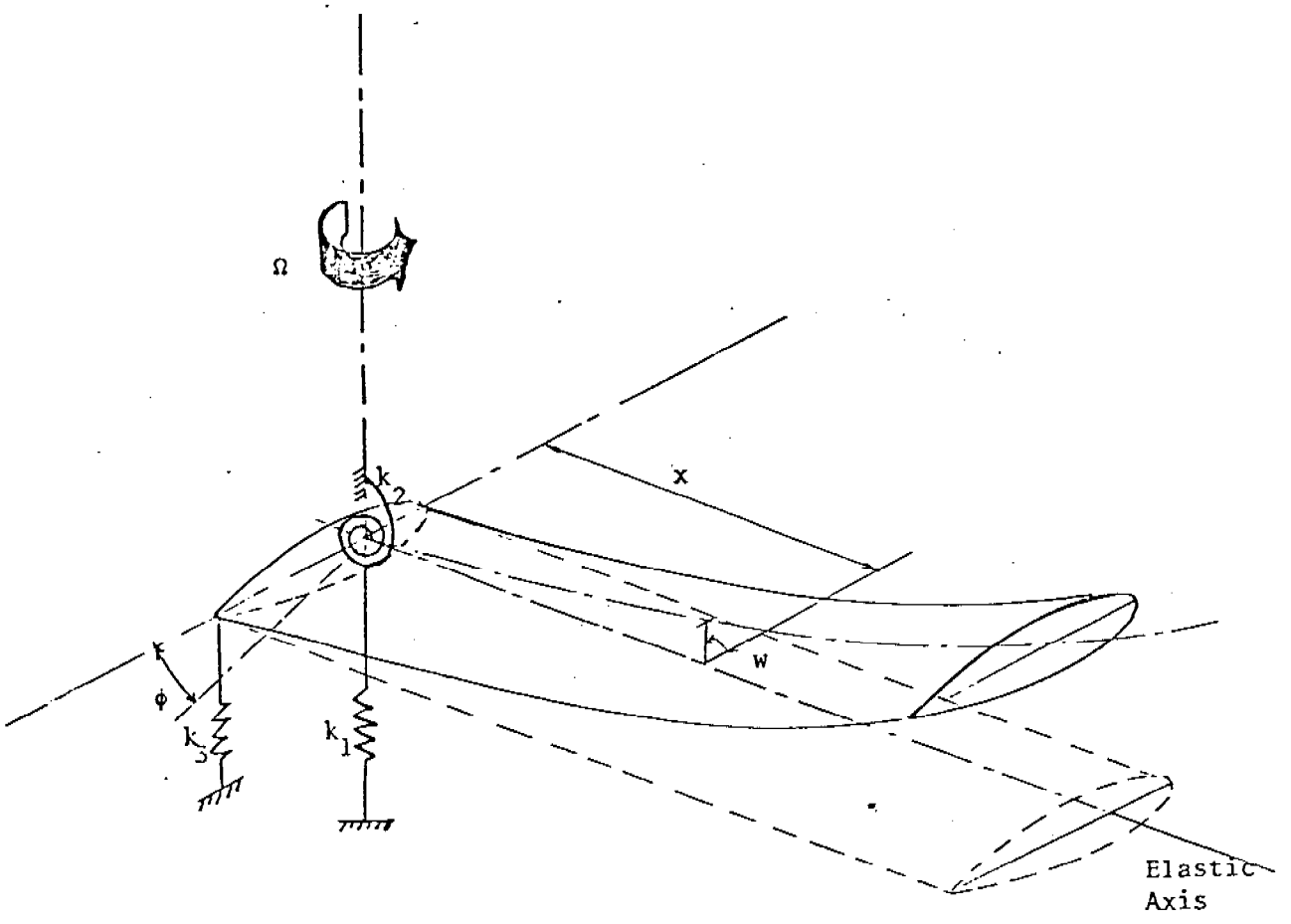
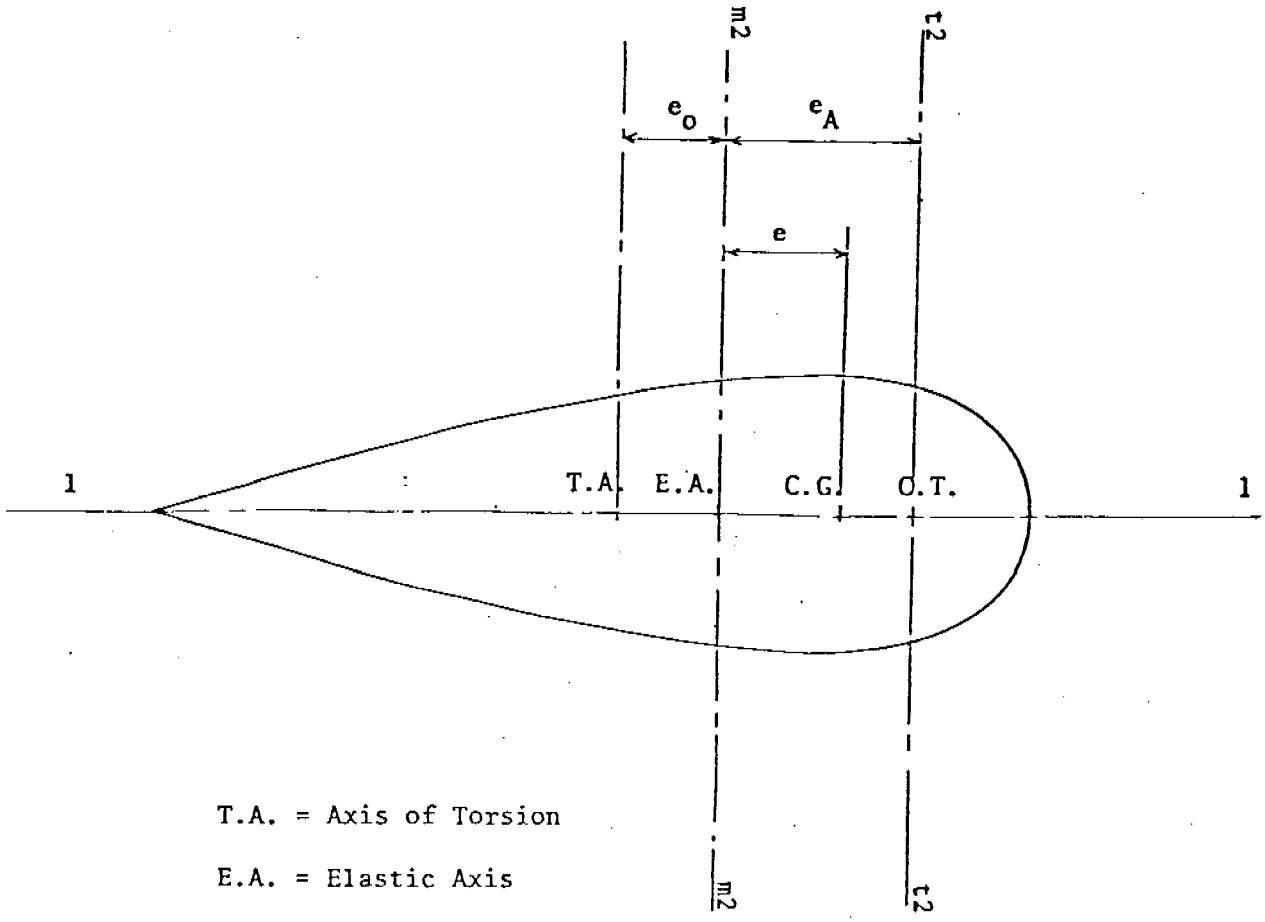


Figure 1. Problem Configuration. (Elastic axis is shown to coincide with the axis of torsion in the figure.)



T.A. = Axis of Torsion

E.A. = Elastic Axis

C.G. = Centre of Gravity

C.T. = Centroid of Tensile Area of a Cross Section

Figure 2. Parameters Related to Off-sets of Various Axis.

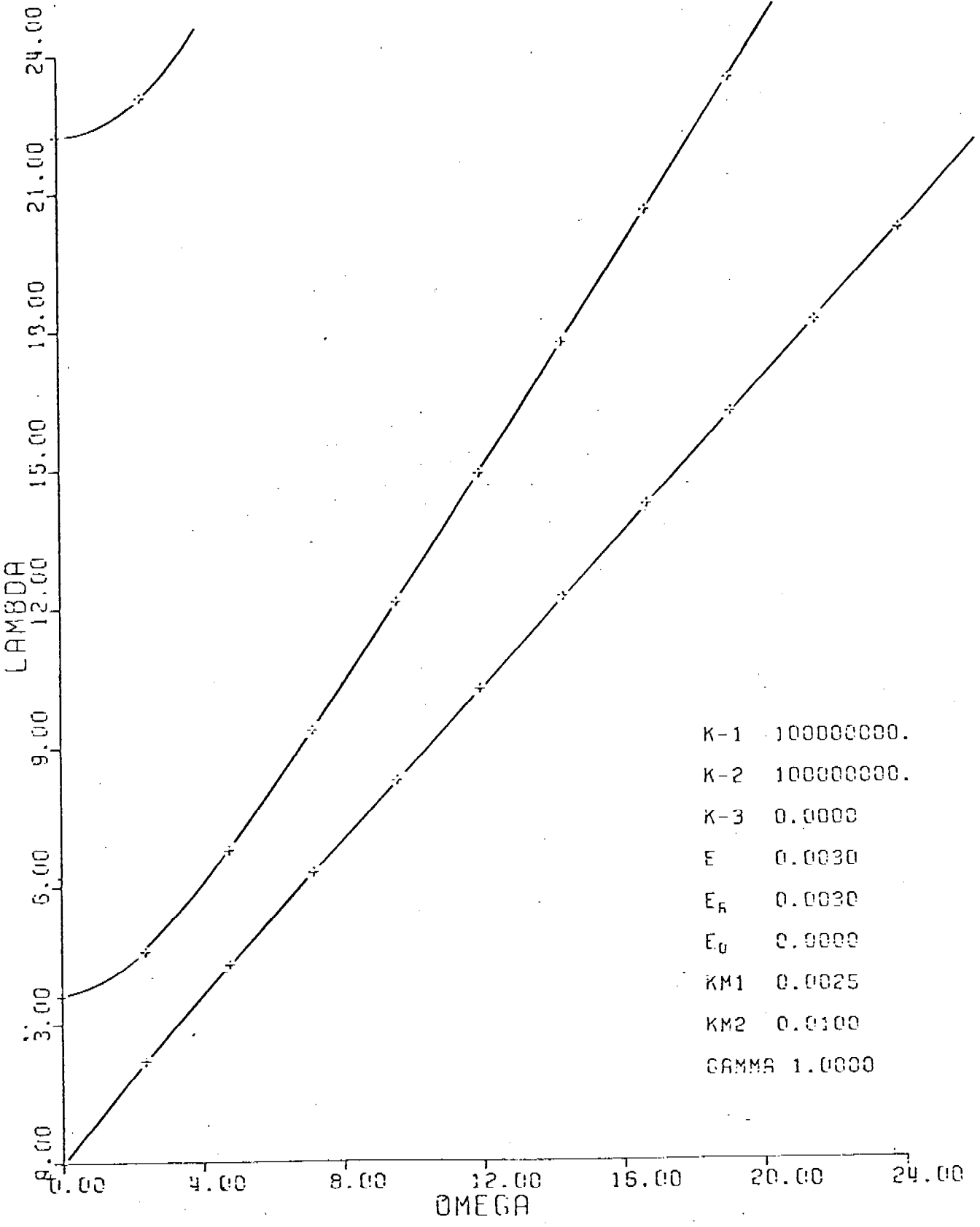


FIGURE 3. Vibration Frequency vs. Speed of Rotation (e_A=0.0030, e=0.0030). No Instability in the Range of Speed Shown.

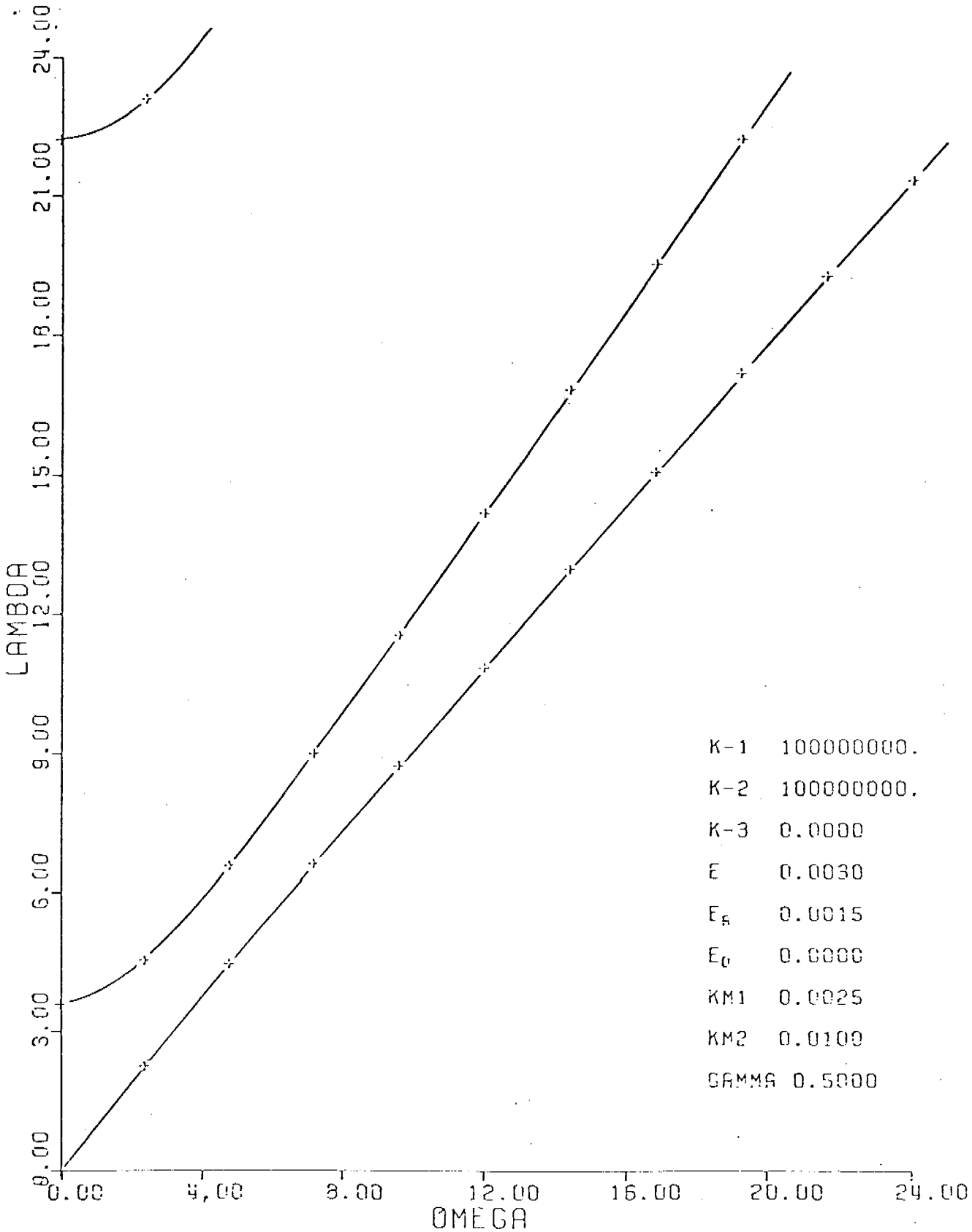


FIGURE 4. Vibration Frequency vs. Speed of Rotation (e_A=0.0015, e=0.0030). No Instability in the Range of Speed Shown.

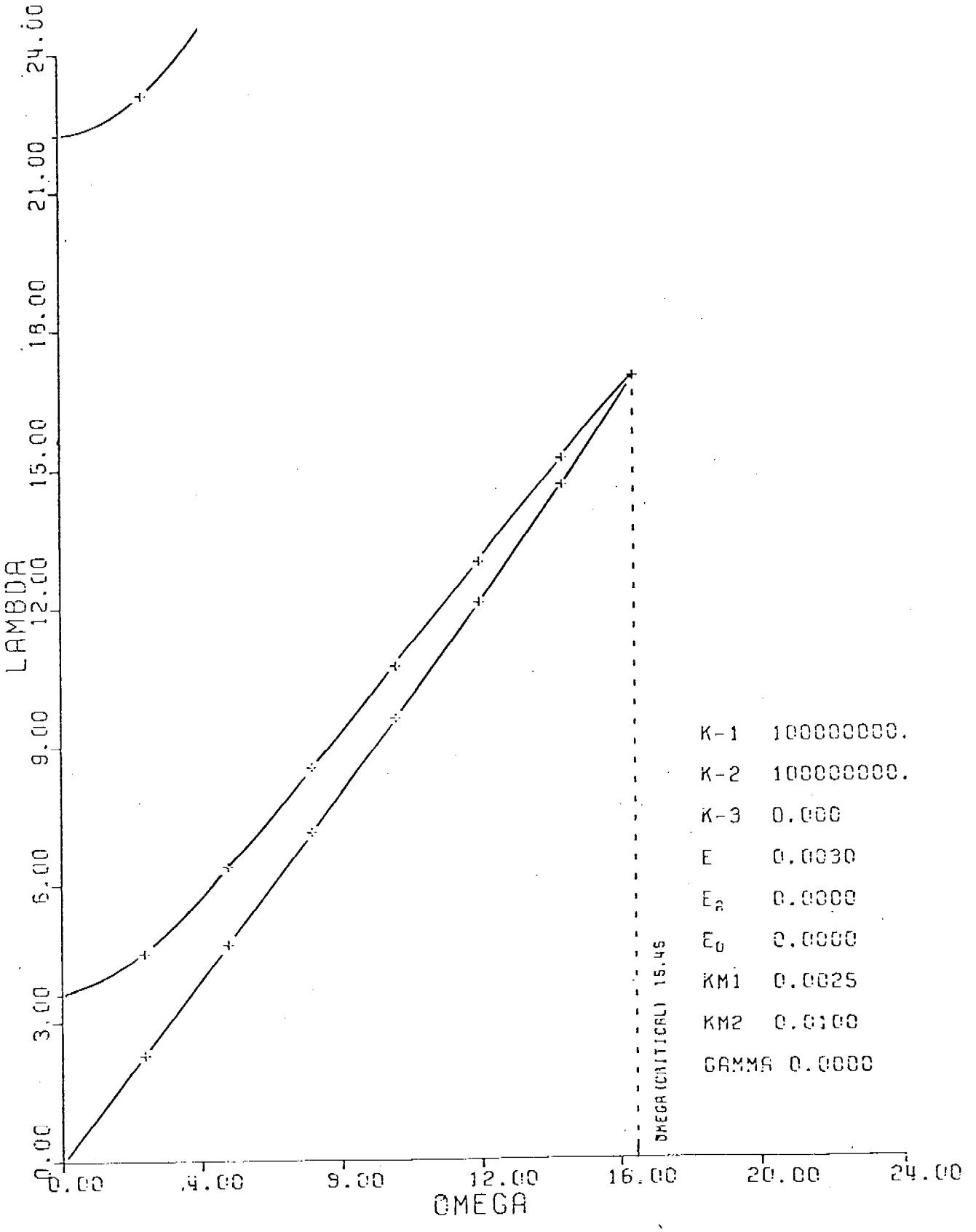


FIGURE 5. Vibration Frequency vs. Speed of Rptation (e_A=0, e=0.0030). Flutter Instability at Ω=16.46.

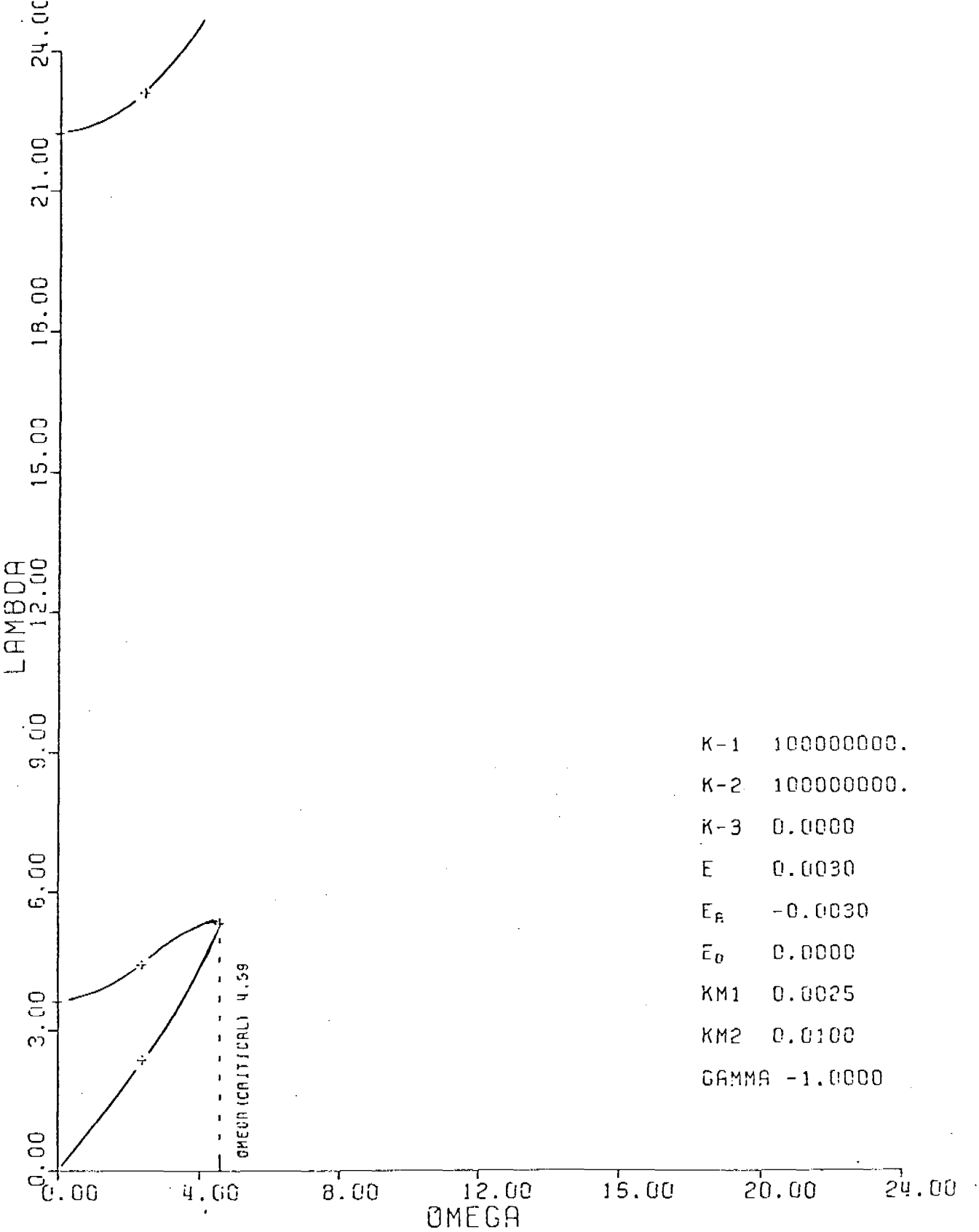


FIGURE 7. Vibration Frequency vs. Speed of Rotation (e_A=-0.0030, e=0.0030). Flutter Instability at Ω=4.59.

EFFECT OF DAMPING AT THE SUPPORT OF A ROTATING BEAM ON VIBRATIONS

J. D. Vasilakis and J. J. Wu
Applied Mathematics and Mechanics Section
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. The paper presents a formulation for the study of damping effects in dynamic structural problems and a specific application. A finite element formulation is first derived from the versatile unconstrained variational approach. The vibration of a rotating beam is used here as a concrete example. Viscous damping terms at the support can be present due to either local deflection or rotation. These terms can obviously affect the frequencies of the rotating beam. They are easily incorporated in the present formulation using the concept of unconstrained variations. Numerical data will be presented to demonstrate the qualitative as well as quantitative effects on the vibratory behavior of this rotating beam due to such damping terms.

1. INTRODUCTION. The applicability of the unconstrained adjoint variational statement in solving nonconservative stability problems has been shown in a series of articles [1-4]. The problems involved the stability of beams or columns subject to concentrated or distributed tangential loads. The problems are solved by finding the variational statement associated with the differential equation and they are rendered unconstrained by incorporating the geometric boundary conditions into the variational statement through the use of Lagrange multipliers. With the variational statement now available, the problem is discretized and solved using finite elements. Various types of external forces and geometric boundary conditions can be handled using the above techniques. It is the purpose of this paper to incorporate into the above-mentioned formulation the effect of support damping and to examine its effect on the solution. The specific problem chosen was that of a rotating cantilever beam of constant cross section.

This problem was chosen for its application to a simplified helicopter blade and although no nonconservative forces are considered the solution technique outlined above is applicable.

The effects of support damping on the vibration response of beams has been investigated by others. Fu and Mentel [5] and Mentel [6] considered support damping due to viscoelastic layers applied to the ends of a beam in its supports. The effect of translational (axial) damping was found to be of the same order in terms of energy dissipation at the supports as that of material damping. Material damping effects were found to stiffen the beam which increased both, the resonance frequency and the energy dissipation at the supports. They also found that rotational motion dominates the damping properties if all parameters are suitably optimized. Ruzicka [7] presented an evaluation of the resonance characteristics of undirectional

vibration isolation systems including directly coupled (Kelvin/Voight) and elastically coupled (Zener) damping elements. His results were mostly for the Zener model and he found that resonant frequencies of vibration isolation systems with viscous damping may increase or decrease with an increase in the viscous damping coefficients depending on the stiffnesses in the system. MacBain and Genin treated support flexibility in a series of papers. Support and material damping was introduced in [8]. The support is viewed as a complex rotational support stiffness based on bounds for the elastic modulus found in their earlier papers. They find that when the support damping constant is an optimum, the support loss factor is also an optimum, and system loss factor reaches a maximum value. This same value of the support loss factor is also that which critically dampens the system in free vibration.

The results presented here show the effects of support damping on the flexural frequencies of vibration of the rotating beam with both deflection and rotation flexibility at the support.

2. PROBLEM STATEMENT. The geometry of the problem is shown in Figure 1. The beam has a constant cross section of area A, density ρ, Young's modulus, E, and moment of inertion, I. The beam rotates about an axis fixed at one end of the beam and is flexibly supported at that end by a deflection spring, k_1, and a rotation spring, k_2. Viscous dashpots, c_1 and c_2, are assumed in parallel to the deflection and rotation springs, respectively. The beam rotates at constant angular velocity, Ω. S(0) represents support reaction.

The differential equation governing the motion is given by [9]

$$u'''' - \frac{\Omega^2 \rho A}{2EI} [(\ell^2 - x^2)u']' + \frac{\rho A}{EI} \ddot{u} = 0 \tag{1}$$

and the boundary conditions are

at x = 0

$$u''(0) - \frac{c_2}{EI} \dot{u}'(0) - \frac{k_2}{EI} u'(0) = 0$$

$$u'''(0) + \frac{c_1}{EI} \dot{u}(0) + \frac{k_1}{EI} u(0) - \frac{S(0)}{EI} u'(0) = 0 \tag{2}$$

at x = ℓ

$$u''(\ell) = 0$$

$$u'''(\ell) = 0 \tag{3}$$

The differential equation and boundary conditions are rewritten using dimensionless variables and parameters defined by the following:

$$\bar{u} = \frac{u}{\ell}\ , \quad \bar{x} = \frac{x}{\ell}\ , \quad Q = \frac{\Omega^2 \rho A \ell^4}{2EI}\ , \quad \bar{t} = \left[\frac{EI}{\rho A \ell^4}\right]^{1/2} t$$

$$\bar{c}_1 = \frac{c_1 \ell}{(EI\rho A)^{1/2}} \qquad \bar{c}_2 = \frac{c_2}{(EI\rho A)^{1/2}\ell} \tag{4}$$

$$\bar{k}_1 = \frac{k_1 \ell^3}{EI} \qquad \bar{k}_2 = \frac{k_2 \ell}{EI}$$

Time is removed by assuming displacements to have the form

$$\bar{u}(\bar{x},\bar{t}) = \bar{u}(\bar{x})e^{\lambda \bar{t}} \tag{5}$$

Then the differential equation becomes (dropping the bar symbol):

$$u'''' - Q[(1 - x^2)u']' + \lambda^2 u = 0 \tag{6}$$

and the boundary conditions

$$x = 0 \begin{cases} u''(0) - (\lambda c_2 + k_2)u'(0) = 0 \\ u'''(0) + (\lambda c_1 + k_1)u(0) - Qu'(0) = 0 \end{cases} \tag{7}$$

$$x = 1 \begin{cases} u''(1) = 0 \\ u'''(1) = 0 \end{cases} \tag{8}$$

The eigenvalues, λ, will be complex,

$$\lambda = \lambda_R + i\lambda_I \tag{9}$$

The frequencies will be given by

$$\omega = \lambda_I \left[\frac{EI}{\rho A \ell^4}\right]^{1/2} \tag{10}$$

and for this problem, $\lambda_R < 0$, i.e., the real component of the eigenvalue is negative and no instabilities should exist.

3. VARIATIONAL STATEMENT. To find the form of the variational statement, the differential equation is multiplied by an arbitrary variation of the adjoint field variable, $\delta v(x)$, and integrated over the beam length. Integration by parts indicates the form of the variational statement and the natural boundary conditions. The geometric boundary conditions are attached with the values of the springs and dashpots playing the role of Lagrange multipliers. The variational statement is finally given by

$$\delta J = 0 \tag{11}$$

where

$$J = \int_0^1 [u''v'' + Q[1 - x^2]u'v' + \lambda^2 uv]dx +$$

$$+ (\lambda c_1 + k_1)u(0)v(0) + (\lambda c_2 + k_2)u'(0)v'(0) \tag{12}$$

Performing the variation of J with respect to u and v, one can arrive at the original boundary value problem as well as its adjoint. In this case the two problems are identical.

4. FINITE ELEMENTS. To solve the problem using finite element techniques, the beam must be divided into segments and the nodes defined. The value for the unknown variable within each element must then be expressed in terms of the nodal values of the function through the use of interpolating shape functions. A global expression, or matrix is then formed and the eigenvalues found.

The procedure begins by taking the variation of Equation (12) and allowing the variations in the problem variable, $\delta u(x)$, to be zero,

$$\int_0^1 [u''\delta v'' + Q(1 - x^2)u'\delta v' + \lambda^2 u\delta v]dx +$$

$$(\lambda c_1 + k_1)u(0)\delta v(0) + (\lambda c_2 + k_2)u'(0)\delta v'(0) = 0 \tag{13}$$

The beam is divided into L elements, letting

$$\xi = L\{x - \frac{i-1}{L}\} \qquad i = 1,2,3...L \tag{14}$$

be the running coordinate in each element. Substituting Eq. (14) into Eq. (13):

$$\sum_{i=1}^{L} \int_0^1 [L^3 u^{(i)''}\delta v^{(i)''} + \frac{Q}{L}(L^2 - [\xi + (i - 1)]^2)u^{(i)'}\delta v^{(i)'}$$

$$+ \frac{\lambda^2}{L} u^{(i)}\delta v^{(i)}]d\xi + (\lambda c_1 + k_1)u^{(1)}(0)\delta v^{(1)}(0) + (\lambda c_2 + k_2)\cdot$$

$$L^2 u^{(1)'}(0)\delta v^{(1)'}(0) = 0 \tag{15}$$

In order that the displacements and their derivatives within an element be expressed in terms of their nodal values, the coordinate vectors

$$\bar{U}^{(i)T} = \{U_1^{(i)} \quad U_2^{(i)} \quad U_3^{(i)} \quad U_4^{(i)}\}$$

and (16)

$$\bar{V}^{(i)T} = \{V_1^{(i)} \quad V_2^{(i)} \quad V_3^{(i)} \quad V_4^{(i)}\}$$

are introduced. $U_1^{(i)}$, $U_2^{(i)}$ represent the displacement and slope at the left end of the ith element and $U_3^{(i)}$ and $U_4^{(i)}$ represent deflection and slope at the right end. A similar interpretation applied to the adjoint coordinate vector $\bar{V}^{(i)}$. The transform is indicated by T.

Hermitian polynomials are used to relate the displacements within an element to its nodal values, hence, the following shape function is assumed,

$$\bar{a}^T(\xi) = \{1 - 3\xi^2 + 2\xi^3 \quad \xi - 2\xi^2 + \xi^3 \quad 3\xi^2 - 2\xi^3 \quad -\xi^2 + \xi^3\} \tag{17}$$

So that

$$u^{(i)}(\xi) = \bar{a}^T(\xi)\bar{U}^{(i)}$$
$$v^{(i)}(\xi) = \bar{a}^T(\xi)\bar{V}^{(i)} \tag{18}$$

Substituting Eq. (18) into Eq. (15)

$$\sum_{i=1}^{L} \bar{U}^{(i)T}\{L^3\bar{C} + [QL - \frac{Q}{L}(i-1)^2]\bar{B} - \frac{Q}{L}\bar{E} - 2(i-1)\frac{Q}{L}\bar{D} + \frac{\lambda^2}{L}\bar{A}\}\delta\bar{V}^{(i)}$$

$$+ (\lambda c_1 + k_1)\bar{U}^{(1)T}\bar{H}\delta\bar{V}^{(1)} + L^2(\lambda c_2 + k_2)\bar{U}^{(1)T}\bar{F}\delta\bar{V}^{(1)} = 0 \tag{19}$$

where

$$\bar{A} = \int_0^1 \bar{a}(\xi)\bar{a}^T(\xi)d\xi, \qquad \bar{E} = \int_0^1 \xi^2\bar{a}'(\xi)\bar{a}^{T}{}'(\xi)d\xi$$

$$\bar{B} = \int_0^1 \bar{a}'(\xi)\bar{a}^{T}{}'(\xi)d\xi, \qquad \bar{F} = \bar{a}'(0)\bar{a}'^{T}(0)$$

$$\bar{C} = \int_0^1 \bar{a}''(\xi)\bar{a}^{T}{}''(\xi)d\xi$$

$$\bar{D} = \int_0^1 \bar{a}'(\xi)\bar{a}^{T}{}'(\xi)d\xi, \qquad \bar{H} = \bar{a}(0)\bar{a}^T(0) \tag{20}$$

Regrouping of (19),

$$\sum \bar{U}^{(i)T}\{\lambda^2 p^{(i)} + \lambda R^{(i)} + S^{(i)}\}\delta\bar{V}^{(i)} = 0 \tag{21}$$

where

$$p^{(i)} = \bar{A}/L \qquad i = 1,2,\ldots L \tag{22}$$

$$R^{(1)} = + c_1\bar{H} + c_2\bar{F}L^2 \qquad i = 1$$

$$R^{(i)} = 0 \qquad i = 2,3,\ldots L \tag{23}$$

$$S^{(1)} = L^3\bar{C} + QL\bar{B} - Q/L\ \bar{E} + k_1\bar{H} + k_2\bar{F}L^2 \qquad i = 1$$

$$S^{(i)} = L^3\bar{C} + QL[1 - \frac{1}{L^2}(i-1)^2]\bar{B} - \frac{Q}{L}\bar{E} - 2(i-1)Q\bar{D}/L \qquad i = 2,3,\ldots L \tag{24}$$

Using certain continuity conditions between the element nodal values

$$\begin{aligned} U_1^{(i)} &= U_3^{(i-1)} & V_1^{(i)} &= V_3^{(i-1)} \\ U_2^{(i)} &= U_4^{(i-1)} & V_2^{(i)} &= V_4^{(i-1)} \end{aligned} \tag{25}$$

One can write

$$\bar{U}^{(T)}\{\lambda^2[P] + \lambda[R] + [S]\}\delta V = 0 \tag{26}$$

where now

$$\bar{U}^{(T)} = \{U_1^{(1)}\ U_2^{(1)}\ U_3^{(1)}\ U_4^{(1)}\ U_3^{(2)}\ U_4^{(2)}\ldots U_3^{(L)}\ U_4^{(L)}\}$$

$$\bar{V}^{T} = \{V_1^{(1)}\ V_2^{(1)}\ V_3^{(1)}\ V_4^{(1)}\ V_3^{(2)}\ V_4^{(2)}\ldots V_3^{(L)}\ V_4^{(L)}\}$$

[P], [R], [S] are N x N matrices (N = 2L + 2). Since δV is arbitrary, the eigenvalue problem reduces to

$$\bar{U}^{(T)}\{\lambda^2[P] + \lambda[R] + [S]\} = 0 \tag{27}$$

for the eigenvalues of the problem.

An existing subroutine was used to find the eigenvalues which required the standard eigenvalue problem form

$$\{[A] + \lambda[I]\}\bar{U} = 0 \tag{28}$$

The equation

$$\{\lambda^2[A] + \lambda[B] + [C]\}\bar{U} = 0 \tag{29}$$

can be reduced to Eq. (28) by defining

$$\bar{W} = \lambda\bar{U} \tag{30}$$

This leads to the matrix equation

$$\begin{bmatrix} [0] & [I] \\ -[A]^{-1}[C] & -[A]^{-1}[B] \end{bmatrix} \begin{Bmatrix} \bar{U} \\ \bar{W} \end{Bmatrix} = \lambda \begin{Bmatrix} \bar{U} \\ \bar{W} \end{Bmatrix} \tag{31}$$

which is in the required format. The drawback here is that the order of the matrix has been doubled. Equation (31), however, is the form used for computing the eigenvalues.

5. RESULTS. Figures 2 and 3 show the effects for zero damping at the support. Figure 2 shows the effect of the rotation spring (k_2) only on the frequency with load as a parameter. The deflection spring is assumed to be infinitely stiff. For Q = 0, the beam is only vibrating and is not rotating. One can see that a stiffening effect occurs, i.e., the vibrating frequencies increase with an increase in the rotation spring constant. The frequencies rapidly approach those for a fully clamped vibrating and rotating beam. These results also fall within the bounds computed by Boyce, DiPrima and Handelman [10]. In Figure 3, the rotation spring is assumed to be infinitely stiff and the effect of varying the deflection spring is shown for different loads. Again, in general, there is a stiffening effect as the deflection spring value increases. For very small values of the deflection spring, the first vibrating frequency decreases slightly for increased loads, although only Q = 0 and 200 are shown. As k_1 is increased, there are cross over points after which higher loads do imply higher frequencies.

Figure 4 shows the effect of rotation damping at the support of a beam having rotation flexibility at the support. The deflection spring is assumed infinitely stiff and the deflection dashpot is zero. The figure is for a specific value of the rotation spring, $k_2 = 1$, and shows the first two eigenvalues for each of two loads, Q = 0 and Q = 100. A stiffening effect is found for increasing damping for $Q \neq 0$. For Q = 0, there is very slight decrease for very small damping values. For smaller rotation spring constants, and zero load frequencies decrease with increased damping as

shown in Figure 4 by the portion of the results for $k_2 = .1$. These results are better shown in Figure 5. The results in Figure 4 are interesting since one would expect a decrease in frequency as damping is increased. Stiffening effects due to damping are found elsewhere [5] and could be due to the manner in which it is introduced in the problem. Figure 4 also shows that the results for the fully damped beam are approached rapidly as damping is increased. Figure 5 shows the vibrating frequencies on a complex plane for a beam with rotation flexibility and damping. The load is zero (non-rotating beam) and the rotation spring is kept at $k_2 = .1$ while the dashpot value changes. The arrows adjacent to the curves show the direction of the values on the curve as damping increases. For zero damping, the results are purely imaginary and are approximately .54 and 15.5 for the 1st and 2nd frequencies. As damping increases, the frequencies become complex with the imaginary components decreasing for the first eigenvalue and increasing for the second. The behavior of the first frequency is interesting. As the damping value increases, the imaginary component vanishes (as also seen in Figure 4) as if the system becomes critically damped. However, the real component can also be followed on the complex plot and the beam appears to vibrate again in this first mode as damping increases further. For sufficiently large damping the frequencies seem to approach those for a beam which is damped at the support. Points on the real axis represent zero motion but move with changes in damping values to points on the real axis where it is intersected by a branch or mode. Figure 6 shows the same results for load $Q = 25$. A final result for rotation flexibility is shown in Figure 7 for $k_2 = 10$. Here the rotation spring is relatively stiff and the fully damped results are rapidly approached with initial effects for near zero dashpot values overshadowed.

The effect of support damping on the frequencies on beams with deflection flexibility are shown in Figures 8-10. A decrease in frequency with increased damping is seen here. Figures 9 and 10 show the effect on the complex plane for $Q = 0$ and $Q = 200$, respectively.

Finally, Figure 11 shows the results of a case when both rotation spring and deflection spring flexibilities are allowed. Little effect on frequency is noted as the rotation spring is varied while the deflection spring and dashpot remain constant in value. An almost parallel increase in frequencies are found when the rotation dashpot is increased in value. The investigation into the response of the beam with all springs and dashpots finite was limited to those shown. A study should be performed to indicate areas where the effects on beam response will be most pronounced.

REFERENCES

1. J. J. Wu, "Column Instability Under Nonconservative Forces, With Internal and External Damping - Finite Element Using Adjoint Variational Principles," Development in Mechanics, Vol. 7, Proceedings of the 13th Midwestern Mechanics Conference.

2. J. J. Wu, "A Unified Finite Element Approach to Column Stability Problems," Developments in Mechanics, Vol. 8, Proceedings of the 14th Midwestern Mechanics Conference.

3. J. J. Wu, "On Adjoint Problems and Variational Principles," Developments in Mechanics, Vol. 8, Proceedings of the 14th Midwestern Mechanics Conference.

4. J. J. Wu, "On Mode Shapes of a Stability Problem," Journal of Sound and Vibration (1976), 46 (1), p. 51-57.

5. C. C. Fu and T. J. Mentel, "Steady State Response of Beams with Translational and Rotational Damping Motions at the Support," NADD Technical Report 60-60, May 1960.

6. T. J. Mentel, "Viscoelastic Boundary Damping of Beams and Plates," Journal of Applied Mechanics, March 1964.

7. J. E. Ruzicka, "Resonance Characteristics of Undirectional Viscous and Coulomb Damped Vibration Isolation Systems," Journal of Engineering for Industry, November 1967.

8. J. C. MacBain and J. Genin, "Energy Dissipation of a Vibration Timoshenko Beam Considering Support and Material Damping," Int. J. Mech. Sci., 1975, Vol. 17, pp. 255-265.

9. N. D. Ham, "Helicopter Blade Flutter," AGAARD Report No. 607, AD756728.

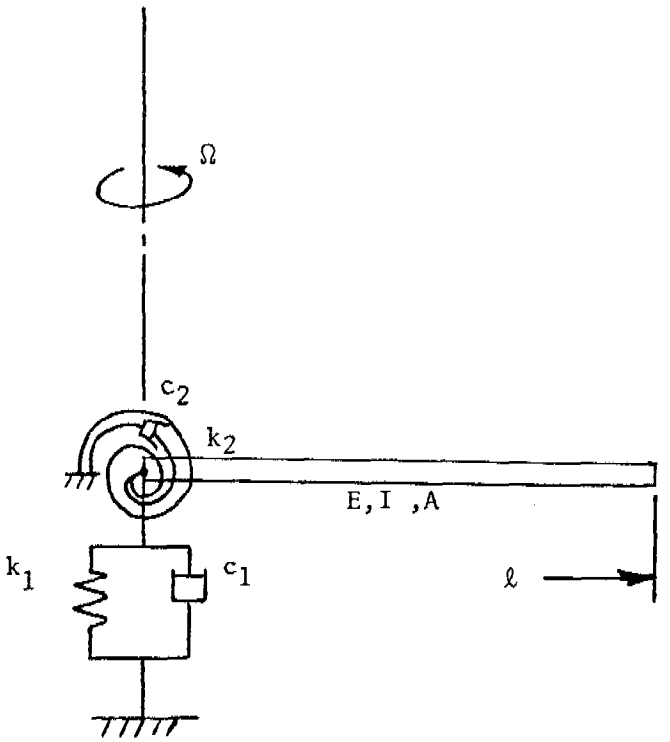
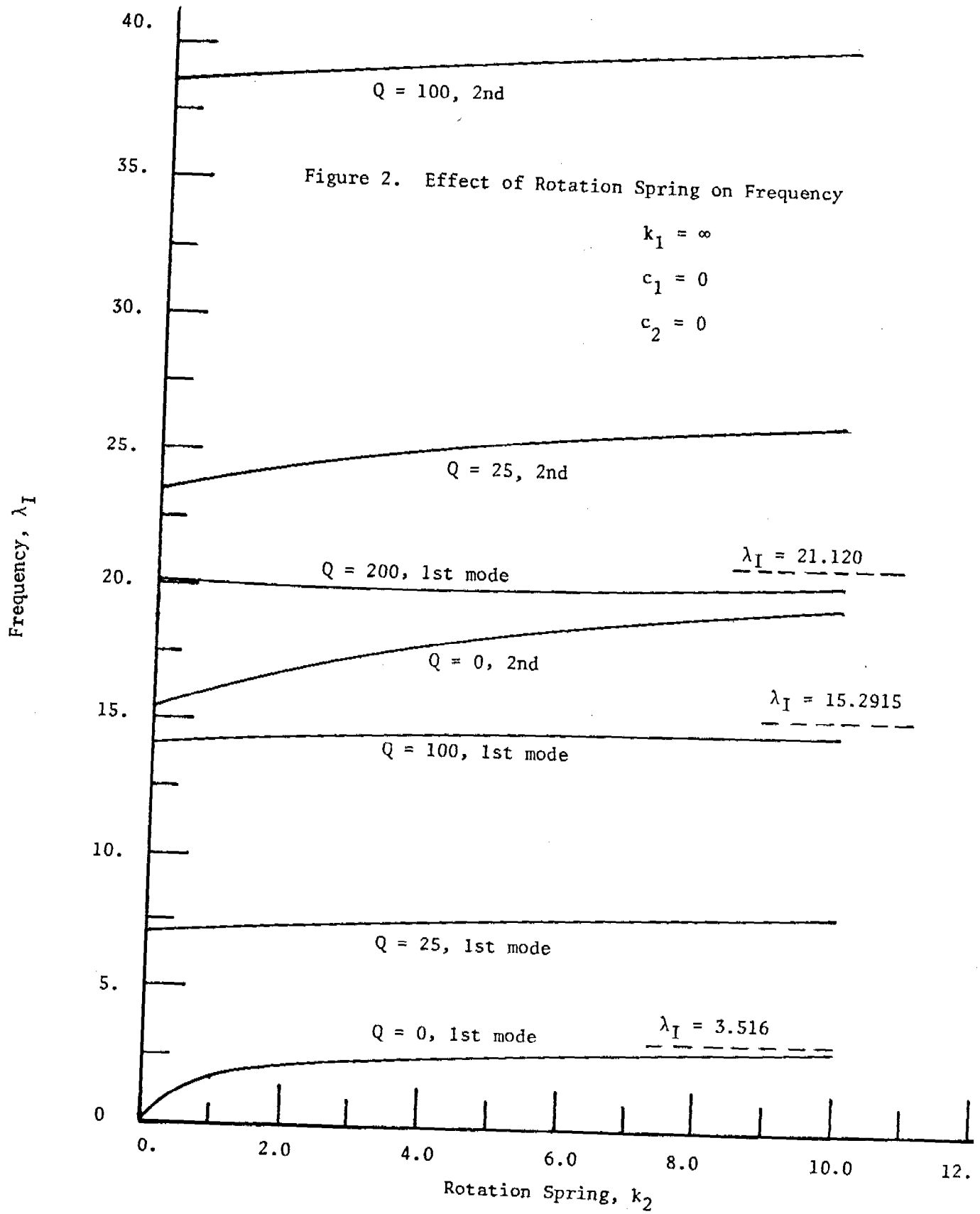


Figure 1. Geometry of Rotating Beam


40.
35.
30.
25.
20.
15.
10.
5.
0
Frequency, λ_I
Q = 100, 2nd
Figure 2. Effect of Rotation Spring on Frequency
$k_1 = \infty$
$c_1 = 0$
$c_2 = 0$
Q = 25, 2nd
Q = 200, 1st mode
$\lambda_I = 21.120$
Q = 0, 2nd
$\lambda_I = 15.2915$
Q = 100, 1st mode
Q = 25, 1st mode
Q = 0, 1st mode
$\lambda_I = 3.516$
0.
2.0
4.0
6.0
8.0
10.0
12.
Rotation Spring, k_2

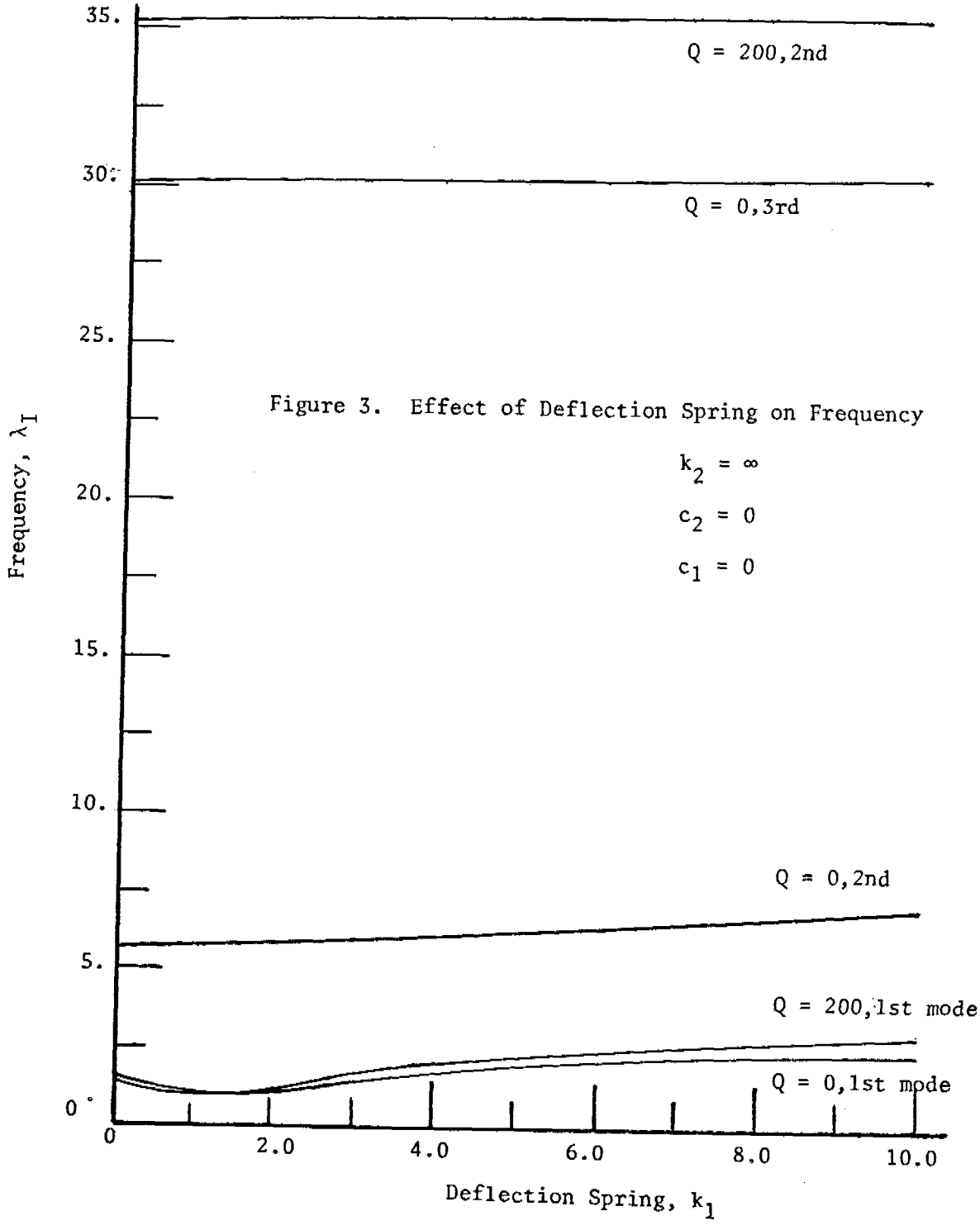


Figure 3. Effect of Deflection Spring on Frequency

$k_2 = \infty$

$c_2 = 0$

$c_1 = 0$

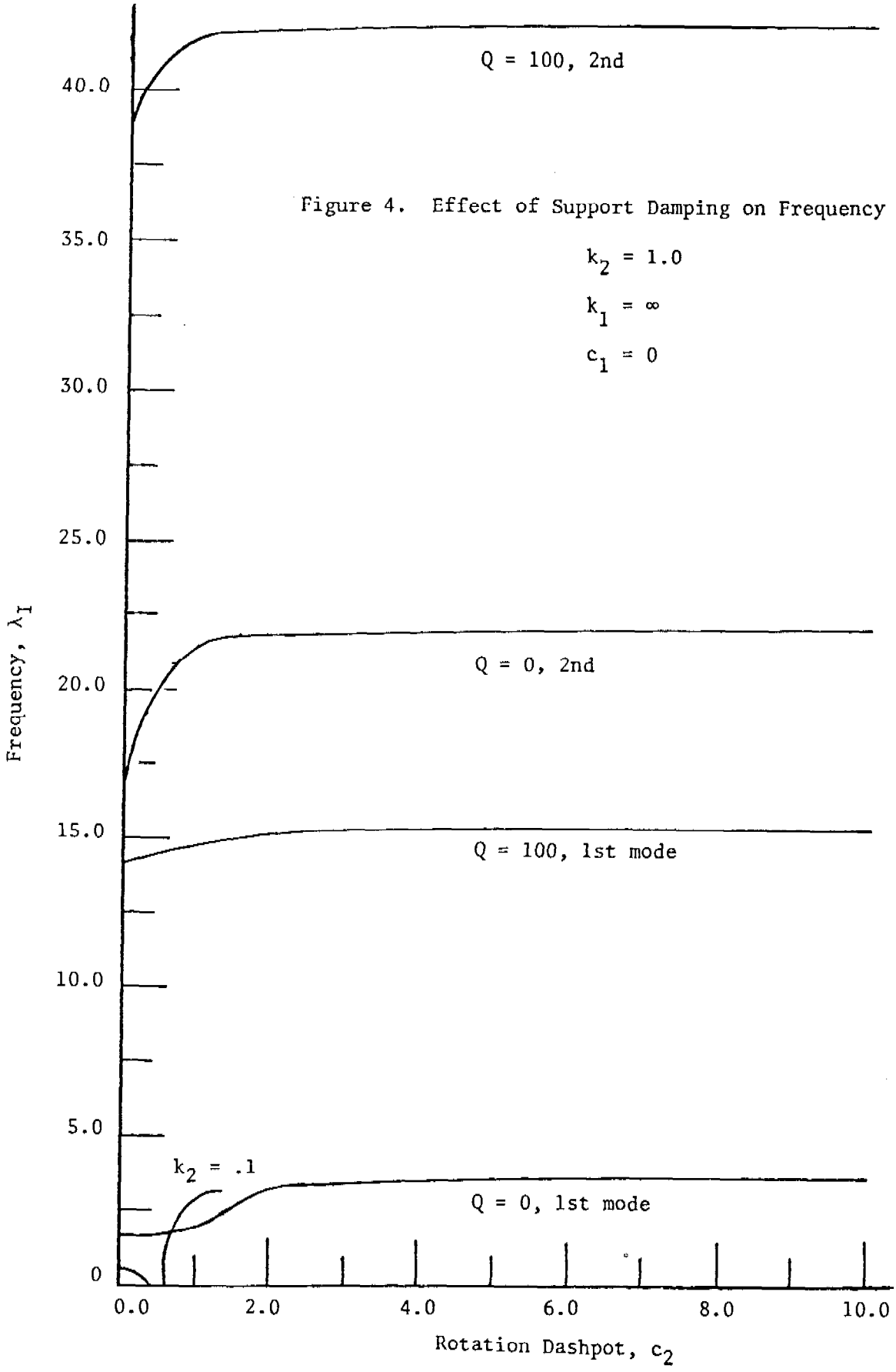


Figure 4. Effect of Support Damping on Frequency

$k_2 = 1.0$

$k_1 = \infty$

$c_1 = 0$

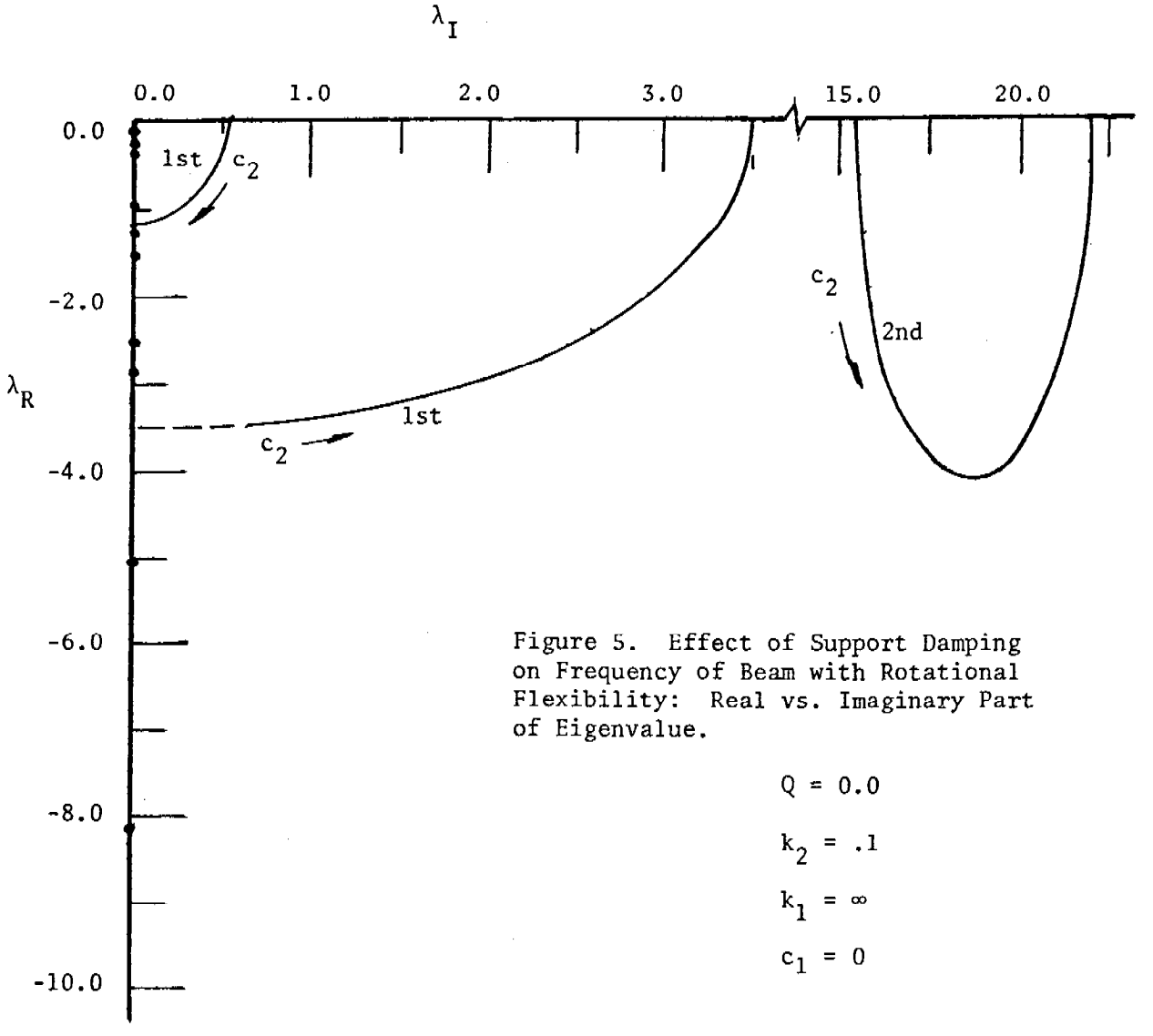


Figure 5. Effect of Support Damping on Frequency of Beam with Rotational Flexibility: Real vs. Imaginary Part of Eigenvalue.

$Q = 0.0$

$k_2 = .1$

$k_1 = \infty$

$c_1 = 0$

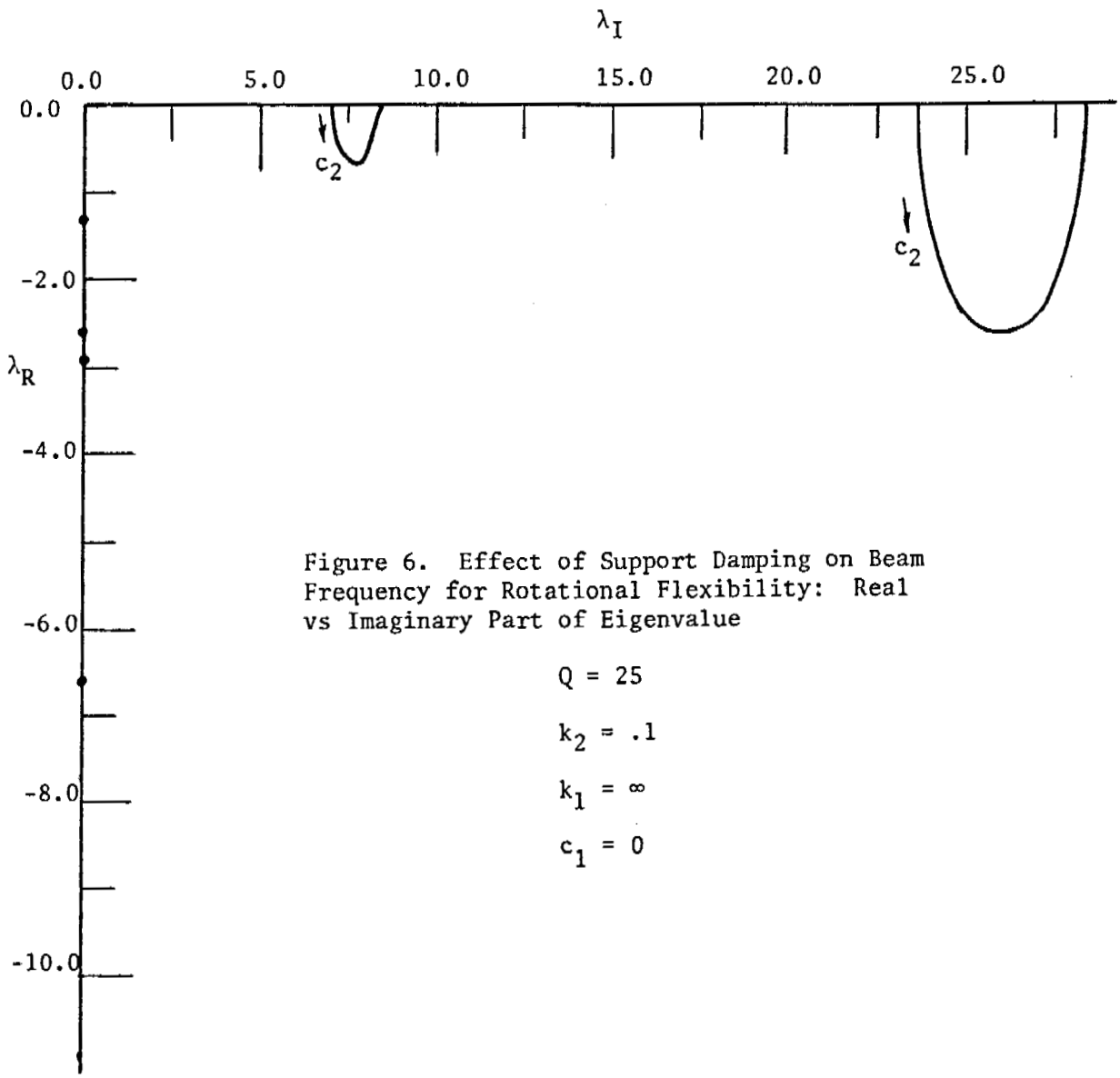


Figure 6. Effect of Support Damping on Beam Frequency for Rotational Flexibility: Real vs Imaginary Part of Eigenvalue

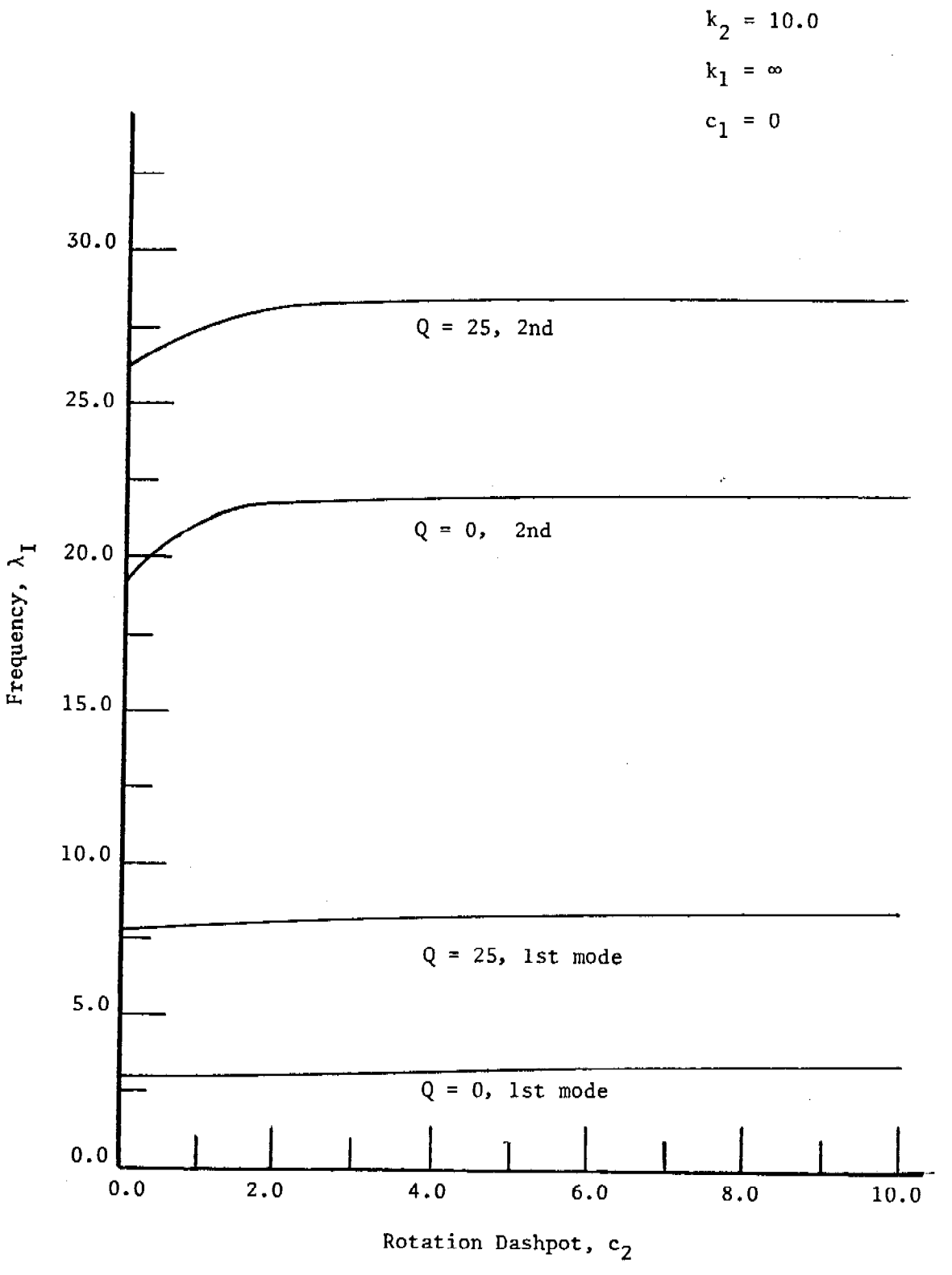
$Q = 25$

$k_2 = .1$

$k_1 = \infty$

$c_1 = 0$

Figure 7. Effect of Support Damping on Frequency



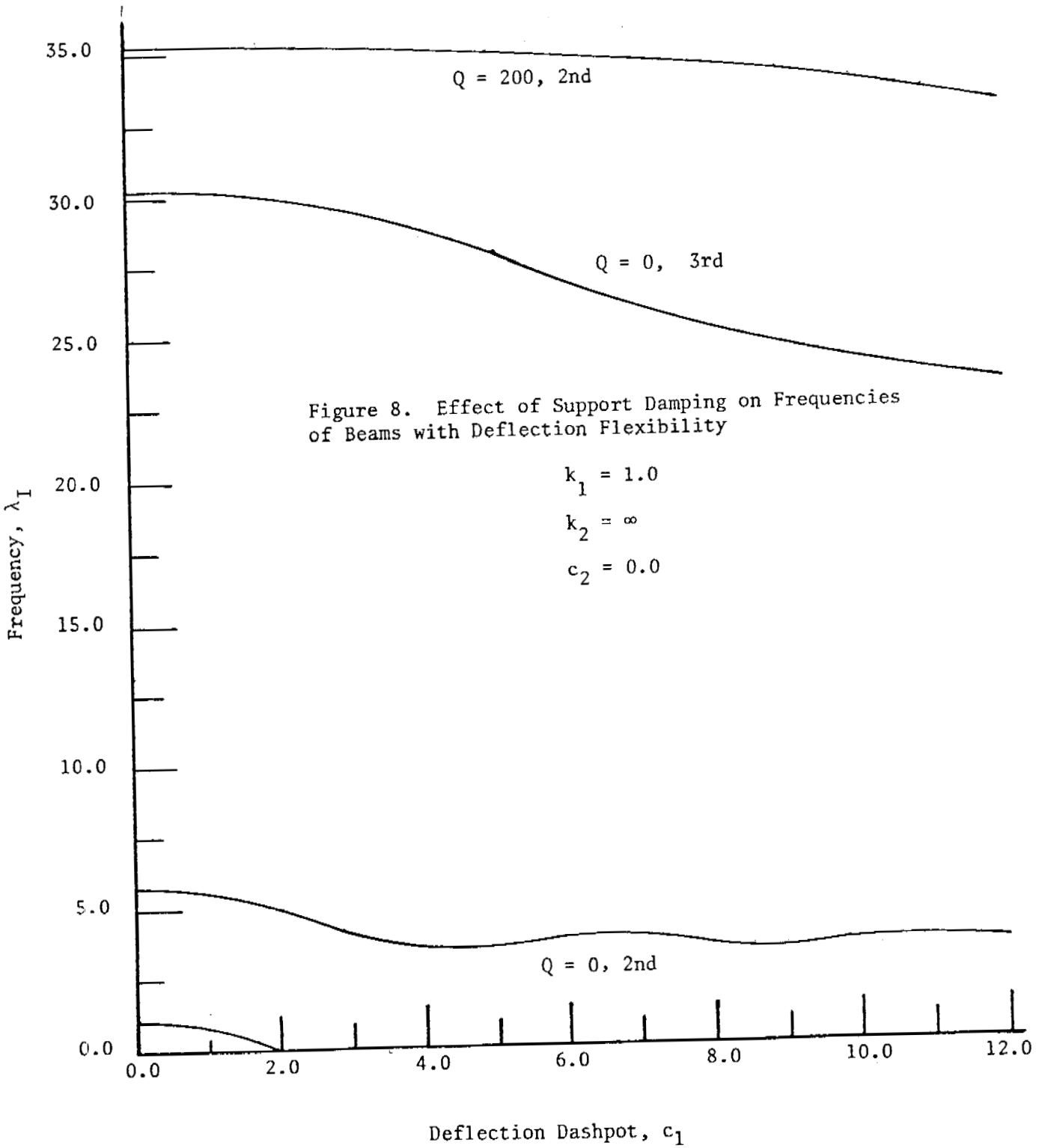


Figure 8. Effect of Support Damping on Frequencies of Beams with Deflection Flexibility

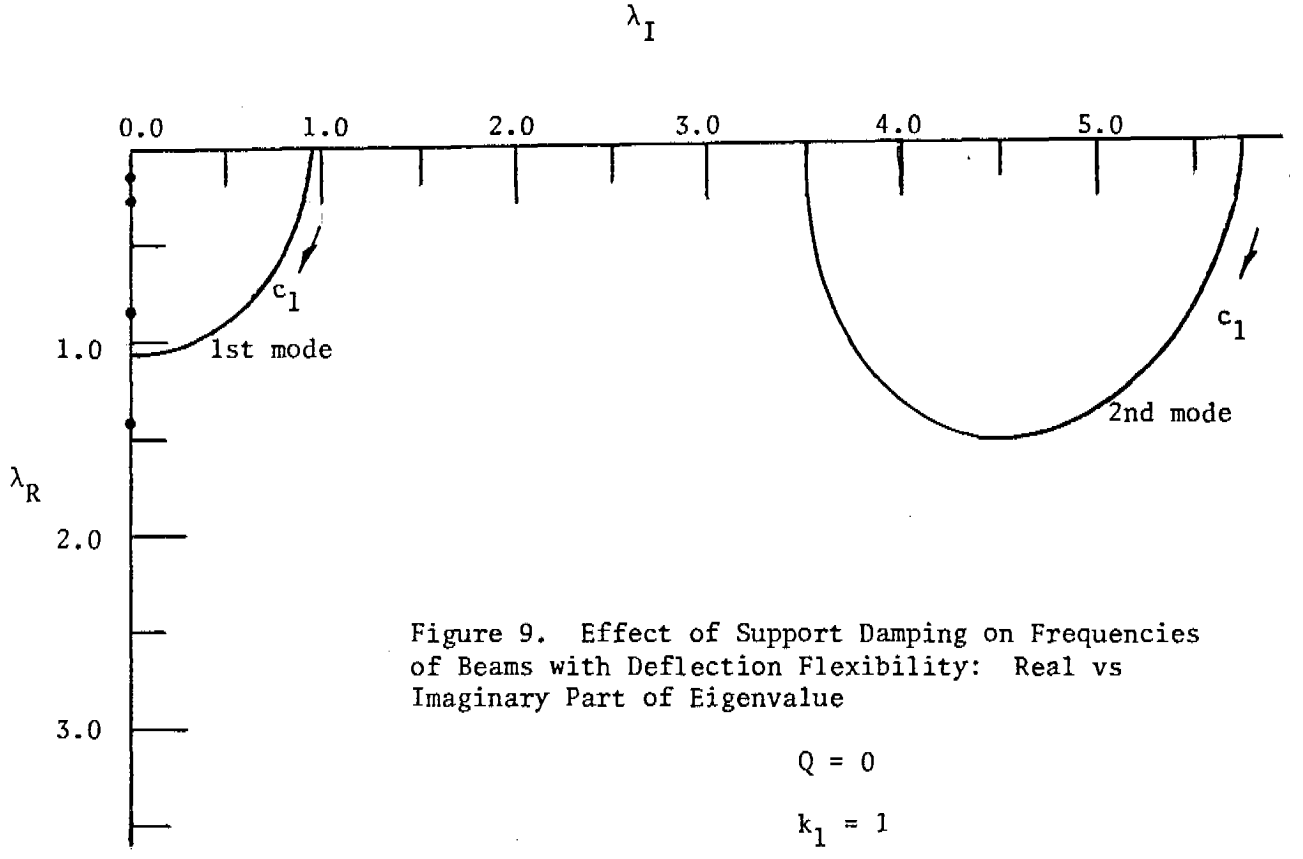


Figure 9. Effect of Support Damping on Frequencies of Beams with Deflection Flexibility: Real vs Imaginary Part of Eigenvalue

$Q = 0$

$k_1 = 1$

$k_2 = \infty$

$c_2 = 0$

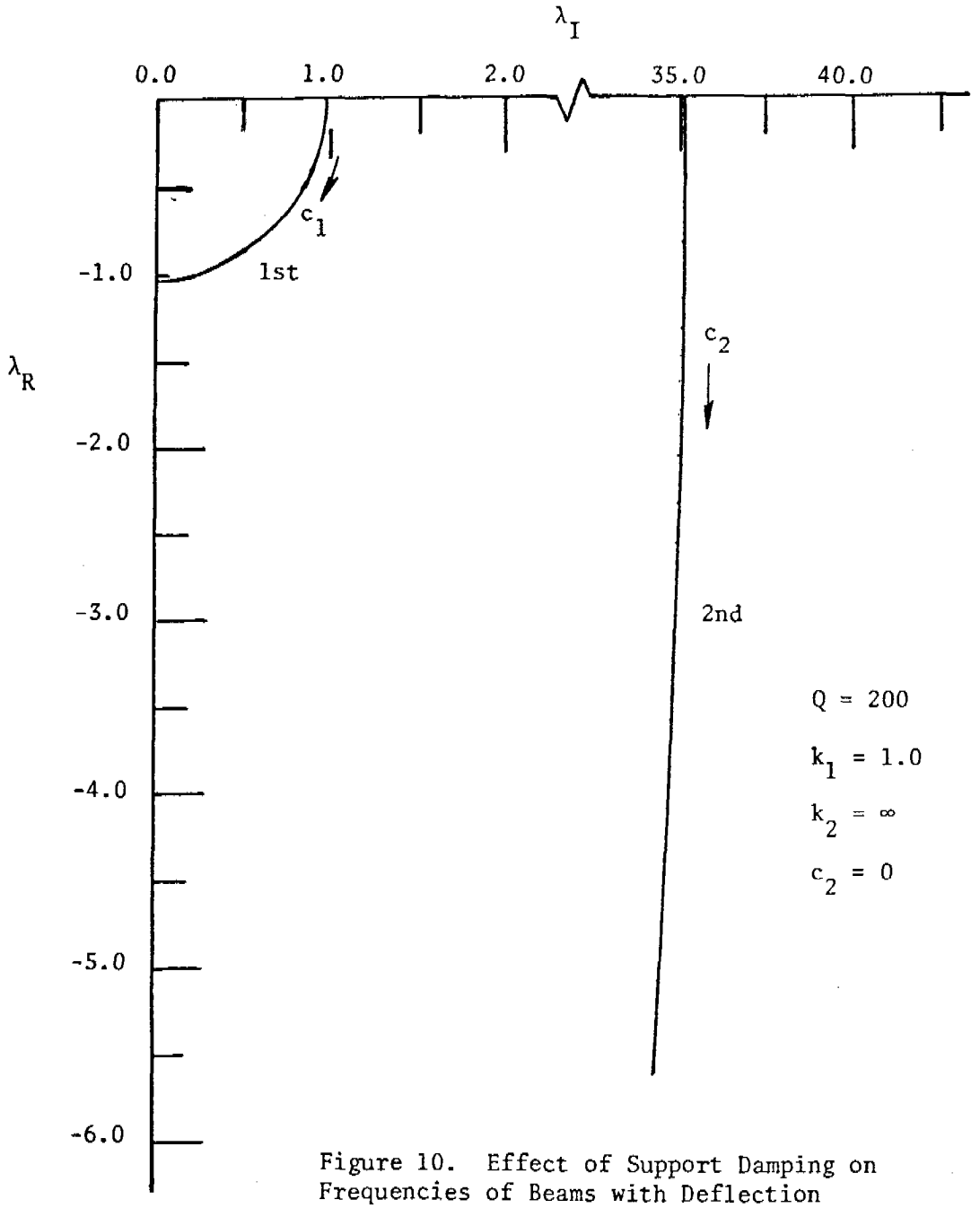


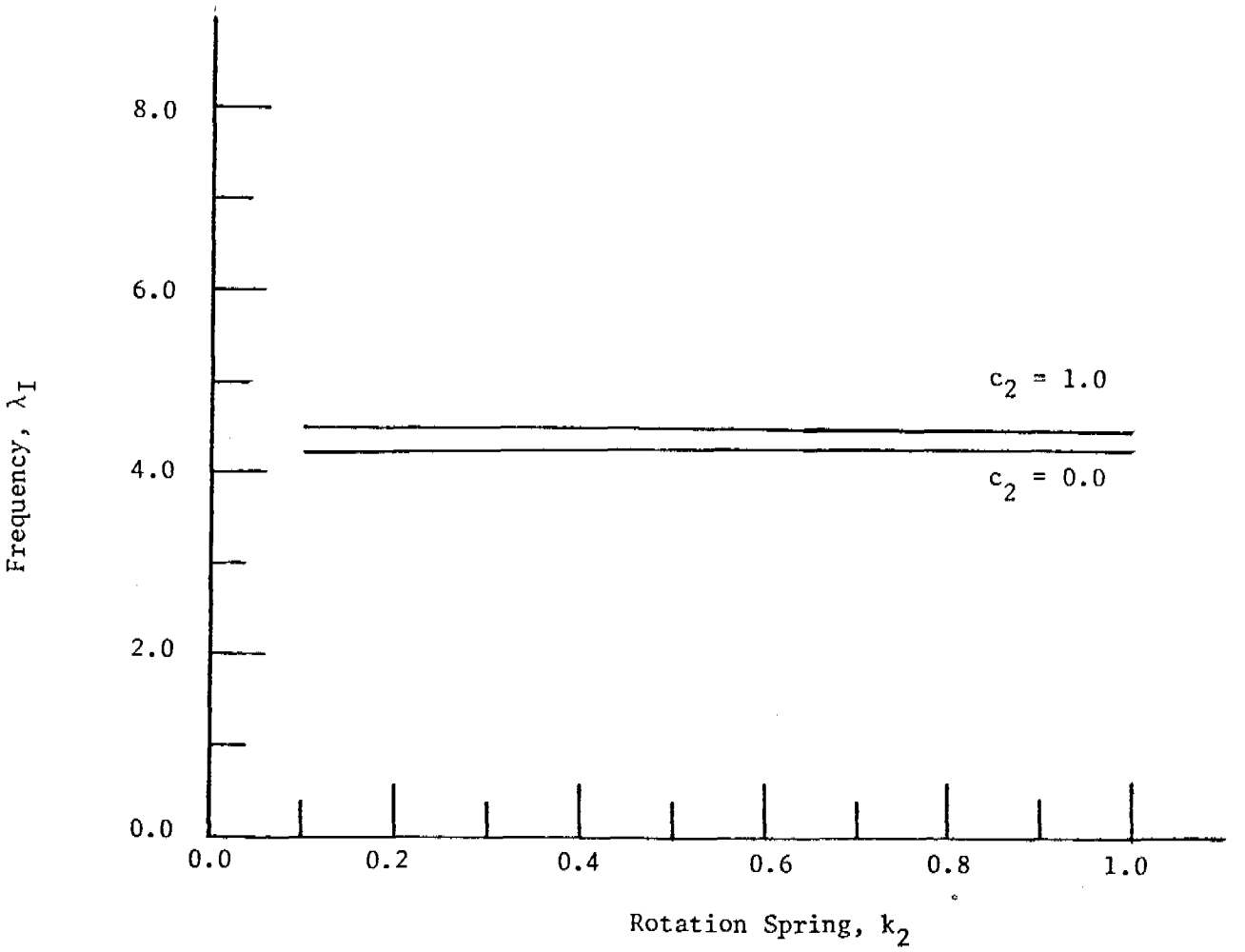
Figure 10. Effect of Support Damping on Frequencies of Beams with Deflection Flexibility: Real vs Imaginary Part of Eigenvalue

Figure 11. Effect of Rotation Spring on Frequencies for Fixed Finite Deflection Spring

$Q = 25.0$

$k_1 = 25.0$

$c_1 = 5.0$



AN EVALUATION PROCEDURE FOR INCOMPLETE GAMMA FUNCTIONS

Walter Gautschi
Mathematics Research Center
University of Wisconsin-Madison
Madison, Wisconsin 53706

ABSTRACT. A computational procedure is developed for evaluating Tricomi's incomplete gamma function $\gamma^*(a,x) = (x^{-a}/\Gamma(a)) \int_0^x e^{-t}t^{a-1}dt$, and the complementary incomplete gamma function $\Gamma(a,x) = \int_x^\infty e^{-t}t^{a-1}dt$, both in the region $x \geq 0$, $-\infty < a < \infty$. Each of these functions can be obtained from the other by means of simple interrelationships. The choice of primary function, i.e., the function to be computed first, will be dictated by considerations of numerical stability and computational convenience. In the strip $0 \leq x \leq 1.5$, $-\infty < a < \infty$, the choice goes to $\gamma^*(a,x)$, which is easily evaluated by Taylor's series. This entails certain difficulties for $\Gamma(a,x)$, when a is very close (or equal) to a nonpositive integer, but these can be dealt with by a careful analysis of the limit behavior of $\Gamma(a,x)$ as $a \to -m$, $m = 0, 1, 2, \ldots$. The function $\gamma^*(a,x)$ continues to serve as primary function in the region $a \geq x \geq 1.5$, where it can be effectively evaluated by a continued fraction due to Perron. In the remaining region $x \geq 1.5$, $a < x$, the primary function is taken to be $\Gamma(a,x)$, and is evaluated by a classical continued fraction of Legendre.

The complete paper is available as MRC Technical Summary Report #1717, February 1977.

A METHOD OF EVALUATING LAPLACE TRANSFORMS WITH SERIES OF COMPLETE OR INCOMPLETE BETA FUNCTIONS

Alexander S. Elder
Emma M. Wineholt
Propulsion Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. In a previous paper factorial series were used to calculate ordinary and modified Bessel functions of the second kind. In the present paper the factorial series is generalized so that Laplace integrals in which the integrand has a branch point at the origin are represented by a series of beta functions. To effect the required transformation, formulas for calculating Stirling numbers of fractional order were derived; these were used in the same manner as the Stirling numbers of integer order are used to calculate the coefficients of a factorial series. Formulas for calculating $K_o(x)$ and $K_1(x)$ have been derived and programmed, using these modified Stirling numbers. Formulas for calculating $I_o(x)$ and $I_1(x)$ have been derived and programmed using series of incomplete beta functions in a similar algorithm. Results for $K_o(x)$ and $K_1(x)$ agree to thirteen significant figures when $x>8$ and for $I_o(x)$ and $I_1(x)$ when $x>15$. The modified Stirling numbers increase very slowly with order and index since gamma functions do not occur in the definition. Consequently no problems with overrun of the electronic computer occurred during the course of the calculations.

1. INTRODUCTION. Factorial series for Bessel functions, confluent hypergeometric functions, and certain other special functions can be used

to check the accuracy of calculations for these functions provided the argument is not too small. Generally, if a function is analytic in the right half plane, including the imaginary axis, and can be represented as a Laplace transform, then a factorial series can be derived which will converge in the right half plane. Buchal and Duffy (1) obtained factorial series for Hankel functions, Coulomb wave functions, and Mathieu functions. These authors also studied the convergence properties of the factorial series for the Hankel functions in considerable detail. Their results showed that factorial series were more accurate than Hankel asymptotic series for the same argument. The analysis was based on Bernoulli polynomials as discussed by Doetsch (2) and Milne - Thomson (3).

The analysis in this paper is based on an algorithm of Wasaw (4) which uses Stirling numbers of the first kind to calculate coefficients for the factorial series. The Stirling numbers of the first kind increase very rapidly with order, eventually obtaining overrun in the electronic computer. Moreover, the factorial series for complex argument is awkward if there is a branch point at the origin. Rosser (5) obtained a generalized factorial series for modified Bessel functions of the second kind by direct manipulation of the Laplace transform and also established the convergence properties. His analysis is quite difficult and requires a separate treatment for each case. In this paper we derive Stirling numbers of the first kind and fractional order, leading to a generalization of Wasaw's algorithm. Overrun in the computer is eliminated by scaling the Stirling numbers of integral and fractional order by omitting the gamma functions in the definition.

A further generalization of factorial series is required if the Laplace integral representing the function is evaluated between finite limits. If a factorial series is regarded as a series of beta functions, it is logical to represent a Laplace integral with finite limits in terms of a series of incomplete beta functions. The new series obtained by this method is convergent even though the corresponding factorial series may be divergent.

2. MODIFIED STIRLING NUMBERS OF THE FIRST KIND AND FRACTIONAL ORDER

Stirling numbers of the first kind are defined as coefficients which occur when a factorial is expanded into a polynomial (6), (7).

$$x\ (x-1)\ (x-2)\ \ldots\ (x-n+1) = \sum_{m=0}^{n} S_n^{(m)}\ x^m \tag{1}$$

Clearly m must be an integer in the above equation. However, a generating function involving logarithms is not subject to this restriction.

$$\left[\ell n\ (1+x)\right]^m = m!\ \sum_{n=m}^{\infty} S_n^{(m)}\ \frac{x^n}{n!}\ ,\ |x|<1 \tag{2}$$

or

$$\left[\ell n\ (1+x)\right]^m = \Gamma(m+1)\ \sum_{n=m}^{\infty} S_n^{(m)}\ \frac{x^n}{\Gamma(n+1)}\ . \tag{3}$$

In order to avoid the gamma function and non-integral indices we define $W_n^{(\nu)}$ by the equation

$$\left[\ell n\ (1+x)\right]^{\nu} = \sum_{k=o}^{\infty} W_k^{(\nu)}\ x^{\nu+k} \tag{4}$$

where ν may take on fractional as well as integral values. When the $W_k^{(\nu)}$ have been calculated, Stirling numbers of the first kind may be obtained from the following equations.

$$S_n^{(m)} = \Gamma(n+1)\ W_{n-\nu}^{(m)} \quad / \Gamma(m+1) \tag{5}$$

where $\nu = m$ (6)

$k = n - \nu$ (7)

To find $W_o^{(\nu)}$, divide Eq (4) by x^{ν} and evaluate the limit of each side of resulting equations as $x \to 0$. We obtain

$$W_o^{(\nu)} = 1. \tag{8}$$

We can also prove that

$$W_k^{(o)} = 0 \qquad , k \geq 1 \tag{9}$$

and $W_o^{(o)} = 1.$ (10)

We now derive a sequence of triangular equations for calculating $W_1^{(\nu)}$, $W_2^{(\nu)}$, ..., $W_\ell^{(\nu)}$ in turn. Let

$$y(x) = [\ell n\ (1+x)]^{\nu} \tag{11}$$

and $$u(x) = (x+1)\ \ell n\ (1+x). \tag{12}$$

Then $$u(x)\ y'(x) = \nu y(x) \tag{13}$$

But $$y(x) = \sum_{k=o}^{\infty} W_k^{(\nu)}\ x^{\nu+k}$$

$$y'(x) = \sum_{k=o}^{\infty} W_k^{(\nu)}\ (\nu+k)\ x^{\nu+k-1} \tag{14}$$

$$u(x) = x + \frac{1}{2}\ x^2 - \frac{1}{6}\ x^3 + \frac{1}{12}\ x^4 - \ldots + (-1)^k\ \frac{x^k}{k(k-1)}\ ,\ k>1 \tag{15}$$

On inserting these series into Eq (13), carrying out the indicated multiplication and equating coefficients of like powers of x on each side of the resulting equation, we find

$$W_o^{(\nu)} = W_o^{(\nu)}$$

$$\frac{\nu+1}{1}\ W_1^{(\nu)} + \frac{1}{2}\ \nu = \nu\ W_1^{(\nu)}$$

$$\frac{\nu+2}{1}\ W_2^{(\nu)} + \frac{\nu+1}{2}\ W_1^{(\nu)} - \frac{\nu}{6} = \nu\ W_2^{(\nu)}$$

On re-arranging these equations we find

$$W_1^{(\nu)} + \frac{1}{2}\nu = 0 \tag{16}$$

$$2\,W_2^{(\nu)} + \frac{\nu+1}{2}\,W_1^{(\nu)} - \frac{\nu}{6} = 0 \tag{17}$$

$$3\,W_3^{(\nu)} + \frac{\nu+2}{2}\,W_2^{(\nu)} - \frac{\nu+1}{6}\,W_1^{(\nu)} + \frac{\nu}{12} = 0 \tag{18}$$

and in general

$$\ell W_\ell^{(\nu)} + \sum_{k=0}^{\ell-1} (-1)^{\ell+k+1} \frac{\nu+k}{(\ell-k)(\ell-k+1)}\,W_k^{(\nu)} = 0\ . \tag{19}$$

To find a recurrence formula involving different orders, differentiate Eq (4) with respect to x and then multiply both sides of the resulting equation by (x+1):

$$\nu\,[\ell n\,(1+x)]^{\nu-1} = \sum_{k=0}^{\infty} W_k^{(\nu)}\,(\nu+k)(x+1)\,x^{\nu+k-1} \tag{20}$$

If we replace ν by $(\nu-1)$ in Eq (4) and multiply both sides of the resulting equation by ν, we find

$$\nu\,[\ell n(1+x)]^{\nu-1} = \sum_{k=0}^{\infty} W_k^{(\nu-1)}\,\nu x^{\nu+k-1} \tag{21}$$

On comparing coefficients of like powers of x in the last two equations we find

$$W_{k+1}^{(\nu)} = [\ \nu\ W_{k+1}^{(\nu-1)} - (\nu+k)\ W_{k}^{(\nu)}\ ] \ /\ [\ \nu+k+1\ ]\ . \tag{22}$$

Hence we can calculate $W_{k+1}^{(\nu+1)}$ from $W_{k+1}^{(\nu)}$ and $W_{k}^{(\nu+1)}$. Higher order numbers can be generated in succession in the same manner. The values of $W_{o}^{(\nu)}$, $W_{1}^{(\nu)}\ldots$, $W_{\ell}^{(\nu)}$ must be calculated from Eq (19) before the recurrence formula given by Eq (22) can be used; a double entry table is finally obtained.

Wasaw uses Schlomlich's definition of factorial coefficients in his development of factorial series (7):

$$x\ (x+1)\ (x+2)\ \ldots\ (x+n-1) = \sum_{m=o}^{n-1} \Gamma_{m}^{n}\ x^{n-m}\ . \tag{23}$$

It follows that the factorial coefficients and Stirling numbers of the first kind are related by the formula

$$\Gamma_{n-m}^{n} = (-1)^{n-m}\ S_{n}^{(m)}\ . \tag{24}$$

We define

$$V_{n-m}^{(m)} = (-1)^{m-n}\ \Gamma(m+1)\ S_{n}^{(m)}\ /\ \Gamma(n+1) \tag{25}$$

and for fractional orders

$$[\ -\ \ell n\ (1-x)\ ]^{\nu} = \sum_{k=o}^{\infty} V_k^{(\nu)}\ x^{\nu+k}. \tag{26}$$

Finally, we obtain the following recurrence formulas in the manner indicated previously:

$$\ell\ V_{\ell}^{(\nu)} - \sum_{k=o}^{\ell-1} \frac{\nu+k}{(\ell-k)(\ell-k+1)}\ V_k^{(\nu)} = 0 \tag{27}$$

$$V_{k+1}^{(\nu)} = [\ \nu\ V_{k+1}^{(\nu-1)} + (\nu+k)\ V_k^{(\nu)}\ ]\ /\ [\nu+k+1] \tag{28}$$

3. GENERALIZED FACTORIAL SERIES

We now derive an extension of Wasaw's algorithm for a Laplace integral to functions with a branch point at the origin. The branch point involves fractional powers, in the same context as Watson's Lemma (8); logarithmic branch points are not considered. Assume

$$F(x) = \int_o^{\infty} f(t)\ e^{-xt}\ dt \tag{29}$$

and let

$$t = -\ \ell n\ (1-u)\ ; \tag{30}$$

then

$$F(x) = \int_0^1 f[-\ell n(1-u)\ ]\ [1-u]^{x-1}du\ . \tag{31}$$

If

$$f(t) = \sum_{n=o}^{\infty} a_n t^{\nu+n},\ \nu>-1,\ 0<t<1 \tag{32}$$

then

$$F(x) = \sum_{n=o}^{\infty} a_n \int_0^1 [-\ell n(1-u)\ ]^{\nu+n}\ [1-u]^{x-1}du. \tag{33}$$

On referring to Eq (26) we see

$$F(x) = \sum_{n=o}^{\infty} \sum_{k=o}^{\infty} a_n V_k^{(\nu+n)} \int_0^1 u^{\nu+k+n}\ (1-u)^{x-1}du\ . \tag{34}$$

Now

$$B\ (\alpha,\beta) = \int_0^1 t^{\alpha-1}\ (1-t)^{\beta-1}dt \tag{35}$$

so that

$$F(x) = \sum_{n=o}^{\infty} \sum_{k=o}^{\infty} a_n\ V_k^{(\nu+n)}\ B\ (k+n+\nu+1,x)\ . \tag{36}$$

We enter the terms in a double entry table; then by summing the diagonals, we obtain the Cauchy sum of the double series (9). Let

$$k+n=\ell \tag{37}$$

and define

$$b_\ell = \sum_{n=o}^{\ell} a_n V_{\ell-n}^{(\nu+n)} \tag{38}$$

then

$$F(x) = \sum_{\ell=o}^{\infty} b_\ell B(\nu+\ell+1,x) \quad . \tag{39}$$

On noting that

$$B(\nu+\ell+1,x) = B(\nu,x) \frac{\nu\ (\nu+1)\ \ldots\ (\nu+\ell)}{(x+\nu)(x+\nu+1)\ldots(x+\nu+\ell)} \tag{40}$$

and using Pochhammer's symbol to represent the factorials, we find

$$F(x) = B(\nu,x) \sum_{\ell=o}^{\infty} b_\ell (\nu)_{\ell+1} / (x+\nu)_{\ell+1} \tag{41}$$

in a formal sense. By analogy with conventional factorial series, the series should converge in a half plane which lies to the right of the imaginary axis. Details of the required analysis will not be considered at this time.

4. ON A SERIES OF INCOMPLETE BETA FUNCTIONS. Since a generalized factorial series is in fact a series of beta functions, it is natural to represent a Laplace integral with finite limits of integration as a series of incomplete beta functions. The lower limit of integration can be taken equal to zero without loss of generality. Assume

$$F(x) = \int_0^\tau e^{-tx} f(t)\, dt, \quad \tau > 0 \tag{42}$$

Let

$$\varepsilon = 1 - e^{-\tau} \tag{43}$$

then

$$F(x) = \int_0^\varepsilon f[-\ell n(1-u)]\, [1-u]^{x-1} du\ . \tag{44}$$

On referring to Eq (26) we find

$$F(x) = \sum_{n=0}^{\infty} \sum_{k=0}^{\infty} a_n V_k^{(\nu+n)} \int_0^\varepsilon u^{\nu+k+n} (1-u)^{x-1} du. \tag{45}$$

Since

$$B_\varepsilon(\alpha,\beta) = \int_0^\varepsilon t^{\alpha-1} (1-t)^{\beta-1} dt \tag{46}$$

we find

$$F(x) = \sum_{n=o}^{\infty} \sum_{k=o}^{\infty} a_n \, V_k^{(\nu+n)} B_{\varepsilon}(\nu+k+n+1,x) \tag{47}$$

or

$$F(x) = \sum_{\ell=o}^{\infty} b_{\ell} \, B_{\varepsilon}(\nu+\ell+1,x) \tag{48}$$

on referring to Eqs (37) and (38).

To compute the incomplete beta function, set

$$\nu+\ell = \delta \tag{49}$$

and use the formula

$$B_{\varepsilon}(\delta+1,x) = B(x,\delta+1) - B_{1-\varepsilon}(x,\delta+1) \tag{50}$$

The beta function on the right side of Eq (50) was expressed in terms of gamma functions, as shown in Eq (51):

$$B(x,\delta+1) = \Gamma(x) \, \Gamma(\delta+1) \, / \, \Gamma(x+\delta+1). \tag{51}$$

The gamma functions were obtained from the subroutine CDLGAM by H. Kuki (10). This subroutine is valid for both real and complex values of x.

The incomplete beta function on the right side of Eq (50) is given by the integral formula

$$B_{1-\varepsilon}(x,\ \delta+1) = \int_0^{1-\varepsilon} t^{x-1}\ (1-t)^{\delta} dt\ . \tag{52}$$

On expanding the binomial factor in the integrand and integrating term by term we find

$$B_{1-\varepsilon}\ (x,\delta+1) = \frac{(1-\varepsilon)^x}{x} - \frac{\delta}{1!}\ \frac{(1-\varepsilon)^{x+1}}{(x+1)}$$

$$+ \frac{\delta(\delta-1)}{2!}\ \frac{(1-\varepsilon)^{x+2}}{(x+2)} - \frac{\delta(\delta-1)(\delta-2)}{3!}\ \frac{(1-\varepsilon)^{x+3}}{(x+3)} + \ldots \tag{53}$$

This series is satisfactory if δ is small, but is subject to round-off error if δ is large and positive. To overcome this difficulty, integrate the right-hand side of Eq (52) repeatedly by parts. We find

$$B_{1-\varepsilon}\ (x,\delta+1) = \frac{(1-\varepsilon)^x}{x}\ \varepsilon^{\delta} + \frac{\delta(1-\varepsilon)^{x+1}}{x(x+1)}\ \varepsilon^{\delta-1}$$

$$+ \frac{\delta(\delta-1)(1-\varepsilon)^{x+2}}{x(x+1)(x+2)}\ \varepsilon^{\delta-2} + \ldots + \frac{\delta(\delta-1)(\delta-2)\ldots(\delta-m)}{x(x+1)(x+2)\ldots(x+m)}\ R_m$$

where

$$R_m = B_{1-\varepsilon}\ (x+m+1,\ \delta-m)\ . \tag{55}$$

We choose m so that

$$1 < \delta - m \leq 2. \tag{56}$$

Then the terms of the series given by Eq (54) are positive when x is real and positive, and consequently the round off error should be small.

The remainder R_m is calculated from Eq (53).

$$R_m = \frac{(1-\varepsilon)^{x+m+1}}{x+m+1} - \frac{(\delta-m-1)(1-\varepsilon)^{x+m+2}}{1!\ (x+m+2)} + \frac{(\delta-m-1)(\delta-m-2)(1-\varepsilon)^{x+m+3}}{2!\ (x+m+3)} - \ldots \tag{57}$$

All the terms of this series after the first term are negative, and decrease rapidly in magnitude. Hence the error in calculating R_m should be small.

5. CALCULATIONS. We calculated the modified Bessel functions $K_o(x)$, $K_1(x)$, $I_o(x)$, and $I_1(x)$ from the integrals:

$$K_o(x) = \frac{\Gamma(\frac{1}{2})\ e^{-x}}{\Gamma(\frac{1}{2})} \int_o^\infty e^{-xt}(t^2+2t)^{-\frac{1}{2}}dt \tag{58}$$

$$K_1(x) = \frac{\Gamma(\frac{1}{2})\ (\frac{1}{2}x)e^{-x}}{\Gamma\left(\frac{3}{2}\right)} \int_o^\infty e^{-xt}(t^2+2t)^{\frac{1}{2}}dt \tag{59}$$

$$I_o(x) = \frac{e^x}{\Gamma(\frac{1}{2})\ \Gamma(\frac{1}{2})} \int_o^2 e^{-xt}(2t-t^2)^{-\frac{1}{2}}dt \tag{60}$$

$$I_1(x) = \frac{(\frac{1}{2}x)\ e^x}{\Gamma(\frac{1}{2})\ \Gamma\left(\frac{3}{2}\right)} \int_o^2 e^{-xt}(2t-t^2)^{\frac{1}{2}}dt. \tag{61}$$

On referring to Eqs (29) and (32), we see that

$$\nu = -\tfrac{1}{2} \tag{62}$$

in Eqs (58) and (60), and

$$\nu = \tfrac{1}{2} \tag{63}$$

in Eqs (59) and (61). Hence the modified Stirling numbers $V_k^{-\frac{1}{2}+m}$ and $V_k^{\frac{1}{2}+m}$ are required for the coefficients of the factorial series. A short table of these numbers is given in Table 1.

Next, the coefficients a_n for the series expansions of $(1+\frac{1}{2}t)^{-\frac{1}{2}}$, $(1-\frac{1}{2}t)^{-\frac{1}{2}}$, $(1+\frac{1}{2}t)^{\frac{1}{2}}$, and $(1-\frac{1}{2}t)^{\frac{1}{2}}$ were calculated from the appropriate recurrence formulas, as shown in Table 2.

The coefficients b_ℓ were calculated from Eq (38) for each of the four cases listed above. In addition, the partial sums

$$C_m = \sum_{\ell=0}^{m} b_\ell \tag{64}$$

were calculated in order to study the convergence of the factorial series. The series $\sum_m^\infty C_m$ must converge if the corresponding factorial series is to converge. This condition is apparently violated for the functions $I_0(x)$ and $I_1(x)$, which shows why the factorial series for these functions apparently diverged. These results are given in Table 3.

TABLE 1. MODIFIED STIRLING NUMBERS, $V_K^{(\nu)}$

ν \ K	0	1	2	3	4	5	
-.5	1.000	-.250	-.073	-.039	-.026	-.019	...
.5	1.000	.250	.135	.091	.068	.054	...
1.5	1.000	.750	.594	.492	.421	.369	...
2.5	1.000	1.250	1.302	1.289	1.253	1.209	...
3.5	1.000	1.750	2.260	2.607	2.844	3.007	...
4.5	1.000	2.250	3.469	4.570	5.538	6.380	...
5.5	1.000	2.750	4.927	7.305	9.742	12.156	...
6.5	1.000	3.250	6.635	10.935	15.924	21.410	...
7.5	1.000	3.750	8.594	15.586	24.616	35.493	...
8.5	1.000	4.250	10.802	21.383	36.411	56.061	...
9.5	1.000	4.750	13.260	28.451	51.966	85.112	...
10.5	1.000	5.250	15.969	36.914	71.999	125.010	...
11.5	1.000	5.750	18.927	46.898	97.291	178.523	...
12.5	1.000	6.250	22.135	58.529	128.687	248.849	...
13.5	1.000	6.750	25.594	71.930	167.092	339.653	...
14.5	1.000	7.250	29.302	87.227	213.476	455.091	...
15.5	1.000	7.750	33.260	104.544	268.869	599.846	...
16.5	1.000	8.250	37.469	124.008	334.365	779.160	...
17.5	1.000	8.750	41.927	145.742	411.121	998.863	...
18.5	1.000	9.250	46.635	169.872	500.356	1265.403	...
19.5	1.000	9.750	51.594	196.523	603.349	1585.880	...
20.5	1.000	10.250	56.802	225.820	721.447	1968.078	...
21.5	1.000	10.750	62.260	257.888	856.053	2420.492	...
22.5	1.000	11.250	67.969	292.852	1008.638	2952.363	...
23.5	1.000	11.750	73.927	330.836	1180.733	3573.708	...

TABLE 2. BINOMIAL COEFFICIENTS FOR SERIES EXPANSION

N	$(1+.5*T)^{-.5}$	$(1+.5*T)^{.5}$	$(1-.5*T)^{-.5}$	$(1-.5*T)^{.5}$
1	C.1000CCCCCCE 01	0.1CCCCCCCCCE 01	0.1000000000E 01	0.1000000C0CE 01
2	C.2500CCCCCCE 0C	-0.25CCCCCCCCE 00	-0.2500000000E 00	0.2500000C0CE 00
3	C.9375CCCCCCE-01	-0.3125CCCCCCE-01	0.9375000000E-01	-0.3125000000E-01
4	C.3906250CCCE-01	-0.78125CCCCCE-02	-0.3906250000E-01	0.7812500000E-02
5	C.1708984375E-01	-0.24414C6250E-02	0.1708984375E-01	-0.244140625CE-02
6	C.7690429688E-02	-0.8544921875E-03	-0.7690429688E-02	0.8544921875E-03
7	C.3524780273E-02	-0.3204345703E-03	0.3524780273E-02	-0.3204345703E-03
8	C.1636505127E-02	-0.1258850098E-03	-0.1636505127E-02	0.1258850098E-03
9	C.7671117783E-03	-0.5114078522E-04	0.7671117783E-03	-0.5114078522E-04
10	C.3622472286E-03	-0.213C866051E-04	-0.3622472286E-03	0.213C866051E-04
11	C.1720674336E-03	-0.9056180716E-05	0.1720674336E-03	-0.9056180716E-05
12	C.8212C9331E-04	-0.3910623491E-05	-0.8212309331E-04	0.3910623491E-05
13	C.3935C64888E-04	-0.171C897777E-05	0.3935064888E-04	-0.1710897777E-05
14	C.1891858119E-04	-0.7567432476E-06	-0.1891858119E-04	0.756743247CE-06
15	C.9121458788E-05	-0.3378318070E-06	0.9121458788E-05	-0.3378318070E-06
16	C.44087C5081E-05	-0.1520243131E-06	-0.4408705081E-05	0.1520243131E-06
17	C.2135466524E-05	-0.6888601689E-07	0.2135466524E-05	-0.6888601689E-07
18	C.1036329342E-05	-0.3140391946E-07	-0.1036329342E-05	0.3140391946E-07
19	C.5037712C81E-06	-0.1439346309E-07	0.5037712081E-06	-0.1439346309E-07
20	C.2452570355E-06	-0.6628568527E-08	-0.2452570355E-06	0.6628568527E-08
21	C.1195628048E-06	-0.3065712944E-08	0.1195628048E-06	-0.3065712944E-08
22	C.5835803568E-07	-0.1423366724E-08	-0.5835803568E-07	0.1423366724E-08
23	C.2851585834E-07	-0.6631594964E-09	0.2851585834E-07	-0.6631594964E-09
24	C.1394797419E-07	-0.3099549820E-09	-0.1394797419E-07	0.309954982CE-09
25	C.6828695697E-08	-0.1452913978E-09	0.6828695697E-08	-0.1452913978E-09

TABLE 3. PARTIAL SUMS, C_M

M	C_M (FOR $I_0(X)$)	C_M (FOR $I_1(X)$)	C_M (FOR $K_0(X)$)	C_M (FOR $K_1(X)$)
1	0.1000000000E 01	0.1000000000E 01	0.1000000000E 01	0.1000000000E 01
2	0.0000000000E 00	0.0000000000E 00	-0.5000000000E 00	0.5000000000E 00
3	0.8333333333E-01	-0.8333333333E-01	-0.4166666667E-01	0.2916666667E 00
4	0.1041666667E 00	-0.1041666667E 00	-0.4166666667E-01	0.2083333333E 00
5	0.1187500000E 00	-0.1180555556E 00	-0.2447916667E-01	0.1616319444E 00
6	0.1329861111E 00	-0.1156250000E 00	-0.1814236111E-01	0.1320312500E 00
7	0.1483775628E 00	-0.1183139054E 00	-0.1392712467E-01	0.1115689071E 00
8	0.1655691964E 00	-0.1208354001E 00	-0.1121135086E-01	0.9658668155E-01
9	0.1849887015E 00	-0.1235905120E 00	-0.9297019245E-02	0.8514439389E-01
10	0.2070163777E 00	-0.1267699119E 00	-0.7891392514E-02	0.7612111033E-01
11	0.2320438054E 00	-0.1304793276E 00	-0.6820430172E-02	0.6882367917E-01
12	0.2605000552E 00	-0.1347898414E 00	-0.5981169633E-02	0.6280056847E-01
13	0.2928672853E 00	-0.1397599840E 00	-0.5308128754E-02	0.5774504858E-01
14	0.3296929445E 00	-0.1454462447E 00	-0.4757997793E-02	0.5344147367E-01
15	0.3716013404E 00	-0.1519085092E 00	-0.4301053015E-02	0.4973381885E-01
	.			
	.			
	.			
85	0.4486264325E 04	-0.2132930456E 03	-0.4049428750E-03	0.8466639314E-02
86	0.5158581027E 04	-0.2421390452E 03	-0.3988377918E-03	0.8367300496E-02
87	0.5932950354E 04	-0.2749495588E 03	-0.3928963534E-03	0.8270263458E-02
88	0.6823459192E 04	-0.3122761345E 03	-0.3871123162E-03	0.8175449148E-02
89	0.7848108406E 04	-0.3547478056E 03	-0.3814797435E-03	0.8082782095E-02
90	0.9027163686E 04	-0.4030820658E 03	-0.3759929875E-03	0.7992190200E-02

The beta functions and incomplete beta functions required in the series expansions for $K_o(x)$, $K_1(x)$, $I_o(x)$, and $I_1(x)$ were calculated from formulas discussed previously. Sample tabulations are shown in Table 4. Finally, the modified Bessel functions were calculated for a limited range of variables. These results are shown in Table 5.

6. RESULTS AND CONCLUSIONS. These series expansions in terms of beta functions and incomplete beta functions were derived in order to check the accuracy of our Bessel function subroutine (11) (12) with independent calculations. The error analysis of our subroutine by theoretical methods would be very difficult, especially for the section involving continued fractions. Hence computational efficiency is a secondary consideration for the new series expansions. Addressing the comment of the reviewer*, we believe our algorithm for the incomplete beta function is as efficient as the continued fractions of Segun (6) when x is large, as Eq (54) can then be used without the remainder. We have not made any specific comparison for small values of x.

Since factorial series are a method of summing certain asymptotic series, they are most effective for large and moderately large values of the argument. The convergence is slow when x is small, so that an excessive number of terms is required. Round-off error may occur in the coefficients b_ℓ and in summing the series. Alternate methods of calculating Bessel functions, such as quadratures, are required when x is small. Subroutines used for checking should not use continued fractions or other procedures used in the subroutine, to insure the calculations are in fact independent.

* See ACKNOWLEDGMENTS

TABLE 4. INCOMPLETE AND COMPLETE BETA FUNCTIONS

X = 15. ε = .86466

$\nu+\ell+1$	INCOMPLETE	COMPLETE
.50	0.4614745534E C0	0.4614745534E C0
1.50	0.1488627592E-01	0.1488627592E-01
2.50	0.1353297811E-02	0.1353297811E-02
3.50	0.1933282586E-03	0.1933282587E-03
4.50	0.3657561650E-04	0.3657561650E-04
5.50	0.8440526882E-05	0.8440526885E-05
6.50	0.2264531600E-05	0.2264531603E-05
7.50	0.6846258310E-06	0.6846258336E-06
8.50	0.2282086089E-06	0.2282086112E-06
9.50	0.8254353824E-07	0.8254354022E-07
10.50	0.3200667713E-07	0.3200667886E-07
11.50	0.1317921920E-07	0.1317922071E-07
12.50	0.5719283140E-08	0.5719284458E-08
13.50	0.2599673602E-08	0.2599674753E-08
14.50	0.1231423877E-08	0.1231424883E-08
15.50	0.6052757576E-09	0.6052766375E-09
16.50	0.3075988334E-09	0.3075996027E-09
17.50	0.1611229288E-09	0.1611236014E-09
18.50	0.8675827411E-10	0.8675886229E-10
19.50	0.4791109612E-10	0.4791161052E-10
20.50	0.2708002559E-10	0.2708047551E-10
21.50	0.1563762750E-10	0.1563802107E-10
22.50	0.9211092752E-11	0.9211437068E-11
23.50	0.5526560982E-11	0.5526862241E-11
24.50	0.3373275674E-11	0.3373539290E-11

TABLE 5. MODIFIED BESSEL FUNCTIONS

X	$I_0(X)$	$I_1(X)$	$K_0(X)$	$K_1(X)$	
15.	0.3396493733E 06	0.3281249220E 06	0.9819536482E-07	0.1014172937E-06	IBF
15.	0.3396493733E 06	0.3281249220E 06	0.9819536482E-07	0.1014172937E-06	SUB
20.	0.4355828256E 08	0.4245497339E 08	0.5741237815E-09	0.5883057970E-09	IBF
20.	0.4355828256E 08	0.4245497339E 08	0.5741237815E-09	0.5883057970E-09	SUB
25.	0.5774566066E 10	0.5657865130E 10	0.3464161562E-11	0.3532778073E-11	IBF
25.	0.5774566066E 10	0.5657865130E 10	0.3464161562E-11	0.3532778073E-11	SUB
30.	0.7816722978E 12	0.7685320389E 12	0.2132477496E-13	0.2167732002E-13	IBF
30.	0.7816722978E 12	0.7685320389E 12	0.2132477496E-13	0.2167732002E-13	SUB
35.	0.1073388185E 15	0.1057941261E 15	0.1331035149E-15	0.1349917834E-15	IBF
35.	0.1073388185E 15	0.1057941261E 15	0.1331035149E-15	0.1349917834E-15	SUB
40.	0.1489477479E 17	0.1470739616E 17	0.8392811000E-18	0.8497131955E-18	IBF
40.	0.1489477479E 17	0.1470739616E 17	0.8392811000E-18	0.8497131955E-18	SUB
45.	0.2083414075E 19	0.2060133462E 19	0.5333456123E-20	0.5392394594E-20	IBF
45.	0.2083414075E 19	0.2060133462E 19	0.5333456123E-20	0.5392394594E-20	SUB
50.	0.2932553784E 21	0.2903078590E 21	0.3410167750E-22	0.3444102227E-22	IBF
50.	0.2932553784E 21	0.2903078590E 21	0.3410167750E-22	0.3444102227E-22	SUB
55.	0.4148785561E 23	0.4110898645E 23	0.2191310218E-24	0.2211142272E-24	IBF
55.	0.4148785561E 23	0.4110898645E 23	0.2191310218E-24	0.2211142272E-24	SUB
60.	0.5894077056E 25	0.5844751588E 25	0.1413897841E-26	0.1425632027E-26	IBF
60.	0.5894077056E 25	0.5844751588E 25	0.1413897841E-26	0.1425632027E-26	SUB
65.	0.8403039846E 27	0.8338148547E 27	0.9154467321E-29	0.9224619528E-29	IBF
65.	0.8403039846E 27	0.8338148547E 27	0.9154467321E-29	0.9224619528E-29	SUB

The generalized factorial series for $K_o(x)$ and $K_1(x)$ yield accurate numerical values when x is only moderately large and the Hankel asymptotic series is not sufficiently accurate. On the other hand, the new series for $I_o(x)$ and $I_1(x)$ have the same range of accuracy as the Hankel asymptotic series, and do not offer any computational advantage. This is probably due to the apparent divergence of the series for the partial sums of b_ℓ.

The generalized factorial series for $K_o(x)$ and $K_1(x)$ are being extended to the complex plane. Programming of generalized factorial series for the ordinary Bessel functions is in progress. Alternate methods of calculating $I_n(x)$ are also being considered, as the results obtained in this paper fell short of our expectations.

The modified Stirling numbers as defined in this paper are more useful for computations involving Wasaw's algorithm than the original Stirling numbers, as problems arising from very large numbers and overrun of the computer registers are entirely avoided. The method of scaling employed in this paper, which merely involves the omission of gamma functions in the definition of Stirling numbers is more effective than the method of scaling used in previous paper by the authors (13).

ACKNOWLEDGMENTS

We acknowledge the assistance of Prof. J.B. Rosser, Mathematics Research Center, during the early stages of this analysis. Dr. M.A. Weinberger, Director of Mathematics and Statistics, Operational Research and Analysis Establishment, Department of National Defence, Canada, noted several

errors during the course of the discussion. Mr. E.L. Leese, Dr. Weinberger's predecessor, reviewed the analysis in detail and corrected several formulas. Dr. R.N. Buchal made several helpful suggestions and later provided a copy of his report.

REFERENCES

1. Robert N. Buchal and Gerald Duffy, "Factorial Series Representations as Check Solutions in the Development of Special Function Subroutines," Argonne National Laboratory Technical Memorandum No. 200, May 1970.

2. Gustav Doetsch, Hanbuch Der Laplace-Transformation, Band II, Birkhauser Verlag, Bosel and Stuggart, 1955, pages 201-232.

3. L.M. Milne - Thomson, The Calculus of Finite Differences, Macmillan and Co., Limited, London, 1933.

4. Wolfgang Wasaw, Asymptotic Expansions for Ordinary Differential Equations, Interscience Publishers, a Division of John Wiley & Sons, Inc., New York, pages 323-331.

5. J. Barkley Rosser, "Factorial Expansions for Certain Bessel Functions", Mathematics Research Center Summary Report No. 1572, November 1975.

6. M. Abramowitz and L.A. Stegun, Handbook of Mathematical Functions, Graphs, and Mathematical Tables. National Bureau of Standards, Washington, D.C., No. 55, Applied Mathematics Series, June 1964. See Formulas 24.1.3, page 824, for discussion of Stirling numbers. Continued fractions for the incomplete beta functions are given in Formulas 26.5.8, page 944.

7. Charles Jordan, Calculus of Finite Differences, Rottig and Romwalter, Sapron, Hungary, 1939. Stirling numbers are discussed in Chapter 4, pages 142-229.

8. E.T. Copson, An Introduction to the Theory of Functions of a Complex Variable, Oxford, 1935, pages 218-219.

9. Konrad Knapp, Infinite Sequences and Series, Dover Publications, New York, 1956, pages 82-85.

10. Hirondo Kuki, Communications of the ACM, Vol 15, No 4, April 1972. The subroutine CDLGAM was extracted from Algorithm 421.

11. Alexander S. Elder, "Formulas for Calculating Bessel Functions of Integral Order and Complex Argument," BRL Report No. 1423, November, 1968.

12. K.L. Zimmerman, A.S. Elder and A.K. Depue, "User's Manual for the BRL Subroutine to Calculate Bessel Functions of Integral Order and Complex Argument," BRL Report being published.

13. A.S. Elder and E.M. Wineholt, "Factorial and Hadamard Series for Bessel Functions of Orders Zero and One", Transactions of the Twenty-Second Conference of Army Mathematicians, Army Research Office, 1977, pp 277-287.

APPROXIMATION OF IRREGULAR SURFACES

Helmut M. Sassenfeld
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico 88002

ABSTRACT

A method is outlined to approximate irregular (empirical) surfaces $z = f(x,y)$ in the real domain, provided z is single-valued when the inverse functions x and y may be multi-valued, using sets of quadratic and linear expressions for constant and variable z values and/or superposition of simple analytically described surfaces. The parameters of the approximation elements are geometrically identifiable and so easily obtained from graphs or numerical values of z. A procedure is given to interpolate between the approximation curves that applies a simplified gradient and insures unambiguous z approximations in the given domain. The interpolation algorithm further contains an associative look-up that considerably reduces the computational effort for highly detailed approximations for points in close vicinity. The method achieves fairly good approximations with much fewer parameters than polynomial approximations and/or grid point data sets. The restriction to easily treatable approximation elements makes for computational effectiveness and analytic simplicity. The method is being used to approximate terrain for simulation; it can effectively be applied to approximate other functions $f(x,y)$ especially those that require large amounts of data in digitized form. Two examples of applications to terrain are given.

INTRODUCTION

The methods to approximate irregular surfaces described here resulted from an attempt to approximate terrain by a limited number of analytical expressions in order to overcome the need for huge amounts of digital terrain data for combat simulation, especially since data of that kind were only available for a few areas. The approximation by analytical functions also reduces the effort of line of sight computation, i.e., determination whether direct line (of sight) between two points on the surface is obstructed by some part of the surface itself. The process of line of sight computation does use a significant amount of computing resources for several combat models.

The methods presented here are de facto empirical, i.e., even in principal any desired accuracy of approximation can be obtained by reapplying the process; the process can become more cumbersome after applying a limited set of approximation elements and then to find more approximation elements to obtain another order of magnitude of accuracy. This does not invalidate the objective, which is to approximate the gross structure of an irregular surface by simple means and not to achieve utmost accuracy.

Two distinct approaches to approximate surfaces are presented here, the contour line approach (Ref 1) and the superposition of elementary surfaces. The first method is more effective for "rugged" surfaces with few macro structure elements. The second has the advantage of greater simplicity and also easiness of extending the approximated surface without recomputing the approximation elements and still achieve continuity beyond the original area of consideration.

The two methods are described and examples given. The process of line of sight determination is also outlined briefly since it is of such basic interest to the current application of the method.

CONTOUR LINE APPROXIMATION

We assume that the surface $z = F(x,y)$ to be approximated is continuous and single valued. The inverse functions $x = g(x,z)$ and $y = g(x,z)$ may be multivalued. Then the surface can be depicted by a set of curves $z_\nu = F(x,y) = \text{const}$ (contour lines). The z_ν may be sets of several distinct curves for a given z_ν. Curves for different z_ν do not intersect. A set of such contour lines is depicted in Fig. 1.

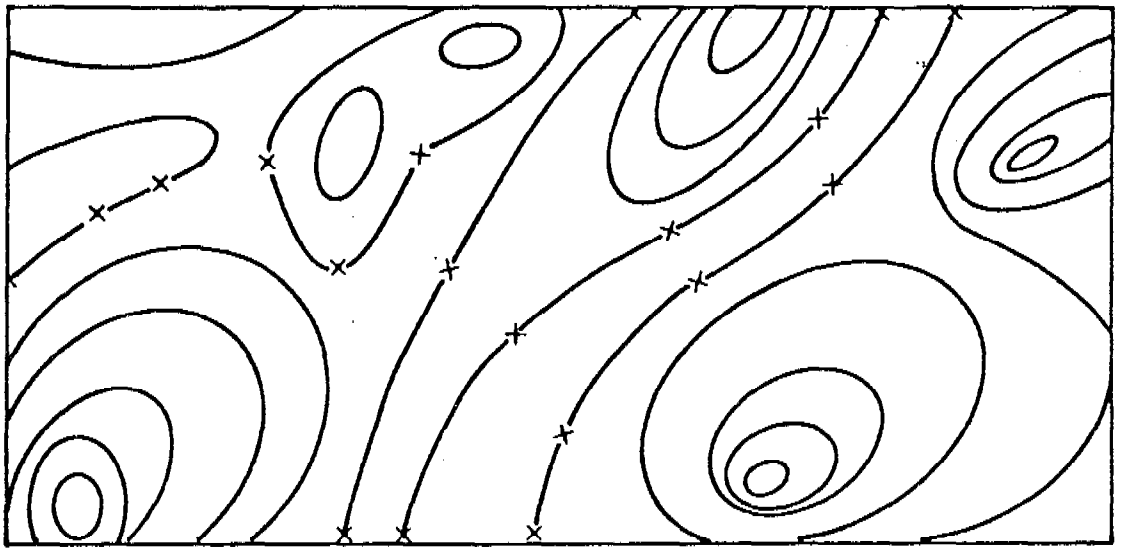


Fig. 1

For ease of description we will refer sometimes to z as elevation.

We approximate the contour lines by sets of quadratic or linear expressions. Frequently contour lines have pronounced indentations or bulges. Such curves can be approximated by defining pseudo contour lines. Pseudo contour lines are analytically simple approximation curves and that have elevation differentials for certain points assigned along their arc. This avoids the use of higher order approximation elements that result in lengthy expressions.

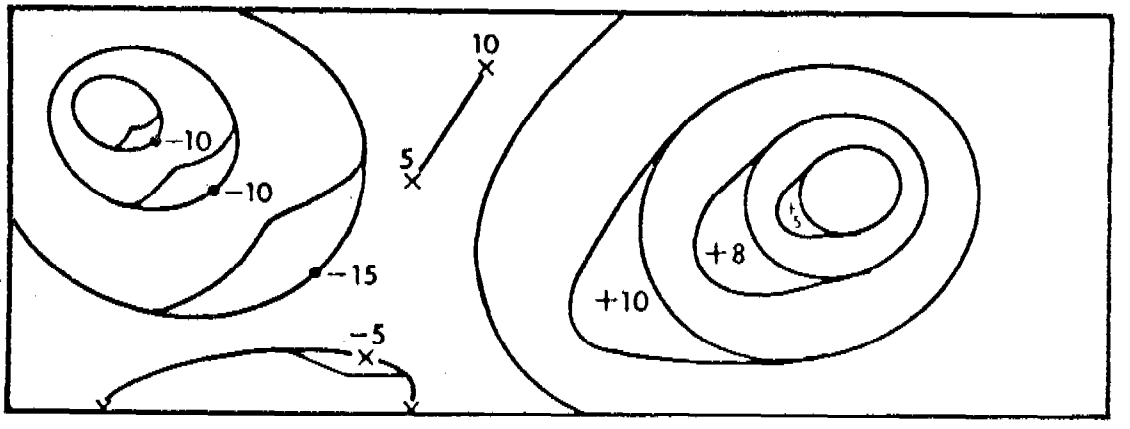


Fig. 2

Approximation Elements

We will restrict the choice of curves to ellipses, parabolas and polygons as elements for the approximation of contour lines and pseudo contour lines. The main reason for this restriction is the ease in obtaining the relevant parameters from given contour lines and also to allow explicit line of sight computations which are important for our application of the method to combat modeling.

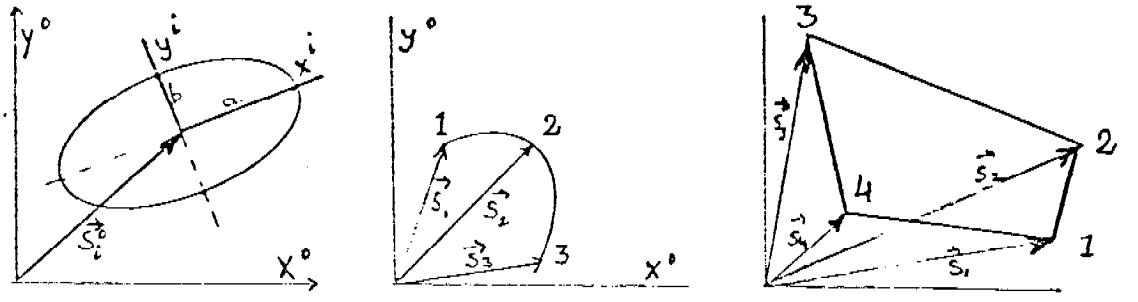


Fig. 3

Ellipses. An ellipse is defined by five parameters, but for practical purposes we will use six parameters (two are not independent), i.e., the coordinate vector to its center and the four parameters of its transformation matrix. Given coordinate vector $\vec{s}_i^{\,o}$ of the center in the k_o coordinate system, then the transformation

$$(1)\quad \vec{s}^{\,i} = (x^i, y^i) = (\vec{s}^{\,o} - \vec{s}_i^{\,o}) \cdot \tau_i \qquad \tau_i = \begin{bmatrix} a^{-1}\cos x^o x^i & b^{-1}\cos x^o y^i \\ a^{-1}\cos y^o x^i & b^{-1}\cos y^o y^i \end{bmatrix}$$

reduces the equation of the ellipse to a unit circle

$$(2)\quad \vec{s}^{\,i} \cdot \vec{s}^{\,i} = 1.$$

It is also true that a point $\vec{s}^{\,i}$ is inside the ellipse if the scalar product is less than one and outside if the scalar product is larger than one, a simple criterion that will be used later.

Parabola. We will use parabolas (Fig. 3) given by three points and defined by the equation

$$(3)\quad \vec{s} = \vec{u}t^2 + \vec{v}t + \vec{s}_2 \qquad \vec{u} = (\vec{s}_1 - 2\vec{s}_2 + \vec{s}_3)/2$$

$$\vec{v} = (\vec{s}_3 - \vec{s}_1)/2$$

The vectors $\vec{u}$ and $\vec{v}$ can be precomputed from the $\vec{s}_\nu$. The dimensionless parameter t is -1, 0, and +1 for $\vec{s}_1$, $\vec{s}_2$, and $\vec{s}_3$ respectively. Note that there are actually three different parabolas possible through three points and the labeling of $\vec{s}_\nu$ determines the "mid-point" and the desired parabola.

Polygons. Polygons are represented by sets of straight lines according to the equations (Fig. 3).

$$(4)\quad \vec{s} = \vec{s}_\nu + (\vec{s}_{\nu+1} - \vec{s}_\nu)t \qquad \nu = 1, 2, 3 \ldots n(\text{mod } n)$$

The points should be labeled cylically. For the sides (or points on the sides) of the polygons is $0 \leq t \leq 1$, a criterion that will be used later to decide whether a point is inside the polygon.

Pseudo Contour Line Approximations

To approximate pseudo contour lines we also need to define a Δz at given points on the curves defined above. As can be seen on Fig. 2. The fact that we have an "indentation" or a "bulge" depends on the sign of Δz. The abruptness of change from a "smooth" contour line depends in addition to the amount of Δz on the closeness of other contour lines. For an ellipse pseudo contour lines can be approximated for four points at the intersections with the coordinate system κ_i by

$$(5)\quad z = z_n + \sum_{\nu=1}^{4} (\vec{g}_\nu \cdot \vec{g} + |\vec{g}_\nu \cdot \vec{g}|)\Delta_\nu / 2(|g_\nu| \cdot |g|)$$

where $\vec{g}$ is the coordinate vector of the point on the circumference for which the elevation is to be established and the g_ν are the unit vectors of the given points with Δ_ν elevation differentials. The sums of the scalar products with their absolute value really indicate that there are never more than two terms of the sums, namely the two $\vec{g}_\nu$ that limit the quadrant in which $\vec{g}$ is located. The interpolation actually uses a Fourier form to avoid the use of transcendental functions.

Procedure

To approximate a surface by elements as defined above one establishes from the contour lines the approximate center points and axes of ellipses (for those contour lines that resemble ellipses) and precomputes the parameters of the transformation matrix. For the other contour lines one chooses either parabolas

through three points or polygons with convenient points. Parabolas can be chained to form any variety of curves. With all derived parameters precomputed one obtains a set of parameters that describes the approximating surface and will be referred to as the structure set. To regenerate the approximate surface one proceeds as outlined in the next chapter and the algorithm.

Interpolation

Once we have established a set of approximation elements and their parameters we can describe the contour lines of a surface as shown in Fig. 1. For a point on the area under consideration we have to compute the z- value (elevation), but since normally a point would not fall on a contour line we have to interpolate between adjacent contour lines. The correct interpolation is along the gradient. To obtain the gradient in a system (set of approximation elements) is equivalent of using an irregular curved linear coordinate system. There are also cases where there are more than two adjoining contour lines (Fig. 1) and the minimum distance must be chosen. Since the gradient and especially its length between a point and an approximated contour line cannot be computed explicitly we are using the simplifications indicated below.

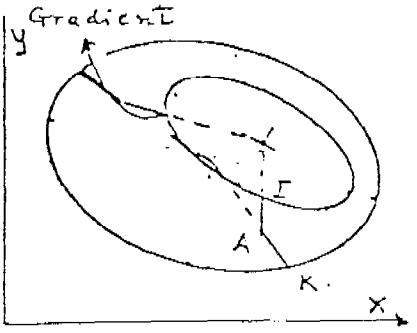


Fig. 4

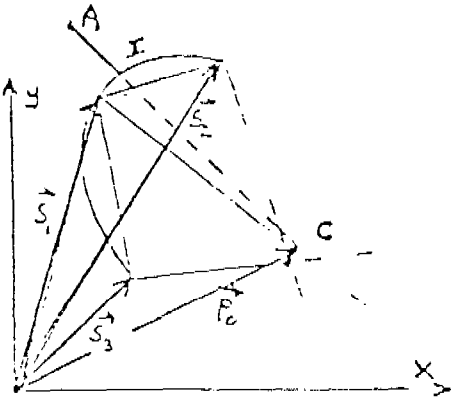


Fig. 5

For ellipses we use the distance between the point A and the periphery of the ellipse along the straight line that goes from the center of the point A (Fig. 4). Using the transformation indicated in (1) we receive for the distance with

$$(6)\quad (\vec{p}_A-\vec{s}_i)\cdot\tau_i = \vec{w}_i \quad \text{and} \quad (\vec{p}_A-\vec{s}_\kappa)\cdot\tau_\kappa = \vec{w}_\kappa$$

$$D^2 = (1-|\vec{w}_i|)^2\cdot(\vec{p}_A-\vec{s}_i)\cdot(\vec{p}_A-\vec{s}_i)$$

and if the respective values of z_ν = const are z_i and z_k this interpolates to

$$(7)\quad z_A = (z_i D_k^m + z_k D_i^m)/(D_i^m + D_k^m) \qquad m = 1 \text{ or } 2$$

For m=1 we have linear interpolation and for m=2 we have pythagorean interpolation which saves the computation of square roots in the numeric process. For confocal ellipses these approximations can be made more accurate, see reference 1.

For parabolas the distance from a point to the parabola is approximated as follows. One uses a center point that, for all practical cases, is inside the parabola and reasonably far away from the parabola itself. Such a point is halfway between the intersections of lines orthogonal to $\vec{s}_3 - \vec{s}_2$ and $\vec{s}_1 - \vec{s}_2$ through the points p_3 and p_1 respectively and the bisector of the angle $p_1\ p_2\ p_3$ as shown in Fig. (5).

The point $\vec{p}_c$ is defined as

$$(8)\quad \vec{p}_c = \vec{s}_2 + (\tilde{s}_{12} + \tilde{s}_{32})\ (|\vec{s}_{12}| + |\vec{s}_{32}|)/2(1 + \tilde{s}_{12}\cdot\tilde{s}_{32})$$

$$\vec{s}_{\mu\nu} = \vec{s}_\mu - \vec{s}_\nu, \quad \tilde{s}_{\mu\nu} = \vec{s}_{\mu\nu}/|\vec{s}_{\mu\nu}|$$

For the distance between point A and the parabola we obtain the approximation

$$(9)\quad D_{Ap}^2 = T_A^2 \vec{p}_{cA}\cdot\vec{p}_{cA} \qquad \vec{p}_{cA} = \vec{p}_c - \vec{p}_A$$

$$T_A = [2\langle\vec{v}\times\vec{u}\rangle t_A + \langle(\vec{s}_2-\vec{p}_A)\times\vec{u}\rangle]/\gamma$$

$$t_A = [\pm D-\beta]/\gamma \qquad D^2 = \langle(\vec{s}_2-\vec{p}_A)\cdot\vec{p}_{ck}\rangle\gamma + \beta^2$$

$$\beta = \langle\vec{p}_{cA}\times\vec{v}\rangle \qquad \gamma = \langle\vec{p}_{cA}\times\vec{u}\rangle$$

$$\langle\vec{a}\times\vec{c}\rangle = a_x c_y - a_y c_x, \quad t_A^2 \le 1$$

If t_A^2 is larger than one the point A lies outside the validly defined portion of the parabola. There are formulas for these pathological cases, reference 2.

The point A is outside the defined area of the parabola if $t_A^2 \leq 1$ and $T_A < 1$. The point A is inside if $t_A^2 \leq 1$ and $1 < T_A$.

Algorithm for surface generation and interpolation.

Given a set of approximation elements $\vec{e}_i$ whose components are identification as to type (ellipse, parabola or polygon), the elevation z_ν, indicators of pseudo or non-pseudo contour line, and all parameters required to describe the approximation element. It is practical to order the set of $\vec{e}_i$ into a matrix as follows beginning with the lowest z_ν (largest area) go the next higher z_ν contained in the same area and forth to the highest contour line contained in the first area. Then build the second group from the lowest z_ν where there were more than one contained in a contour line. Continue until all sets are used. For concave portions of the surfaces one should start from the contour line that covers the largest area and go up and down similar to the above. The order in which the groups (sets of $\vec{e}_i$) are counted is irrelevant. The ordering in the above fashion is computationally expedient.

To obtain the z coordinate for a given point A one proceeds as follows. Check the $\vec{e}_i$ (starting with the first group) for the element that contains the point A, using the respective formulas above. If such an $\vec{e}_\rho$ is found one proceeds to the next higher $e_{\rho+1}$ of the same group until one finds the $\vec{e}_\sigma$ that do not contain it. One interpolates then between $\vec{e}_\sigma$ and $\vec{e}_{\sigma-1}$. If the next higher elevations are in another group that contains the point, one follows that group.

When large numbers of points are needed that are close together a simple adaptive procedure can be used. The highest element that contained the previous point is saved and the next point is tested against it first. If the new point satisfies the test no further search is needed. If the new point does not satisfy the same test with $\vec{e}_i$, one proceeds to the next lower one and starts a full search if necessary. If the consecutive points to be analyzed are close together the searches are very short therefore saving considerable computer time. (See also Ref. 1)

APPROXIMATION BY SURFACE SUPERPOSITION

In this approach to approximate a given surface, a set of surfaces is superimposed in such a way that for any given point P(x,y) the highest of one or more overlapping surfaces defines the result surface. This results in a unique surface as long as the element surfaces are convex with respect to the base plane z=0. If concave surface elements are used a limitation with respect to z must be given such that the "ends" do not protrude (Fig. 7).

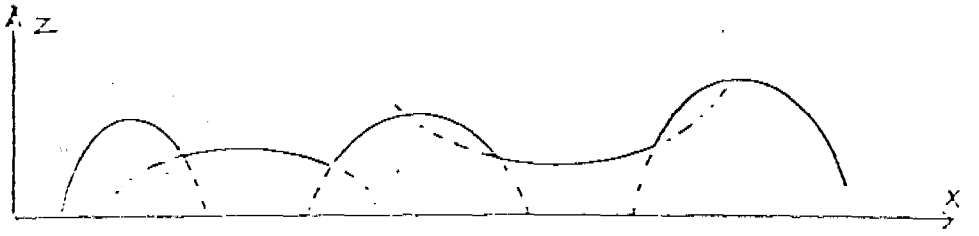


Fig. 7

In principle, any kind of analytically describable surface can be used as an approximation element, however, as in the case of contour line approximation we will restrict ourselves to quadratic or polyquadratic[1] surfaces. First, the general shape of such surfaces if very clear and therefore more suitable for empirical composition of a surface, and second using quadratic surfaces results in explicit formulas for line of sight determination.

Given a set of surfaces $z_i = f_i(x,y)$ then the resulting surface is defined as

$$(10)\quad \mathrm{Max} f_i(x,y) = Z(x,y) \qquad i = 1, 2, 3 \ldots n$$

$$\text{with}\quad f_k(x,y) = z_k^* \qquad k \subset i$$

[1]Polyquadratic in this case is quadratic.

for those $f_k(x,y)$ that are concave and have the limit z_k^*. All values of x, y and z must be real values. Whenever an $f_i(x,y)$ produces a complex z value for a given point (x_A,y_A) the value is to be excluded from the maximum search (10).

To approximate a given surface by this method one can structure the search for the proper set of $f_i(x,y)$ iteratively.

If $z = F(x,y)$ is the given surface to be approximated then

$$(11)\quad Z^{(\nu+1)} = \max_i f_i^{(\nu+1)}(x,y) = F(x,y) - \sum_\nu Z^{(\nu)}$$

$$\text{and}\quad F(x,y) = \sum_{\nu=1}^{m} \max_i f_i^{(\nu)}(x,y) + \varepsilon_m(x,y) \qquad Z^{(o)} \equiv 0$$

where $\varepsilon_m(x,y)$ is the remaining error after the m-th approximation. This method is convergent in the general case considering a broad choice of $f_i(x,y)$ and that F(x,y) is real, one-valued and continuous over the considered area. In practice, however, with successive iterations, the structure of the residue surface tends to gain complexity due to the repeated differencing process. Therefore, it is generally better to change the first and second set of $f_i(x,y)$ rather than trying to attain higher accuracy by adding more sets of $f_i^{(\nu)}(x,y)$ which will also increase the computational effort. This is commensurate with the objective of the approximation method to achieve a reasonable approximation with relatively few parameters, but not to achieve utmost accuracy with the addition of a large number of base data.

APPROXIMATION SURFACES

As stated before, any non-singular continuous function can be used as an approximation element. For practical reasons we will restrict the choice to quadratic (polyquadratic) surfaces. We will use surface elements that are expressed by the following equation

$$(12)\quad \pm x^2/a^2 = y^2/b^2 + (z/c)^p \qquad p = 1 \text{ or } 2$$

The plus signs and p=2 represent an ellipsoid, p=1 an elliptical paraboloid. The minus signs describe similar surfaces but they are concave with respect to the x-y plane, see Fig. 9.

The basic form (12) is achieved by a coordinate transformation similar to (1), however three dimensionally. We will assume that the coordinate systems of the approximation elements have their

z-axes parallel to the κ_e coordinate system, hence the approximate directional cosines are zero. We can therefore handle the approximation elements with the same two-dimensional vector $\vec{s}_i^o$ as in equation (1).

$$(13)\quad \vec{s}^i = (\bar{x},\bar{y}) = (\vec{s}^o - \vec{s}_i^o)\cdot\tau_i$$

and the approximation surface is defined as

$$(14)\quad Z(x^o,y^o) = \max_i f_i(x,y) = \max_i c_i(1\pm\vec{s}^i \cdot \vec{s}^i)^{1/\rho}$$

To utilize the approximation elements the relevant parameters $\vec{s}_i^o$ and the coefficients of the matrix τ_i have to be determined. Since one knows only the surface that needs approximating these parameters have to be derived from the given set of contour lines. The relationships for maximum elevation and minimum elevation to be considered by an element are evident from Fig. 8 and Fig. 9. With

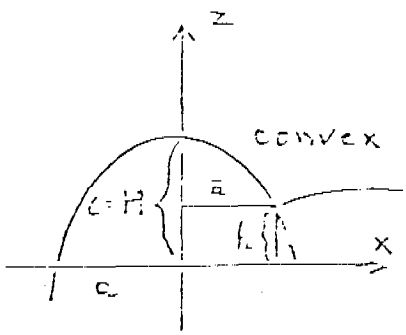


Fig. 8

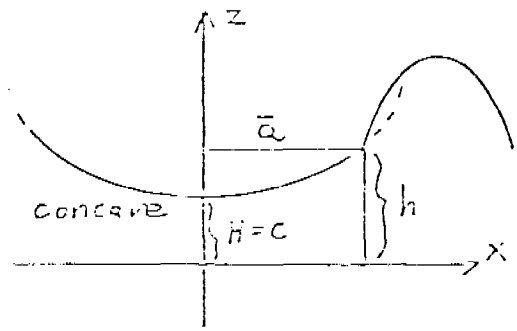


Fig. 9

equation (12) this leads to

$$(15)\quad a^2 = \pm\bar{a}^2H^\rho/(H^\rho - h^\rho) \qquad \rho = 1 \text{ or } 2$$

The minus sign applies to concave surface elements. The same formula applies for b in the y^i-z plane. Note that the plane depicted in Fig. 8 and Fig. 9 is the transformed κ_j coordinate system going through the center of the ellipsoid or other approximation element. When convex and concave surface elements are used it is important to note that curvature of the concave element must be less than all adjacent ones, see also Fig. 6. A sufficient criterion for this is

$$\mathrm{Max}(\bar{a}_c^2/a_c^2, \bar{b}_c^2/b_c^2) \quad < \quad \mathrm{Min}(\bar{a}_\nu^2/a_\nu^2, \bar{b}_\nu^2/b_\nu^2)$$

and

$$\mathrm{Max}(c_c/a_c, c_c/b_c) \quad < \quad \mathrm{Min}(c_\nu/a_\nu, c_\nu/b_\nu)$$

for all adjacent surfaces(index ν) and the "axes" of the concave surface (index c).

This criterion is cumbersome to use in its exactness. For practical purposes it is usually quite obvious whether the concave surface is flatter than the adjoining surfaces.

Polyquadratic Surfaces

If in equation (12) $\rho=1/2$ the approximation surface is

$$(16) \quad Z(x^o, y^o) = \underset{i}{\mathrm{Max}} f_i \ (x,y) = \underset{i}{\mathrm{Max}}[L_i + (1-\vec{S}^i \cdot \vec{S}^i)^\lambda]$$

$$2 \leq \lambda \qquad \vec{S}^i \cdot \vec{S}^i \leq 1$$

The profile of such surface in the x-z plane is shown in Fig. 10. For $\vec{S}^i \cdot \vec{S}^i < 0$ we have a partially concave surface and its extent is not easily evident when used in actual approximations.[1] The larger λ the steeper the curve. For practical purposes it is best to restrict the choice to $\lambda=2$ and $\lambda=4$. As Fig. 10 shows one can use L to move the point of zero curvature further up, as long as L<h.
The relation between a and $\bar{a}$ is given by equation (17).

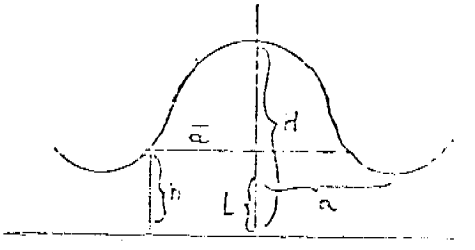


Fig. 10

[1] We are talking here about obvious shape without having done a point by point computation of course any point of that surface can be computed exactly if desired.

(17) $a^2 = \bar{a}^2(H-L)^\rho/[(H-L)^\rho - (h-L)^\rho]$ $\quad c = H-L$

The formula for b^2 is identical to (17) with the respective H and b.

LINE OF SIGHT DETERMINATION

For applications of surface approximations to terrain in conjunction with combat models it is of primary importance to know whether the straight line between two points on the surface is obstructed (intersected) by other parts of the surface. Because that computation has to be repeated numerous times it is imperative that the computation be simple and short.

For contour line approximation the algorithm is briefly as follows.* The parametric representation of the line between two points with the position vectors $\vec{P}_A$ and $\vec{P}_B$ is

$$\vec{p} = \vec{P}_A + (\vec{P}_B - \vec{P}_A)T.$$

This equation is transformed according to equation (1) and then T is determined by using equation (2) and similarily by equations (3) or (4) with respect to other approximation elements. If T_1 or T_2 of the intersection is between $0<T<1$ and $z = z_A + (z_B-z_A)T_{1,2}$ is less than the elevation of the intersecting contour line, the line of sight is obstructed. If none of the approximation elements (contour lines) is obstructing, line of sight exists. This involves a search algorithm. It can be shortened for numerical purposes by starting the search for an adjacent point by checking results (obstruction element) of the last point first, since it is likely that the same contour obstructs again (Ref 1).

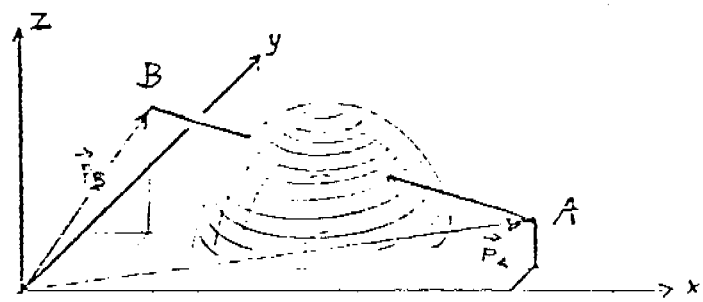


Fig.11

*Detailed explanation is given in Ref 2.

For superpositioned surface the line of sight can be determined as follows using the ellipsoid as an example

With $\vec{p}_A = (x_A, y_A, z_A)$ and $\vec{p} = \vec{p}_A + (\vec{p}_B - \vec{p}_A)T$ and $\vec{s}_E = (s_x^o, s_y^o, s_z^o)$

one obtains

$$\vec{q} = \textcircled{H}_i \cdot (\vec{p} - \vec{s}_E) \quad , \quad \textcircled{H} = \begin{bmatrix} \tau_i & 0 & \\ & & 0 \\ 0 & 0 & C^{-1} \end{bmatrix} \quad , \quad \vec{s}^i \cdot \vec{s}^i = 1$$

for T_1 and T_2 $\qquad T_{1,2} = (\pm D - \beta)/\alpha$

$$D^2 = \alpha(1-\gamma) + \beta^2, \ \alpha = q_{BA}^2 \quad , \quad \beta = \vec{q}_A \cdot \vec{q}_{BA} \quad , \quad \gamma = \vec{q}_A^2$$

The conditions for line of sight are:

$D^2 \leq 0$	No Obstruction
$D^2 > 0$ either T, or both $0<T<1$	Obstruction

For elliptical paraboloids the same conditions apply, however α, β and γ are different with respect to the contributing z- components.

For polyquadratic surfaces enveloping elliptical paraboloids are used as a first approximation, which is then refined, if necessary, for certain conditions. A detailed discussion is given in Ref 2; it goes beyond the scope of this paper.

RESULTS

The methods of surface approximation have been applied to some real terrain. Fig. 12 shows a graph of contour line approximations.

A comparison of digital data and various degrees of approximation by surface superposition was made for a terrain. The results are shown in Table 1 below. Note that the given values for the mean

Approx. Elements	Mean	Std. Deviation
6	2.5%	5.7%
12	0.5%	4.1%
16	0.1%	3.7%
19	0.2%	3.6%

TABLE 1. Steinbach Terrain Comparison

and standrad deviation are taken from the difference between approximation and digital data, and related to the maximum elevation. The approximation uses only about 1/40th of the data that the digital

representation requires to achieve the deviation of only 3.6% in the above case.

Figure 13 and Figure 14 show a relief and a plot of contour lines for the Steinbach Terrain, this case was also used in the comparison shown in Table I.

Figure 15 shows a line of sight map of the same terrain, i.e., all blank parts of the terrain are visible from the observation point(*). The digits indicate elevation intervals in multiples of 50 meters.

REFERENCES

1. H. M. Sassenfeld, An Approach to Terrain Representation for Application to Combat Simulation Models. Proc. AORS XV 1976.

2. H. M. Sassenfeld, Terrain Representation by Surface Superposition. TRASANA Report. March 1977.

3. The Tank Weapon System Report AR 69-4(U) Systems Res. Group. Ohio State U, for US Combat Dev. Command Armor Agg, Ft. Knox.

4. S. H. Perry and C. J. Needles, Parameters Station of Terrain in Combat Analysis. Proc. AORS XV 1976.

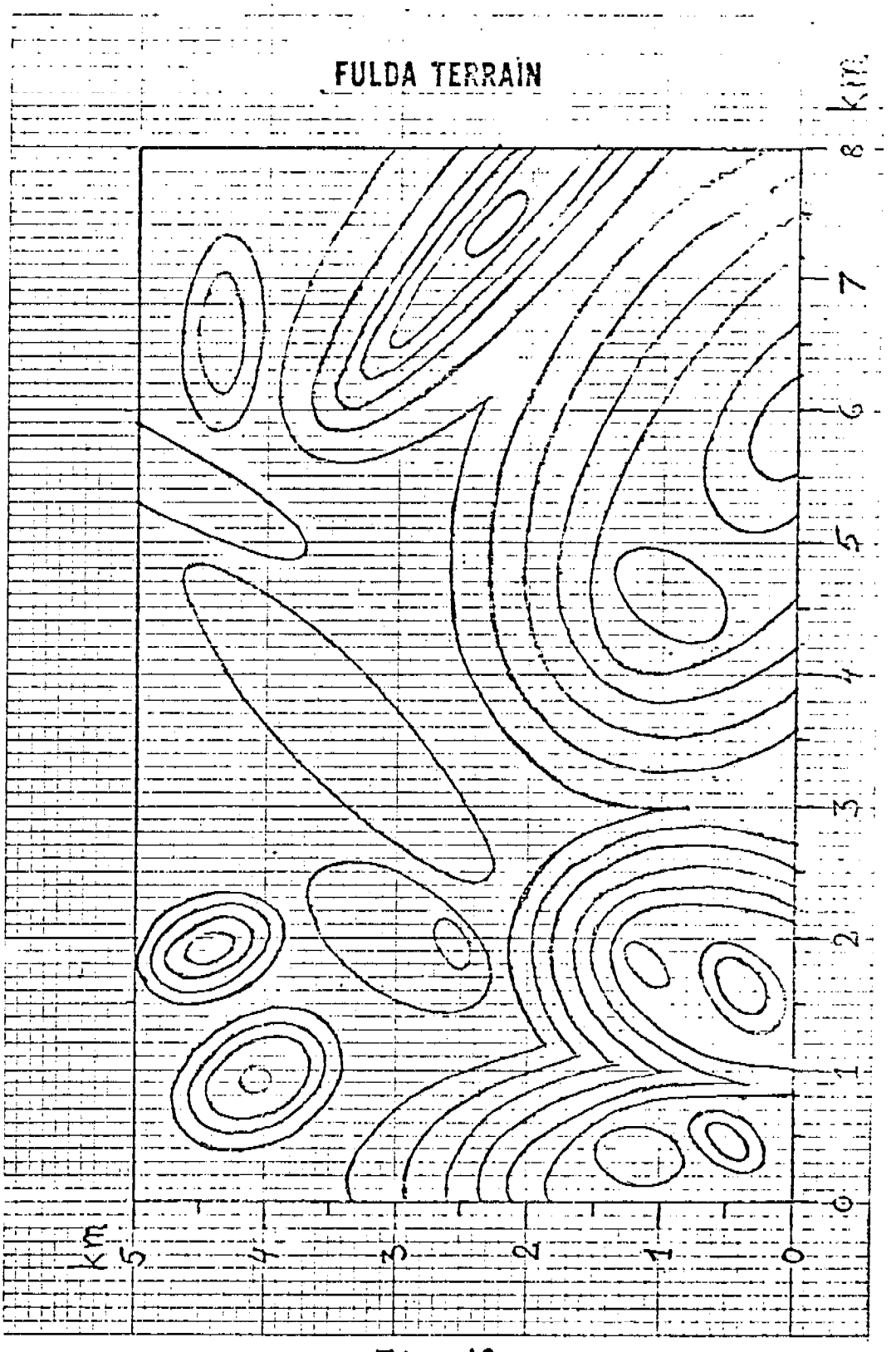


Fig. 12

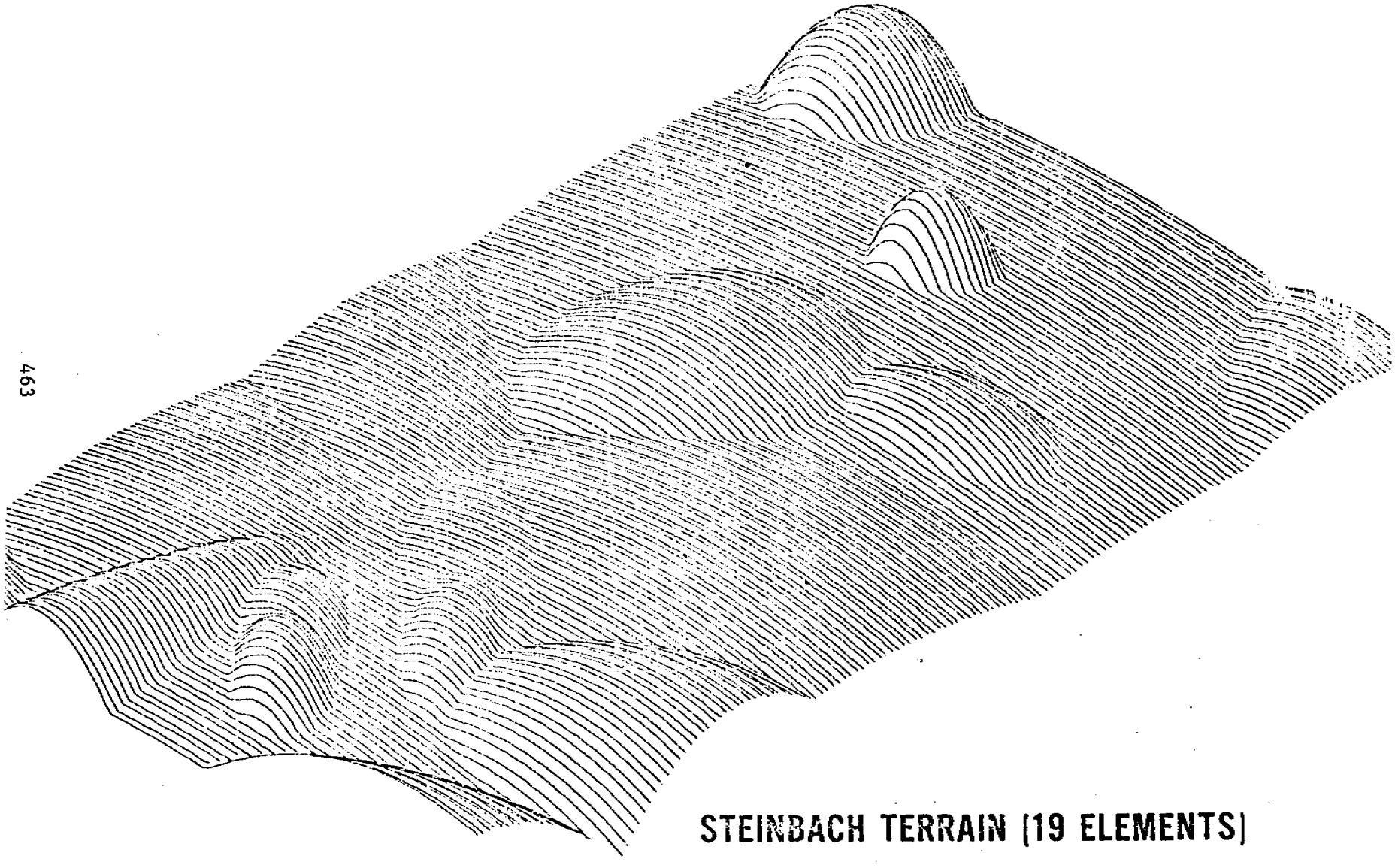


Figure 13

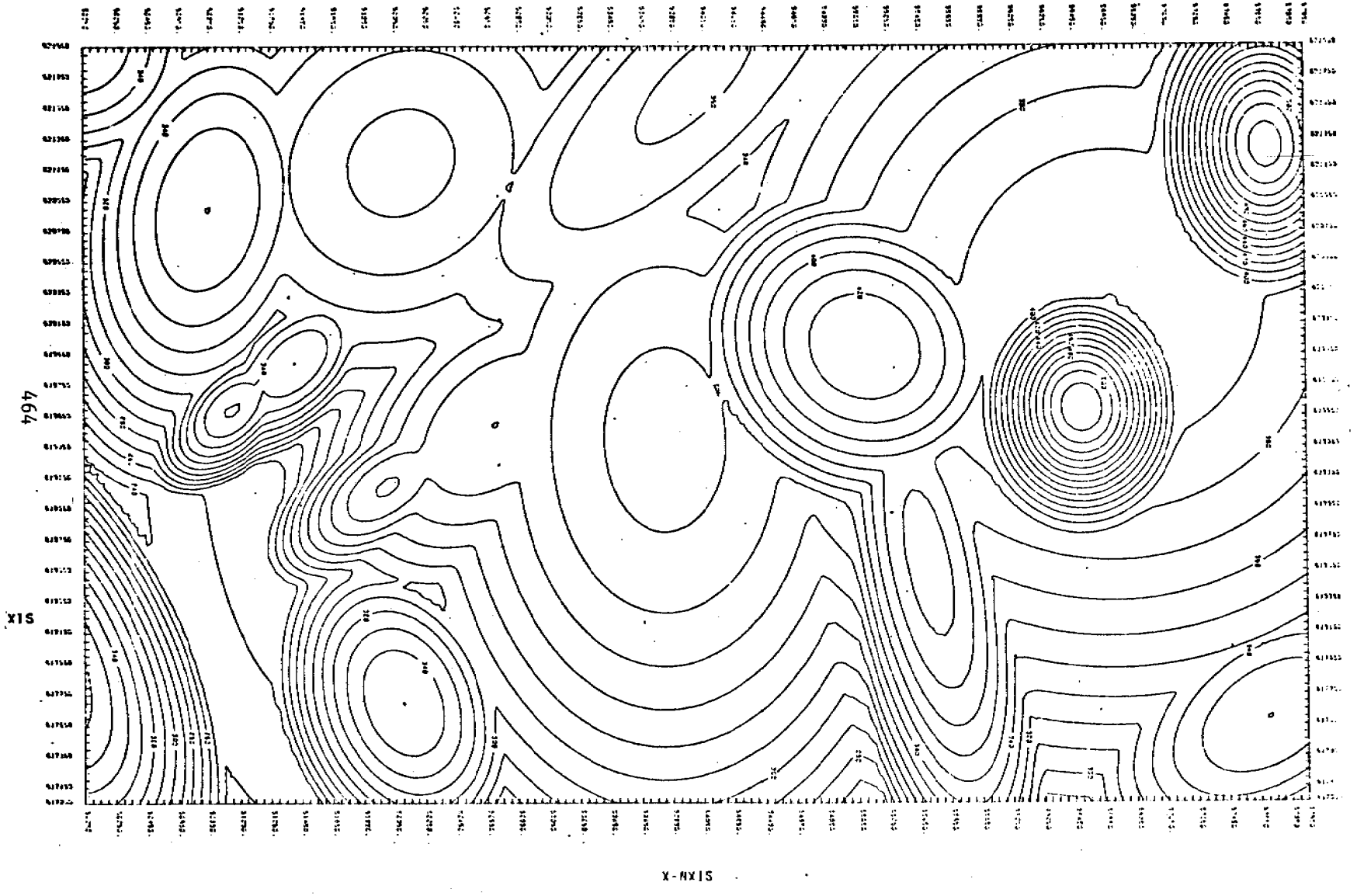


Figure 14 Contour Lines of Approximated Terrain

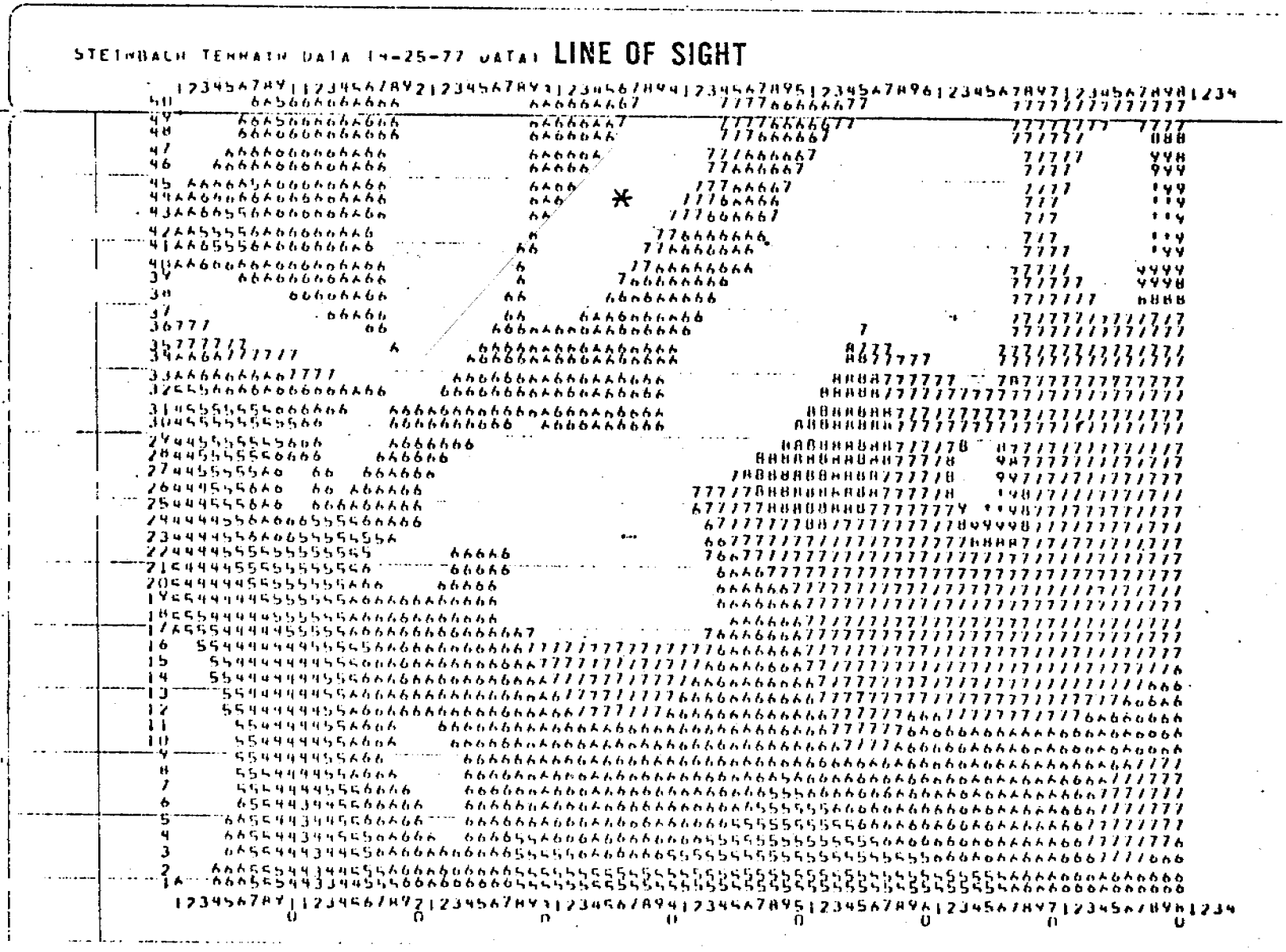


Fig. 15

GROUP THEORETIC METHODS IN BIFURCATION THEORY

D.H. Sattinger
School of Mathematics
Minneapolis, MN 55455

I. Bifurcation at Multiple Eigenvalues.

Suppose we want to investigate the bifurcation of solutions of a system of equations

$$G(\lambda,u) = 0 \tag{1}$$

in the neighborhood of a known solution (λ_o,u_o). Assume $L_o = G_u(\lambda_o,u_o)$ is a Fredholm operator of index 0 and that $n = \dim \ker G_u(\lambda_o,u_o) > 1$. Then by the Lyapounov - Schmidt method[4] we can reduce the bifurcation problem to that of solving a system of algebraic equations

$$F_i(\lambda, z_1,\ldots,z_n) = 0 \qquad i = 1,\ldots,n \ . \tag{2}$$

If $G(\lambda,u)$ is an analytic operator then the F_i are also analytic.

In practice the computation of even the lowest order terms of the F_i is a non-trivial matter, especially if (1) is a particularly complicated

system of equations, as in elasticity or fluid mechanics. Moreover, given n equations in n unknowns virtually anything can happen in the way of solution sets and their stabilities, and the algebraic problem can be quite complex. There is a natural assumption we can make in such problems which not only allows us to bypass the numerical difficulties inherent in the Lyapounov - Schmidt procedure but which also provides us with a general approach to bifurcation at multiple eigenvalues and with a way of classifying bifurcation points at multiple eigenvalues. I will assume that the mapping G is <u>covariant</u> with respect to a transformation group $\mathcal{G}$. That is, let T_g be a representation of $\mathcal{G}$:

$$T_{g_1 g_2} = T_{g_1} T_{g_2}$$

and assume that

$$H_1 : \quad T_g G(\lambda, u) = G(\lambda, T_g u) \quad . \tag{3}$$

This is a natural assumption in physical theories and is a mathematical expression of the axiom that the equations of mathematical physics be independent of the observer.

From (3) it follows that

$$T_g G_u(\lambda, u) = G_u(\lambda, T_g u) T_g$$

so if u_o is a solution which happens to be invariant under the entire group $\mathcal{G}$, $T_g u_o = u_o$, we have

$$T_g L_o = L_o T_g$$

where $L_o = G_u(\lambda_o, u_o)$. Therefore $N_o = \ker L_o$ is invariant under T_g , and $T_g\big|_{N_o}$ is a finite dimensional representation of $\mathcal{G}$. Let me write the bifurcation equations (2) in the form $F(\lambda, v) = 0$ where $v \in N_o$ and $F: \mathbb{C} \times N_o \to N_o$.

Theorem 1. If $G(\lambda, u)$ is covariant and $T_g u_o = u_o$ then so is F: $T_g F(\lambda, v) = F(\lambda, T_g v)$. (See [4])

Let me now expand F in a power series in v:

$$F(\lambda, v) = A(\lambda)v + B_2(\lambda, v, v) + B_3(\lambda, v, v, v) + \dots .$$

Then we must have

$$T_g A(\lambda) = A(\lambda) T_g \tag{4}$$

$$T_g B_2(\lambda, v, w) = B_2(\lambda, T_g v, T_g w) \quad . \tag{5}$$

Throughout this talk I will make the assumption

H_2: N_o is irreducible under T_g .

From (4) and H_2 it follows (by Schur's lemma [3]) that

$$A(\lambda) = \sigma(\lambda) I$$

where I is the identity. Suppose for convenience $\lambda_o = 0$ and $\sigma(\lambda) = C_1\lambda + C_2\lambda^2 + \ldots$. Then by various scaling arguments[4] the bifurcation problem can be reduced to an analysis of the equations

$$\lambda w = B_k(w) \tag{6}$$

where B_k is the first nonvanishing term in F , homogeneous of degree k . Equations (6) are called the reduced bifurcation equations. It can be shown ([5] Theorem 7.2) that the stability of the bifurcating solutions can be determined to lowest order from an analysis of the Jacobian of (6) at a solution.

The group theoretic approach, then, is to compute the lowest nonvanishing term B_k , find all solutions of (6), and determine their stability in the neighborhood of the branch point.

II. Bifurcation in the presence of $O(3)$.

Today I will show how the above program can be carried out in a specific case, namely when the group in question is $O(3)$, the group of rotations of the sphere. There are a number of classical bifurcation problems in which this situation arises, for example the buckling of a perfectly uniform spherical shell, or the onset of convection in a spherical mass. The latter problem and its possible connection with convection in the earth's mantle and plate tectonics has been discussed recently by F. Busse.[1]

Our first task is to construct the 2nd or 3rd order covariant mappings $B(v,w)$ or $B(u,v,w)$. It is important to keep in mind that these are completely symmetric mappings: $B(v,w) = B(w,v)$.

The irreducible reps of $SO(3)$ are of dimension $2\ell+1$, $\ell = 0,1,\ldots$ and are denoted by D^ℓ. To get quadratic mappings we consider the tensor product $D^\ell \otimes D^\ell$ acting on $N \otimes N$. Let us recall the Clebsch-Gordon series ([3], p. 233)

$$D^\ell \otimes D^\ell = D^{2\ell} \oplus D^{2\ell+1} \oplus \ldots \oplus D^0 \tag{7}$$

This means that $N \otimes N$ decomposes into a direct sum of invariant irreducible subspaces V^ℓ:

$$N \otimes N = V^{2\ell} \oplus V^{2\ell-1} \oplus \ldots \oplus V^\ell \oplus \ldots \oplus V^0$$

and precisely one of these subspaces, namely V^ℓ, transforms like D^ℓ.

Since we require

$$B(D^{\ell}v\ ,\ D^{\ell}w) = D^{\ell}\ B(v,w)$$

this is precisely the one we want.

Now in the above decomposition $V^{2\ell}$ contains symmetric tensors, $V^{2\ell-1}$ anti-symmetric tensors, and so forth. Accordingly V^{ℓ} will be symmetric iff ℓ is even . Therefore

<u>for odd ℓ there is no quadratic term</u>

and we must go to cubic terms.

Let's consider the case of even ℓ . I will construct the mapping B using Lie algebra methods which are well known in the theory of angular momentum coupling in quantum mechanics. Since B can be assumed to be completely symmetric we can work with polynomials and write

$$F_i(z_1,\ldots,z_n) = \sum_{j,k} G_{ijk}\ z_j\ z_k \quad .$$

We begin with the infinitesimal generators of the rotation group

$$J_1\ ,\ J_2\ ,\ J_3 \quad .$$

These satisfy the commutation relations

$$[J_i,\ J_j] = \epsilon_{ijk}\ J_k\ ,$$

where ε_{ijk} is the completely anti-symmetric tensor. Now put

$$J^{\pm} = \pm J_2 + i J_1 \qquad J^3 = -i J_3 \ .$$

These operators satisfy the commutation relations

$$[J^+, J^-] = 2 J^3 \ , \qquad [J^3, J^{\pm}] = \pm J^{\pm} \ . \tag{8}$$

The operators $J^{\pm}$ are the familiar ladder operators of quantum mechanics. Let N be an irreducible (real) invariant vector space which transforms like D^{ℓ} under $SO(3)$. Then the complexified space $N + iN$ has a basis f_m such that (see [3], Chapter 7)

$$\begin{aligned} J_3 f_m &= m f_m \\ J_{\pm} f_m &= \beta_{\pm m} f_{m \pm 1} \\ \overline{f}_m &= (-1)^{\ell - m} f_{-m} \end{aligned} \tag{9}$$

where

$$-\ell \leq m \leq \ell \quad \text{and} \quad \beta_m = \sqrt{(\ell - m)(\ell + m + 1)}$$

This basis may be constructed using the commutation relations (8). We now represent N as the vector space of linear polynomials in $z_{-\ell}, \ldots, z_{\ell}$, where the variables z_m act as the f_m in (9). Denote

by $K[z_{-\ell},\ldots,z_{\ell}]$ the ring of polynomials in the independent variables $z_{-\ell},\ldots z_{\ell}$. K is isomorphic to the algebra of symmetric tensors over N . (There is a natural correspondence between tensors over N and multilinear transformations of N into itself; See [4]) . We now extend the operators J_3 and $J_{\pm}$ to be <u>derivations</u> over K :

$$J(\alpha f + \beta g) = \alpha\, Jf + \beta\, J_g$$

$$J(fg) = f\, J_g + (Jf)_g \qquad (10)$$

where f , $g \in K$ and α and β are scalars. It is natural to extend the J's in this way since they are Lie derivatives.

I apologize if I have lost the reader with what may seem to be meaningless algebraic abstractions; but now, using (9) and (10), I am ready to present a simple algorithm for the construction of the polynomials $F_m(z_{-\ell},\ldots z_{\ell})$. The procedure should be familiar to those readers who have studied angular momentum coupling in quantum mechanics. We require the F_m to transform in the same way as the z_m :

$$J_3\, F_m = m\, F_m \qquad J_{\pm}\, F_m = \beta_{\pm m}\, F_{m\pm 1} \quad .$$

Consider first the quadratic terms in F_m . (The linear term of F_m is a scalar times z_m ; since the representation is irreducible the linear term must be a scalar multiple of the identity.) For quadratic terms

$$J_3(z_j z_k) = (J_3 z_j) z_k + z_j (J_3 z_k)$$

$$= j\, z_j z_k + k\, z_j z_k$$

$$= (j+k) z_j z_k \quad .$$

Since F_m is to be a sum of quadratic terms and $J_3 F_m = m F_m$ we require $(j+k) = m$. So

$$F_m(z_{-\ell}, \ldots z_\ell) = \sum_{m_1+m_2=m} G_{m_1 m_2 m}\, z_{m_1} z_{m_2} \quad .$$

In particular (when ℓ is even)

$$F_\ell = G_0 z_\ell z_0 + G_1 z_\ell z_{\ell-1} + \ldots + G_{\ell/2} (z_{\ell/2})^2 \quad .$$

Furthermore $J_+ F_\ell = \beta_\ell F_\ell = 0$, and this condition gives us a set of linear equations for the coefficients $G_0, \ldots, G_{\ell/2}$. For example, in the case $\ell = 2$ we have

$$F_2 = a\, z_2 z_0 + b\, z_1^2$$

$$J_+ F_2 = a \beta_0 z_2 z_1 + 2b\, z_1 z_2 \beta_1$$

$$= (a \beta_0 + 2b \beta_1) z_1 z_2 = 0 \ ,$$

$$a \beta_0 + 2b \beta_1 = 0 \quad .$$

Once F_ℓ is known we get $F_{\ell-1}$ from

$$J_- F_\ell = \beta_{\ell-1} F_{\ell-1}$$

and so forth. In this way we can construct all the F_m's .

This procedure extends immediately to higher order terms. For example to get 3rd order terms we write

$$F_\ell = \sum_{i+j+k=\ell} A_{ijk} z_i z_j z_k$$

and apply

$$J_+ F_\ell = 0$$

to get a linear system of equations for the A_{ijk} . For $\ell = 1$ there is only one solution but for $\ell = 3$ there are two independent solutions. In fact, the condition $J_+ F_\ell = 0$ leads to five equations in seven unknowns.

For the quadratic case the general mapping is given in terms of the Clebsch-Gordon coefficients

$$F_m(z_{-\ell}, \ldots z_\ell) = \sum_{m_1+m_2=m} C(\ell,m;\ell_1 m_2 \mid \ell,m) z_{m_1} z_{m_2}$$

or the Wigner 3-j coefficients

$$= (-1)^m \sum_{m_1+m_2=m} \begin{pmatrix} \ell & \ell & \ell \\ m_1 & m_2 & -m \end{pmatrix} z_{m_1} z_{m_2} \quad .$$

Much is known about the $3-j$ symbols, since they are of prime importance in atomic spectroscopy. Here is a generating function for them ([3], p. 261)

$$\frac{(x_3-x_1)^{j_1-j_2+j_3}\,(x_2-x_3)^{-j_1+j_2+j_3}\,(x_1-x_2)^{j_1+j_2-j_3}}{\sqrt{(j_1+j_2-j_3)!\;(j_1-j_2+j_3)!\;(-j_1+j_2+j_3)!\;(1+j_1+j_2+j_3)}}$$

$$= \sum_{-j_i \le m_i \le j_i} \frac{x_1^{j_1+m_1}\,x_2^{j_2+m_2}\,x_3^{j_3+m_3}\,\begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix}}{\sqrt{(j_1+m_1)!\;(j_1-m_1)!\;(j_2+m_2)!\;(j_2-m_2)!\;(j_3+m_3)!\;(j_3-m_3)!}}$$

III. Special Results.

Let me discuss some special results which can be obtained for low ℓ: $\ell = 1,2,3,4$.

1. $\ell = 1$ (vector fields). The reduced bifurcation equations are

$$\lambda z_1 = a z_1(z_o^2 - 2 z_1 z_{-1})$$

$$\lambda z_o = a z_o(z_o^2 - 2 z_1 z_{-1})$$

$$\lambda z_{-1} = a z_{-1}(z_o^2 - 2 z_1 z_{-1}) \quad .$$

For real solutions we require $\bar{z}_m = (-1)^m z_{-m}$. The entire (non-trivial) solution set is

$$z_o^2 + 2|z_1|^2 = \frac{\lambda}{a} \, .$$

The parameter a depends on the specific physical problem. We must have $\lambda/a > 0$ for bifurcation, so the bifurcation is supercritical if $a > 0$ and subcritical if $a < 0$. By rescaling we can assume $\lambda/a = 1$. Then the orbit of solutions is

$$z_o = \cos\theta \ , \ z_1 = \frac{\sin\theta}{\sqrt{2}} e^{i\varphi} \ , \ z_{-1} = -\frac{\sin\theta}{\sqrt{2}} e^{-i\varphi} \quad .$$

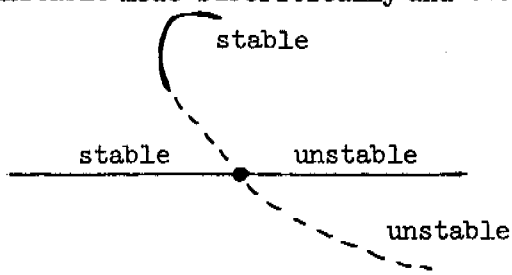
These solutions are all axisymmetric. They are orbitally stable when they appear supercritically and unstable when they appear below criticality.

2. $\ell = 2$ (second order tensors)

Again we get only axisymmetric solutions (Busse [1]) . This time the branch is transcritical. The eigenvalues of the Jacobian of the reduced equations are

$$3,0,-1,0,3 \quad .$$

The two zeroes are a consequence of the rotational invariance of the equation and the fact that the orbit is two dimensional. There is one unstable mode subcritically and two unstable modes supercritically. We therefore have in this case the possibility of hard buckling (snap-through instability) .



3. $\ell = 3$. The interesting thing about this case is that there are two independent covariant mappings viz.

$$F_3 = z_3(z_o^2 - 2\, z_1 z_{-1} + 2\, z_2 z_{-2} - 2\, z_3 z_{-3})$$

and

$$F_3 = 9\sqrt{\frac{60}{7}}\, z_3^2\, z_{-3} - 9\sqrt{\frac{60}{7}}\, z_3\, z_2\, z_{-2}$$

$$+ 3\sqrt{\frac{60}{7}}\, z_3\, z_1\, z_{-1} - 3\sqrt{10}\, z_2\, z_1\, z_o$$

$$+ \frac{30}{\sqrt{7}}\, z_2^2\, z_{-1} + \sqrt{7}\, z_1^3 \quad .$$

This means the bifurcation equations take the form

$$\lambda z = A F(z) + B G(z)$$

where the parameters A and B depend on the external physical constants of the problem. This situation occurs in the Bénard problem and gives rise to mechanisms for pattern selection[5], [6] .

4. $\ell = 4$. (Busse[1]) There are many solutions to the bifurcation equations in this case. Two special ones are

1) axisymmetric solutions: $z_{\pm 1} = \ldots = z_{\pm 4} = 0$, $z_0 \neq 0$. The eigenvalues of the Jacobian are

$$-1, 0, 0, \frac{20}{9}, \frac{20}{9}, \frac{10}{3}, \frac{10}{3}, \frac{-5}{9}, \frac{-5}{9} \quad .$$

The axisymmetric solutions are thus unstable on both sides of criticality, with 3 unstable subcritical modes.

2) octahedral symmetry

$$z_4 = \frac{5}{\sqrt{14}} = z_{-4} \qquad z_0 = \sqrt{5} \qquad z_{\pm 1} = z_{\pm 2} = z_{\pm 3} = 0 \quad .$$

This solution was found by Busse. The eigenvalues are

$$0, 0, 0, -1, \frac{20}{7}, \frac{20}{7}, \frac{20}{7}, \frac{5}{7}, \frac{5}{7} \quad .$$

This solution thus has one unstable subcritical mode.

Here are the quadratic polynomials for the case $\ell = 4$:

$$F_4 = \frac{1}{\sqrt{5}} z_4 z_0 - \frac{1}{\sqrt{2}} z_3 z_1 + \frac{3}{2\sqrt{14}} z_2^2$$

$$F_3 = \frac{1}{\sqrt{2}} z_4 z_{-1} - \frac{3}{2\sqrt{5}} z_3 z_0 + \frac{1}{\sqrt{14}} z_2 z_1$$

$$F_2 = \frac{3}{\sqrt{14}} z_4 z_{-2} - \frac{1}{\sqrt{14}} z_3 z_{-1} - \frac{11}{14\sqrt{5}} z_2 z_0 + \frac{3}{7\sqrt{2}} z_1^2$$

$$F_1 = \frac{1}{\sqrt{2}} z_4 z_{-3} + \frac{1}{\sqrt{14}} z_3 z_{-2} - \frac{6}{7\sqrt{2}} z_2 z_{-1} + \frac{9}{7\sqrt{20}} z_1 z_0$$

$$F_0 = \frac{1}{\sqrt{5}} z_4 z_{-4} + \frac{3}{2\sqrt{5}} z_3 z_{-3} - \frac{11}{14\sqrt{5}} z_2 z_{-2} - \frac{9}{14\sqrt{5}} z_1 z_{-1} + \frac{9}{14\sqrt{5}} z_0^2 \quad .$$

The others may be obtained from the relation

$$F_{-m}(z_\ell, \ldots z_{-\ell}) = F_m(z_{-\ell}, \ldots z_\ell) \quad .$$

$\ell=6,8$. Busse [1] also found special solutions for these cases too. Busse used an extremum principle to determine the physically relevant solution. I will discuss that principle in the final section. It indicates that the axisymmetric solutions are not the physically relevant ones. Busse conjectures that for $\ell=6$ the relevant solution has the symmetry of a dodecahedron.

IV. Gradient Structure of the Bifurcation Equations.

The extremum principle I referred to above is the following. For ℓ even the reduced bifurcation equations can be written in the form

$$\lambda\, z_m = a\, F_m(z_{-\ell} \cdots z_\ell) \tag{11}$$

where

$$F_m(z_{-\ell} \cdots z_\ell) = \sum_{m_1+m_2+m=0} (-1)^m \begin{pmatrix} \ell & \ell & \ell \\ m_1 & m_2 & -m \end{pmatrix} z_{m_1}\, z_{m_2} \quad .$$

Consider the function

$$p(z) = \frac{1}{3} \sum_{-\ell}^{\ell} F_m\, \bar{z}_m = \frac{1}{3} \sum_{-\ell}^{\ell} (-1)^m F_m\, z_{-m}$$

$$= \frac{1}{3} \sum_{m_1+m_2+m_3=0} \begin{pmatrix} \ell & \ell & \ell \\ m_1 & m_2 & m_3 \end{pmatrix} z_{m_1} z_{m_2} z_{m_3} \quad .$$

For ℓ even the Wigner coefficients are completely symmetric and therefore

$$F_m(z_{-\ell}, \ldots z_\ell) = \frac{\partial p}{\partial z_m} \quad .$$

Consequently our reduced bifurcation equations have a gradient structure, and this fact is independent of the structure of the original equations $G(\lambda, u)$: It is a purely group theoretic result.

Therefore our bifurcation equations (11) are the Euler-Lagrange equations for the minimax problem

$$\begin{aligned} &\text{Min} \quad p(z) \\ &|z| = 1 \end{aligned}$$

where

$$|z|^2 = \sum_{-\ell}^{\ell} z_m \bar{z}_m = \sum_{-\ell}^{\ell} (-1)^m z_m z_{-m} \quad .$$

The function p is a third order invariant for $O(3)$. That is $p(T_g z) = p(z)$ for all $g \in O(3)$. In terms of the infinitesimal generators this is equivalent to

$$J^{\pm} p = J_3 p = 0 \quad .$$

The norm $|z|^2$ is the second order invariant.

Leon Green (School of Mathematics, University of Minnesota) and I have succeeded in casting the bifurcation problem in a slightly different way. Consider the Clebsch-Gordon series

$$D^{\ell/2} \otimes D^{\ell/2} = D^{\ell} \oplus D^{\ell-1} \oplus \ldots \oplus D^{0} \tag{12}$$

and the associated representation

$$u_g A = D^{\ell/2}(g)\ A\ D^{\ell/2}(g^{-1})$$

on $(\ell+1) \times (\ell+1)$ symmetric matrices A. This representation is unitary relative to the inner product

$$(A,B) = \text{tr } A\, B^* \quad .$$

Furthermore, the third order invariant (there is only one) is

$$P(A) = \frac{1}{3} \text{tr } A^2\, A^* \quad .$$

Now the highest weight space, the one that transforms like D^ℓ in (12), consists of symmetric tensors $(A = A^+)$ so we may rephrase our bifurcation problem as

$$\text{Min } \frac{1}{3} \text{tr } A^3$$

subject to

$$\text{tr } A^2 = 1 \quad \text{and} \quad \text{tr } A\, B_j = 0$$

where the B_j are symmetric matrices which lie in the lower weight invariant subspaces. In particular, $\text{tr } A\, I = \text{tr } A = 0$. For $\ell = 2$ we get the bifurcation equations

$$A^2 = \lambda A + \gamma I \quad .$$

So far we have only been able to apply this approach in the case $\ell = 2$. (which we have solved completely); but it is interesting because of its similarity to L. Michel's approach to symmetry breaking problems in physics. [2] .

References

1. F. Busse, "Patterns of Convection in Spherical Shells", J. Fluid Mech. (1975), 72, 67-85.

2. L. Michel, "Les Brisures Spontanées de Symétrie en Physique", Journal de Physique, Tome 36, Nov 1975, pp. C7-41, C7-51.

3. W. Miller, Symmetry Groups and their Applications, Academic Press, New York, 1972.

4. D. Sattinger, "Group representation theory and branch points of nonlinear functional equations", SIAM Jour. Math. Anal.

5. _____, "Group representation theory, bifurcation theory, and pattern formation", Jour. Funct. Anal.

6. _____, "Selection Mechanisms for Pattern Formation", Arch. Rat. Mech. Anal.

ORDINARY DIFFERENTIAL EQUATIONS IN INFINITE DIMENSIONS AND ACCRETIVE OPERATORS

Michael G. Crandall
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, Wisconsin 53706

ABSTRACT. In the last ten years a nonlinear theory of evolution problems which has applications in classical ordinary differential equations, nonlinear diffusion problems, Stefan problems, control problems as well as many other areas has been developed. This is the sense in which this theory has applications and the basic results of the theory are described.

INTRODUCTION. There is a nonlinear "theory" of evolution problems which provides existence, uniqueness and continuous dependence results for a spectrum of problems ranging from some in classical ordinary differential equations to others involving the heat equation, the wave equation, nonlinear diffusion, a single conservation law, the Stefan problem, quasi-variational inequalities of evolution as well as many more. An expository introduction to this subject is given in [5] while [1] provides a development in depth. However, these sources both need updating in view of recent developments.

The current paper is intended to supplement these sources by summarizing the basic results while incorporating some recent advances from the literature. Section 1 consists of a brief informal discussion of the relationship of the abstract results of Section 2 to applications. Section 2 summarizes the results of interest rather tersely - it should be read as a supplement to [5] or [1].

1. Orientation. In classical ordinary differential equations one studies initial-value problems of the form

$$\text{(1)} \qquad \begin{cases} \dfrac{du}{dt} + A(u) = 0 \\ u(0) = u_0 \end{cases}$$

where A maps a subset $D(A)$ of $\mathbb{R}^N$ into $\mathbb{R}^N$. Under mild conditions it is proved that (1) has a unique solution on some time interval and that this solution depends nicely on A and u_0 in various senses. These results imply that models met in many applications are "well-posed". In these models u represents the state of the system under consideration. For example, u might be a list of numbers giving the positions and velocities of a finite collection of particles. We are going to discuss some existence, uniqueness and continuous dependence results for (1) where A is a mapping in some Banach space. The relationship of these results to applications is very much the same as in the classical case. That is, many applied models involving partial differential equations can be abstracted in the form (1) where A is an operator in an infinite dimensional Banach space (rather than in some $\mathbb{R}^N$) and A will satisfy the assumptions made in the next section. Thus the abstract theory provides us with the ability to think about a wide spectrum of problems simultaneously as well as basic facts about their solutions. However, just as the classical well-posedness theory does not give any detailed information about, for example, the solutions of the three body problem, the theory discussed here does not give detailed information about particular solutions of particular problems.

We begin by sketching how particular problems in partial differential equations can be regarded as special cases of the abstract Cauchy problem

$$CP(A,f,u_0): \begin{cases} \frac{du}{dt} + A(u) = f(t) \\ u(0) = u_0 \end{cases} .$$

In $CP(A,f,u_0)$, A maps its domain $D(A)$, which is a subset of a Banach space X, into X (or $A:D(A) \subseteq X \to X$), $f:[0,T] \to X$ is a strongly integrable function, $T > 0$ is regarded as fixed hereafter and $u_0 \in X$.

As an example, we choose the nonlinear diffusion problem

$$(NDP) \begin{cases} (DE)\ u_t - (k(u)u_x)_x = f(x,t), & x \in (0,1),\ t > 0 , \\ (BC)\ u_x(0,t) = u_x(1,t) = 0, & t > 0 , \\ (IC)\ u(x,0) = u_0(x), & x \in (0,1) . \end{cases}$$

In (NDP) we may regard the unknown function $u(x,t)$ as the temperature of a one-dimensional rod. The rod has insulated ends, a temperature-dependent conductivity $k(u)$ and is being heated externally. Let us think of (NDP) as an equation describing how the whole temperature field varies as t varies. That is, we will regard (NDP) as telling us the rate of change of the function $t \to u(\cdot,t)$ which assigns to the time t the "state" of the rod-namely, the temperature field at that time. We write $u(t)$ for this temperature field. The ways of thinking "$t \to u(t)$" and the classical "$(x,t) \to u(x,t)$" are related by $u(t)(x) = u(x,t)$. A good state space for this problem is $X = L^1(0,1)$. The requirement that $u(t) \in L^1(0,1)$ simply corresponds to the heat energy in the rod being finite at the time t. We rewrite the equation $u_t - (k(u)u_x)_x = f(x,t)$ in the following way: Replace u_t by $\frac{du}{dt}$ since we now are thinking of u as a function of t whose values are in X. In the same spirit we replace $-(k(u)u_x)_x$ by $A(u)$ where $A(v) = -(k(v)v')'$. In words, to compute $A(v)$, where v is a function of x, differentiate v, multiply the result by $-k(v)$ and differentiate again. The expression $f(x,t)$ will likewise be replaced by $f(t)$ where $f(t) \in X$ for each t. Now the equation $u_t - (k(u)u_x)_x = f(x,t)$ is abbreviated to $du/dt + A(u) = f(t)$.

The boundary conditions (BC) are handled by incorporation into the domain of A. Set

$$D(A) = \{v \in X : v \text{ and } v' \text{ are absolutely continuous and } v'(0) = v'(1) = 0\} .$$

That is, every function in $D(A)$ satisfies the zero flux condition at $x = 0$ and $x = 1$. With these identifications the abstract Cauchy problem $CP(A,f,u_0)$ contains all the information (NDP) contained. This rewriting of (NDP) in the above form attains significance only when we have information about the solutions of $CP(A,f,u_0)$ which is of interest for (NDP). Such information is the topic of the next section.

Section 2. One of the most basic and general approaches to the solvability of $CP(A,f,u_0)$ involves approximation by implicit difference schemes. An implicit difference approximation to $CP(A,f,u_0)$ on $[0,T]$ is defined by a partition $\{0 = t_0 < t_1 < \cdots < t_n \leq T\}$ of $[0,t_n]$, a corresponding sequence $\{g_i\}_{i=1}^n \subseteq X$ and a

starting value $x_0 \in X$. A solution of the approximation defined by $\{t_0 < t_1 < \cdots < t_n\}$, $\{g_i\}_{i=1}^n$ and x_0 is a piecewise constant function $v:[0,t_n] \to X$ satisfying $v(t) = v(t_i)$ for $t \in (t_{i-1},t_i)$ and

$$(1) \qquad \begin{cases} v(0) = x_0 \ , \\ \dfrac{v(t_i) - v(t_{i-1})}{t_i - t_{i-1}} + A(v(t_i)) = g_i \quad \text{for} \quad i = 1,2,\cdots,n \ . \end{cases}$$

The implicit difference approximation is ε-approximate to $CP(A,f,u_0)$ on $[0,T]$ if

$$(2) \qquad \begin{cases} \text{(i)} \quad 0 \le T - t_n \le \varepsilon, \ t_i - t_{i-1} \le \varepsilon \quad \text{for} \quad i = 1,2,\cdots,n \ , \\ \text{(ii)} \quad \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} \| f(\tau) - g_i \| d\tau \le \varepsilon \ , \\ \text{(iii)} \quad \| x_0 - u_0 \| \le \varepsilon \ . \end{cases}$$

The function $v:[0,t_n] \to X$ is an ε-difference approximate solution of $CP(A,f,u_0)$ on $[0,T]$ if it is a solution of an ε-approximate implicit difference scheme on $[0,T]$. We have:

Theorem 1 (Convergence of solutions of difference approximations): Assume there is an $\omega \in \mathbb{R}$ such that

$$(3) \qquad \begin{cases} \| x - \hat{x} + \lambda(A(x) - A(\hat{x})) \| \ge (1 - \lambda\omega) \| x - \hat{x} \| \\ \text{for} \quad x,\hat{x} \in D(A) \quad \text{and} \quad \lambda > 0 \ . \end{cases}$$

Let $u_0 \in \overline{D(A)}$, $T > 0$ and v_k be an ε_k-difference approximation to $CP(A,f,u_0)$ on $[0,T]$ where $\lim_{k\to\infty} \varepsilon_k = 0$. Then there is exactly one continuous function $u:[0,T] \to X$ such that $\lim_{k\to\infty} \| v_k(t) - u(t) \| = 0$ uniformly on compact subsets of $[0,T]$.

Assumption (3) is the only restriction on A, replacing the (Lipschitz) continuity and/or compactness one is accustomed to. Here some simple examples for orientation: If A is Lipschitz continuous with constant L, then $\omega = L$ works; if $X = \mathbb{R}$ and A is nonincreasing then $\omega = 0$ works; if X is a Hilbert space and A is linear, perhaps unbounded, self-adjoint and $A \ge 0$, or A is skew-adjoint, then $\omega = 0$ works. If (3) holds one says that $A + \omega I$ is accretive or $-(A + \omega I)$ is dissipative. Some differential operators and spaces in which they have accretive realizations are: (a) $A(u) = -\Delta u$ in L^p, $1 \le p \le \infty$, (b) $A(u) = -\Delta(u^\alpha)$, $\alpha > 0$, in L^1 and H^{-1}, (c) $A(u) = -(\Delta u)^\alpha$, $\alpha > 0$, in L^∞, (d) $A(u) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}(g_i(u))$ in L^1, (e) $A(u) = -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\left| \frac{\partial u}{\partial x_i} \right|^{q-1} \frac{\partial u}{\partial x_i} \right)$ for $q \ge 1$ in L^p, $1 \le p \le \infty$, (f) $A(u) = g(\text{grad } u)$ in L^∞. Theorem 1 is proved in [6].

Assuming that the hypotheses of the above theorem are satisfied, we denote the limiting function u whose existence is asserted by $K(A,f,u_0)$. One simply defines $u = K(A,f,u_0)$ to be the "solution" of $CP(A,f,u_0)$ on $[0,T]$ if $K(A,f,u_0)$ "exists" (i.e. the hypotheses of the theorem hold; in particular (3) holds and ε_k-difference approximate solutions v_k with $\varepsilon_k \to 0$ exist).

It would seem natural to discuss the relationship between the notion "$K(A,f,u_0)$" of solution of $CP(A,f,u_0)$ and more traditional ideas of a solution, which require the existence of derivatives, next. However, it is more convenient to turn to the question of when $K(A,f,u_0)$ exists first.

Everywhere below we assume that (3) holds and $f:[0,T] \to X$ is strongly integrable. The simplest condition which guarantees the existence of $K(A,f,u_0)$ is:

$$(4) \qquad \begin{cases} \text{There exists } \lambda_0 > 0 \text{ such that } R(I+\lambda A) = X \text{ for } 0 < \lambda < \lambda_0, \\ \text{where } R(I+\lambda A) \text{ is the range of } I+\lambda A \text{ .} \end{cases}$$

Indeed, then every implicit difference approximation to $CP(A,f,u_0)$ on $[0,T]$ with $\max(t_i - t_{i-1}) < \lambda_0$ has a solution. For, given $v(t_{i-1})$ and g_i, we can (by (4)) choose $v(t_i)$ so that $v(t_i) + (t_i - t_{i-1})A(v(t_i)) = v(t_{i-1}) + (t_i - t_{i-1})g_i$ and hence satisfy (1). If $\varepsilon > 0$, then $x_0 = u_0$ satisfies (2)(iii) and we can satisfy (2)(i) and (ii) with $t_n = T$ by choosing a step function $g:[0,T] \to X$ satisfying $\int_0^T \|f(\tau)-g(\tau)\| d\tau < \varepsilon$ and letting $\{t_1,t_2,\cdots,t_n\}$ be the nodes of g (with points added as needed to achieve $t_i - t_{i-1} \leq \varepsilon$) and $g_i = g(t_i-)$. If $f = 0$ we can choose $g_i = 0$ and $t_i = iT/n$ for n large enough so that $T/n \leq \varepsilon$. If A is accretive (i.e. (3) holds with $\omega = 0$) and $R(I + \lambda A) = X$ for $\lambda > 0$, then A is called m-accretive. This case frequently occurs in applications.

Much more subtle conditions than $R(I + \lambda A) = X$ guarantee the existence of ε-difference approximate solutions. For example, if $f = 0$ one has:

Theorem 2. Let

$$\liminf_{\lambda \to 0} \lambda^{-1} \text{ distance } (R(I+\lambda A),x) = 0 \quad \text{for} \quad x \in \overline{D(A)} \ .$$

Then $K(A,0,u_0)$ exists for every $u_0 \in \overline{D(A)}$.

This theorem is proved in [8]. Requiring that $K(A,f,u_0)$ exist for every integrable $f:[0,T] \to X$ and $u_0 \in \overline{D(A)}$ implies that $R(I+\lambda A) = X$ for $\lambda > 0$, and $\lambda\omega < 1$ (if the graph of A is closed). The proof of this fact involves showing that $K(A,f,u_0)$ exists whenever $u_0 \in D(A)$ and

$$(5) \qquad \begin{cases} f(t) \in \{y: \liminf_{\lambda \to 0} \lambda^{-1} \text{ distance } (R(I+\lambda A),x+\lambda y) = 0 \\ \qquad \text{for } x \in \overline{D(A)}\} \quad \text{for almost all } t \in (0,T) \ , \end{cases}$$

generalizing Theorem 2. See [8] concerning a generalization of Theorem 2 to cover a range of quite different possibilities.

We now discuss the relationship between $K(A,f,u_0)$ and more classical notions of solutions of $CP(A,f,u_0)$. A function $u:[0,T] \to X$ is a strong solution of $CP(A,f,u_0)$ on $[0,T]$ if: (i) $u(0) = u_0$, (ii) there is an integrable function $w:[0,T] \to X$ such that $u(t)-u(s) = \int_s^t w(\tau)d\tau$ for $t,s \in [0,T]$ (this condition on u will be abbreviated to $u \in W^{1,1}(0,T:X)$ and implies u' exists and $u' = w$ a.e.), and (iii) $u'(t) + A(u(t)) = f(t)$ a.e. $t \in [0,T]$. We have:

Theorem 3. (a) Let $CP(A,f,u_0)$ have a strong solution u on $[0,T]$. Then $K(A,f,u_0)$ exists and $u = K(A,f,u_0)$.

(b) Let $K(A,f,u_0)$ exist and $u = K(A,f,u_0) \in W^{1,1}(0,T;X)$. If the graph of A is closed and (5) holds, then u is a strong solution of $CP(A,f,u_0)$.

(c) Let $u = K(A,f,u_0)$ exist, $u_0 \in D(A)$ and $f:[0,T] \to X$ be of bounded variation. Then u is Lipschitz continuous.

(d) In addition to the conditions of (c), assume that X is reflexive and the graph of A is closed. Then u is a strong solution of $CP(A,f,u_0)$ on $[0,T]$.

The proof of Theorem 3(a) uses the fact that integrable functions can be well approximated by Riemann sums. See [7, Sections 4 and 10] for a more general result. Similarly, [7, Section 10] can be adapted in order to prove (b). The result of (c) can be deduced from the estimate

$$(6)\quad e^{-\omega t}\|u(t) - v(t)\| - e^{-\omega s}\|u(s) - v(s)\| \leq \int_s^t e^{-\omega\tau}\|f(\tau) - g(\tau)\|\,d\tau$$

which holds whenever $u = K(A,f,u(0))$, $v = K(A,g,v(0))$ and $0 \leq s \leq t \leq T$. Finally, (d) follows from the fact that if X is reflexive, then $u:[0,T] \to X$ is in $W^{1,1}(0,T;X)$ iff u is absolutely continuous and (b).

The final type of result we want to consider is the dependence of $K(A,f,u_0)$ on the data A,f,u_0. Inequality (6) implies that $K(A,f,u_0)$ is Lipschitz continuous in f and u_0 uniformly for A satisfying (3). Varying A also, one has:

Theorem 4. Let $\{A_n\}_{n=0}^{\infty}$ be a sequence of operators satisfying (3) with the same ω and $R(I+\lambda A_n) = X$ for small $\lambda > 0$. If $y \in X$ and $x_n + A(x_n) \ni y$ for $n = 0,1,2,$ implies $\lim_{n\to\infty} x_n = x_0$, then $\lim_{n\to\infty} K(A_n,f,u_{n0}) = K(A,f,u_{00})$ uniformly on $[0,T]$ provided that $u_{n0} \in \overline{D(A_n)}$ and $\lim_{n\to\infty} u_{n0} = u_{00}$.

Thus $K(A,f,u_0)$ depends continuously on the initial data, the forcing term and the the equation. For example, within the context of our nonlinear heat flow model, one could interpret changing the conduction coefficient $k(u)$ as varying A, and then deduce that the solution of (NDP) depends continuously on $k(u)$. A proof of Theorem 4 can be found in [2]. Recently, this continuous dependence result (in a somewhat different form, see [4]) was used to explain in a simple way a numerical method for the Stefan problem ([3]).

REFERENCES

[1] V. Barbu, Nonlinear Semigroups and Differential Equations in Banach Spaces, Noordhoff, Leyden, The Netherlands, 1976.

[2] Ph. Benilan, Equations d'evolution dans un espace de Banach queleonque et applications, Thesis, University of Paris, Orsay, 1972.

[3] A. Berger, H. Brezis, J. Rogers, to appear.

[4] H. Brezis and A. Pazy, Convergence and approximations of semigroups of nonlinear operators in Banach spaces, J. Functional Analysis 9 (1972), 63-74.

[5] M. G. Crandall, An introduction to evolution governed by accretive operators, Dynamical Systems, Vol. 1, An International Symposium, Academic Press, New York, 1976, 131-165.

[6] M. G. Crandall and L. C. Evans, On the relation of the operator $\partial/\partial s + \partial/\partial \tau$ to evolution governed by accretive operators, Israel J. Math. 21 (1975), 261-278.

[7] L. C. Evans, Nonlinear evolution equations in an arbitrary Banach space, Israel J. Math. 26 (1977), 1-42.

[8] M. Pierre, Un théorème général de génération de semi-groupes nonlinéaires, Israel J. Math. 23 (1976), 189-199.

HARMONIC FUNCTIONS ON REGIONS WITH REENTRANT CORNERS, PART I

J. Barkley Rosser
Mathematics Research Center
University of Wisconsin-Madison
Madison, Wisconsin 53706

ABSTRACT. It has been known for quite a while that if a function $u(x,y)$ harmonic in a region with reentrant corners, there are almost certainly infinite discontinuities of the first derivative of u in the neighborhood of the reentrant corner (or corners). Simple examples are for an L-shaped region or T-shaped region. Some instances of these have been treated by conformally mapping the region into the interior of a rectangle. Attempts to solve the problem as first posed by a finite difference scheme or a finite element scheme will usually give poor approximations near any reentrant corner because the finite differences or finite elements have large truncation errors when a first derivative is infinite. When conformal mapping is tried, the conformal maps are usually only approximate, and similar errors arise, for more or less similar reasons.

In view of recent work giving convergent expansions for u in the neighborhood of reentrant corners (see "Calculation of Potential in a Sector, Part I," by J. Barkley Rosser, MRC TSR #1535) one can now give accurate solutions for such problems. Some experiments with such regions are reported.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

1. Background. A function $u(x,y)$ is said to be harmonic if

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \ . \tag{1.1}$$

We approximate (1.1) by the familiar difference formula

$$u(x + h,y) + u(x - h,y) + u(x,y + h) + u(x,y - h) - 4u(x,y) \cong 0 \ . \tag{1.2}$$

The error involves fourth derivatives of u. We suppose that these exist and are reasonably well behaved. This is usually the case, so for a long time it has been customary to seek an approximate solution for (1.1) by solving the set of linear equations resulting from using various values of (x,y) in (1.2). If h is small, so that the error in (1.2) is small, the solution of (1.2) gives good approximations to the values of u at a set of grid points.

Unfortunately, if h is small, then one has a very large number of linear equations to solve, and the labor of computation is very great. Until the advent of the computer, one compromised by using a fairly large value of h, to curtail the calculation, but one had to be content with not a very good approximation.

One effort to use computers to improve this situation is embodied in item [1] of the Bibliography, by Kantorovich, Krylov, and Chernin. If one has values for u prescribed on the boundary of a rectangle, the tables in [1] allow one to get fairly quickly the solution of (1.2) inside the rectangle. (This can now be done more quickly by means of the Fast Fourier Transform, so that [1] is now obsolete.)

To show the effectiveness of their tables, the authors of [1] undertook to find the solution of (1.2) inside an L-shaped region (see Fig. 1). They prescribed values for u around the boundary, and used the Schwarz alternating procedure. Specifically, they first guessed values along CF. Using these with the boundary conditions, the tables gave the solution of (1.2) inside the rectangle ABFG. In particular, they gave values along HC. Using these with the boundary conditions, the tables gave the solution of (1.2) inside the rectangle HDEG. This led to a better guess for the values along CF. Using the better guess, the process was repeated. After a modest number of repetitions, the procedure converged to give about six decimal accuracy.

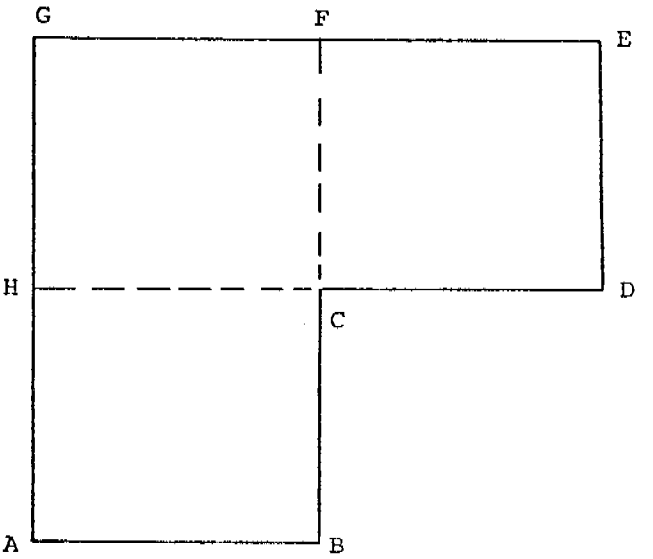


Figure 1

This was the solution of (1.2) inside the L-shaped region. As noted, the error of (1.2) as an approximation to (1.1) depends on fourth derivatives of u. In [2] and [3], which appeared six to four years before [1], Wasow and Lehman had shown that in the neighborhood of a reentrant corner (such as C in Fig. 1) one should expect to have an unbounded first derivative. With an unbounded first derivative, one cannot expect good behavior from fourth derivatives. So, in the neighborhood of C in Fig. 1, one should expect the solution of (1.2) (which was obtained in [1]) to be a poor approximation to the solution of (1.1).

2. Two lemmas. To get some comprehension of the results of Wasow and Lehman in [2] and [3], we use two lemmas, which we state here without proof.

Lemma 2.1. Let $a < b$. Let $f(a) = f(b) = 0$. Let $f(x)$ have almost everywhere a second derivative for $a \leq x \leq b$ which is of bounded variation. Then the Fourier series for $f(x)$ in the interval $a \leq x \leq b$,

$$f(x) = \sum_{m=1}^{\infty} D_m \sin \frac{\pi m(x - a)}{b - a} \tag{2.1}$$

with

$$D_m = \frac{2}{b - a} \int_a^b f(x) \sin \frac{\pi m(x - a)}{b - a} \, dx \, , \tag{2.2}$$

converges very rapidly, and a large number of the D_m can be calculated very quickly by means of the Fast Fourier Transform.

By "converges very rapidly" is meant that $|D_m|$ goes to zero at least of the order of m^{-3}. Thus one can truncate the series on the right of (2.1) after 500 terms and reasonably expect to get from six to eight significant decimal places correct. And the Fast Fourier Transform will enable one to calculate the needed 500 coefficients very quickly. The reasoning to establish this lemma is given in pp. 6-8 of [4].

Lemma 2.2. Let $u(x,y)$ be harmonic in a region, part of the boundary of which is a straight line segment. Let $u(x,y) = f(s)$ on this straight line segment, where s is length along the segment. Let $f(s)$ and its first n derivatives be continuous, and let the $(n + 1)$-st derivative be bounded and continuous except at a set of points of measure zero. Then each partial derivative of $u(x,y)$ of order $\leq n$ has a continuous extension to the straight line boundary.

Thm. 2.3 on p. 27 of [5] states this for a special case. The truth of the lemma in general follows easily from the special case.

In the present report, we shall confine our attention to the case where the prescribed values of u around the boundary are quite smooth; say that the third derivative is bounded and continuous except at a set of measure zero. It is planned to write a sequel to [5] explaining how to handle a variety of irregularities along the boundary. Certain sorts of irregularities that could occur along the boundary can be "removed" by the methods given on pp. 221-222 of [6]. So it does not seem unduly restrictive to confine our attention to harmonic functions $u(x,y)$ which are quite smooth around the boundary. In the present report we do so.

Let $u(x,y)$ be such a harmonic function in the L-shaped region of Fig. 2.

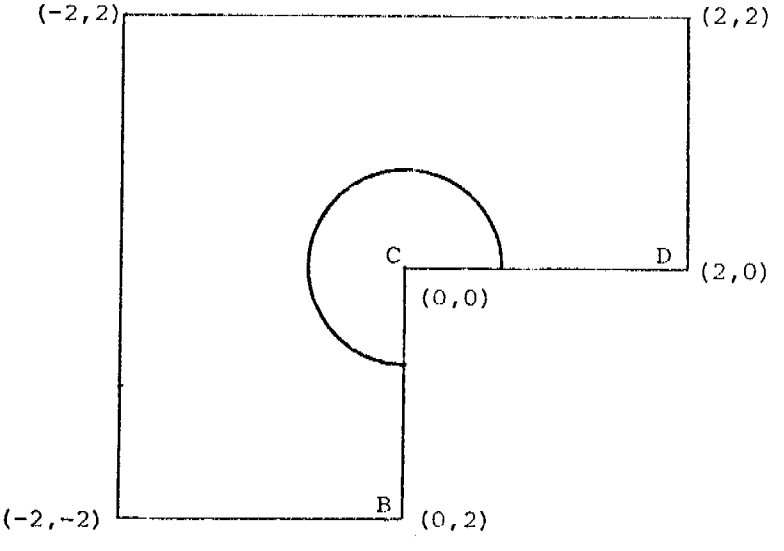


Figure 2

Indeed we will shortly specialize to prescribing that on the boundary in Fig. 2 we will have

$$u(x,y) = \frac{1}{2}\,\ell n\{(x+1)^2 + (y-1)^2\}\ . \tag{2.3}$$

In Fig. 2 we have shown three-quarters of a circle of radius A and center at the origin. We undertake to determine the behavior of $u(x,y)$ inside the three-quarters circle.

Choose $u^{(s)}(x,y)$ a function which in the interior of the figure is harmonic in the neighborhood of the sides BC and CD, including the three-quarters circle (one can take A quite small if need be), and which takes the same values along BC and CD that are prescribed for $u(x,y)$. Instructions for finding such a function $u^{(s)}(x,y)$ are set forth in [5]. If we have prescribed the particular boundary conditions (2.3) for $u(x,y)$, such a function is

$$\frac{1}{4}\,\mathcal{R}\{3\,\ell n(z+1-i) + \ell n(z+1+i) + \ell n(z-1-i) - \ell n(z-1+i)\}\ , \tag{2.4}$$

where we have taken

$$z = x + iy\ . \tag{2.5}$$

It can easily be verified that the right side of (2.4) satisfies (2.3) along the entire x-axis and the entire y-axis. Also (2.4) is harmonic except at the four points $z = \pm 1 \pm i$. To keep the three-quarters circle inside the region where (2.4) is harmonic, it suffices to take $A < \sqrt{2}$. We choose such a value for A, and proceed.

Since $u(x,y)$ and $u^{(s)}(x,y)$ take the same values along BC and CD, we conclude that along these two segments

$$u(x,y) - u^{(s)}(x,y) \tag{2.6}$$

is zero.

Measure the angle θ as usual, counterclockwise around the origin from CD. On the three-quarters circle, we will have (2.6) a function of θ only, since we have fixed A. Call this $f(\theta)$. As $u(x,y)$ and $u^{(s)}(x,y)$ take values along BC and CD that are differentiable an infinite number of times, it follows from Lemma 2.2 that $f(\theta)$ is infinitely differentiable as θ approaches $0+$ and $(3\pi/2)-$. For $0 < \theta < 3\pi/2$, $f(\theta)$ is the difference of two harmonic functions, and hence is infinitely differentiable. So for $0 \le \theta \le 3\pi/2$, $f(\theta)$ has derivatives of all orders.

As (2.6) is 0 on BC and CD, we have $f(0) = f(3\pi/2) = 0$. So we take $a = 0$ and $b = 3\pi/2$ in Lemma 2.1, and conclude that $f(\theta)$ has a rapidly converging Fourier series expansion for $0 \le \theta \le 3\pi/2$; put θ for x in (2.1). Because of the rapid convergence, we see that

(2.7) $$\sum_{m=1}^{\infty} D_m \left(\frac{r}{A}\right)^{2m/3} \sin \frac{2m\theta}{3}$$

is a harmonic function for $0 < r < A$ and $0 < \theta < 3\pi/2$. It equals $f(\theta)$ for $r = A$, and is 0 for $\theta = 0$ or $\theta = 3\pi/2$. But (2.6) satisfies these same conditions. As a harmonic function is uniquely determined in a region by its values around the boundary, we must have (2.6) equal to (2.7) inside and on the three-quarters circle. Solving for $u(x,y)$, we must have $u(x,y)$ equal to the sum of (2.4) and (2.7) inside and on the three-quarters circle.

For the particular boundary conditions which we have chosen (see (2.3)) we have $D_1 \neq 0$ in (2.7). So if we fix a value of θ, $0 < \theta < 3\pi/2$, and approach the origin along that ray, (2.7) will have an infinite first derivative. As (2.4) is harmonic inside the entire circle of radius A and center at the origin (we took $A < \sqrt{2}$), it has well behaved derivatives of all orders at C. So $u(x,y)$ must have an infinite derivative as r approaches zero.

Of course, if it had turned out that $D_1 = D_2 = D_4 = D_5 = 0$, then $u(x,y)$ would have had well behaved fourth derivatives. But it did not turn out that way. In [2] and [3], Wasow and Lehman made a study of the asymptotic behavior of harmonic functions near reentrant corners. Their studies were quite general, covering curved boundaries and a wide variety of conditions. The series they got were only asymptotic, but series like (2.7) were typical (except that (2.7) converges). Indeed, we are lucky with our particular problem, in that our $u^{(s)}(x,y)$ is harmonic in the neighborhood of the corner. More generally, $u^{(s)}(x,y)$ contributes additional complications, such as terms involving logarithms.

In view of this, one wonders why the authors of [1] managed to get such good results near the reentrant corner. This came about as follows. In order to be able to check if their procedure was giving the right answers, they took a problem in which the answers were known. They chose $u(x,y)$ a function that was well behaved over a much larger region than that shown in Fig. 1. From it, they read values around the boundary, and proceeded to solve, getting back $u(x,y)$ of course. Since they started with a function that was well behaved over a large region, including the reentrant corner, they insured that $D_1 = D_2 = D_4 = D_5 = 0$ in (2.7). So of course they had well behaved fourth derivatives, and (1.2) was an excellent approximation to (1.1), and their answers agreed closely with the true values. Very comforting for them, but very misleading for the reader. Had they used the boundary conditions (2.3), their answers would have been very poor near C. On the other hand, they probably did not have a way to get the correct answers for the boundary conditions (2.3), and so would not have known if they had good answers or not.

3. The solution inside a rectangle. This brings us to the crucial question. How does one get correct answers with boundary conditions like (2.3)? First we have to have a technique for carrying out a solution inside a rectangle, which we now explain. Given a rectangle with smooth boundary conditions prescribed around its perimeter, how does one determine a $u(x,y)$ which is harmonic in the interior and takes the prescribed values on the perimeter?

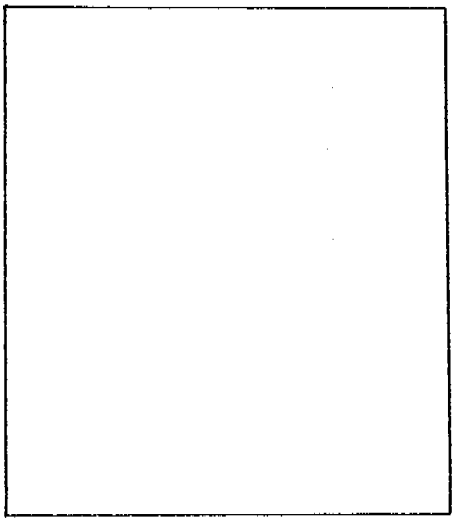


Figure 3

Consider

$$u^*(x,y) = u(x,y) - A - Bx - Cy - Dxy \,. \tag{3.1}$$

This is harmonic, and has smooth boundary conditions. It is easy to choose A, B, C, and D so that $u^*(x,y)$ takes the value zero at each corner of the rectangle. Along the top let $u^*(x,y) = f(x)$, for $a \leq x \leq b$. Our choice of A, B, C, and D assures that $f(a) = f(b) = 0$. Also, as we were assuming smooth boundary conditions, let us say that that assures that $f(x)$ has almost everywhere a second derivative for $a \leq x \leq b$ which is of bounded variation. So $f(x)$ satisfies the conditions of Lemma 2.1. We get its Fourier expansion, (2.1), with the D_m defined by (2.2). Consider

$$u_t(x,y) = \sum_{m=1}^{\infty} D_m \frac{\sinh \frac{\pi m(y-c)}{b-a}}{\sinh \frac{\pi m h}{b-a}} \sin \frac{\pi m(x-a)}{b-a} \,, \tag{3.2}$$

where h is the height of the rectangle and c is the value of y at the bottom of the rectangle. Clearly $u_t(x,y)$ is harmonic. It is zero along the left side of the rectangle $(x = a)$, it is zero along the right side of the rectangle $(x = b)$, it is zero along the bottom of the rectangle $(y = c)$, and it equals $f(x)$ along the top of the rectangle; that is, on the top it agrees with $u^*(x,y)$. We carry out an analogous construction for each of the other three sides of the rectangle, and add together the resulting four series. Since the sum agrees with $u^*(x,y)$ on the entire perimeter, it has to be equal to $u^*(x,y)$ throughout the rectangle. Then we determine $u(x,y)$ from (3.1).

Armed with this technique, let us return to the problem of Fig. 1. Values of $u(x,y)$ have been prescribed around the boundary (for example, see (2.3)). We guess values along CF. With the boundary conditions, this gives values

around the perimeter of the rectangle ABFG. As described just above, we get a harmonic function inside this rectangle which takes the prescribed boundary conditions. It gives us values along HC. With these and the boundary conditions, we have values around the perimeter of the rectangle HDEG. From these, we get values in the interior, including along the line CF. This will be an improvement over our first guess.

We repeat the process. In an actual calculation, with the conditions (2.3), it took about fifteen iterations for convergence. However, because the Fast Fourier Transform gets the D_m very quickly, the calculation to convergence did not take very long. However, it did not converge to the $u(x,y)$ we were seeking. Recall that in Lemma 2.1, it was required that $f(x)$ have a second derivative of bounded variation. But the $u(x,y)$ defined by conditions (2.3) has an infinite first derivative as one approaches C along CF.

This seems too bad. However, the procedure we just described is not without value. In fact, it will be the one we will use in the end, but with a slight modification. Our difficulty (refer to Fig. 1) is that, along the lines BF and HD, the function $u(x,y)$ that we are trying to determine does not have almost everywhere a second derivative of bounded variation. If we should try this procedure on a $u(x,y)$ which does have almost everywhere a second derivative of bounded variation along the lines BF and HD, we would succeed admirably in determining that $u(x,y)$, and in terms of rapidly converging Fourier series. All we need for $u(x,y)$ is to know its values around the boundary, and to be assured that it is sufficiently smooth along the lines BF and HD.

4. A slight modification. Recall that the $u(x,y)$ we are seeking to determine equals (2.4) plus (2.7) inside and on the three-quarters circle, and that (2.4) is very smooth along the lines BF and HD. Because of this, we will show that

$$u(x,y) - D_1\left(\frac{r}{A}\right)^{\frac{2}{3}} \sin\frac{2\theta}{3} - D_2\left(\frac{r}{A}\right)^{\frac{4}{3}} \sin\frac{4\theta}{3} \tag{4.1}$$

has a second derivative of bounded variation along both the lines BF and HD. Along BC and CD, (4.1) equals the right side of (2.3), which is very smooth. Inside and on the three-quarters circle, (4.1) equals (2.4) plus the remainder of the series (2.7), which is smooth enough. And from the three-quarters circle out to F or H, (4.1) is the sum of three harmonic functions out to a straight line border along which their boundary values are infinitely differentiable; by Lemma 2.2, all derivatives exist continuously out to the border.

If we could somehow determine the values of D_1 and D_2, we could determine (4.1) by the procedure of the previous section. We certainly can determine the values of (4.1) around the boundary; we had had to choose a value of A before the values of D_1 and D_2 could be defined (in fact, we had chosen $A = 1$ for our calculation), and the values of $u(x,y)$ are given by (2.3). Also, as we have just carefully ascertained, (4.1) has a second derivative of bounded variation along the lines BF and HD. Being given the values of D_1 and D_2, we could then calculate $u(x,y)$ from (4.1).

So we are faced with the problem of determining D_1 and D_2.

We remind the reader that a computer operates linearly. To calculate (4.1) by the procedure of the previous section, we would get the same numerical answers by either of the two following procedures.

(1) Apply the procedure to the total function (4.1).

(2) Apply the procedure first to $u(x,y)$, getting some Fourier expansions S^I, then apply the procedure to

$$\left(\frac{r}{A}\right)^{\frac{2}{3}} \sin\frac{2\theta}{3}, \tag{4.2}$$

getting some Fourier expansions S^{II}, then apply the procedure to

$$\left(\frac{r}{A}\right)^{\frac{4}{3}} \sin\frac{4\theta}{3}, \tag{4.3}$$

getting some Fourier expansions S^{III}, and finally combine the various Fourier expansions into

$$S^I - D_1 S^{II} - D_2 S^{III}. \tag{4.4}$$

Although S^I will be a poor representation of $u(x,y)$, as we observed in the previous section, and S^{II} and S^{III} will be poor representations of (4.2) and (4.3), for similar reasons, the combination (4.4) will be an excellent representation of (4.1), since the linearity of the computer assures that it comprises the same numbers that one would get by applying the procedure of the previous section to the entirety of (4.1).

With no more ado, let us proceed to determine S^I, S^{II}, and S^{III}. Considering D_1 and D_2 as two (as yet) unknown parameters, we can take (4.4) as representing (4.1). Subtracting (2.4) from (4.4), we will have a representation of

$$u(x,y) - u^{(s)}(x,y) - D_1\left(\frac{r}{A}\right)^{\frac{2}{3}} \sin\frac{2\theta}{3} - D_2\left(\frac{r}{A}\right)^{\frac{4}{3}} \sin\frac{4\theta}{3}. \tag{4.5}$$

That is, using (4.4) minus (2.4), we can actually calculate values of (4.5) at any points of the L-shaped region of Fig. 2, except that the values will come out as linear combinations of D_1 and D_2.

Observe that (4.5) is zero on both the lines BC and CD, since (2.6) was. So, by the same method that we used to get the expansion (2.7) for (2.6) inside and on the three-quarters circle, we can get an expansion like (2.7) for (4.5). Obviously, this expansion has to consist of (2.7) with the first two terms deleted. So, when we take $m = 1$ and 2 in (2.2) to get D_1 and D_2, we must

get the value zero. But, as the values of $f(x)$ in (2.2) are taken from (4.4) minus (2.4), the numerical quadratures to determine (2.2) must yield linear combinations of D_1 and D_2. Putting these linear combinations equal to zero for $m = 1$ and $m = 2$ gives us two simultaneous linear equations for D_1 and D_2. We solve these. Putting the solutions into (4.4) gives Fourier expansions for (4.1). But now we know D_1 and D_2, and so can calculate $u(x,y)$ from (4.1).

5. Acknowledgements. In a Part II, we will report numerical results for the problem considered above, and results for other problems that can be handled by similar techniques. I wish to express my gratitude to Gershon Kedem for his assistance with these activities. He carried out the needed programming, and supervised the actual calculations. He also suggested simplifications, and helped me get my thoughts in order and keep track of the details.

BIBLIOGRAPHY

[1] L. V. Kantorovich, V. I. Krylov, and K. Ye Chernin, "Tables for the numerical solution of boundary value problems of the theory of harmonic functions," Ungar Publishing Co., New York, 1963.

[2] Wolfgang Wasow, "Asymptotic development of the solution of Dirichlet's problem at analytic corners," Duke Mathematical Journal, vol. 24 (1957), pp. 47-56.

[3] R. Sherman Lehman, "Developments at an analytic corner of solutions of elliptic partial differential equations," Journal of Mathematics and Mechanics, vol. 8 (1959), pp. 727-760.

[4] J. Barkley Rosser, "Fourier series in the computer age," MRC Technical Summary Report #1401, February 1974. Also appeared as Brunel University Report TR/43, May 1974. Also appeared in "Transactions of the Twentieth Conference of Army Mathematicians," 1974, Army Research Office, Box CM, Durham, N.C.

[5] J. Barkley Rosser, "Calculation of potential in a sector, Part I," MRC Technical Summary Report #1535, May 1975.

[6] W. E. Milne, "Numerical solution of differential equations," John Wiley and Sons, Inc., 1960.

ADAPTIVE ACCELERATION OF SSOR FOR SOLVING LARGE LINEAR SYSTEMS

Vitalius Benokraitis
Ballistic Modeling Division
U. S. Army Ballistic Research Laboratory

ABSTRACT. Symmetric successive overrelaxation (SSOR) for solving large, sparse systems of linear equations involves the estimation of a parameter ω. An adaptive procedure is outlined for improving the estimates for ω and the spectral radius $S(S_\omega)$ of the iteration matrix S_ω. These estimates are then used in the SSOR method with Chebyshev acceleration. The objective is to achieve convergence in only a few more iterations than would be required if the best possible values of ω and $S(S_\omega)$ were used from the outset. The method is applied to obtain finite difference solutions of a number of generalized Dirichlet problems. In certain cases, the number of iterations required using the adaptive procedure increases like $h^{-\frac{1}{2}}$, where h is the mesh size.

1. INTRODUCTION. We shall be concerned with iteratively determining the N-vector u of a large, sparse linear system

$$(1) \quad Au = b$$

where A is a real, symmetric, positive definite matrix of order N and b is a given N-vector. Such systems arise in the finite difference solution of elliptic boundary value problems. Particularly, we shall develop an adaptive scheme based on the symmetric SOR (SSOR) iterative method with Chebyshev acceleration. Related work which has recently appeared includes Axelsson (1972), Hayes and Young (1977) and Young (1974a, 1974b, 1974c, 1977).

2. BASIC METHOD. By defining

$$B = I - D^{-1}A = L + U$$

$$c = D^{-1}b$$

where D = diag (A) and L and U are strictly lower and upper triangular matrices, respectively, we may replace the system (1) by

$$u = Bu + c.$$

The SSOR method (Sheldon (1955)) is then defined by forming a single SSOR iteration from a forward SOR iteration followed by a backward SOR iteration; that is, for n = 0, 1, 2, ... we set

$$(2)\qquad \begin{cases} u^{(n+\frac{1}{2})} = \omega(Lu^{(n+\frac{1}{2})} + Uu^{(n)} + c) + (1-\omega)u^{(n)} \\ u^{(n+1)} = \omega(Lu^{(n+\frac{1}{2})} + Uu^{(n+1)} + c) + (1-\omega)u^{(n+\frac{1}{2})} \end{cases}$$

where $u^{(0)}$ is an arbitrary initial approximation to the solution u, and ω is a real relaxation parameter such that $0<\omega<2$.

Elimination of $u^{(n+\frac{1}{2})}$ in (2) gives

$$u^{(n+1)} = S_\omega u^{(n)} + k_\omega$$

where

$$S_\omega = U_\omega L_\omega$$

$$= \left\{(I-\omega U)^{-1}\ (\omega L + (1-\omega)I)\right\}\left\{(I-\omega L)^{-1}(\omega U+(1+\omega)I)\right\}$$

$$k_\omega = \omega(2-\omega)\ (I-\omega U)^{-1}\ (I-\omega L)^{-1}c$$

Note that L_ω corresponds to the familiar SOR iteration matrix. The backward SOR operator U_ω is defined analogously.

If storage for an extra N-vector is provided, the work required for one SSOR iteration may be reduced to about the work necessary for a single SOR iteration. The work-saving technique is due to Niethammer (1964) and is described in Benokraitis (1974, 1976) and Young (1977). The method has been rediscovered by Conrad and Wallach (1977).

The SSOR method converges if $S(S_\omega)$, the spectral radius of the iteration matrix S_ω, is less than 1, which holds if $0<\omega<2$ and A is positive definite. The rate of convergence is governed by the ordering of the equations and by the parameter ω. Assuming the natural ordering, Young (1974a, 1974b, 1977) has shown that for a certain discrete generalized Dirichlet problem one can choose a "good" ω depending on bounds for the eigenvalues of B and LU so that the SSOR method converges with the same order-of-magnitude as the SOR method. For a finite difference discretization with mesh size h, the number of iterations required for both methods increases like h^{-1}.

Therefore, even by employing Niethammer's work-saving scheme, there is little justification for using SSOR. However, the SSOR method can be accelerated by an order-of-magnitude by means of Chebyshev semi-iteration since the eigenvalues of the matrix S_ω are real and nonnegative. (Chebyshev semi-iteration was first studied by Varga (1957) and Golub and Varga (1961).) This approach is precluded for SOR with optimum $\omega = \omega_b$ since many of the eigenvalues of L_{ω_b} are complex. (See Varga (1957) and Young (1971).) Also, there is no improvement when semi-iteration is applied to SOR with $1<\omega<\omega_b$. (See Kincaid (1974).) For accelerating the Gauss-Seidel method (SOR with $\omega = 1$), see Sheldon (1959) and Young (1971).

3. ACCELERATED METHOD. The optimum semi-iterative method based on SSOR, denoted by SSOR-SI, is defined by

$$(3) \qquad u^{(n+1)} = \rho_{n+1}\left\{\bar{\rho}\,(S_\omega u^{(n)} + k_\omega) + (1-\bar{\rho})\,u^{(n)}\right\} + (1-\rho_{n+1})\,u^{(n-1)}$$

Here

$$\bar{\rho} = \frac{2}{2-S(S_\omega)}$$

$$\rho_1 = 1$$

$$\rho_2 = (1 - \sigma^2/2)^{-1}$$

$$\rho_{n+1} = (1 - \frac{\sigma^2\rho_n}{4})^{-1} \quad , n = 2, 3, \ldots$$

where

$$\sigma = \frac{S(S_\omega)}{2-S(S_\omega)}$$

In order to apply the SSOR-SI method, we must determine the two parameters ω and $S(S_\omega)$. Some a priori methods for obtaining these parameters are discussed by Habetler and Wachspress (1961), Evans and Forrington (1963), Young (1974a, 1974b, 1977), and Benokraitis (1974, 1976). Since finding these parameters may take as much work as solving the original problem, we are led to consider adaptive techniques which approximate the parameters and at the same time improve the solution of the linear system.

4. FOUNDATION FOR ADAPTIVE METHOD. We begin by characterizing the eigenvalues of S_ω in terms of certain inner products. This result is due to Habetler and Wachspress (1961). See also Young (1977).

THEOREM 1. Let λ be an eigenvalue of S_ω where $0<\omega<2$ and let v be an associated eigenvector. Then λ may be represented by

$$\lambda = 1 - \omega(2-\omega)\frac{1-\alpha}{1-\omega\alpha+\omega^2\beta} = \phi(\omega,v) \tag{4}$$

where

$$\begin{cases} \alpha = \dfrac{(v,DBv)}{(v,Dv)} \\ \beta = \dfrac{(v,DLUv)}{(v,Dv)} \end{cases} \tag{5}$$

THEOREM 2. The representation $\phi(\omega,v)$ given by (4) for any vector $v \neq 0$ is a Rayleigh quotient with respect to the vector

$$w = (I-\omega\tilde{U})D^{\frac{1}{2}}v$$

and the symmetric matrix

$$\bar{S}_\omega = (I-\omega\tilde{U})D^{\frac{1}{2}}S_\omega D^{-\frac{1}{2}}(I-\omega\tilde{U})^{-1}$$

where

$$\tilde{U} = D^{\frac{1}{2}}UD^{-\frac{1}{2}} .$$

That is,

$$\phi(\omega,v) = \frac{(w,\bar{S}_\omega w)}{(w,w)} .$$

Furthermore, $\bar{S}_\omega$ is similar to S_ω and

$$\phi(\omega,v) \leq S(\bar{S}_\omega) = S(S_\omega). \tag{6}$$

Proof: See Benokraitis (1974).

We emphasize that (6) holds for any nonzero vector v, not just for eigenvectors of S_ω. However, the closer we approach a fundamental eigenvector, the closer $\rho(S_\omega)$ we shall be able to determine $S(S_\omega)$ from $\phi(\omega,v)$ given by (4). Therefore, it would be fortunate if somehow we could determine the fundamental eigenvector without deviating from the path of improving the approximate solution of (1). A clue leading to the desired situation is contained in the following theorem (cf. Young (1974c)).

THEOREM 3. The pseudo-residual vector

$$\delta^{(n)} = S_\omega u^{(n)} + k_\omega - u^{(n)} \tag{7}$$

where $u^{(n)}$ is the latest SSOR-SI iterate, satisfies

$$\delta^{(n)} = P_n(S_\omega)\delta^{(0)} \quad . \tag{8}$$

Here

$$P_n(S_\omega) = \frac{T_n\left(\dfrac{2S_\omega}{S(S_\omega)} - 1\right)}{T_n\left(\dfrac{2}{S(S_\omega)} - 1\right)} .$$

where $T_n(x)$ is the n^{th} degree Chebyshev polynomial defined by the three-term recurrence relation

$$T_0(x) = 1, \; T_1(x) = x$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \; n \geq 1 \quad .$$

Proof: See Benokraitis (1974).

We note that if $P_n(S_\omega)$ is replaced by S_ω^n then (8) reminds us of the power method for computing the dominant eigenvector. With this motivation, the next theorem comes as no surprise.

THEOREM 4. The pseudo-residual vector $\delta^{(n)}$ given by (7) converges in direction to the eigenvector associated with the eigenvalue $S(S_\omega)$ as n tends to infinity.

Proof: See Benokraitis (1974). Also compare with Diamond (1971) and Hageman (1972).

By Theorem 4 it is possible, then, to determine the fundamental eigenvector with a little additional effort by computing the pseudo-residual vector. However, since

$$\delta^{(n)} = \underbrace{S_\omega u^{(n)} + k_\omega}_{\text{one SSOR iteration}} - \underbrace{u^{(n)}}_{\text{latest SSOR-SI iteration}} = \tilde{u}^{(n)} - u^{(n)}$$

and since

$$\tilde{u}^{(n)} = S_\omega u^{(n)} + k_\omega$$

must be computed as part of the next SSOR-SI iteration $u^{(n+1)}$ (see (3)), the pseudo-residual vector is essentially obtained as a byproduct of applying the SSOR-SI method.

5. ADAPTIVE METHOD. By using Theorems 1, 2, and 4 as a foundation we are able to present the basic structure for an adaptive procedure. For detailed descriptions of this method and several possible variations for the adaptive acceleration of SSOR, see Benokraitis (1974).

We state the steps of the "algorithm" in outline form with a synopsis of the controlling theorem(s) in the heading. We use the word "algorithm" loosely, since admittedly much is left unspecified.

I. Theorem 2. For any $v \neq 0$, $\phi(\omega,v) \leq S(S_\omega)$.

1. Choose, v_1, $v_2 \neq 0$.

2. Observe

   a. $\phi_1 = \phi(\omega,v_1) \leq S(S_\omega)$

   b. $\phi_2 = \phi(\omega,v_2) \leq S(S_\omega)$

   c. $\Phi_1(\omega) = \max_{v_1,v_2} (\phi_1, \phi_2) \leq S(S_\omega)$

3. Minimize $\Phi_1(\omega)$ with respect to ω to obtain estimate ω_1.

4. Choose $\Phi_1(\omega_1)$ as $S_E(S_{\omega_1})$, an estimate of $S(S_{\omega_1})$. (The situation is depicted in Figure 1.)

II. Theorem 4. The pseudo-residual vector $\delta^{(n)}$ converges in direction to dominant eigenvector v.

1. Set i = 1

2. Iterate n times with SSOR-SI with parameters ω_i, $S_E(S_{\omega_i})$, test for for convergence

3. Compute

$$\delta^{(n)} = (S_\omega n^{(n)} + k_\omega) - u^{(n)}$$

which approaches dominant eigenvector.

4. Check if parameters should be changed.

a. If $\phi(\omega_i, \delta^{(n)}) \leq S_E(S_{\omega_i})$ do not change parameters. Go to II.2

b. If $\phi(\omega_i, \delta^{(n)}) > S_E(S_{\omega_i})$ continue to step III to change parameters.

III. Theorems 1, 2, 4. As $\delta^{(n)}$ approaches dominant eigenvector, $\phi(\omega_i, \delta^{(n)})$ approaches $S(S_{\omega_i})$.

1. Set $v_{i+2} = \delta^{(n)}$

$$\phi_{i+2} = \phi(\omega, v_{i+2})$$

2. Observe $\Phi_{i+1}(\omega) = \max_{v_k, k=1, \ldots, i+2} (\phi_1, \phi_2, \ldots, \phi_{i+2}) \leq S(S_\omega)$

3. Minimize $\Phi_{i+1}(\omega)$ with respect to ω to obtain next estimate ω_{i+1}.

4. Choose $\Phi_{i+1}(\omega_{i+1})$ as $S_E(S_{\omega_{i+1}})$, an estimate of $S(S_{\omega_{i+1}})$. Set i = i+1. Go to II.2. Process is continued until convergence.

We briefly discuss how to choose n in step II.2. Here we make use of the average and asymptotic average rates of convergence for the SSOR-SI method (Young (1971)). A strategy which produces acceptable results is to choose n so that the average rate of convergence after n iterations is 90% of the asymptotic average rate of convergence. The convergence rates are computed using the latest estimate of $S(S_\omega)$. That is, n is chosen to be the least n which satisfies

$$-\frac{1}{n}\log\frac{2r^n}{1+r^{2n}} \geq .9\ (-\log r)$$

where

$$r = \frac{1 - \sqrt{1-S_E(S_\omega)}}{1 + \sqrt{1-S_E(S_\omega)}}.$$

A word about the additional work required in the adaptive algorithm is in order. Mainly, the added expense comes in changing the parameters. This involves the computation of α and β, two quotients of inner products in the formula for $\phi(\omega,v)$ given in (4)-(5). For problems of the type discussed in Section 6, to compute α and β requires approximately $28J^2$ arithmetic operations if the mesh size is $h = 1/J$. One SSOR-SI iteration requires approximately $39J^2$ operations. Therefore, four parameter changes are approximately equivalent to three SSOR-SI iterations in terms of work performed. Since no more than four parameter changes were required for the problems considered, the number of iterations for the adaptive algorithm should effectively be increased by about three iterations.

6. NUMERICAL EXAMPLES. We present results for a sample of the generalized Dirichlet problems considered. The results are given in graphic form in Figures 2-4. In each figure, we give the differential equation, the region considered and the boundary values. We replace the differential equation by a 5-point symmetric difference equation (see Young (1977)).

The number of iterations required for varying mesh sizes is recorded for optimum, adaptive, and estimated SSOR-SI parameters. In the adaptive case, the subscript on the number of iterations indicates the number of parameter changes required. The estimated parameters are the values of Young (1977) which depend on bounds for the eigenvalues of B and LU. (For Problem 3, the results for the estimated parameters are not given since an excessive number of iterations are required.) The slopes s of the lines indicate that the number of iterations required for convergence increases like h^{-s}.

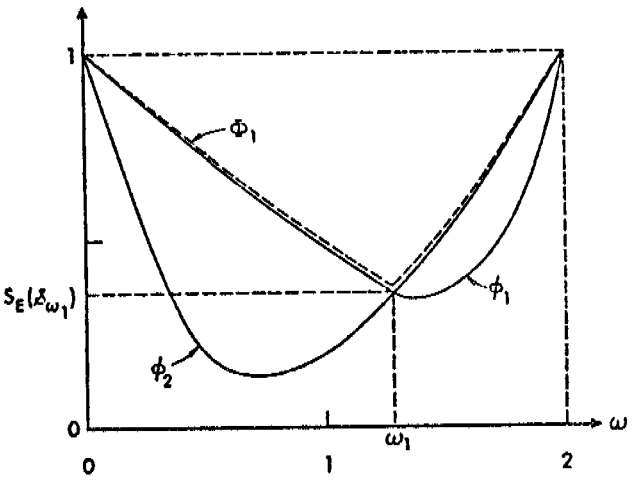


Fig. 1. Determination of First Approximation of ω and $S(S_\omega)$

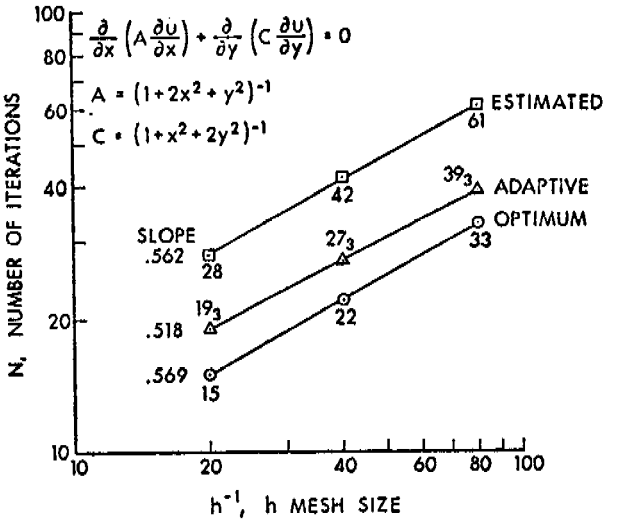


Fig. 2. Prob. 1. Unit Square w/Zero Boundary Values, Except Unity on Side Y = 0

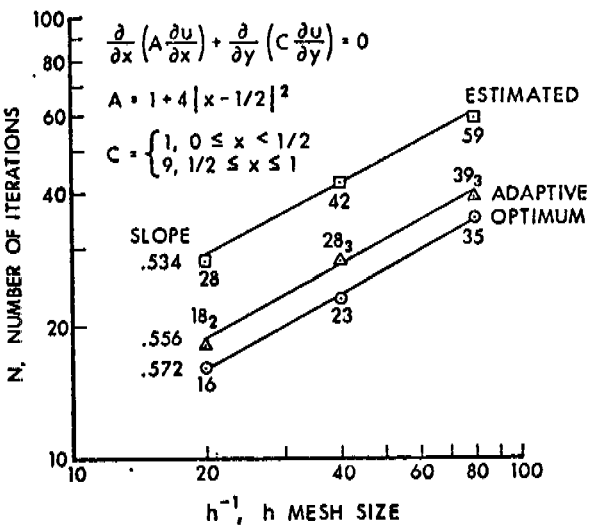


Fig. 3. Prob. 2. Unit Square w/Zero Boundary Values, Except Unity on Side Y = 0

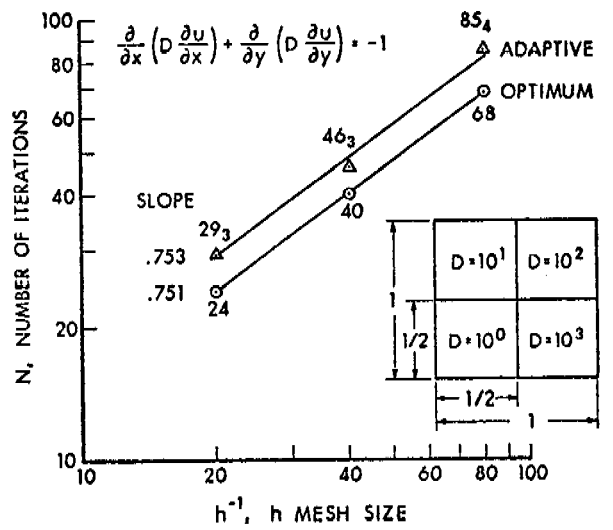


Fig. 4. Prob. 3. Unit Square w/Zero Boundary Values

For smooth and some discontinuous coefficients (Problems 1 and 2), the number of iterations required behaves like $h^{-\frac{1}{2}}$, an order-of-magnitude better than SOR or SSOR. For cases involving higher discontinuity (Problem 3), the behavior is like $h^{-3/4}$.

ACKNOWLEDGEMENT

This is a small part of a thesis written under the direction of Professor David M. Young, whose guidance is gratefully acknowledged.

REFERENCES

1. Axelsson, O. (1972), "A Generalized SSOR Method," BIT 13, 443-467.

2. Benokraitis, V.J. (1974), "On the Adaptive Acceleration of Symmetric Successive Overrelaxation," Doctoral Thesis, University of Texas, Austin.

3. Benokraitis, V. (1976), "An Improved Iterative Method For Optimizing Symmetric Successive Overrelaxation," in ARO Report 76-3, Proceedings of the 1976 Army Numerical Analysis and Computers Conference, 133-140.

4. Conrad, V. and Y. Wallach (1977), "A Faster SSOR Algorithm," Numer. Math. 27, 371-372.

5. Diamond, M. A. (1971), "An Economical Algorithm for the Solution of Finite Difference Equations," Report UIUC DCS-R-71-492, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois.

6. Evans, D. J. and C. V. D. Forrington (1963), "An Iterative Method for Optimizing Symmetric Successive Overrelaxation," Comput. J. 6, 271-273.

7. Golub, G. H. and R. S. Varga (1961), "Chebyshev Semi-iterative Methods, Successive Overrelaxation Iterative Methods, and Second-order Richardson Iterative Methods," Numer. Math., Parts I and II, 3, 147-168.

8. Habetler, G. J. and E. L. Wachspress (1961), "Symmetric Successive Overrelaxation in Solving Diffusion Difference Equations," Math. Comp. 15, 356-362.

9. Hageman, L. A. (1972), "The Estimation of Acceleration Parameters for the Chebyshev Polynomial and the Successive Overrelaxation Iteration Methods," Report WAPD-TM-1038, Bettis Atomic Power Laboratory, Westinghouse Electric Corp., Pittsburgh, Pennsylvania.

10. Hayes, L. J. and D. M. Young (1977), "The Accelerated SSOR Method for Solving Large Linear Systems: Preliminary Report," Report CNA-123, Center for Numerical Analysis, The University of Texas, Austin, Texas.

11. Kincaid, D. R. (1974), "On Complex Second-degree Iterative Methods," SIAM J. Numer. Anal. 11, 211-218.

12. Niethammer, W. (1964), "Relaxation bei Komplexen Matrizen," Math. Zeitsch. 86, 34-40.

13. Sheldon, J. W. (1955), "On the numerical Solution of Elliptic Difference Equations," Math. Tables Aids Comput. 9, 101-112.

14. Sheldon, J. W. (1959), "On the Spectral Norms of Several Iterative Processes," J. Assoc. Comput. Mach. 5, 39-46.

15. Young, D. M. (1971), Iterative Solution of Large Linear Systems, Academic Press, New York.

16. Young, D. M. (1973a), "On the Accelerated SSOR Method for Solving Elliptic Boundary Value Problems," in Lecture Notes in Mathematics, A. Dold and B. Eckmann (eds.), Vol. 363, Conference on the Numerical Solution of Differential Equations, Dundee 1973, (G. A. Watson, ed.), Springer-Verlag, New York, 195-206.

17. Young, D. M. (1974b), "Solution of Linear Systems of Equations," in Numerical Solutions of Partial Differential Equations, J.G. Gram (ed.), D. Reidel Pub. Col., Holland, 35-54 (Proceedings of Conference "Advanced Study Institute on Numerical Solution of Partial Differential Equations," Kjeller, Norway, Aug. 20-24, 1973).

18. Young, D. M. (1974c), "Stopping Criteria and Adaptive Parameter Determination for Iterative Methods for Solving Large Linear Systems," Proceedings of the Gatlinburg VI Symposium on Numerical Algebra, Munchen, Germany, December 15-22, 1974.

19. Young, D. M. (1977), "On the Accelerated SSOR Method for Solving Large Linear Systems," Advances in Mathematics 23, 215-271.

APPLICATION OF MACSYMA IN THE SOLUTION OF BOUNDARY VALUE PROBLEMS

Elizabeth Cuthill and L. Kenton Meals
Computation, Mathematics and Logistics Department
David W. Taylor Naval Ship Research and Development Center
Bethesda, Maryland

ABSTRACT

MACSYMA (Project MAC SYmbolic MAnipulation System) is used to develop a number of solutions for a sample linear boundary value problem, and results are compared.

A brief outline of MACSYMA capabilities is given followed by a general description of the class of problems treated, and the specific boundary value problem of this class used to exemplify the application of MACSYMA.

A brief overview of the approach to solution with MACSYMA and a MACSYMA demonstration of this approach for one approximate solution of the sample problem is given.

Ten approximations to the solution of the boundary value problem obtained using MACSYMA, are compared with the true solution by means of MACSYMA-generated error curves.

INTRODUCTION

MACSYMA, Project MAC Symbolic Manipulation System [1], [2] was developed under the sponsorship of the Advanced Research Projects Agency of the Dept. of Defense by the Mathlab Group of the Laboratory for Computer Science (formerly Project MAC) at the Massachusetts Institute of Technology where it is installed and maintained. It is, however, available to a wide community of users via ARPANET, a national Defense Communications Agency operated computer network. MACSYMA is an interactive system that handles numeric as well as symbolic manipulation. Many applications of MACSYMA are cited in [3].

More than ten years ago a project was established in the Applied Mathematics Laboratory (now the Computation, Mathematics, and Logistics Department) of the David Taylor Model Basin (now the David W. Taylor Naval Ship Research and Development Center) to exploit the combined capabilities of digital computers for carrying out extensive calculations and for manipulating mathematical operators in symbolic form. The symbolic manipulation system used was FORMAC [4], the computer system was the IBM 7090, and the problem area was the solution of boundary and initial value problems, especially those arising in mathematical physics.

Two general approaches to the solution of such problems are:

a. The determination of fixed points of operators by means of iterative techniques.

b. The determination of a set of parameters which will minimize in some sense the difference between the desired solution and an approximating function. The approximating function will involve parameters and independent variables. Collocation methods, least squares methods, finite difference methods, methods based on Taylor series expansion, Galerkin's method, and many others can be viewed as being of this type.

Some computer routines that were developed to apply the second of the above techniques to linear problems are given in [5]. These routines permitted the needed equations to be stated symbolically as part of the input (differential equations, equations required to describe auxiliary conditions, etc.), together with a numerical specification of the regions in which they were to be applied. The symbolic form of the approximating function to be used could also be part of the input. Then an approximate solution function was generated, which minimized an error norm in the least squares sense. This included collocation as a special case. The routines

developed proved very flexible and useful for curve fitting as well as for approximating solutions of problems involving differential equations despite the limitations of the computers then in use. It is the purpose of this paper to describe more recently developed MACSYMA programs of the same family.

This paper contains, in addition to a brief overview of some related efforts, a description of the mathematical approach used in the computer programs plus some examples illustrating their performance. The accuracy of the solutions obtained is estimated.

A recent review by Eason [6] on least squares methods for solving problems involving partial differential equations contains an extensive bibliography on this subject. Among the conclusions of the review is: one of the "major factors discouraging more wide spread use" of least squares methods seems to be "the presumption that least squares is clumsy to apply". We hope that our experience will help to dispell that notion.

The MACSYMA System

MACSYMA has been described as an automated mathematical co-worker. This characterization finds justification in the fact that MACSYMA can do many mathematical operations, both symbolic and numeric, very rapidly when the appropriate instructions are issued by its user.

Figure 1 lists some of the mathematical capabilities of MACSYMA along with features that enhance its utility. See the MACSYMA Reference Manual [1] for more details.

Figure 2 is a reproduction of a sequence of instructions and MACSYMA responses illustrating the symbolic integration, differentiation, and simplification of a rational expression. Also included is the Taylor expansion in X about 0 of $\sqrt{1+x}$ through the term containing X^5. Instructions typed by the user are underlined. Note that % references the preceding expression.

The General Problem

The problem area chosen for the application of MACSYMA is that of [5] which can be characterized in general as the solution of

$$TV-g=0 \qquad (1)$$

where V is an element of a subset F of a linear space R

Capabilities of MACSYMA

- Variable precision fixed and floating point arithmetic
- Two-dimensional display of expressions
- Rational function manipulation
- Trigonometric function manipulation
- Power series manipulation
- Matrix manipulation
- Integration subsystems
- Factorization of polynomials
- Pattern matching
- Noncommutative operations
- Tensor manipulation
- Laplace and inverse Laplace transforms
- Special functions
- Plotting
- Series summation
- Definition of functions
- String and list processing
- FORTRAN output
- Interactive capabilities
- Continuing system maintenance

Figure 1

```
(C9) D2;
                                          2 X
(D9)                                    ------
                                          3
                                         X  + 1

(C10) INTEGRATE(%,X);

SIN FASL DSK MACSYM being loaded
loading done

SCHATC FASL DSK MACSYM being loaded
loading done
                                                 2 X - 1
                    2                      ATAN(-------)
              LOG(X  - X + 1)                   SQRT(3)     LOG(X + 1)
(D10)      2 (--------------- + ------------- - ----------)
                     6                 SQRT(3)           3

(C11) DIFF(%,X);
                         2                   2 X - 1           1
(D11)     2 (----------------- + -------------- - ---------)
                          2               2               3 (X + 1)
                (2 X - 1)            6 (X  - X + 1)
            3 (---------- + 1)
                    3

(C12) RATSIMP(%);
                                          2 X
(D12)                                   ------
                                          3
                                         X  + 1

(C13) TAYLOR(SQRT(1+X),X,0,5);

HAYAT FASL DSK MACSYM being loaded
loading done
                               2     3      4      5
                       X    X     X    5 X    7 X
(D13)/T/          1 + - - -- + -- - ---- + ---- + . . .
                       2    8    16    128    256
```

Figure 2

T is an operator defined in a domain F of R and maps F uniquely into R.
g is a specified element of R.

The solution is approximated by a function W of a set of parameters $a_0, a_1, \ldots a_p$ so that some norm of the difference of TV and TW designated by

$$|TV-TW| \tag{2}$$

is minimized.

More specifically, let

$$L_0(V(x), x) = g(x) \qquad \alpha \leq x \leq \beta \tag{3}$$

represent linear differential equation with x on the real interval $[\alpha, \beta]$ with linear homogeneous boundary conditions.

$$L_1(V(x), x) = 0 \text{ at } X = \alpha \tag{4}$$

$$L_2(V(x), x) = 0 \text{ at } X = \beta \tag{5}$$

Assume a solution of the form

$$V = \sum_{i=1}^{p} a_i f_i(x) \tag{6}$$

where the f_i are functions which may or may not satisfy the boundary conditions, i.e.

$$L_1(f_i(\alpha),\alpha) = L_2(f_i(\beta),\beta) = 0 \qquad i = 1,2,\ldots.p \tag{7}$$

Define

$$\begin{aligned} \varepsilon_k(a,x) &= L_k\left(\sum_{i=1}^{p} a_i f_i(x)\right) - g(x) \\ &= \sum_{i=1}^{p} a_i L_k(f_i(x)) - g(x) \qquad \text{for } k = 0,1,2 \end{aligned} \tag{8}$$

The problem is then to find an approximate solution in the form (6) to (3) with boundary conditions (4), (5) where the a's are chosen so that an appropriate linear combination of the norm of $\epsilon_k(a,x)$ in the appropriate domains will be minimized.

Solution techniques

Two approaches to the solution, the Weighted Least Squares method and the Galerkin method were used and the results compared.

The Weighted Least Squares method can be expressed as

$$\min_{a} \int_{\alpha}^{\beta} \omega(x)\epsilon^2(a,x)dx = \min_{a} \int_{\alpha}^{\beta} \left(\sum_{i=1}^{p} a_i L(f_i(x)) - g(x)\right)^2 \omega(x)\,dx \quad (9)$$

or equivalently, solve for a the systems of p linear equations in the a's given by

$$\frac{\partial}{\partial a_k}\left(\int_{\alpha}^{\beta} \omega(x)\epsilon^2(a,x)dx\right)=0 \qquad (k=1,2,.....,p) \quad (10)$$

which reduces to the system

$$\sum_{i=1}^{p} a_i \int_{\alpha}^{\beta} \omega(x)L(f_i(x))L(f_k(x))dx = \int_{\alpha}^{\beta} \omega(x)g(x)L(f_k(x))dx \quad (11)$$

$$(k = 1,2,.....,p)$$

In a similar way the Galerkin method can be expressed as the system of equations given by

$$\int_{\alpha}^{\beta} \epsilon(a,x)f_k(x)dx = 0 \quad (k=1,2,...,p) \quad (12)$$

where the $f_k(x)$ satisfy the boundary conditions,

or equivalently

$$\sum_{i=1}^{p} a_i \int_{\alpha}^{\beta} L_0(f_i(x))f_k(x)dx = \int_{\alpha}^{\beta} g(x)f_k(x)dx \quad (13)$$

$$(k=1,2,.....,p)$$

Note the similarity of form in the expressions for Weighted Least Squares and Galerkin methods when $\omega(x)=1$.

The Specific Example

The specific example chosen as a demonstration problem for the class is the following:

$$\frac{d^2V}{dx^2} + 2x\frac{dV}{dx} + (1-x^2)V = 1-x^2 \quad \text{for } 0 \leq x \leq 1 \tag{14}$$

with boundary conditions

$$\frac{dV}{dx}(0) = V(1) = 0 \tag{15}$$

The exact solution is known to be

$$V = 1 - e^{\frac{1}{2}(x^2-1)} \tag{16}$$

Several approaches to finding approximate solutions to equation (14) subject to boundary conditions (15) were taken. An appraisal of accuracy of each approach to the solution of this problem was made by comparing numerical results with (16) evaluated at selected points on the interval [0,1].

Several variations of the Least Squares solution and the Galerkin solution were programmed and carried out in MACSYMA Using an approximating function of the form

$$V = \sum_{i=1}^{n} a_i(1-x^{i+1}) \tag{17}$$

Solutions were obtained for n=4 using the continuous least squares and the continuous Galerkin methods i.e. solving equations (11) and (13) respectively for $\alpha=0$, $\beta=1$. Solutions using approximating function (17) with n=4 were also obtained by the discrete least squares method where the integral form 0 to 1 was replaced by a weighted sum over the set of points 0(.05).95, first by a sum with uniform weight then by a sum with weights of 1/2 at the boundaries so that a more accurate approximation to the integral is used.

In the discrete least squares solution the system of linear equations in the function coefficients, a_i, was carried out by the Gram-Schmidt orthonormalization process applying a technique used by Cuthill [5] and described by Davis and Rabinowitz [7]. The discrete least squares solution was also carried out using the matrix inversion embodied in MACSYMA for comparison.

It should be noted that the approximating functions (17) automatically satisfy the boundary conditions (15) for the problem.

A second approximating function of the form

$$V = \sum_{i=1}^{n} a_i x^{i-1} \tag{18}$$

was used and coefficients were computed using the least squares method. Since the boundary conditions (15) are not automatically satisfied by (18) it was necessary to give special attention to computing the coefficients in such a way that the boundary conditions would be satisfied. The problem was set up so that the appropriate boundary equations were being fit at X=0 and X=1 and that equation (14) was being fit on the set of points .05(.05).95.

Two procedures were then used to enforce the boundary conditions. One of these was to assign a weight of 10 to each boundary point and a weight of 1 to all other points and use a weighted least squares solution.

The second procedure used was that of considering the boundary equations with the approximations substituted at the boundary points as a system of equations to be satisfied exactly and the differential equation with the approximating functions substituted at points internal to the interval, to be an overdetermined system of linear equations to be solved by the uniformly weighted least squares method. This is equivalent to a constrained least squares problem. These assumptions were effected by expressing the composite of these two systems as a matrix equation and partitioning the matrices appropriately. This led to two matrix equations that were solved simultaneously. The format used required that the boundary equations be given first.

Solution Procedure with MACSYMA

A general input procedure for setting up initial and boundary value problems in MACSYMA was developed and followed for the specific example chosen. Figure 3 outlines this procedure. Symbols used in MACSYMA to identify the various entities are given where applicable.

A MACSYMA demonstration of one of the programs following this procedure is given in the appendix. Exact rational arithmetic was used in these calculations to demonstrate the power of MACSYMA. The calculations could as easily have been done in floating point arithmetic.

INPUT PROCEDURE for LINEAR INITIAL and BOUNDARY VALUE PROBLEMS

For all problems specify

The number of equations	NEQ
The equations	EQ[I]
The number of approximating functions	NF
The functions	FN[I]

For Discrete Problems specify

The number of points	NP
The points	X[I]
For each point, the equation to be used	K[I]

For Continuous Problems specify

The Region of Integration
The Domain of applicability for each equation

For all constrained problems specify

The number of constraints	NEX[I]
Identify the constraining equations	

Figure 3

Comparison of Results: The solution of the differential equation and comparisons of the various approximations to the solution are presented in graphical form using the plotting capability of MACSYMA. The plots were obtained on a Tektronix 4014-1 terminal with a 4631 Hard Copy Unit. Figure 4 is a graph of the exact solution function and Figures 5, 6, and 7 are error curves for the approximating functions obtained.

Table 1 is a MACSYMA printout of ten polynomials obtained as approximations to the solution of the differential equation. Table 2 identifies each of the polynomials with its source and gives the symbol used to mark the corresponding error curve in Figure 5, 6, or 7.

$$\underline{\text{FASOL(X)}} = \frac{5739188167112147\ (1 - X^{5})}{10273089659885638 2} + \frac{13838424092109221\ (1 - X^{4})}{24655415183725531 68} + \frac{10542903198955897\ (1 - X^{3})}{30819268979656914 6} + \frac{79531480872800984 5\ (1 - X^{2})}{26710033115702659 32}$$

$$\underline{\text{FGASOL(X)}} = 0.052227298\ (1 - X^{5}) + 0.0123990587\ (1 - X^{4}) + 0.0308086944\ (1 - X^{3}) + 0.2980095\ (1 - X^{2})$$

$$\underline{\text{FAPV(X)}} = 0.051606234\ (1 - X^{5}) + 0.0157067273\ (1 - X^{4}) + 0.0262761153\ (1 - X^{3}) + 0.30004945\ (1 - X^{2})$$

$$\underline{\text{FLSAPV(X)}} = \frac{4572728389014839909919683375932834394810400000 00\ (1 - X^{5})}{8860824139815566982202625578080094691643409693381} + \frac{1391769033042574477704801403669306933880425400 00\ (1 - X^{4})}{8860824139815566982202625578080094691643409693381} + \frac{2328259823867569012612044877645843790090334200 00\ (1 - X^{3})}{8860824139815566982202625578080094691643409693381} + \frac{2658686008733236245532269471955192887452194692250\ (1 - X^{2})}{8860824139815566982202625578080094691643409693381}$$

Table 1 - Approximating Functions

$$\underline{FHAPV(X)} = 0.052707125\ (1 - X^5) + 0.0125995092\ (1 - X^4)$$
$$+ 0.0292615863\ (1 - X^3) + 0.29891991\ (1 - X^2)$$

$$\underline{FULS1(X)} = - 0.055080894\ X^5 - 5.7278722E\text{-}3\ X^4 - 0.035697244\ X^3$$
$$- 0.29656702\ X^2 + 4.7926395E\text{-}4\ X + 0.392789636$$

$$\underline{FWLS(X)} = - 0.055173752\ X^5 - 5.6203716E\text{-}3\ X^4 - 0.03595056\ X^3$$
$$- 0.29639734\ X^2 + 4.7991139E\text{-}5\ X + 0.39311364$$

$$\underline{FCLS(X)} = - 0.055184639\ X^5 - 5.60691386E\text{-}3\ X^4 - 0.035980111\ X^3$$
$$- 0.296378005\ X^2 + 0.393149674$$

$$\underline{FSOLFT(X)} = 0.050618228\ (1 - X^5) + 0.015670215\ (1 - X^4)$$
$$+ 0.028779681\ (1 - X^3) + 0.29837708\ (1 - X^2)$$

$$\underline{FPOLFT(X)} = - 0.053396624\ X^5 - 7.159276E\text{-}3\ X^4 - 0.038121192\ X^3$$
$$- 0.29397657\ X^2 - 8.0931693E\text{-}4\ X + 0.393478915$$

Table 1 (continued)

Symbol for Error Function	Approximating Function	Source of Approximating Function
1	FASOL(x)	Least Squares, continuous, with $(1-x^{i+1})$
2	FGASOL(x)	Galerkin, continuous, with $(1-x^{i+1})$
3	FAPV(x)	Gram-Schmidt Least Squares with $(1-x^{i+1})$ with integral replaced by sum over points 0(.05).95
4	FLSAPV(x)	Ordinary Least Squares with $(1-x^{i+1})$ using rational arithmetic. Integral replaced by sum over points 0(.05).95
5	FHAPV(x)	Weighted Gram-Schmidt Least Squares with $(1-x^{i+1})$ using weight of 1/2 at x=0 and weight of 1 at all other points. Integral replaced by sum over 0(.05)1.0
6	FULS1(x)	Gram-Schmidt Least Squares with (x^{i-1}) Integral replaced by sum over 0(.05)1.0
7	FWLS(x)	Weighted Gram-Schmidt Least Squares with (x^{i-1}). Boundary conditions weighted 10, other points weighted 1. Integral replaced by sum over points 0(.05)1.0
8	FCLS(x)	Gram-Schmidt constrained Least Squares with (x^{i-1}). Boundary conditions forced. Integral replaced by sum over .05(.05).95
9	FSOLFT(x)	Polynomial fit of the exact solution, FN(x), using $(1-x^{i+1})$
10	FPOLFT(x)	Polynomial fit of the exact solution, FN(x), using (x^{i-1})

Table 2 - Identification of Functions

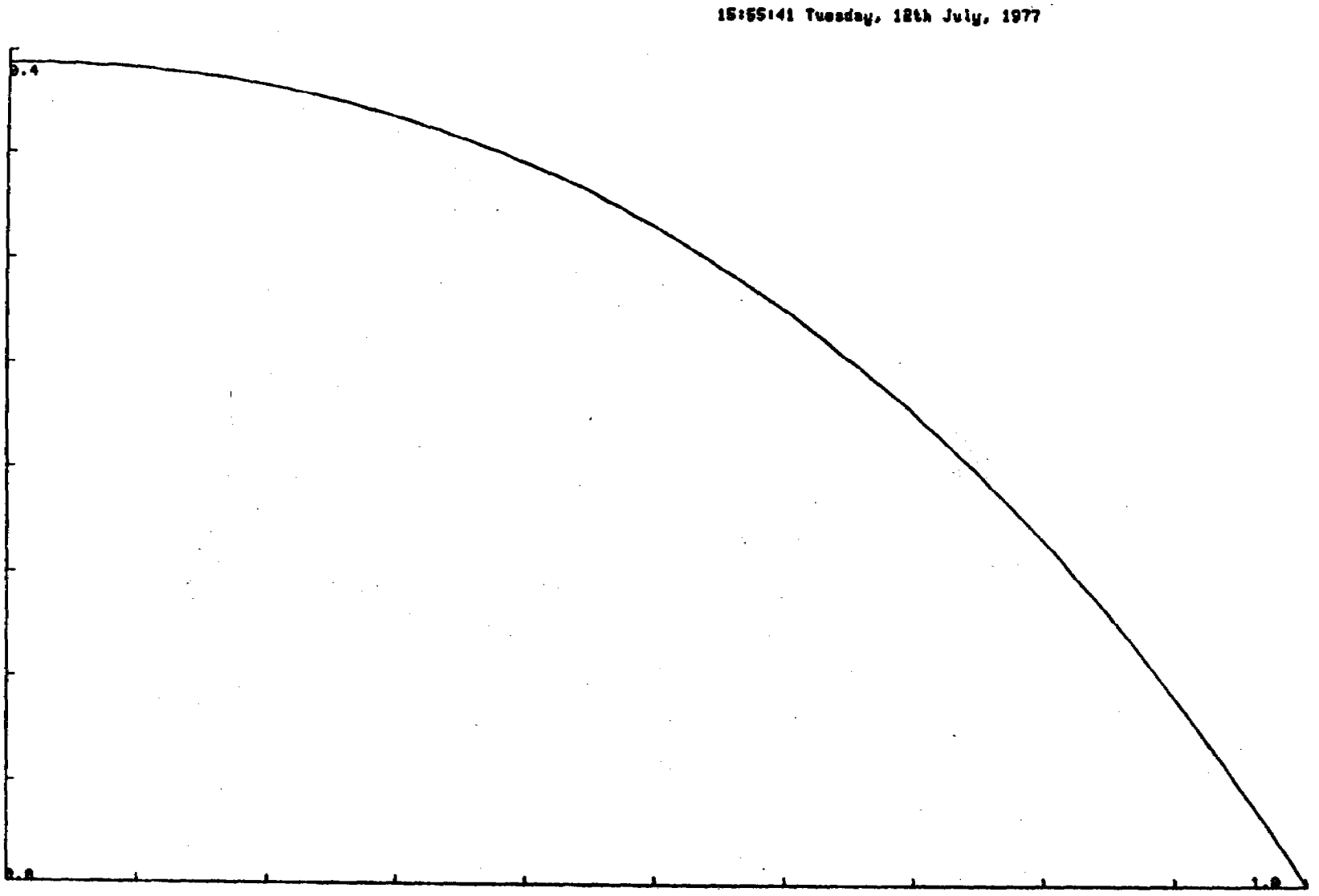


Figure 4 - $FN(x) = 1 - e^{1/2(x^2 - 1)}$

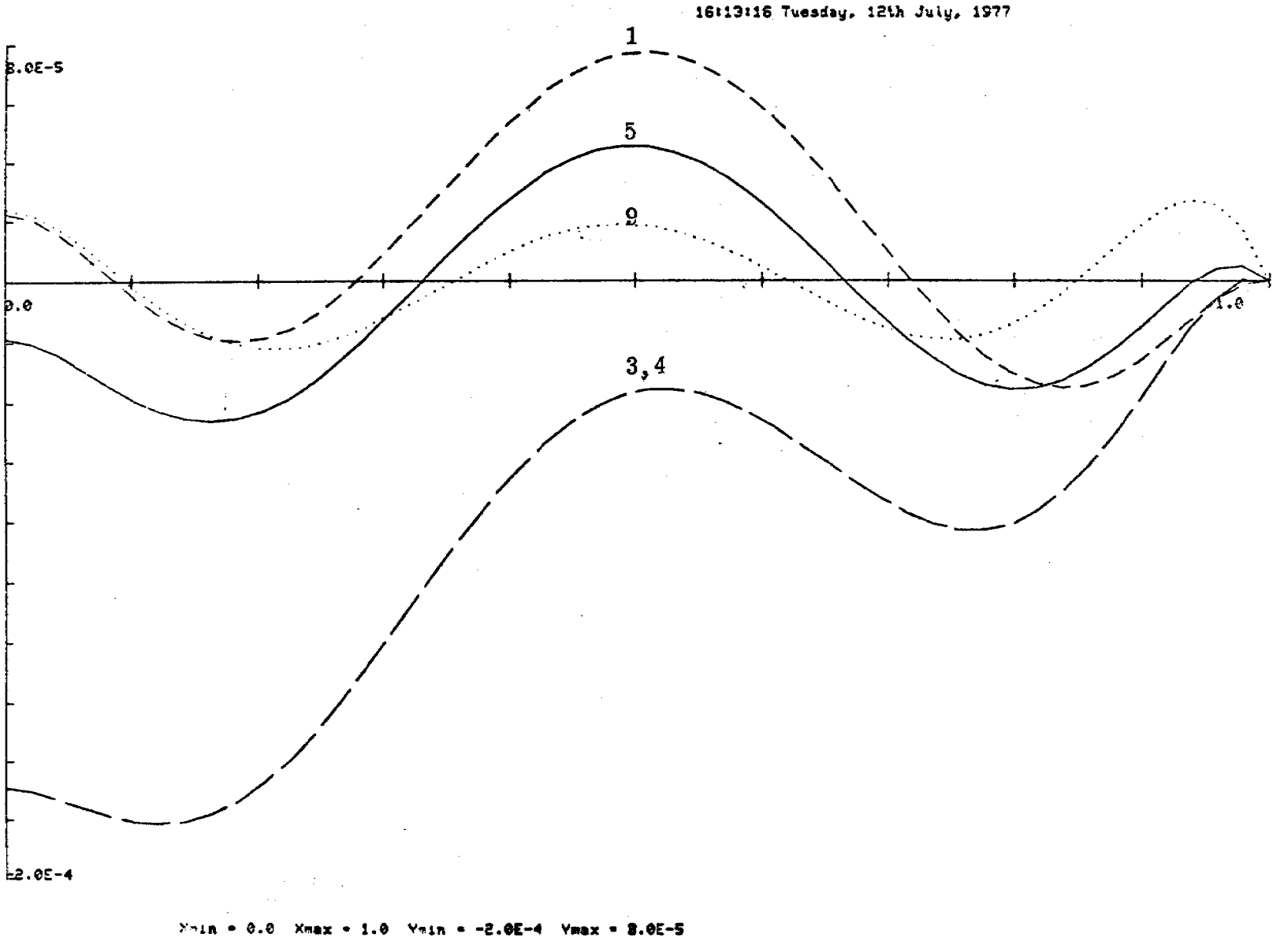

16:13:16 Tuesday, 12th July, 1977
8.0E-5
1
5
9
0.0
1.0
3,4
-2.0E-4
Xmin = 0.0 Xmax = 1.0 Ymin = -2.0E-4 Ymax = 8.0E-5

Figure 5

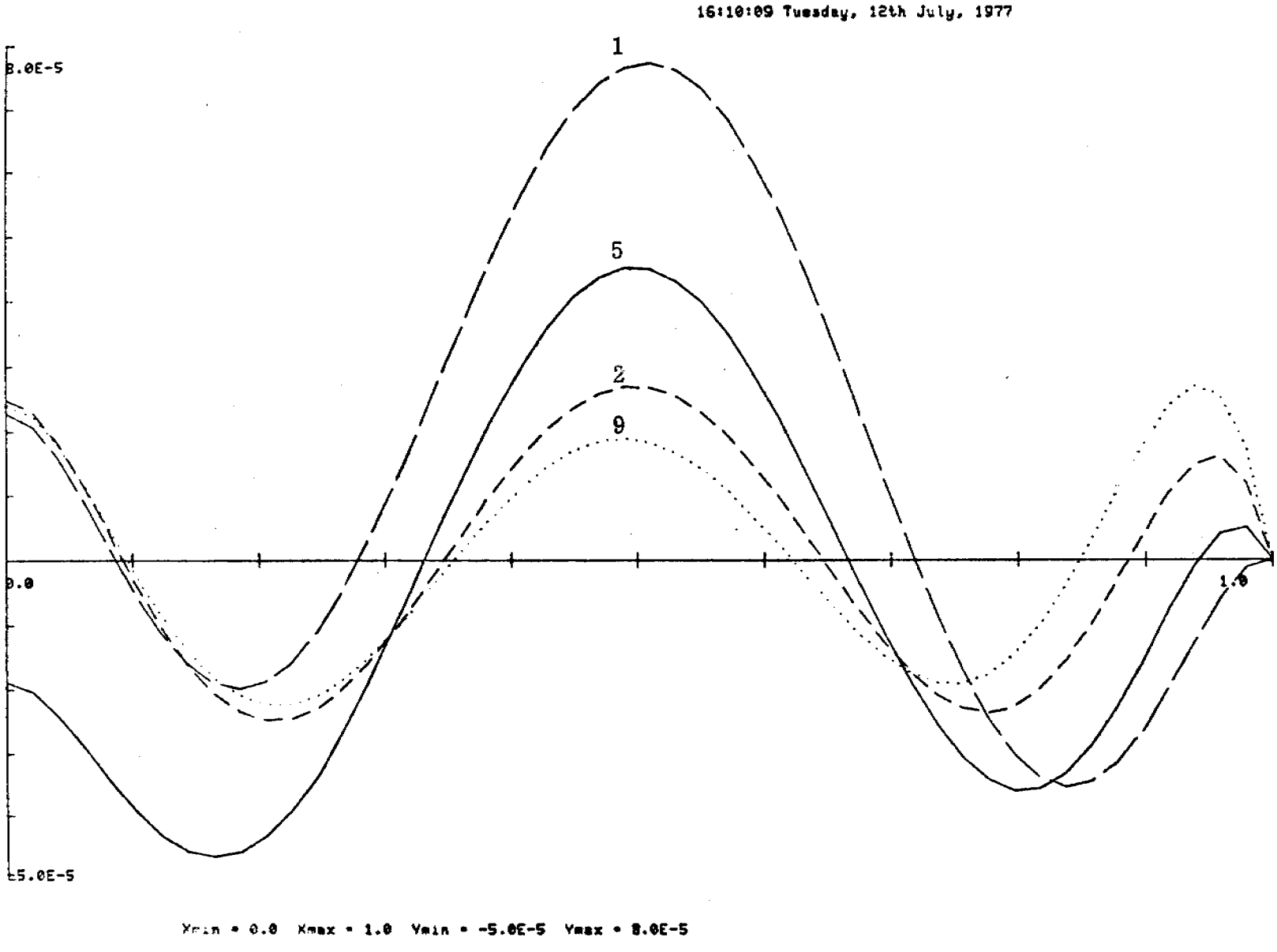

16:10:09 Tuesday, 12th July, 1977
8.0E-5
1
5
2
9
0.0
1.0
-5.0E-5
Xmin = 0.0 Xmax = 1.0 Ymin = -5.0E-5 Ymax = 8.0E-5

Figure 6

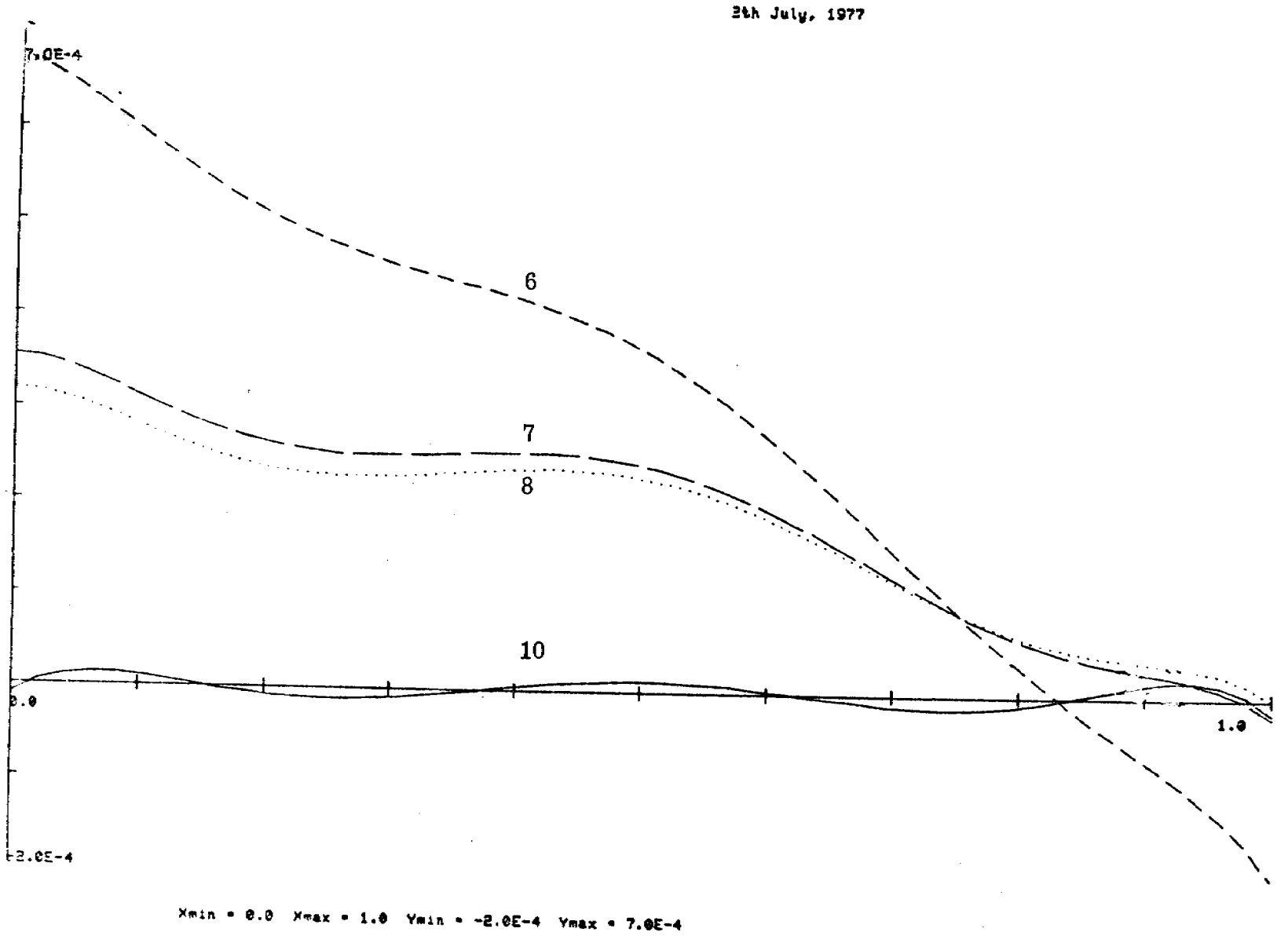

24th July, 1977
7.0E-4
6
7
8
10
0.0
1.0
-2.0E-4
Xmin = 0.0 Xmax = 1.0 Ymin = -2.0E-4 Ymax = 7.0E-4

Figure 7

REFERENCES

[1] MACSYMA Reference Manual, The Math Lab Group, Project MAC, MIT, Version 8, November, 1975.

[2] Papers by J. Moses, W.A. Martin, and R.J. Fatemanin, the ACM Proceedings of the Second Symposium on Symbolic and Algebraic Manipulation, Los Angeles, Calif., March 1971.

[3] L.K. Meals, "MACSYMA - A Resource for the Navy Laboratory Computer Network", David W. Taylor Naval Ship Research and Development Center, Computation, Mathematics, and Logistics Department, Departmental Report CMLD-77-04 (January 1977)

[4] E. Bond, et al, FORMAC, an Experimental Formula Manipulation Compiler, in Assoc. for Computing Machinery Proceedings of the 19th National Conference, Aug 1964, ACM, New York.

[5] E. Cuthill, "A FORMAC Program for the Solution of Linear Boundary and Initial Value Problems", Presented at ACM Symposium on Symbolic and Algebraic Manipulation, Washington, D.C., 30 Mar 1966. Abstract published in Communications of the ACM, 9, p. 550, 1966.

[6] E.D. Eason, A review of least-squares methods for solving partial differential equations, International Journal for Numerical Methods in Engineering, Vol. 10, 1021-1046 (1976).

[7] P. Davis and P. Rabinowitz, Advances in Orthonormalizing Computation, in Advances in Computers, 2, (F. Alt and M. Rubinoff, eds), Academic Press, New York, 1961.

APPENDIX: MACSYMA demonstration of the solution of

$$V(1 - x^2) + 2x\frac{dV}{dx} - \frac{d^2V}{dx^2} = 1 - x^2$$

by the continuous Least Squares method using approximating functions of the form $(1 - x^n)$ for $2 \leq n \leq 7$

```
(C4) DEPENDENCIES(V(X));
TIME= 1 MSEC.
(D4)                                  [V(X)]

(C5) EQ:-DIFF(V,X,2)+2*X*DIFF(V,X)+(1-X*X)*V=1-X*X;
TIME= 19 MSEC.
```

$$(D5) \qquad V\,(1 - X^2) + 2\,\frac{dV}{dX}\,X - \frac{d^2V}{dX^2} = 1 - X^2$$

```
(C6) FN[I,J]:=1-X**(I+1);
TIME= 2 MSEC.
```

$$(D6) \qquad FN_{I,\,J} := 1 - X^{I+1}$$

```
(C7) MFN:GENMATRIX(FN,NF,1);

GENMAT FASL DSK MAXOUT being loaded
loading done
TIME= 40 MSEC.
```

$$(D7) \qquad \begin{bmatrix} 1 - X^2 \\ 1 - X^3 \\ 1 - X^4 \\ 1 - X^5 \\ 1 - X^6 \\ 1 - X^7 \end{bmatrix}$$

```
(C8) FOR I THRU NF DO( EF[I]:EV(LHS(EQ),V:FN[I,1],DIFF),DISPLAY(EF[I])),
```

$$EF_1 = (1 - X^2)^2 - 4\,X^2 + 2$$

$$EF_2 = -6\,X^3 + (1 - X^2)\,(1 - X^3) + 6\,X$$

$$EF_3 = -8\,X^4 + (1 - X^2)\,(1 - X^4) + 12\,X^2$$

$$EF_4 = -10\,X^5 + (1 - X^2)\,(1 - X^5) + 20\,X^3$$

$$EF_5 = -12\,X^6 + (1 - X^2)\,(1 - X^6) + 30\,X^4$$

$$EF_6 = -14\,X^7 + (1 - X^2)\,(1 - X^7) + 42\,X^5$$

```
TIME= 166 MSEC.
(D8)                                  DONE
```

```
(C9) FOR I THRU NF DO FOR J THRU I DO INT[I,J]:INTEGRATE(EF[I]*EF[J],X,0,
1);

DEFINT FASL DSK MACSYM being loaded
loading done

LIMIT FASL DSK MACSYM being loaded
loading done

RESIDU FASL DSK MACSYM being loaded
loading done

SIN FASL DSK MACSYM being loaded
loading done

SCHATC FASL DSK MACSYM being loaded
loading done
TIME= 8981 MSEC.
(D9)                                    DONE

(C10) FOR I THRU NF DO FOR J THRU I-1 DO INT[J,I]:INT[I,J];
TIME= 89 MSEC.
(D10)                                   DONE

(C11) MINT:GENMATRIX(INT,NF,NF);
TIME= 21 MSEC.
```

$$
(D11)\quad
\begin{bmatrix}
\frac{1196}{315} & \frac{167}{56} & \frac{2008}{1155} & \frac{223}{840} & \frac{20204}{15015} & \frac{853}{280} \\
\frac{167}{56} & \frac{67909}{13860} & \frac{3917}{630} & \frac{2587901}{360360} & \frac{219493}{27720} & \frac{72931}{8580} \\
\frac{2008}{1155} & \frac{3917}{630} & \frac{69616}{6435} & \frac{277}{18} & \frac{896248}{45045} & \frac{122677}{5040} \\
\frac{223}{840} & \frac{2587901}{360360} & \frac{277}{18} & \frac{1093003}{45045} & \frac{1858349}{55440} & \frac{12545701}{291720} \\
\frac{20204}{15015} & \frac{219493}{27720} & \frac{896248}{45045} & \frac{1858349}{55440} & \frac{36916532}{765765} & \frac{1763651}{27720} \\
\frac{853}{280} & \frac{72931}{8580} & \frac{122677}{5040} & \frac{12545701}{291720} & \frac{1763651}{27720} & \frac{47470367}{554268}
\end{bmatrix}
$$

```
(C12) ING[I,J]:=INTEGRATE(EF[I]*RHS(EQ),X,0,1);
TIME= 1 MSEC.
(D12)                ING      := INTEGRATE(EF  RHS(EQ), X, 0, 1)
                       I, J                  I
```

```
(C13) MING:GENMATRIX(ING,NF,1);
TIME= 1170 MSEC.
```

(D13)

$$\begin{bmatrix} \frac{44}{35} \\ \\ \frac{179}{120} \\ \\ \frac{104}{63} \\ \\ \frac{53}{30} \\ \\ \frac{6428}{3465} \\ \\ \frac{77}{40} \end{bmatrix}$$

```
(C14) AA:(MINT^^-1).MING;
TIME= 2798 MSEC.
```

(D14)

$$\begin{bmatrix} \frac{1413800313119993398289004985979855}{4663686919507035939586142897627209} \\ \\ \frac{216398780655926336708535282158 08}{13991060758521107818758428692881627} \\ \\ \frac{957375469919753387681595437768384}{13991060758521107818758428692881627} \\ \\ \frac{77614519628445509702807812245248}{4663686919507035939586142897627209} \\ \\ -\frac{8706480312721328582243575598 0758}{13991060758521107818758428692881627} \\ \\ \frac{46286096750200933496768518476544}{4663686919507035939586142897627209} \end{bmatrix}$$

```
(C15) AAN:EV(AA,NUMER);
TIME= 10 MSEC.
```

(D15)

$$\begin{bmatrix} 0.30315078 \\ \\ 1.54669316E-3 \\ \\ 0.068427654 \\ \\ 0.016642309 \\ \\ -6.222888E-3 \\ \\ 9.924786 44E-3 \end{bmatrix}$$

```
(C16) ASOL:AA.MFN;
TIME= 18 MSEC.
                                                 7
       46286096750200933496768518476544 (1 - X )
(D16) --------------------------------------------
         4663686919507035939586142897627209

                                              6
  87064803127213285822435755980758 (1 - X )
- --------------------------------------------
      13991060758521107818758428692881627

                                              5
  77614519628445509702807812245248 (1 - X )
+ --------------------------------------------
      4663686919507035939586142897627209

                                               4
  957375469919753387681595437768384 (1 - X )
+ --------------------------------------------
      13991060758521107818758428692881627

                                              3
  21639878065592633670853528215808 (1 - X )
+ --------------------------------------------
      13991060758521107818758428692881627

                                                 2
  1413800313119993398289004985979855 (1 - X )
+ ----------------------------------------------
        4663686919507035939586142897627209


(C26) ASOLR:EV(ASOL,RATSIMP);
TIME= 576 MSEC.
                                                 7
(D26) - (138858290250602800490305555429632 X

                                                6                                   5
 - 87064803127213285822435755980758 X  + 232843558885336529108423436735744 X

                                                4                                   3
 + 957375469919753387581595437768384 X  + 21639878065592633670853528215808 X

                                                   2
 + 4241400939359980194867014957939565 X  - 5505053333354052259995757160108375)

/13991060758521107818758428692881627
```

```
(C27) FOR I:0 THRU 1 STEP 1/10 DO
   (APP:EV(ASOL,X:I),AP:EV(APP,NUMER),
    SO:EV(1-EXP((X^X-1)/2),X:I,NUMER),
     ER:SO-AP,
     PRINT("X=",I,"APPROX=",AP,"ASOLN=",SO,"ERROR=",ER));
X= 0 APPROX= 0.39346933 ASOLN= 0.39346934 ERROR= 1.1175871E-8
   1
X= -- APPROX= 0.39042927 ASOLN= 0.3904291 ERROR= - 1.71363353E-7
   10
   1
X= - APPROX= 0.38121639 ASOLN= 0.3812166 ERROR= 2.12341547E-7
   5
   3
X= -- APPROX= 0.36555166 ASOLN= 0.36555204 ERROR= 3.7625432E-7
   10
   2
X= - APPROX= 0.34295328 ASOLN= 0.34295318 ERROR= - 1.00582838E-7
   5
   1
X= - APPROX= 0.312711194 ASOLN= 0.312710725 ERROR= - 4.6938658E-7
   2
   3
X= - APPROX= 0.27385114 ASOLN= 0.273850955 ERROR= - 1.86264515E-7
   5
   7
X= -- APPROX= 0.22508315 ASOLN= 0.2250835 ERROR= 3.4831464E-7
   10
   4
X= - APPROX= 0.164729524 ASOLN= 0.16472979 ERROR= 2.6449561E-7
   5
   9
X= -- APPROX= 0.090627265 ASOLN= 0.09062706 ERROR= - 2.0582229E-7
   10
X= 1 APPROX= 0 ASOLN= 0 ERROR= 0
TIME= 1908 MSEC.
(D27)                                DONE


(C28) ERR:INTEGRATE(EV((LHS(EQ)-RHS(EQ))**2,V:ASOLR,DIFF,RATSIMP),X,0,1);
TIME= 1458 MSEC.
                          3467350112145544592486756548
(D28)                     -----------------------------------
                          692557507546794837028542202976405365

(C29) ERR**.5;
TIME= 2 MSEC.
(D29)                              7.0757247E-5

TIME= 4284 MSEC.
(D30)                               BATCH DONE

(C31)
```

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

MOVING-WEIGHTED-AVERAGE SMOOTHING EXTENDED TO THE EXTREMITIES OF THE DATA

T. N. E. Greville

Technical Summary Report #

ABSTRACT

A symmetrical moving weighted average (MWA) for smoothing observational data which may be regarded as equally spaced measurements of a function of one variable has the form

$$u_x = \sum_{j=-m}^{m} c_j \, y_{x-j} \quad , \tag{1}$$

where y_x is an observed value, u_x is the corresponding smoothed value, and the c_j are real coefficients whose sum is unity, with $c_{-j} = c_j$. This process does not yield smoothed values of the first m and the last m observations unless additional data are available. A natural method is suggested for extending the smoothing to the extremities of the data.

If (1) is exact for polynomials up to the degree $2s - 1$, it can be written in the form

$$u_x = [1 - (-1)^s \delta^{2s} q(E)] y_x \quad ,$$

where δ is the finite difference taken centrally, E is defined by $Ef(x) = f(x + 1)$, and

$$q(E) = \sum_{j=-m+s}^{m-s} q_j \, E^j$$

for some coefficients q_j. If $q(z)$ has no zero on the unit circle, there is a Laurent expansion

$$[q(z)]^{-1} = \sum_{j=-\infty}^{\infty} h_j \, z^j$$

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

convergent in an annulus containing the unit circle.

We regard the overall smoothing process as a matrix-vector operation

$$u = Gy \ , \tag{2}$$

where u and y are vectors of N components and G is symmetric with rows, except for the first m and the last m, that merely reflect the application of (1). We determine the first m and the last m rows by taking

$$G = I - K^T DK \ ,$$

where K is the matrix of $N - s$ rows and N columns that transforms a vector into the vector of sth finite differences of its components, and D is the symmetric matrix of order $N - s$ whose inverse is the Toeplitz matrix $T = (t_{ij}) = (t_{i-j})$, with $t_{i-j} = h_{i-j}$.

The same vector u can be obtained by a computational short-cut. Let $p(z)$ be the monic polynomial of degree $m - s$ whose zeros are those zeros of $q(z)$ lying within the unit circle, and let

$$a(z) = (z - 1)^s \, p(z) = z^m - \sum_{j=1}^{m} a_j \, z^{m-j} \ .$$

Then, if the range of x is from A to B, recursively calculate fictitious extended values y_x for $x = A - 1, \ A - 2, \ \ldots, \ A - m$ by

$$y_x = \sum_{j=1}^{m} a_j \, y_{x+j} \ .$$

Similarly, calculate extended values for $x = B + 1, \ B + 2, \ \ldots, \ B + m$ recursively by

$$y_x = \sum_{j=1}^{m} a_j \, y_{x-j} \ .$$

Finally, apply (1) to the entire sequence of observed and extended values to obtain smoothed values for $x = A, \ A + 1, \ \ldots, \ B$.

Schoenberg (1946) defined the characteristic function of (1) as

$$\phi(t) = \sum_{j=-m}^{m} c_j \, e^{ijt} \ ,$$

and calls an MWA a smoothing formula if

$$-1 \leq \phi(t) \leq 1 \quad ,$$

with some ambiguity as to whether the inequalities should be strict for $0 < t < 2\pi$. It is shown here that the limit $\lim_{n \to \infty} G^n$ exists for all $N > 2m$ if and only if $-1 \leq \phi(t) < 1$ for $0 < t < 2\pi$.

If $0 \leq \phi(t) < 1$ for $0 < t < 2\pi$, (2) is equivalent to the minimization of

$$(u - y)^T (u - y) + (Ku)^T HKu \quad ,$$

where $H = (D^{-1} - KK^T)^{-1}$ is positive definite. This generalizes the Whittaker (1923) smoothing process.

EXPLANATION

The use of a moving weighted average of $2m + 1$ terms to smooth equally spaced observations of a function of one variable does not yield smoothed values of the first m and the last m observations, unless additional data beyond the range of the original observations are available. Using Toeplitz matrices, Laurent series, and analogies to the Whittaker smoothing process, we develop a natural method of extending the smoothing to the extremities of the data.

AMS(MOS) Subject Classification -

Key Words - Smoothing, Toeplitz matrix, Laurent series, Moving weighted average

Work Unit Number 2 - Mathematical Methods

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

1. INTRODUCTION

A time-honored method of smoothing equally spaced observations of a function of one variable to remove or reduce unwanted irregularities is the moving weighted average (MWA). An example is Spencer's 15-term average (Macaulay 1931; Henderson 1938), which can be expressed in the form

$$u_x = \frac{1}{320}(-3y_{x-7} - 6y_{x-6} - 5y_{x-5} + 3y_{x-4} + 21y_{x-3} + 46y_{x-2} + 67_{x-1} + 74y_x$$

$$+ 67y_{x+1} + 46y_{x+2} + 21y_{x+3} + 3y_{x+4} - 5y_{x+5} - 6y_{x+6} - 3y_{x+7}) , \qquad (1.1)$$

where y_x is the observed value corresponding to the argument x, and u_x is the corresponding adjusted value. Actuarial writers commonly refer to such smoothing of data as "graduation."

More generally (Schoenberg 1946) a symmetrical MWA is of the form

$$u_x = \sum_{j=-m}^{m} c_j y_{x-j} , \qquad (1.2)$$

where m is a given positive integer and the real coefficients c_j are such that $c_{-j} = c_j$ and

$$\sum_{j=-m}^{m} c_j = 1 .$$

Such averages have a long history, that includes some eminent names, but the literature concerning them is little known in the general mathematical community. Among the early writers on the subject was the Italian astronomer G. V. Schiaparelli (1866), who is chiefly remembered for his observations of the planet Mars. The majority of publications in this area have appeared in English and Scottish actuarial journals starting with John Finlaison in 1829 (see Maclean 1913). Probably the first writer to make a systematic investigation of such averages was the American mathematician E. L. De Forest (1873, 1875, 1876, 1877). His work, published in obscure places, was rescued from total oblivion largely through the efforts of Hugh H. Wolfenden (1892-1968), who also made important contributions to the subject (Wolfenden 1925). E. T. Whittaker (1923) suggested an alternative method of

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

smoothing, which has been widely employed, especially by actuaries, and will be referred to extensively later, because of numerous analogies to the MWA procedure. The first writer to apply sophisticated mathematical tools to the study of these averages was I. J. Schoenberg (1946, 1958, 1953), who introduced the notion of the characteristic function of an MWA, and utilized it to formulate a criterion for judging whether a given average can properly be called a "smoothing formula." This criterion will be discussed in Section 10.

2. THE PROBLEM OF SMOOTHING NEAR THE EXTREMITIES OF THE DATA

When MWA's have been used by actuaries, the argument x is usually age (of a person) in completed years. When they are used for smoothing economic time series, x denotes the position of a particular observation in a time sequence. The latter area of application appears to stem largely from the work of Frederick R. Macaulay (1931), who was the son of an actuary.

In either case, a serious disadvantage of the method is that it does not produce adjusted values for arguments too near the extremities of the data. For example, suppose Spencer's 15-term average is used to smooth monthly data extending from 1970 through 1976. The formula does not give smoothed values for the first 7 months of 1970 or the last 7 months of 1976 unless data can be obtained for the last 7 months of 1969 and the first 7 months of 1977. Clearly, acquisition of data extending farther into the past is less of a problem than acquisition of future data.

Actuaries in North America seem to have largely abandoned the use of MWA's in favor of Whittaker's method, which does not have the disadvantage described. It is likely that British actuaries may still use these averages to some extent. They appear to be currently employed by economic and demographic statisticians (Shiskin, Young, and Musgrave 1967).

Various suggestions have been made (De Forest 1877, Miller 1946, Greville 1957, 1974a) for dealing with the problem of adjustment of data near the extremities, but none of them have won general acceptance. De Forest's (1877, p. 110) suggestion is so relevant to the subject of the present paper that it is worth quoting in full:

"As the first m and the last m terms of the series cannot be reached directly by the formula [of $2m + 1$ terms], the series should be graphically extended by m terms at both ends, first plotting the observations on paper as ordinates, and then extending

the curve along what seems to be its probable course, and measuring the ordinates of the extended portions. It is not necessary that this extension should coincide with what would be the true course of the curve in those parts. The important part is that the m terms thus added, taken together with the $m + 1$ adjacent given terms, should follow a curve whose form is approximately algebraic and of a degree not higher than the third."

Elsewhere (Greville 1974a) I have proposed extrapolating the observed data by fitting a least-squares cubic to the first $m + 1$ values and a similar cubic to the last $m + 1$ observations. This is very much in the spirit of De Forest's suggestion; it is not a long step from graphic to algebraic extrapolation.

Another approach (Greville 1957) regards the adjustment process as a matrix-vector operation. We write

$$u = Gy ,$$

where y is the vector of observed values, u is the corresponding vector of adjusted values, and G is a square matrix. If a specified symmetrical MWA of $2m + 1$ terms is to be used wherever possible, then the nonzero elements of G, except for the first m and the last m rows, are merely the weights in the moving average, these weights moving to the right as one proceeds down the rows of the matrix. In the first m and the last m rows special unsymmetrical weights, determined in some appropriate manner, must be inserted. The matrix approach and the extrapolation are not wholly unrelated, since the final results of the extrapolation approach can be expressed in matrix form.

It is the purpose of the present paper to show that when a <u>given</u> MWA is being employed, there is a natural, preferred method of extending the adjustment to the extremities of the data, strongly suggested by the mathematical properties of the weighted average. This natural method of extension seems to have eluded previous writers on the subject, as indeed it eluded me during the many years I have thought about the matter. The preferred method of extension has the interesting property that it can be arrived at either through the matrix approach or the extrapolation approach. In the latter case, one must employ a special extrapolation formula uniquely determined by the given MWA. Though the two approaches appear to be quite different, they will be shown in Section 9 to be mathematically equivalent, and they will give identical results except for rounding error.

In my own thinking I arrived at the procedure first through the matrix approach, guided largely by extensive analogies to the Whittaker process (which is most conventiently expressed in matrix terms). It was only later that I became aware that identical results could be obtained by means of an extrapolation algorithm. Though the matrix approach provides far greater insight into the rationale behind the procedure, the extrapolation approach is simpler computationally. Therefore, we shall first describe and illustrate the extrapolation algorithm, and shall then motivate and justify the procedure by means of the matrix approach.

The extrapolation approach is merely a computational short cut, and nearly always the extended values obtained by its use are highly unrealistic if regarded as extrapolated values of the function under observation. This fact is irrelevant, but has seriously "turned off" some users. Hereafter I shall therefore avoid the use of the words "extrapolate" and "extrapolation," and shall speak of "extension," "extended values," and "intermediate values."

It is emphasized that the procedure to be described (or any other procedure for completing the graduation) is recommended for use only when additional data extending beyond the range of the original data are not available.

3. THE EXTENSION ALGORITHM

A weighted average of the form (1.2) will be called exact for the degree r if it has the property that, in case all the observed values y_{x-j} in (1.2) should happen to be the corresponding ordinates of some polynomial $P(x - j)$ of degree r or less, then

$$u_x = y_x = P(x) \ . \tag{3.1}$$

In other words, an average that is exact for the degree r reproduces without change polynomials of degree r or less. If the weights are symmetrical, r must be odd, and we may write $r = 2s - 1$. This implies that $r < 2m + 1$, and therefore $s \leq m$.

For a simple (unweighted) average, $r = 1$. For the overwhelming majority of MWA's used in practice, $r = 3$. The preference for cubics has a long history. De Forest (1873, p. 281) suggests that "a curve of the third degree, which admits a point of inflexion ... is ... better adapted than the common parabola to represent the form of a series whose second difference changes its sign."

We shall use the notation of the calculus of finite differences, wherein E is the "displacement operator" defined by

$$Ef(x) = f(x + 1) ,$$

and δ is the "central difference" operator defined by

$$\delta f(x) = f(x + \tfrac{1}{2}) - f(x - \tfrac{1}{2}) ,$$

so that

$$\delta^2 f(x) = f(x + 1) - 2f(x) + f(x - 1) .$$

If the weighted average (1.2) is exact for the degree $2s - 1$, it can be written in the form

$$u_x = [1 - (-1)^s \delta^{2s} q(E)]y_x , \tag{3.2}$$

where $q(E)$ is of the form

$$q(E) = \sum_{j=-m+s}^{m-s} q_j E^j \tag{3.3}$$

with $q_{-j} = q_j$. In a typical smoothing formula $q(E)$ has only positive coefficients, but this is not necessarily the case. If $q(z)$ is multiplied by z^{m-s} to eliminate negative exponents, the resulting polynomial is of degree $2m - 2s$. Because of the symmetry of the coefficients, it is a reciprocal polynomial. In other words, if r is a zero of the polynomial, it follows that r^{-1} is a zero. We shall make the assumption that this polynomial has no zero on the unit circle. If it does have such zeros, the extension of the smoothing process to the extremities of the data is undefined.

Let $p(z)$ denote the polynomial of degree $m - s$ with leading coefficient unity whose zeros are the $m - s$ zeros of $z^{m-s} q(z)$ located within the unit circle. In general, some or all of these zeros are complex, but they must occur in conjugate pairs, so that $p(z)$ has real coefficients. Now we define a polynomial $a(z)$ of degree m and its coefficients a_j by

$$a(z) = (z - 1)^s p(z) = z^m - \sum_{j=1}^{m} a_j z^{m-j} . \tag{3.4}$$

Suppose the given data consist of $N = B - A + 1$ given values extending from $x = A$ to $x = B$. We assume that $N \geq 2m + 1$, so that at least one smoothed value is obtained

by direct application of the given MWA. Then we obtain m intermediate values to the left of $x = A$ by successive application of the recurrence

$$y_x = \sum_{j=1}^{m} a_j \, y_{x+j} \quad . \tag{3.5}$$

Similarly, m intermediate values to the right of $x = B$ will be obtained by the analogous recurrence

$$y_x = \sum_{j=1}^{m} a_j \, y_{x-j} \quad .$$

Finally, application of the symmetrical MWA of $2m + 1$ terms to the $N + 2m$ observed and intermediate values gives adjusted values u_x for $x = A, A + 1, \ldots, B$.

For example, Spencer's 15-term formula (1.1) can be expressed in the form (3.2) with $s = 2$, where

$$q(E) = \frac{1}{320} \left(3E^{-5} + 18E^{-4} + 59E^{-3} + 137E^{-2} + 242E^{-1} + 318 + 242E + 137E^{2}\right.$$
$$\left. + 59E^{3} + 18E^{4} + 3E^{5}\right) .$$

Using a computer program to find the zeros of $z^5 \, q(z)$, constructing the polynomial $P(z)$, and finally applying the formula (3.4), we obtain for Spencer's 15-term formula

$$a(z) = z^{7} - .961572z^{6} - .372752z^{5} - .015904z^{4} + .123488z^{3} + .125229z^{2}$$
$$+ .075887z + .025624 \quad .$$

The coefficients are rounded to the nearest sixth decimal place, except that the final digits of the coefficients of z^3 and z^2 have been adjusted by one unit to make the sum of the coefficients exactly zero.

Note that in the trivial case $s = m$, $q(z)$ is a constant and $p(z)$ is unity. Thus the algorithm reduces to extrapolation of the observed data by sth differences (i.e., by fitting a polynomial of degree $s - 1$ to the first s observations).

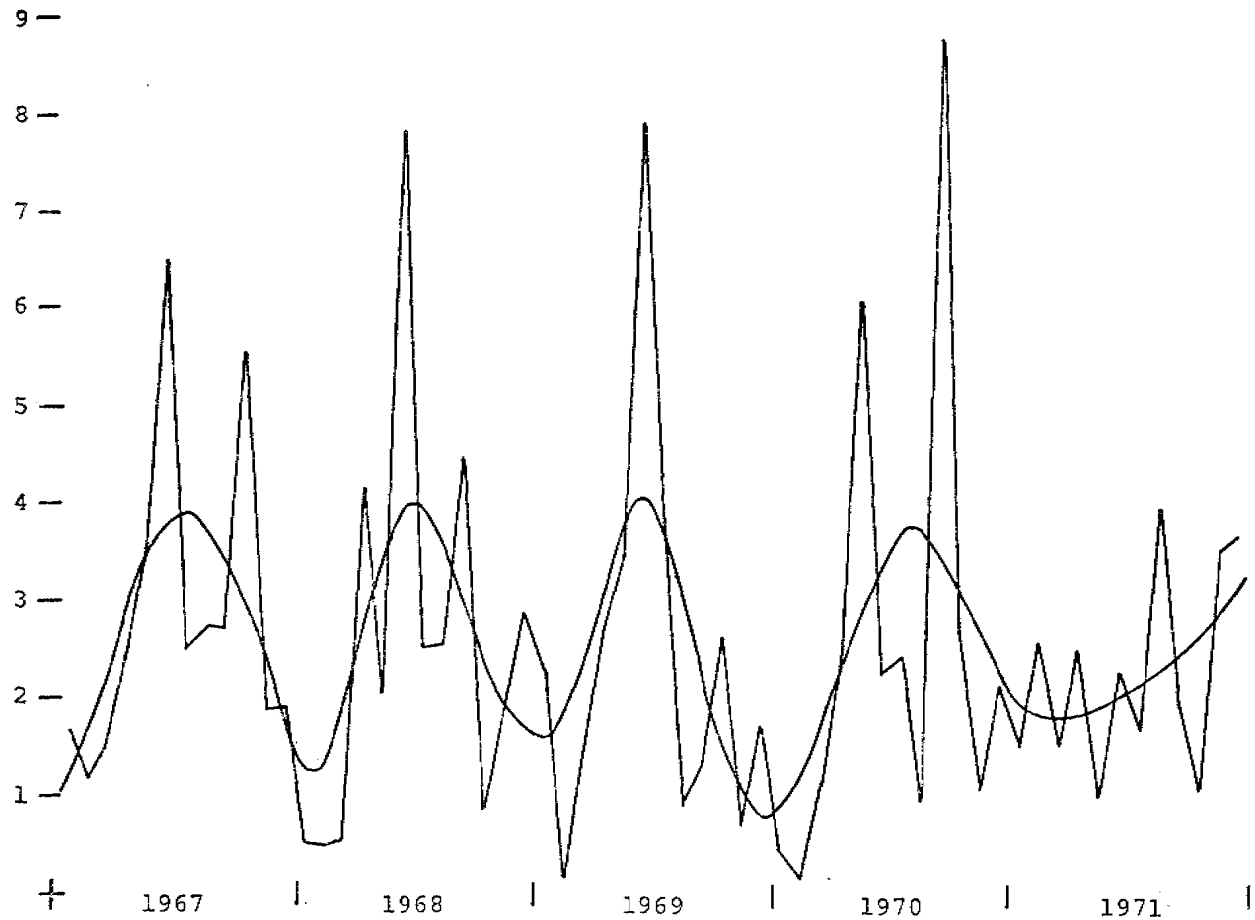
As a numerical illustration, Spencer's 15-term average has been applied to some meteorological data. Table 1 and Figure A show the observed and graduated values of monthly precipitation in Madison, Wisconsin in the years 1967-71. No adjustment has been made for the unequal length of the months.

1. Monthly Precipitation (Inches), Madison, Wisconsin, 1967-71.

Year	Month	Observed Value	Graduated Value	Year	Month	Observed Value	Graduated Value
1967	January	1.63	1.11	1969	July	4.28	3.81
	February	1.17	1.63		August	0.96	3.17
	March	1.49	2.24		September	1.35	2.33
	April	2.57	2.88		October	2.65	1.56
	May	3.53	3.42		November	0.70	1.06
	June	6.46	3.74		December	1.66	0.82
	July	2.51	3.85	1970	January	0.44	0.90
	August	2.71	3.75		February	0.16	1.25
	September	2.68	3.42		March	1.17	1.78
	October	5.52	2.92		April	2.53	2.39
	November	1.83	2.31		May	6.09	2.94
	December	1.89	1.69		June	2.26	3.37
1968	January	0.56	1.31		July	2.42	3.63
	February	0.49	1.36		August	0.97	3.69
	March	0.59	1.87		September	8.82	3.50
	April	4.18	2.69		October	2.65	3.20
	May	2.02	3.49		November	1.06	2.74
	June	7.82	3.91		December	2.12	2.28
	July	2.54	3.92	1971	January	1.48	1.94
	August	2.58	3.54		February	2.59	1.76
	September	4.45	2.97		March	1.52	1.74
	October	0.85	2.45		April	2.42	1.81
	November	1.74	1.99		May	0.98	1.93
	December	2.89	1.64		June	2.27	2.02
1969	January	2.26	1.56		July	1.65	2.13
	February	0.18	1.81		August	3.96	2.24
	March	1.47	2.35		September	1.87	2.40
	April	2.72	3.13		October	1.30	2.63
	May	3.45	3.81		November	3.48	2.84
	June	7.96	4.05		December	3.64	3.28

SOURCE: Observed values from U. S. Department of Commerce, National Oceanic and Atmospheric Administration, Environmental Data Service, Local Climatological Data, Annual Summary with Comparative Data, Madison, Wisconsin, 1972, National Climatic Center, Asheville, N. C., 1973.

A. Observed and Graduated Values of Monthly Precipitation, Madison, Wisconsin, 1967-1971



For the convenience of the user, the weighted-average coefficients and the intermediate-value coefficients for those averages that appear to be in common use or are found in publications accessible to me are given in the next section in Tables 2 and 3. The reader who is more interested in the justification of the procedure and the rationale behind it may skip at once to Section 5.

4. TABLES OF MOVING-AVERAGE AND EXTENSION COEFFICIENTS

Tables 2 and 3 show the coefficients in the MWA and the corresponding extension coefficients (that is, c_j and a_j) for 21 weighted averages that have appeared in the literature. Table 2 is devoted to the class of averages known to actuaries as minimum-R_3 formulas and to economic statisticians as "Henderson's ideal" formulas. They are discussed more fully in Section 7. The values in Table 2 are shown to six decimal places. In both instances, a few final digits have been adjusted by one unit to make the sum exactly unity. The moving-average coefficients are given to the nearest sixth decimal place except for the slight adjustments mentioned; rounding error in the computation of the extension coefficients may have introduced further small errors in some instances.

Table 3 is concerned with 11 moving averages derived by various writers on an ad hoc basis and known by the names of their originators. The source notes for this table do not attempt to cite the earliest publication of the formula in question, but merely indicate a convenient reference where it can be found. All these averages are exact for cubics except Hardy's, which is exact only for linear functions. The coefficients in the averages of Table 3 are rational fractions with relatively small denominators, and the user will probably find it convenient to use as weights the integers in the numerators of the coefficients, dividing by the common denominator as the final step. The column headings, therefore, are c_j multiplied by the common denominator.

In both Tables 2 and 3 advantage has been taken of the symmetry of the coefficients c_j to reduce the length of the columns by approximately one-half. The manner of using the tables may be illustrated by taking Spencer's 15-term average as an example. Equation (1.1) shows the calculation of the moving averages. The intermediate values y_x for $x = A - 1$ to $A - 7$ are calculated successively by the formula

$$y_x = .961572y_{x+1} + .372752y_{x+2} + .015904y_{x+3} - .123488y_{x+4} - .125229y_{x+5}$$

$$- .075887y_{x+6} - .025624y_{x+7} .$$

The intermediate values for $x = B + 1$ to $B + 7$ are calculated by the identical formula except that the "+" signs in the subscripts are changed to "-" signs.

The extension procedure drastically reduces the number of values that need to be tabulated for a given weighted average, and makes it possible, for example, to give complete information about 21 such averages in the reasonably compact Tables 2 and 3. However, the user who intends to apply a single weighted average to many data sets may prefer to tabulate the atypical elements of the smoothing matrix G for that weighted average, and so avoid the extra step of calculating the intermediate values. For the benefit of such users, a method of calculating the atypical rows of G will now be described. Justification of the procedure will be given in Section 9 (see equation (9.10)). We observe that the nonzero elements in each row of G except the first m and the last m rows are merely the coefficients c_j of the MWA centered about the diagonal element. The elements in the first m rows of G, except for the first m columns, follow from the symmetry of G, and if $G = (g_{ij})$, we have

$$g_{ij} = c_{j-i} .$$

This leaves only the square submatrix of order m in the upper left corner to be calculated. Let c denote the constant $-q_{m-s}/p_{m-s}$, where $-p_{m-s}$ is the term free of z in the polynomial $p(z)$, and let $A_1 = (a_{ij})$ denote the square matrix of order m given by

$$a_{ij} = \begin{cases} 0 & \text{for } i > j \\ 1 & \text{for } i = j \\ -a_{j-i} & \text{for } i < j \end{cases}$$

Then the required submatrix in the upper left corner of G is given by

$$I - cA_1^T A_1 .$$

The similar submatrix in the lower right corner of G contains the same elements, but with the order of both rows and columns reversed.

2. Moving-Average Coefficients (c_j) and Extension Coefficients (a_j) of Minimum-R_3 ("Henderson's Ideal") Averages of 5 to 23 Terms Exact for Cubics

	Number of Terms									
	5		7		9		11		13	
j	c_j[a]	a_j	c_j[a]	a_j	c_j[a]	a_j	c_j[a]	a_j	c_j[a]	a_j
0	.559440		.412588		.331140		.277944		.240058	
1	.293706	2	.293706	1.618034	.266557	1.352613	.238693	1.160811	.214337	1.016301
2	-.073426	-1	.058741	-.236068	.118470	.114696	.141268	.281079	.147356	.360880
3			-.058741	-.381966	-.009873	-.287231	.035723	-.140968	.065492	-.021625
4					-.040724	-.180078	-.026792	-.204545	0	-.160909
5							-.027864	-.096377	-.027864	-.138330
6									-.019350	-.056317

[a]Calculated by formula (7.5).

2. Moving-Average Coefficients (c_j) and Extension Coefficients (a_j) of Minimum-R_3 ("Henderson's Ideal") Averages of 5 to 23 Terms Exact for Cubics (continued)

	Number of Terms									
	15		17		19		21		23	
j	c_j[a]	a_j	c_j[a]	a_j	c_j[a]	a_j	c_j[a]	a_j	c_j[a]	a_j
0	.211542		.189232		.171266		.156470		.144060	
1	.193742	.903661	.176390	.813444	.161691	.739580	.149136	.678000	.138318	.625880
2	.145904	.397295	.141112	.410885	.134965	.412090	.128423	.406495	.121949	.397207
3	.082918	.064751	.092293	.124932	.096658	.166162	.097956	.193174	.097395	.212501
4	.024028	-.100710	.042093	-.043456	.054685	.005097	.063038	.046016	.068303	.075236
5	-.014134	-.135445	.002467	-.110644	.017474	-.078255	.029628	-.046290	.038933	-.015313
6	-.024499	-.094424	-.018640	-.106213	-.008155	-.099972	.003119	-.084020	.013430	-.063927
7	-.013730	-.035128	-.020370	-.065896	-.018972	-.081843	-.012896	-.084711	-.004948	-.078737
8			-.009961	-.023052	-.016601	-.047103	-.017614	-.063086	-.014527	-.070064
9					-.007378	-.015756	-.013455	-.034444	-.015687	-.048977
10							-.005570	-.011134	-.010918	-.025714
11									-.004278	-.008092

[a] Calculated by formula (7.5).

3. Moving-Average Coefficients (c_j) and Extension Coefficients (a_j) of Selected Moving Averages

	Macaulay[a]		Spencer 15-Term[b]		Woolhouse[c]		Hardy[d]		Higham[e]		Karup[f]	
j	$864c_j$	a_j	$320c_j$	a_j	$125c_j$	a_j	$120c_j$	a_j	$125c_j$	a_j	$625c_j$	a_j
0	182		74		25		24		25		125	
1	171	.919760	67	.961572	24	.885108	22	.739988	24	.859550	114	.820240
2	127	.393023	46	.372752	21	.421982	17	.386211	18	.399283	87	.402924
3	72	.055273	21	.015904	7	.028721	10	.124325	10	.087040	53	.114622
4	17	-.113111	3	-.123488	3	-.076050	4	-.023648	3	-.072738	21	-.047133
5	-17	-.140462	-5	-.125229	0	-.107285	0	-.080087	0	-.104527	0	-.102491
6	-19	-.084512	-6	-.075887	-2	-.092723	-2	-.079459	-2	-.093953	-8	-.091791
7	-10	-.029971	-3	-.025624	-3	-.059753	-2	-.049327	-2	-.055312	-9	-.060239
8							-1	-.018003	-1	-.019343	-6	-.028636
9											-2	-.007496

[a] Macaulay 1931, p. 55, footnote 2.

[b] Macaulay 1931, p. 55; Henderson 1938, p. 53.

[c] Henderson 1938, p. 53.

[d] Henderson 1938, p. 53; Benjamin and Haycocks 1970, p. 238.

[e] Henderson 1938, p. 53.

[f] Henderson 1938, p. 53.

3. Moving-Average Coefficients (c_j) and Extension Coefficients (a_j) of Selected Moving Averages (continued)

	Andrews[g]		Spencer 21-Term[h]		Hardy Wave-Cutting[i]		Vaughan Formula A[j]		Kenchington[k]	
j	10080c_j	a_j	350c_j	a_j	65c_j	a_j	1440c_j	a_j	385c_j	a_j
0	1688		60		5		182		45	
1	1579	.700747	57	.729724	5	.480996	179	.593256	44	.527740
2	1325	.406808	47	.408707	6	.368708	170	.396409	41	.370688
3	950	.179749	33	.167281	7	.267940	149	.230238	36	.236445
4	551	.027155	18	.009255	7	.166506	115	.096761	30	.128638
5	225	-.054586	6	-.069703	6	.072964	72	-.000857	22	.043118
6	-4	-.083701	-2	-.091513	4	-.008222	29	-.060076	13	-.018390
7	-124	-.078256	-5	-.076165	1	-.075454	-5	-.083321	5	-.053902
8	-135	-.054368	-5	-.049051	-1	-.097387	-26	-.079596	-1	-.067080
9	-110	-.031120	-3	-.022502	-2	-.089039	-29	-.056662	-5	-.064844
10	-61	-.012428	-1	-.006033	-2	-.062016	-19	-.028557	-6	-.050323
11					-1	-.024996	-6	-.007595	-5	-.032035
12									-3	-.015626
13									-1	-.004429

[g] Andrews and Nesbitt 1965, p. 18.

[h] Macaulay 1931, p. 51; Henderson 1938, p. 53.

[i] Benjamin and Haycocks 1970, p. 239.

[j] Vaughan 1933, p. 437.

[k] Henderson 1938, p. 53.

5. THE WHITTAKER GRADUATION PROCESS

It is not the purpose of this paper to consider the Whittaker (1923; see also Henderson 1924) graduation process in detail. However, since the natural method of extension of MWA graduation to the extremities of the data was arrived at primarily on the basis of analogies to the Whittaker method, the latter must be described sufficiently to make these analogies clear. The objective of the Whittaker process is to choose graduated values u_j $(j = A, A + 1, \ldots, B)$ in such a way as to minimize the quantity

$$\sum_{j=A}^{B} W_j (u_j - y_j)^2 + g \sum_{j=A}^{B-s} (\Delta^s u_j)^2 \quad , \tag{5.1}$$

where the weights W_j, the positive constant g, and the positive integer s are chosen a priori by the user. The solution is most conveniently expressed in matrix notation as follows (Greville 1957, 1974a). Let W denote the diagonal matrix of order N whose successive diagonal elements are the W_j, let u and y be defined as in Section 2, and let K denote the rectangular matrix of $N - s$ rows and N columns that transforms a vector v into the vector of sth finite differences of its components. Clearly the non-zero elements of K are binomial coefficients of order s with alternating signs (Greville 1974a). Then, the expression (5.1) can be written in the form

$$(u - y)^T W(u - y) + g(Ku)^T Ku \quad , \tag{5.2}$$

where the superscript T denotes the transpose. It is easily seen (Greville 1974a) that (5.2) is smallest when u satisfies

$$(W + gK^T K)u = Wy \quad . \tag{5.3}$$

It is not difficult to show (Greville 1957, 1974a) that the matrix in the left member of (5.3) is nonsingular (in fact, positive definite) and therefore

$$u = (W + gK^T K)^{-1} Wy \quad .$$

The remaining discussion will be limited to the so-called "Type A" case, in which all the weights W_j are taken equal to unity, as this case has the greatest similarity to MWA graduation. Here $W = I$ (the identity), and it is easily verified (Noble 1969, p. 147) that

$$(I + gK^T K)^{-1} = I - K^T(g^{-1} I + KK^T)^{-1} K . \qquad (5.4)$$

If the entire process of graduation, by whatever method or criterion, including data near the ends, is conceived in terms of matrix-vector multiplication (Greville 1957), so that

$$u = Gy \qquad (5.5)$$

for some matrix G, (5.4) suggests that it may be reasonable to consider matrices G of the form

$$G = I - K^T DK \qquad (5.6)$$

for some square matrix D and some order of differences s .

6. MATRIX DEVELOPMENT OF THE NATURAL METHOD OF COMPLETING THE GRADUATION

We suppose that N equally spaced observed values y_j $(j = A, A + 1, \ldots, B)$ are to be graduated primarily by means of a given symmetrical MWA of $2m + 1$ terms of the form (1.2), that is exact for the degree $2s - 1$. We assume that $N > 2m$. In other words, graduated values u_j for $j = A + m, A + m + 1, \ldots B - m$ will be calculated from the given weighted average. This requirement fixes the elements of the matrix G of (5.5) and (5.6) with the exception of the first m and the last m rows. The nonzero elements of each of the remaining $N - 2m$ rows will be merely the weights in the moving average with the middle weight on the diagonal in each case.

Our determination of the elements of the first m and the last m rows of G will be based on the general requirement that these rows shall not be something extra grafted onto the main part of the matrix, but shall be an integral part of an overall matrix having a well defined structure, this structure having the greatest possible analogy to that of the corresponding matrix for the Whittaker process. We shall try to show that this general requirement leads almost inexorably to the following three assumptions about G for the MWA case:

(i) G is symmetric and of the form (5.6);

(ii) D is such that D^{-1} exists and is a Toeplitz matrix (to be defined presently);

(iii) the elements of $D^{-1} = (d'_{ij})$ are given by

$$d'_{ij} = h_{i-j} \quad , \tag{6.1}$$

where $h_j = h_{-j}$ is a coefficient in the Laurent expansion

$$h(z) = [q(z)]^{-1} = \sum_{j=-\infty}^{\infty} h_j z^j \quad , \tag{6.2}$$

convergent in an annulus containing the unit circle.

These three assumptions (together with the assumption stated in the first paragraph of this section about the rows of G other than the first m and the last m) uniquely determine G. The three assumptions require extensive discussion, explanation, and comment, on which we now embark.

Since analogy to the Whittaker process is to have the highest priority, and (5.4) shows that G for that process is clearly symmetric and of the form (5.6), these being very basic structural properties, there can be little question about assumption (i). This assumption implies that G is a diagonal band matrix of band width $2m + 1$, and its elements are now determined except for a square submatrix of order m in the upper left corner and a similar submatrix in the lower right corner. It also implies that D is symmetric and is a diagonal band matrix of band width $2m - 2s + 1$.

It may be mentioned here that there is one basic, unavoidable difference between the Whittaker process and the MWA process. This is that, while in the MWA process (with the natural extension) G is a diagonal band matrix, in the Whittaker process it is the inverse of such a matrix. In consequence of this difference, the Whittaker process is "global" (each graduated value depending on <u>all</u> the observed values), while MWA is "local" (each graduated value depending only on a few neighboring observed values). This distinction carries over to the related matrix D, which, in the Whittaker process, is not a diagonal band matrix but the inverse of such a matrix (of band width $2s + 1$); from (5.4),

$$D^{-1} = g^{-1} I + KK^T \quad . \tag{6.3}$$

Assumption (i) fixes the elements of D except for those in a square submatrix of order $m - s$ in the upper left corner and a similar submatrix in the lower right corner. Reverting to the expression $q(E)$ of (3.2) and (3.3), we have $D = (d_{ij})$, with

$$d_{ij} = q_{i-j} \quad , \tag{6.4}$$

except within the two submatrices mentioned.

We define a Toeplitz matrix (see Trench 1974) as one in which all the elements on any diagonal line extending downward and to the right are equal. In other words, $T = (t_{ij})$ is a Toeplitz matrix when

$$t_{ij} = t_{i-j}$$

for all i and j.

It is easily verified that D^{-1} for the Whittaker process, given by (6.3), is a Toeplitz matrix. In fact, if $D^{-1} = (d'_{ij})$,

$$d'_{ij} = (-1)^{i-j} \binom{2s}{s+i-j} + g^{-1} \delta_{ij} \quad ,$$

where δ_{ij} is a Kronecker symbol.

Now, it is clear that the Toeplitz property is a very striking and obvious property of those matrices which possess it. Thus, in pursuit of our goal of maximum analogy between the Whittaker and MWA processes, we would wish, if at all possible, to make D^{-1} a Toeplitz matrix in the MWA case. Accordingly, let $D^{-1} = (d'_{ij})$ with $d'_{ij} = d'_{i-j}$ for all i and j. Since D is symmetric, D^{-1} is symmetric and $d'_{-j} = d'_{j}$. Consider the series

$$f(z) = \sum_{j=-\infty}^{\infty} d'_j z^j \tag{6.5}$$

(which may or may not converge). Because of (6.4) this series is a "reciprocal" of $q(z)$ at least in the formal sense that if $q(z)$ and $f(z)$ are formally multiplied together, the product is unity.

The latter fact does not uniquely determine the series (6.5). In order to achieve a unique determination, we invoke a further analogy between the MWA and Whittaker processes. We require that in the MWA case this series converge in some region of the complex plane. The corresponding series for the Whittaker case is finite, and therefore converges everywhere.

Now, a Laurent series like (6.5), if it converges anywhere, converges in an annulus.

Because of the symmetry of the coefficients, if it converges for $z = z_0$, it converges for $z = z_0^{-1}$. Therefore, the annulus of convergence, if it exists, contains the unit circle. Moreover, $[q(z)]^{-1}$ has a Laurent expansion (6.2) convergent in an annulus containing the unit circle if and only if $q(z)$ has no zero on the unit circle.

Thus, assumption (iii) is the only possible assumption consistent with assumptions (i) and (ii) that satisfies the requirement that (6.5) converge in some part of the plane, and assumption (iii) implies that $q(z)$ has no zeros on the unit circle. The prohibition against such zeros of $q(z)$ was previously alluded to in Section 3, and further reasons for insisting on it will be given in Section 10.

In reality, the part of assumption (ii) that asserts the nonsingularity of D is redundant, because it is shown in Section 9 that if a Toeplitz matrix $T(= D^{-1})$ is constructed in accordance with assumption (iii), then the square submatrices of order $m - s$ in the upper left and lower right corners of D can be chosen so that $DT = I$.

In the typical case D is a matrix of nonnegative elements (this is true in the Whittaker case), but this is not a requirement. (It is not true of Hardy's formula.)

The matrix-vector formulation does not lead at once to a convenient method for calculating the graduated values near the ends of the data. It will be shown in Section 9 to be equivalent to the extension algorithm described in Section 3, and also to the method of calculating the atypical elements of G described in Section 4.

7. SPECIAL CLASSES OF MOVING AVERAGES

Of particular interest are those moving averages known to actuaries as minimum-R_3 formulas and to economic statisticians as "Henderson's ideal" formulas. For a given number of terms $2m + 1$, this is the average (1.2), exact for the third degree, for which the quantity

$$\sum_{j=-m-3}^{m} (\Delta^3 c_j)^2 \tag{7.1}$$

is smallest (with the understanding that $c_j = 0$ for $|j| > m$). The "smoothing coefficient" R_3 is defined as the quantity obtained by dividing (7.1) by 20 and taking the square root. The divisor 20 is chosen because this is the value of (7.1) for the trivial case of (1.2) in which $c_0 = 1$ and $c_j = 0$ for $j \neq 0$.

The rationale for minimizing (7.1) may be explained as follows (Greville 1974a). If, for some x, u_x, u_{x+1}, u_{x+2}, and u_{x+3} are given by (1.2), which is the case for $x = A + m$ to $B - m - 3$, inclusive, then

$$\Delta^3 u_x = - \sum_{j=-m-3}^{m} (\Delta^3 c_j) y_{x+j+3} \quad . \tag{7.2}$$

It has been customary to regard the smallness (in absolute value) of the third differences of the graduated values as an indication of smoothness. Therefore (7.2) suggests that smoothness is encouraged by making the quantities $\Delta^3 c_j$ numerically small, and minimizing (7.1) is a way of doing this. The formula corresponding to (7.2) for a general order of differences is

$$\Delta^s u_x = (-1)^s \sum_{j=-m-s}^{m} (\Delta^s c_j) y_{x+j+s} \quad , \tag{7.3}$$

and the general formula for R_s is

$$R_s^2 = \sum_{j=-m-s}^{m} (\Delta^s c_j)^2 / \binom{2s}{s} \quad . \tag{7.4}$$

There is some question whether Henderson's contribution warrants attaching his name to the "ideal" weighted averages. De Forest (1873) treated extensively the formulas that minimize R_4. The concept of choosing the coefficients c_j in order to minimize R_3 seems to have been first mentioned by G. F. Hardy (1909). These averages were fully discussed by Sheppard (1913) slightly earlier than by Henderson (1916). However, Henderson does seem to have been the first to give an explicit formula for the coefficient c_j in the weighted average minimizing R_3 (Henderson 1916, p. 43; Macaulay 1931, p. 54; Henderson 1938, p. 60; Miller 1946, p. 71; Greville 1974a, p. 18). If we write $k = m + 2$, so that the weighted average has $2k - 3$ terms, the formula is

$$c_j = \frac{315[(k-1)^2 - j^2](k^2 - j^2)[(k+1)^2 - j^2](3k^2 - 16 - 11j^2)}{8k(k^2 - 1)(4k^2 - 1)(4k^2 - 9)(4k^2 - 25)} \quad . \tag{7.5}$$

Weighted averages that minimize R_s have been discussed from other points of view by Wolfenden (1925), Schoenberg (1946), and Greville (1966, 1974b).

Also deserving of special mention are the averages (exact for cubics) that minimize

R_0, sometimes called "formulas of maximum weight" or "Sheppard's ideal" formulas. These are sometimes applied to physical measurements when the errors of observation can be regarded as random "white noise" (see discussion of "reduction of error" in Section 8). The weights are given by

$$c_j = \frac{3(3m^2 + 3m - 1) - 15j^2}{(2m - 1)(2m + 1)(2m + 3)} .$$

Weighting coefficients (c_j) and extension coefficients (a_j) for minimum-R_3 (Henderson's ideal) averages of 5, 7, ..., 23 terms are given in Table 2.

8. COMPARISON WITH OTHER METHODS. PRACTICAL CONSIDERATIONS

If a symmetrical MWA exact for the degree $2s - 1$ is being used to smooth the main part of the data, it can easily be deduced, either from the extension algorithm described in Section 3 or from the matrix formulation of (5.5) and (5.6) that the unsymmetrical weightings proposed for smoothing the first m and the last m observations are exact only for the degree $s - 1$. For example, all the averages represented in Tables 2 and 3 with the exception of Hardy's are exact for cubics, and therefore their extensions to values near the ends are exact only for linear functions. Hardy's weighted average is exact for linear functions and its extension only for constants.

The Whittaker process has a similar property. At a sufficient distance from the ends of the data, polynomials of degree $2s - 1$ are "almost" reproduced by that process. In support of this rather loose statement the following heuristic argument is advanced. For the Whittaker process

$$G = (I + gK^T K)^{-1} = I - gGK^T K .$$

Thus, if y is the vector of observed values, the vector of corrections to these values is

$$- gGK^T Ky .$$

Now, the nonzero elements of $K^T K$, with the exception of the first s and the last s rows, are binomial coefficients of order $2s$ with alternating signs. Therefore the components of $K^T Ky$, except for the first s and the last s, are $(2s)$th differences of those of y (or their negatives if s is odd). Thus, if y is a vector of ordinates of

a polynomial of degree $2s - 1$, $K^T Ky$ is a vector of zeros except for the first s and the last s components. The components of $GK^T Ky$ are graduated values of those of $K^T Ky$, and therefore should be very small at some distance from the extremities of the data. Finally, multiplication by g, even though g is typically large, should give small corrections at a sufficient distance from the ends of the data.

Some users may consider the reduction in degree of exactness near the ends of the data a disadvantage of the natural method of extension. Before I became aware of the natural method, I had proposed (Greville 1974a) a different method of extension (already mentioned in Section 2) that does not have this particular disadvantage (though it has other shortcomings). This involves extrapolation by a polynomial of degree $2s - 1$ fitted by least squares to the first $m + 1$ observations. A similar polynomial is fitted to the last $m + 1$ observations for extrapolation at the other end of the data. There may be a gain in simplicity in using a single method of extrapolation for all symmetrical weighted averages, the particular extrapolated values depending only on the number of terms in the main formula. However, there is a loss in that the extension method is no longer tailored to the particular symmetrical average used.

Like the natural method of extension, the method using extrapolation by least squares can be collapsed into a single matrix G. When this is done, the diagonal band character of the smoothing matrix is maintained, but the symmetry is lost. Though the matrix approach is less convenient for computational purposes, the differences between the two methods are best elucidated by comparing the first m rows of the respective matrices G. This is done in Tables 4 and 5 for the case of the 9-term "ideal" formula. Here $m = 4$, but for convenience the fifth row is also shown. Its elements would be repeated in the subsequent rows, moving successively to the right, until we come to the last four rows. While an average of as few as 9 terms would seldom be used in practice, this is a convenient illustration.

As previously indicated, the first m rows and the last m rows of G may be regarded as exhibiting unsymmetrical weighted averages which are to be used near the ends of the data to supplement the symmetrical average used elsewhere. The coefficients that appear in the last m rows are the same as those in the first m rows, but the order is

reversed, both horizontally and vertically. It should be noted that the coefficients in the supplemental averages depend only on those of the underlying symmetrical average. They do not depend on N, the number of observations in the data set (which is the order of G).

The coefficients in the supplemental weighted averages based on least-square extrapolation, exhibited in Table 5, show two undesirable features. These are negative coefficients of substantial numerical magnitude, and successive waves of positive and negative coefficients as one proceeds from left to right along the rows. The number of such waves would increase as the number of terms in the underlying formula increases.

In striking contrast is the character of the coefficients of the natural extension. Like the coefficients in the underlying symmetrical formula, each row exhibits a peak in the vicinity of the main diagonal of the matrix, tapering off to a single group of negative coefficients of reduced size near the edge of the diagonal band.

In the least-squares method only a very small correction is made to the initial observed value. The corresponding correction in the natural method is more substantial.

The "second-difference correction" is the coefficient of the second-difference term when the formula is expressed in terms of increasing orders of differences in the form

$$u_x = y_x + c\,\Delta^2 y_{x-h} + \cdots \quad .$$

The coefficient c does not depend on the subscript $x - h$, in which there is some freedom of choice. For the formulas based on least-squares extrapolation, which are exact for cubics, the fourth-difference correction is similarly defined.

Some writers (Miller 1946, Wolfenden 1942, Greville 1974a) have regarded the observed values y_x as the sum of "true" values U_x and superimposed random errors e_x. If it is assumed that the errors e_x for different x are uncorrelated, and have zero mean and constant variance σ^2 for all x, then the variance of the error in the smoothed value u_x is $R_0^2\,\sigma^2$, where R_0^2 is obtained by taking $s = 0$ in (7.4). Thus, R_0 may be interpreted as the ratio of reduction in the standard deviation of error that results from application of the weighted average.

While the assumptions underlying the preceding analysis may be questioned, nevertheless a good case can be made that, for any weighted average, R should be less than unity. Since R_0^2 is the sum of the squares of the coefficients in the average, R_0 can

4. Matrix Elements for the Natural Extension of the 9-Term Minimum-R_3 Smoothing Formula, with Second-Difference Correction and R_o Value for Each Supplemental Formula

	j										
i	1	2	3	4	5	6	7	8	9	Second-Difference Correction	R_o
1	.773854	.305888	.025938	-.064956	-.040724	0	0	0	0	-.4133	.8360
2	.305888	.360106	.270804	.113799	-.009873	-.040724	0	0	0	.1457	.5579
3	.025938	.270804	.357131	.278254	.118470	-.009873	-.040724	0	0	.1931	.5429
4	-.064956	.113799	.278254	.338473	.266557	.118470	-.009873	-.040724	0	.0744	.5441
5	-.040724	-.009873	.118470	.266557	.331140	.266557	.118470	-.009873	-.040724	0	.5322

5. Matrix Elements for the Least-Squares Extension of the 9-Term Minimum-R_3 Smoothing Formula with Fourth Difference Correction and R_o Value for Each Supplemental Formula

	j										
i	1	2	3	4	5	6	7	8	9	Fourth-Difference Correction	R_o
1	.985350	.058600	-.087900	.058600	-.014650	0	0	0	0	-.01465	.9928
2	.025386	.857731	.315214	-.345889	.188282	-.040724	0	0	0	-.01534	.9962
3	-.206335	.652571	.412341	.048375	.240395	-.009873	-.040724	0	0	-.41580	.8369
4	-.140189	.232497	.299136	.241547	.299136	.118470	-.009873	-.040724	0	-.68194	.5717
5	-.040724	-.009873	.118470	.266557	.331140	.266557	.118470	-.009873	-.040724	-.75525	.5322

never be less than the maximum of the absolute values of the coefficients. Thus, an average cannot be considered satisfactory if the absolute value of any coefficient is equal to or greater than unity.

As indicated in Section 7, it has long been customary to regard a graduation as smooth if the third differences of the graduated values are small in absolute value. If $G = (g_{ij})$, we have

$$u_{A+i-1} = \sum_{j=1}^{N} g_{ij}\, y_{A+j-1} ,$$

and therefore

$$\Delta^s u_{A+i-1} = \sum_{j=1}^{N} y_{A+j-1}\, \Delta_i^s g_{ij} , \tag{8.1}$$

where the subscript of Δ indicates that the differences are taken with respect to i (i.e., down the columns of the matrix). If one avoids the corner submatrices, the nonzero elements g_{ij} in (8.1) are merely coefficients in the underlying symmetrical average, and (8.1) reduces to (7.3). This was the rationale underlying the derivation of the minimum-R_s averages.

Of course, if G is symmetric, it makes no difference whether the differences are taken horizontally or vertically. When the symmetry of G is not assumed, care must be exercised. Many years ago (Greville 1947, 1948) I published what purported to be coefficients in supplemental averages to be used near the ends of the data in conjunction with minimum-R_3 and minimum-R_4 symmetrical averages. The symmetry of G was not assumed, and I made the error of deriving the unsymmetrical coefficients by minimizing their third differences taken horizontally. The tables in question are therefore based on an incorrect assumption. Further it may be mentioned in passing that in the 1947-8 formulation the diagonal band character was not maintained, since the supplemental averages contained the full $2m + 1$ terms.

Table 6 shows, for the natural and the least-squares extensions of the 9-term minimum-R_3 formula, those third differences of the matrix elements, taken vertically, that involve elements of the first five rows. The entries in the fifth row of Table 6 would

6. Third Differences of Matrix Elements for the Natural and Least-Squares Extensions of the 9-Term Minimum-R_3 Smoothing Formula

	j								
i	1	2	3	4	5	6	7	8	9
					Natural Extension				
1	-.000046	.075817	-.006665	-.064956	-.077748	.025917	.112299	-.040724	0
2	.130190	.101036	.084297	-.027899	-.103248	-.077748	.025917	.112299	-.040724
3	-.098634	.059488	.112348	.055964	-.045662	-.103248	-.077748	.025917	.112299
4	-.057216	-.021246	.066051	.095915	.045662	-.045662	-.103248	-.077748	.025917
5	.040724	-.112299	-.025917	.077748	.103248	.045662	-.045662	-.103248	-.077748
					Least-Squares Extension				
1	-.430376	.789377	.095655	-.709595	.157447	.025917	.112299	-.040724	0
2	-.264548	.392618	.142871	-.257320	-.033365	-.077748	.025917	.112299	-.040724
3	-.092060	.033815	.119784	.091815	-.069850	-.103248	-.077748	.025917	.112299
4	.018017	-.139944	.045169	.192841	.013083	-.045662	-.103248	-.077748	.025917
5	.040724	-.112299	-.025917	.077748	.103248	.045662	-.045662	-.103248	-.077748

be repeated in subsequent rows, moving successively to the right. Casual inspection of the table shows that the third differences are numerically smaller for the natural extension. All of these third differences are less than 0.14 in absolute value. Two of those for the least-squares extension exceed 0.7 in absolute value.

It is instructive to compare the natural extension with the least-squares extension for the numerical example of Section 3. Though neither extension is recommended for use when additional data are available beyond the range of the original data set, nevertheless it may be of interest, purely for purposes of illustration, to choose a numerical example in which such additional data are available, and this has been done.

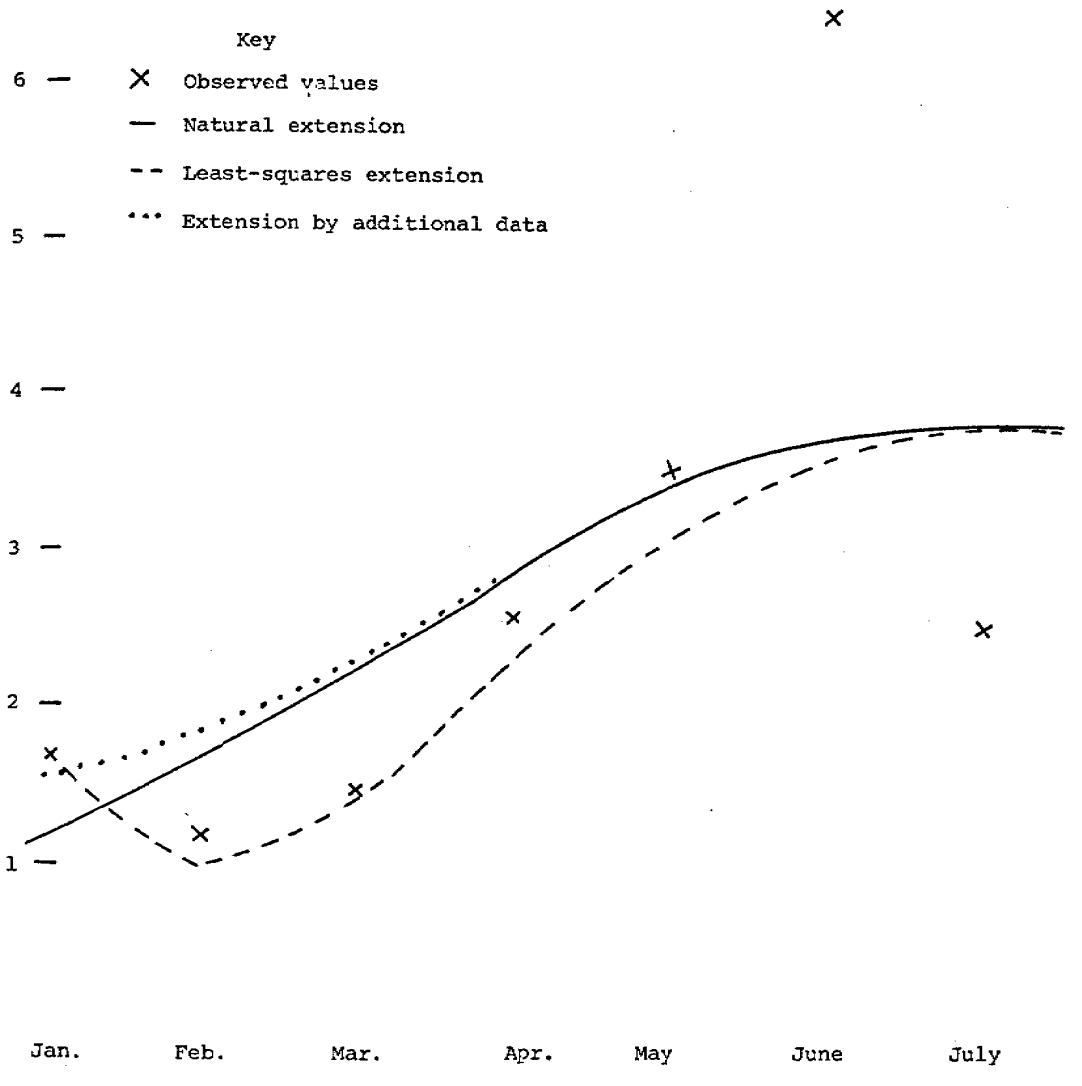
Table 7 and Figures B and C show, for the first seven months of 1967 and the last seven months of 1971, the observed values of precipitation in Madison, Wisconsin, and the graduated values obtained by (i) natural extension of Spencer's 15-term average, (ii) least-squares extension of the same average, and (iii) use of additional data. It will be noted that the least-squares extension is strongly constrained toward each of the two terminal observations (January 1967 and December 1971). This may be explained by the fact that all the values y_x in (1.2) that enter into the calculation of these graduated values are included in either the $m + 1$ observations to which the least-squares cubic was fitted or the m extrapolated values obtained from the same cubic. On the other hand, the natural extension and the least-squares extension are very close together at the interface with the graduated values calculated in the standard manner. Thus, for the months of July 1967 and June 1971, all but one of the values y_x entering into the computation (1.2) are identical for the two methods.

For the months closer to the interface the graduated values obtained by introducing additional data are close to those of the natural extension. This is because the supplemental unsymmetrical averages produced by the natural extension (unlike those of the least-squares extension) give relatively small weight to the observations more remote from the one being graduated (as does the underlying symmetrical formula). For example, the values for the natural extension and those obtained by the use of additional data are indistinguishable in Figure B for April to July 1967. In the last months of 1971 the deviation is greater because the first two months of 1972 were exceptionally dry. This could not have been predicted from the data for preceding months.

7. Extension of 15-Term Spencer Graduation of Madison Precipitation Data to First Seven and Last Seven Months by Different Methods

Year and Month	Observed Value	Extension of Graduation by Natural Method	Least-Squares Cubic	Additional Data
1967				
January	1.63	1.11	1.62	1.56
February	1.17	1.63	0.98	1.84
March	1.49	2.24	1.37	2.29
April	2.57	2.88	2.32	2.85
May	3.53	3.42	3.07	3.36
June	6.46	3.74	3.61	3.70
July	2.51	3.85	3.82	3.84
1971				
June	2.27	2.02	2.00	2.05
July	1.65	2.13	2.03	2.23
August	3.96	2.24	2.00	2.39
September	1.87	2.40	1.97	2.51
October	1.30	2.63	2.08	2.50
November	3.48	2.84	2.58	2.31
December	3.64	3.28	3.85	2.04

B. Observed and Graduated Values of Monthly Precipitation, Madison, Wisconsin, January to March, 1967



C. Observed and Graduated Values of Monthly Precipitation, Madison, Wisconsin, July to December, 1971

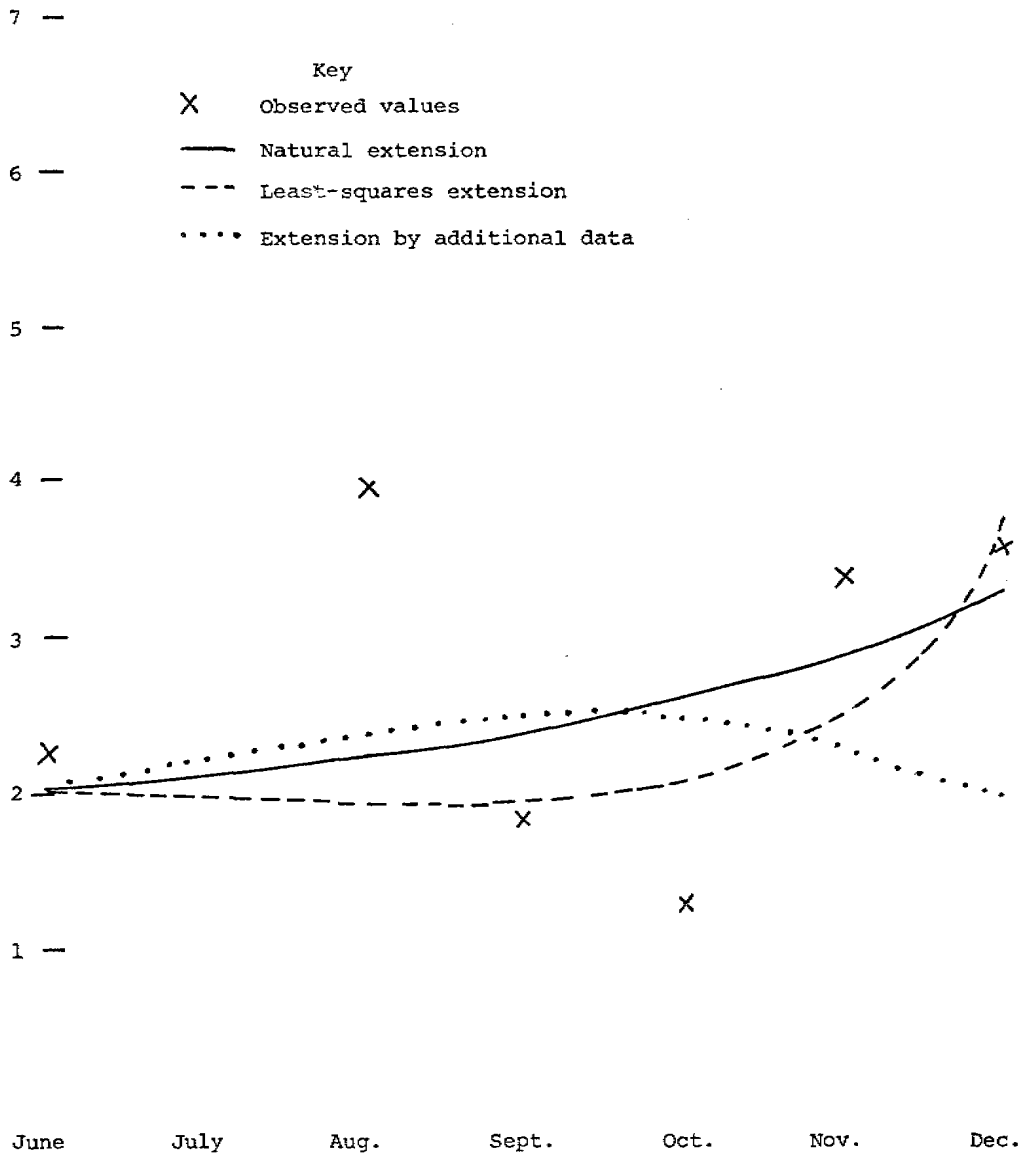


Table 8 gives certain parameters for the various symmetrical weighted averages that have been mentioned previously. The column headed "Error" requires explanation. This is the error committed when the formula in question is used to "smooth" a polynomial of degree four. This naturally tends to increase with the number of terms in the formula. Both R_0 and R_3 tend to decrease with increasing number of terms. Though the "ideal" formulas have been derived to minimize R_3, they tend to produce small values of R_0 as well. In only one instance (Vaughan) does a "name" formula have a smaller R_0 than the ideal formula of the same number of terms. The late Hubert Vaughan was an unusually keen analyst of MWA smoothing.

It may be mentioned in passing that some writers (e. g., Henderson 1938) call the reciprocal of R_0^2 the "weight" and the reciprocal of R_3 the (smoothing) "power."

9. PROOF OF EQUIVALENCE OF THE MATRIX AND INTERMEDIATE-VALUE APPROACHES

Though this proof involves only elementary mathematics, it is fairly long and complicated, and is therefore organized in the form of three lemmas and a theorem.

Let

$$p(z) = z^{m-s} - \sum_{j=1}^{m-s} p_j \, z^{m-s-j} ,$$

where $p(z)$ is the polynomial defined in Section 3 whose zeros are the zeros of $q(z)$ located inside the unit circle.

Lemma 9.1. The quantities h_j of (6.2) satisfy the recurrence

$$h_j = \sum_{\ell=1}^{m-s} p_\ell \, h_{j-\ell} \tag{9.1}$$

for all positive j .

Proof. In an annular region containing the unit circle, we have

$$h(z) \, q(z) = 1 .$$

But, for a suitable (nonvanishing) constant c ,

$$q(z) = cp(z) \, p(z^{-1}) . \tag{9.2}$$

In fact, $c = -q_{m-s}/p_{m-s}$. Therefore,

8. Parameters of the Symmetrical Weighted Averages Listed in Tables 2 and 3

Designation	Number of Terms	R_o	R_3	Error
Minimum-R_3 (Henderson's ideal):	5	.7045	.2735	$-.073\delta^4$
	7	.5971	.1147	$-.29\delta^4$
	9	.5323	.0581	$-.76\delta^4$
	11	.4865	.0331	$-1.57\delta^4$
	13	.4515	.0204	$-2.88\delta^4$
	15	.4234	.0134	$-4.85\delta^4$
	17	.4002	.0095	$-7.64\delta^4$
	19	.3806	.0066	$-11.4\delta^4$
	21	.3636	.0048	$-16.5\delta^4$
	23	.3488	.0036	$-23.1\delta^4$
Macaulay	15	.4273	.01657	$-4.52\delta^4$
Spencer	15	.4389	.01659	$-3.86\delta^4$
Woolhouse	15	.4602	.0654	$-5.4\delta^4$
Hardy	17	.4059	.0105	$\frac{1}{12}\delta^2 - 3.70\delta^4$
Higham	17	.4127	.0179	$-6.4\delta^4$
Karup	19	.4036	.0095	$-7.8\delta^4$
Andrews	21	.3707	.00628	$-14.9\delta^4$
Spencer	21	.3784	.00626	$-12.6\delta^4$
Hardy, wave-cutting	23	.3332	.0154	$-48.8\delta^4$
Vaughan A	23	.3415	.0050	$-26.6\delta^4$
Kenchington	27	.3202	.0031	$-22.4\delta^4$

$$cp(z)\ p(z^{-1})\ h(z) = 1\ . \tag{9.3}$$

Now, $[p(z)]^{-1}$ has an expansion in negative powers of z, with exponents not greater than $-m + s$, whose region of convergence contains the unit circle. Call it $b(z)$. Then,

$$cp(z^{-1})\ h(z) = b(z)\ ,$$

from which it follows that (9.1) holds for all positive j, and the proof is complete.

Let D_{11} denote the (unknown) square submatrix of order $m - s$ in the upper left corner of D. Let $P = (p_{ij})$ be a matrix of $m - s$ rows and $2m - 2s$ columns defined by

$$p_{ij} = \begin{cases} 0 & \text{for } i > j \\ 1 & \text{for } i = j \\ -p_{j-i} & \text{for } 0 < j - i \leq m - s \\ 0 & \text{for } j - i > m - s\ . \end{cases}$$

Let P be partitioned in the form $[P_1\ P_2]$, where P_1 and P_2 are square. Let $T = (t_{ij})$ denote the Toeplitz matrix of order $N - s$ defined by

$$t_{ij} = h_{i-j}\ .$$

In other words, T is D^{-1} under assumption (iii) of Section 6.

Lemma 9.2. T is nonsingular and equal to D^{-1} if D is completed by assigning

$$D_{11} = cP_1^T P_1\ ,$$

together with a corresponding assignment of the square submatrix of order $m - s$ in the lower right corner of D.

Proof. Note that if we try to form the product DT, all elements of the product that do not involve the missing elements in the corners of D have the correct values (0 or 1) . We shall focus on the upper left corner; similar considerations apply to the lower right corner. The lemma will be proved if it can be shown that the product DT is indeed the identity if D_{11} (and its counterpart at the lower right) is chosen in the manner indicated.

Let D_{12} denote the square submatrix of D of order $m - s$ immediately to the

right of D_{11}, let T_{11} be the submatrix corresponding to D_{11} in the upper left corner of T, and T_{12} the one immediately to its right. By symmetry the square submatrix of order $m - s$ immediately below T_{11} is T_{12}^T. The product DT will be the identity if D_{11} is such that

$$D_{11}\, T_{11} + D_{12}\, T_{12}^T = I \tag{9.4}$$

(and if a similar relation holds in the lower right corner).

It is easily verified that

$$D_{12} = cP_1^T P_2 \,. \tag{9.5}$$

Let $\tilde{D}_{11} = (\tilde{d}_{ij})$ be a square matrix of order $m - s$ defined by

$$\tilde{d}_{ij} = q_{i-j} \,.$$

The reader will note that replacement of D_{11} by $\tilde{D}_{11}$ (and a corresponding replacement in the lower right corner) would make D a Toeplitz matrix. It follows from the definition of $h(z)$ in (6.2) and the Toeplitz character of the matrices involved that

$$D_{12}^T\, T_{12} + \tilde{D}_{11}\, T_{11} + D_{12}\, T_{12}^T = I \,. \tag{9.6}$$

(The reader may think of the block immediately below D_{11} as moved up to the left of D_{11}, and the block T_{12} moved to a position immediately above T_{11}.) It is clear from (9.6) that (9.4) will be satisfied if

$$D_{12}^T\, T_{12} + \tilde{D}_{11}\, T_{11} = D_{11}\, T_{11} \,. \tag{9.7}$$

It is easily verified that

$$\tilde{D}_{11} = [P_2^T\, P_1^T] \begin{bmatrix} P_2 \\ P_1 \end{bmatrix} = P_2^T P_2 + P_1^T P_1 \,.$$

Substitution of this result and (9.5) in the left member of (9.7) gives

$$c(P_2^T P_1\, T_{12} + P_2^T P_2\, T_{11} + P_1^T P_1\, T_{11}) \,. \tag{9.8}$$

But

$$P_1\, T_{12} + P_2\, T_{11} = P \begin{bmatrix} T_{12} \\ T_{11} \end{bmatrix} = 0$$

by Lemma 9.1. Thus (9.8) reduces to $cP_1^T P_1 T_{11}$, and (9.7) is satisfied if $D_{11} = cP_1^T P_1$. This completes the proof.

Let L denote the m by N submatrix of $I - G = K^T DK$ consisting of the first m rows, and let $A = (a_{ij})$ be the m by N matrix defined by

$$a_{ij} = \begin{cases} 0 & \text{for } i > j \\ 1 & \text{for } i = j \\ -a_{j-i} & \text{for } 0 < j - i \leq m \\ 0 & \text{for } j - i > m \,, \end{cases} \tag{9.9}$$

where the coefficients a_j were defined in (3.4). Let A_1 denote the square submatrix of A consisting of the first m columns.

Lemma 9.3.

$$L = cA_1^T A \,. \tag{9.10}$$

Proof. Let D_1 denote the submatrix of D consisting of the first m rows, and let K_{11} denote the square submatrix of order m in the upper left corner of K. Then it follows from the placement of zeros in K^T that

$$L = K_{11}^T D_1 K \,.$$

Let $\hat{P}$ denote an m by $N - s$ matrix with the elements defined as in P (following the proof of Lemma 9.1). It is easily verified that

$$A = \hat{P}K \,.$$

Let $\hat{P}_1$ denote the square submatrix of $\hat{P}$ consisting of the first m columns. Then

$$A_1^T = K_{11}^T \hat{P}_1^T \,.$$

Thus,

$$cA_1^T A = cK_{11}^T \hat{P}_1^T \hat{P}K \,.$$

But it follows from the proof of Lemma 9.2 that $c\hat{P}_1^T \hat{P} = D_1$, and so

$$cA_1^T A = K_{11}^T D_1 K = L \,,$$

as required for the proof of the lemma.

Theorem 9.1. The extension method of Section 3 and the matrix formulation of Section 6 are equivalent.

Proof. Let A_2 denote the submatrix of A consisting of the $(m+1)$th to $(2m)$th columns and let

$$\hat{A} = [A_1 \; A_2] \quad .$$

Let $y^{(0)}$ denote the vector of the m intermediate values obtained from the observations by (3.5), let $y^{(1)}$ and $y^{(2)}$, respectively, denote the vectors of the first m observations and the $(m+1)$th to $(2m)$th observations, and let $\hat{y}$ denote the vector consisting of $y^{(0)}$ followed by $y^{(1)}$. Then, the extension method requires $\hat{A}\hat{y} = 0$, or

$$A_1 y^{(0)} = -A_2 y^{(1)} \quad , \tag{9.11}$$

Let $\hat{G} = (\hat{g}_{ij})$ be the square matrix of order m defined by $\hat{g}_{ij} = c_{i-j}$ (where the coefficients c_j were defined in (1.2)), and let G_{12} be the submatrix of G formed from the first m rows and the $(m+1)$th to $(2m)$th columns. Then the vector of the first m graduated values from the matrix formulation is, by Lemma 9.3,

$$y^{(1)} - cA_1^T(A_1 y^{(1)} + A_2 y^{(2)}) \quad . \tag{9.12}$$

By the extension method, the corresponding vector is

$$G_{12}^T y^{(0)} + \hat{G} y^{(1)} + G_{12} y^{(2)} \quad . \tag{9.13}$$

But, since $G = I - K^T DK$,

$$\hat{G} = I - c[A_2^T \; A_1^T] \begin{bmatrix} A_2 \\ A_1 \end{bmatrix} = I - c[A_1^T A_1 + A_2^T A_2] \quad ,$$

and

$$G_{12} = -cA_1^T A_2 \quad .$$

Thus, (9.13) reduces to

$$y^{(1)} - c[A_2^T A_1 y^{(0)} + A_1^T A_1 y^{(1)} + A_2^T A_2 y^{(1)} + A_1^T A_2 y^{(2)}] \quad .$$

The substitution (9.11) reduces this to (9.12), as required.

We note in passing that the computational short cut involving extended values has an analogue in the case of Whittaker smoothing. Especially in actuarial literature, the Whittaker smoothing process is sometimes called the difference-equation method because the difference equation

$$u_x + (-1)^s \delta^{2s} u_x = y_x \tag{9.14}$$

holds for $x = A + s,\ A + s + 1,\ \ldots,\ B - s$. It was pointed out by Aitken (1926) that (9.14) is satisfied for $x = A,\ A + 1,\ \ldots,\ B$ if we introduce at each end of the data set s extrapolated values of both y_x and u_x satisfying the conditions

$$u_x = y_x \qquad (x = A - j,\ x = B + j;\ \ j = 1, 2, \ldots, s)\ ,$$

$$\Delta^s u_x = 0 \qquad (x = A - j,\ x = B - j;\ \ j = 1, 2, \ldots, s)\ .$$

However, this observation is not helpful from a computational point of view. The attempt to utilize it merely increases the order of the linear system to be solved from N to $N + 2s$.

10. THE CHARACTERISTIC FUNCTION AND SCHOENBERG'S DEFINITION OF A SMOOTHING FORMULA

Schoenberg (1946) defined the characteristic function of the MWA (1.2) as

$$\phi(t) = \sum_{j=-m}^{m} c_j e^{ijt}\ . \tag{10.1}$$

For a symmetrical MWA this is a real function of the real variable t, and can be expressed in the alternative form

$$\phi(t) = \sum_{j=-m}^{m} c_j \cos jt\ .$$

It is periodic with period 2π and equal to unity for $t = 2\pi n$ for all integers n.

The effect of MWA's in eliminating or reducing certain waves has been noted (Elphinstone 1951, Hannan 1970). If the input to the smoothing process is a sine wave, which may be represented in the form

$$y_x = C \cos(rx + h)\ , \tag{10.2}$$

it can be shown by simple algebraic manipulation that

$$u_x = y_x \, \phi(2\pi/P) \quad ,$$

where $P = 2\pi/r$ is the period of y_x. Thus, if $\phi(2\pi/P) = 0$, the wave is annihilated by the smoothing process; the amplitude is severely reduced if it is close to zero. Thus MWA smoothing is related to the "filtering" processes considered by Wiener (1949) and others.

Schoenberg (1946) defined a smoothing formula as an MWA whose characteristic function $\phi(t)$ satisfies the condition

$$|\phi(t)| \leq 1 \tag{10.4}$$

for all t. Later (Schoenberg 1948, 1953) he suggested the stronger condition

$$|\phi(t)| < 1 \qquad (0 < t < 2\pi) \quad . \tag{10.5}$$

C. Lanczos (see Schoenberg 1953) pointed out that condition (10.4) is obtained by requiring that every simple vibration (10.2) be diminished in amplitude by the transformation (1.2). The results of Section 6 of the present paper suggest an alternative definition of a smoothing formula. Using the subscript N to emphasize the fact that the order of G is the number of observations in the data set, we may say that (1.2) is a smoothing formula if

$$G_N^{\infty} = \lim_{n \to \infty} G_N^n \tag{10.6}$$

exists for all N. Schoenberg (1953, footnote 3) suggested a relationship between (10.4) and the conditions for existence of the infinite power of a matrix (Oldenburger 1940, Dresden 1942), but he did not elaborate the connection. We shall show that the existence of the limit (10.6) for all N is equivalent to a condition intermediate between (10.4) and (10.5). The following lemma will help to elucidate the situation.

Lemma 10.1. For a given τ in $(0, 2\pi)$, $\phi(\tau) = 1$ if and only if $q(\tau) = 0$.

Proof. From (3.1), (3.2), and (10.1) it follows that

$$\phi(t) = 1 - (-1)^s \, (2i \sin \tfrac{1}{2}t)^{2s} \, q(e^{it}) = 1 - (4\sin^2 \tfrac{1}{2}t)^s \, q(e^{it}) \quad . \tag{10.7}$$

Since $\sin \frac{1}{2}\tau \neq 0$, the lemma is established.

There are two ways in which equality can hold in (10.4), namely $\phi(t) = 1$ and $\phi(t) = -1$, and the situation is different in the two cases. Lemma 10.1 shows that if $\phi(t) = 1$, $q(z)$ has a zero on the unit circle and consequently G_N is not defined. No such problem arises if $\phi(t) = -1$. We are therefore led to the intermediate condition

$$-1 \leq \phi(t) < 1 \qquad (0 < t < 2\pi) , \tag{10.8}$$

which we shall show to be equivalent to the existence of (10.6).

Lemma 10.2. If (10.8) holds, D is positive definite.

Proof. From (9.2) we obtain

$$q(1) = c[p(1)]^2 .$$

It follows from (10.7) and (10.8) that $q(e^{it})$ is positive for $0 < t < 2\pi$. Since it is a continuous function of t, it is nonnegative for $t = 0$: that is, $q(1)$ is nonnegative. By the definition of $p(z)$, $p(1) \neq 0$, and $c = -q_{m-s}/p_{m-s}$ does not vanish. Therefore $q(1)$ is positive and c is positive.

Let the expansion of $b(z)$ of Section 9 be given by

$$b(z) = \sum_{j=m-s}^{\infty} b_j z^{-j} . \tag{10.9}$$

It follows from (9.3) that on the unit circle

$$b(z)\, b(z^{-1}) = ch(z) . \tag{10.10}$$

Substitution of (10.9) gives

$$\sum_{\ell=m-s}^{\infty} b_\ell\, b_{\ell+j} = ch_j \qquad (j = \ldots, -1, 0, 1, \ldots) . \tag{10.11}$$

We note that the coefficients b_j satisfy the difference equation

$$b_j = \sum_{\ell=1}^{m-s} p_\ell\, b_{j-\ell} \qquad (j = m - s + 1, m - s + 2, \ldots) .$$

If $r_1, r_2, \ldots, r_{m-s}$ are the zeros of $p(z)$, it follows that

$$b_j = \sum_{k=1}^{m-s} \alpha_k\, r_k^j$$

for some constant coefficients α_k. Therefore the series in the left member of (10.11) is absolutely convergent.

Let $B = (b_{ij})$ denote the matrix of $N - s$ rows and a denumerable infinity of columns given by

$$b_{ij} = \begin{cases} 0 & (i > j) \\ b_{m-s+j-i} & (i \leq j) \end{cases} , \tag{10.12}$$

It follows from (10.11) and (6.1) that

$$D^{-1} = cBB^T . \tag{10.13}$$

The structure of the right member of (10.13) shows that D^{-1} is nonnegative definite; since it is nonsingular, it is positive definite, and consequently D is positive definite, as required.

Let $\psi(t)$ be defined by

$$\psi(t) = 1 - \phi(t) .$$

Then (10.8) is equivalent to

$$0 < \psi(t) \leq 2 \qquad (0 < t < 2\pi) . \tag{10.14}$$

Let ψ_{max} denote the maximum value of $\psi(t)$, and let $\tilde{A} = (a_{ij})$ be the square matrix of order N defined by (9.9). Let $|v|$ denote the Euclidean norm of a vector v.

Lemma 10.3. If v is any vector of N real components and $\psi(t)$ is positive in $(0, 2\pi)$,

$$|\tilde{A}v|^2/|v|^2 \leq c^{-1} \psi_{max} . \tag{10.15}$$

Proof. Let v be an arbitrary vector of N real components, with jth component v_j, and let

$$V(t) = \sum_{j=1}^{N} v_j e^{ijt} \tag{10.16}$$

be the characteristic function of v. Then, if $a(z)$ is the polynomial defined by (3.4).

$$a(e^{it}) V(t) = \sum_{j=1}^{N+m} w_j e^{ijt} ,$$

where, for $j \geq m$, w_j is the $(j-m)$th component of $\tilde{A}v$. Moreover, it follows from (3.4), (9.2), and (10.7) that

$$\psi(t) = ca(e^{it})\, a(e^{-it}) \quad . \tag{10.17}$$

By Parseval's formula (Schoenberg 1946)

$$|v|^2 = \frac{1}{2\pi} \int_0^{2\pi} |V(t)|^2 \, dt \quad . \tag{10.18}$$

Similarly, in view of (10.17),

$$|\tilde{A}v|^2 \leq \frac{1}{2\pi} \int_0^{2\pi} |a(e^{it})|^2 \, |V(t)|^2 \, dt = \frac{c^{-1}}{2\pi} \int_0^{2\pi} \psi(t) \, |V(t)|^2 \, dt \quad . \tag{10.19}$$

From (10.18) and (10.19), (10.15) follows.

It is easily verified that the symmetric matrix $\tilde{A}^T \tilde{A}$ is indentical with $F = I - G$ except for the elements of the square submatrix in the lower right corner. In fact, the elements in the lower right corner of $\tilde{A}^T \tilde{A}$ are such that the entire matrix becomes a Toeplitz matrix if the first m rows and the first m columns are deleted. It follows that F can be obtained from $c\tilde{A}^T \tilde{A}$ by subtracting a square matrix of order N whose elements are all zero except for the square submatrix of order m in the lower right corner.

In fact, if A_1 and A_2 are defined as in the proof of Theorem 9.1, it is easily verified that the square submatrix of order m in the lower right corner of $c\tilde{A}^T \tilde{A}$ is given by

$$c[A_2^T \; A_1^T] \begin{bmatrix} A_2 \\ A_1 \end{bmatrix} = c(A_2^T A_2 + A_1^T A_1) \quad ,$$

while the corresponding submatrix of F is $cA_1^T A_1$. If, therefore, $\bar{A}$ is defined as the square matrix of order N having A_1 in the lower right corner and zeros everywhere else, we have

Lemma 10.4.

$$F = I - G = c(\tilde{A}^T \tilde{A} - \bar{A}^T \bar{A}) \quad .$$

Before stating the theorem that is the main result of this section, we point out (Oldenberger 1940, Dresden 1942) that, for a given matrix C, $\lim_{n \to \infty} C^n$ exists if and only

if either all eigenvalues of C lie within the unit circle, or else 1 is a simple zero of the minimum polynomial of C and all other eigenvalues lie within the unit circle. As multiplication by G leaves unchanged vectors whose components are successive equally spaced ordinates of a polynomial of degree $s - 1$ or less, it is clear that 1 is an eigenvalue. As G is symmetric, all its eigenvalues are real, and all zeros of its minimum polynomial (including 1) are simple. Therefore the limit (10.6) exists if and only if all eigenvalues of G other than 1 are strictly between -1 and 1.

Theorem 10.1. The limit G_N^∞ exists for all N if and only if the characteristic function $\phi(t)$ satisfies the condition

$$-1 \leq \phi(t) < 1 \qquad (0 < t < 2\pi) . \tag{10.8}$$

Proof. Let (10.8) hold, and recall that (10.8) is equivalent to (10.14). Now we shall consider a particular value of N and, for convenience, drop the subscript N. The eigenvalues of F are obtained by subtracting from 1 those of G. We need to show, therefore, that the nonzero eigenvalues of F are positive and do not exceed 2.

Since $F = K^T DK$ and D is positive definite by Lemma 10.2, F is nonnegative definite. Therefore its nonzero eigenvalues are positive. Let v be an arbitrary nonzero real vector and consider the Rayleigh quotient,

$$r = \frac{v^T Fv}{v^T v} = c\,\frac{v^T \tilde{A}^T \tilde{A}v}{v^T v} + c\,\frac{v^T \bar{A}^T \bar{A}v}{v^T v}$$

by Lemma 10.4. Since $\bar{A}^T \bar{A}$ is nonnegative definite, we have

$$r \leq c|\tilde{A}v|^2/|v|^2 \leq \psi_{max}$$

by Lemma 10.3. Therefore (10.14) gives $r \leq 2$. Since the spectral radius is the maximum value of the Rayleigh quotient, this completes the first part of the proof.

We shall prove the converse by showing that if (10.8) fails, the limit (10.6) does not exist for some N. There are two ways in which (10.8) can fail. Either $\phi(t)$ may be equal to 1, or $|\phi(t)|$ may exceed 1. In the former case, as was pointed out immediately following Lemma 10.1, G is not defined.

We consider therefore the case in which $|\phi(t)| > 1$ for some $t = \tau$, and let v be the N-vector whose jth component is

$$v_j = \exp[i\tau(j - \frac{N+1}{2})] \ . \tag{10.20}$$

Using an asterisk to denote the conjugate transpose, we have $\bar{v}_j \, v_j = 1$, and therefore

$$v^* \, v = N \ . \tag{10.21}$$

Except for the first m and the last m components we have

$$(Gv)_j = \phi(\tau) \, v_j \ ,$$

and so

$$v^* \, Gv = \phi(\tau)(N - 2m) + C \ , \tag{10.22}$$

where C denotes the contribution of the first m and the last m components. Because of the symmetry of both the matrix elements and the vector components, C is real. Since all the vector components have absolute value 1, an upper bound to C is the sum of the absolute values of the elements in the first m and the last m rows of G. Call this C_1. We recall that C_1 does not depend on N.

Now choose N sufficiently large so that

$$N > \frac{C_1 + 2m|\phi(\tau)|}{|\phi(\tau)| - 1} \ .$$

Then

$$N[|\phi(\tau)| - 1] > C_1 + 2m|\phi(\tau)| \ ,$$

and it follows that

$$(N - 2m) \, |\phi(\tau)| > N + |C| \ .$$

Consequently,

$$|(N - 2m) \, \phi(\tau) + C| > N \ ,$$

and therefore, by (10.21) and (10.22),

$$|r| = \left|\frac{v^* \, Gv}{v^* \, v}\right| > 1 \ .$$

It follows that the spectral radius of G is greater than unity, and the proof is complete.

It is easily verified that G^{∞}, when it exists, is the orthogonal projector on the eigenspace of G associated with the eigenvalue 1, that is, the space of N-vectors whose components are successive equally spaced ordinates of polynomials of degree $s - 1$ or less.

11. SMOOTHING FORMULAS IN THE STRICT SENSE AND AN OPTIMAL PROPERTY

At an early stage of the investigations underlying this paper I was trying to explain the natural extension of the MWA graduation to my colleague, I. J. Schoenberg, whose work plays such an important role therein, and he asked me (I thought with a slight show of impatience) "What does it minimize?" My answer was that it doesn't minimize anything, but is just a natural way of extending the MWA graduation to the ends of the data. This was too simplistic an answer, for we shall now show that it does in fact minimize "something."

In a slightly more general form of the Whittaker smoothing method (Greville 1957) one minimizes the sum of the squares of the departures of the smoothed values from the observed values plus a designated positive definite quadratic form in the sth differences of the smoothed values. In other words, one minimizes

$$(u - y)^T (u - y) + (Ku)^T HKu \ ,$$

where H is a given positive definite matrix of order $N - m$. Minimization of this expression leads to the equation

$$(I + K^T HK)u = y \ ,$$

which has a unique solution for u since $I + K^T HK$ is positive definite. I showed (Greville 1957) that this graduation method has the interesting property that if roughness (opposite of smoothness) is measured by the term $(Ku)^T HKu$, smoothness is always increased by the graduation. By Theorem 5.22 of Noble (1969),

$$(I + K^T HK)^{-1} = I - K^T(H^{-1} + KK^T)^{-1} K \ .$$

The last expression is of the form (5.6) and suggests that the use of an MWA with the natural extension might be regarded as a generalized Whittaker smoothing process if

$$D = (H^{-1} + KK^T)^{-1} \ .$$

Solving for H gives

$$H = (D^{-1} - KK^T)^{-1} \quad . \tag{11.1}$$

We are led to inquire, therefore, under what conditions an MWA is such that the right member of (11.1) is positive definite. Clearly H is positive definite if and only if the Toeplitz matrix

$$H^{-1} = D^{-1} - KK^T \tag{11.2}$$

is positive definite.

Schoenberg (1946, p. 53) remarks that it is desirable for an efficient smoothing formula, one that achieves adequate smoothness without producing unnecessarily large departures from the observed values, to have its characteristic function satisfy the stronger condition

$$0 \leq \phi(t) \leq 1 \quad .$$

This remark seems to have been little noted in the years since its publication. We shall call an MWA a smoothing formula in the strict sense if its characteristic function satisfies the condition

$$0 \leq \phi(t) < 1 \qquad (0 < t < 2\pi) \quad , \tag{11.3}$$

and we shall show that (11.2) is positive definite for all N if and only if (11.3) holds.

Theorem 11.1. Let (10.8) hold. Then $Q = D^{-1} - KK^T$ is positive definite for all N if and only if the MWA is a smoothing formula in the strict sense.

Proof. Let (11.3) hold, let v be an arbitrary nonzero real N-vector, and consider the Rayleigh quotient,

$$r = \frac{v^T Qv}{v^T v} = (c|B^T v|^2 - |K^T v|^2)/|v|^2 \tag{11.4}$$

where B is given by (10.12). Let $V(t)$ be defined by (10.16). Then, by Parseval's formula,

$$|B^T v|^2 = \frac{1}{2\pi} \int_0^{2\pi} |e^{-(m-s)it}\, b(e^{-it})|^2 \, |V(t)|^2 \, dt$$

$$= \frac{c^{-1}}{2\pi} \int_0^{2\pi} |h(e^{it})|\,|V(t)|^2 \, dt \tag{11.5}$$

by (10.10). Moreover, again by Parseval's formula,

$$|K^T v|^2 = \frac{1}{2\pi} \int_0^{2\pi} (4\sin^2 \tfrac{1}{2}t)^s \, |V(t)|^2 \, dt \quad . \tag{11.6}$$

It was shown in the proof of Lemma 10.2 that $q(e^{it})$ is positive, and therefore $h(e^{it}) = [q(e^{it})]^{-1}$ is positive, for $0 < t < 2\pi$. By means of (11.5), (11.6), and (10.7), (10.4) gives

$$r = \frac{1}{2\pi} \int_0^{2\pi} [h(e^{it}) - (4\sin^2 \tfrac{1}{2}t)^s] \, |V(t)|^2 \, dt$$

$$= \frac{1}{2\pi} \int_0^{2\pi} h(e^{it}) \, \phi(t) \, |V(t)|^2 \, dt > 0 \, ,$$

since any zeros of $\phi(t)$ constitute a set of measure zero. Since r is positive for arbitrary nonzero v, Q is positive definite.

To prove the converse we shall suppose that (11.3) does not hold and show that r is negative for some N and some v. Because of the hypothesis that (10.8) holds, $\phi(t)$ is less than 1. We suppose, therefore, that $\phi(t) < 0$ for some $t = \tau$ in $(0, 2\pi)$. Let v be given by (10.20). By an argument similar to that used in the proof of Lemma 10.2, it is easily shown that the series (6.2) for $h(z)$ is absolutely convergent for $z = 1$. Thus, for any small positive quantity ϵ, there exists a positive integer M such that

$$\sum_{j=M+1}^{\infty} |h_j| < \frac{1}{2} \epsilon \quad .$$

Thus, for $N > 2M$, the jth component of Qv, for $j = M + 1, M + 2, \ldots, N - M$, is v multiplied by a quantity less than

$$h(e^{i\tau}) + \epsilon - (4\sin^2 \tfrac{1}{2}\tau)^s \, . \tag{11.7}$$

By (10.7).

$$h(e^{i\tau}) - (4\sin^2 \tfrac{1}{2}\tau)^s = h(e^{i\tau}) \, \phi(\tau) \, . \tag{11.8}$$

It was shown in the proof of Lemma 10.2 that (10.8) implies $h(e^{i\tau}) > 0$. Thus, (11.8) is negative. Choose ϵ sufficiently small so that (11.7) is negative.

As in the proof of Theorem 10.1, we find that $v^* v = N$, while

$$v^* Qv \leq (N - 2M)[h(e^{i\tau})\ \phi(\tau) + \epsilon\ ] + C\ , \qquad (11.9)$$

where C denotes the contribution of the first M and the last M rows. As in the proof of Theorem 10.1, it follows from symmetry that C is real. Let

$$\eta = \sum_{j=-\infty}^{\infty} \left| h_j - (-1)^{s-j} \binom{2s}{s-j} \right| ,$$

where $\binom{2s}{j}$ is understood to vanish for negative j or $j > 2s$. Then

$$C \leq |C| \leq 2M\eta ,$$

since all components of v have absolute value unity. Now take

$$N > \frac{2M[h(e^{i\tau})\ \phi(\tau) + \epsilon - \eta]}{h(e^{i\tau})\ \phi(\tau) + \epsilon} . \qquad (11.10)$$

Note that both numerator and denominator of the right member of (11.10) are negative. From (11.10) we obtain

$$(N - 2M)[h(e^{i\tau})\ \phi(\tau) + \epsilon\ ] < -2M\eta ,$$

and (11.9) then gives $v^* Qv < 0$, so that Q is not positive definite. This completes the proof of the theorem.

It is easy to construct an MWA that is a smoothing formula in the strict sense. A trivial example is the formula

$$u_x = \frac{1}{17}(-y_{x-2} + 4y_{x-1} + 11y_x + 4y_{x+1} - y_{x+2}) .$$

However, none of the weighted averages in general use fall in this class. In particular, using the properties of Jacobi polynomials, I have shown elsewhere (Greville 1966) that the characteristic functions of all minimum-R_s averages assume negative values in $(0, 2\pi)$. Thus no such formula is a smoothing formula in the strict sense.

There is, however, one family of moving averages, mentioned in the literature but not in general use, that are smoothing formulas in the strict sense. Elsewhere (Greville 1966) I have considered the limiting case of the minimum-R_s formulas as s tends to infinity. In finite-difference form, the minimum-R_∞ MWA of $2m + 1$ terms, exact for the degree $2s - 1$, is

$$u_x = \mu^{2(m-s+1)} \sum_{j=0}^{s-1} (-4)^{-j} \binom{m-s+j}{j} \delta^{2j} y_x \quad ,$$

where the operator μ is defined by

$$\mu f(x) = \tfrac{1}{2}[f(x + \tfrac{1}{2}) + f(x - \tfrac{1}{2})] \quad ,$$

so that $\mu^2 = 1 + \frac{1}{4}\delta^2$. The characteristic function is

$$\phi(t) = (\cos \tfrac{1}{2}t)^{m-s+1} \sum_{j=0}^{s-1} \binom{m-s+j}{j} \sin^{2j} \tfrac{1}{2}t \quad ,$$

which is nonnegative, with a single zero of order $m - s + 1$ at $t = \pi$.

By a tour de force it is possible to show that, for an MWA that is not a smoothing formula in the strict sense, but whose characteristic function satisfies (10.8), the natural extension does nevertheless "minimize something." For the given MWA, let $-\rho$ denote the minimum value of $\phi(t)$, and let γ be chosen so that $0 < \gamma \leq (1 + \rho)^{-1}$. Then $1 - \gamma(1 + \rho) \geq 0$. Let a modified MWA, $\tilde{u}_x$ be obtained by taking

$$\tilde{u}_x = \gamma u_x + (1 - \gamma) y_x \quad . \tag{11.11}$$

Clearly this is an MWA of the form (1.2),

$$\tilde{u}_x = \sum_{j=-m}^{m} \tilde{c}_j \, y_{x-j} \quad ,$$

with $\tilde{c}_0 = \gamma c_0 + 1 - \gamma$ and $\tilde{c}_j = \gamma c_j$ for $j \neq 0$. The modified MWA is a smoothing formula in the strict sense, and its graduation matrix is $\tilde{G} = I - K^T \tilde{D} K$, with $\tilde{D} = \gamma D$.

The modified graduation minimizes the quantity

$$(\tilde{u} - y)^T (\tilde{u} - y) + (K\tilde{u})^T \tilde{H} K\tilde{u} \quad , \tag{11.12}$$

where

$$\tilde{H} = (\tilde{D}^{-1} - KK^T)^{-1} \tag{11.13}$$

is positive definite. Using (11.11) and (11.13) to express (11.12) in terms of the original graduated values, we find that the quantity minimized is

$$(u - y)^T (u - y) + [u + (\gamma^{-1} - 1)y]^T K^T \hat{Q}^{-1} K[u + (\gamma^{-1} - 1)y] \quad . \tag{11.14}$$

where

$$\hat{Q} = Y^{-1} D^{-1} - KK^T$$

is positive definite. Thus, the total smoothing operation including the "tails," based on an MWA that is a smoothing formula, but not in the strict sense, does in fact minimize the expression (11.14). Using statistical terminology, this expression may therefore be regarded as a "loss function," but in that context is difficult to interpret and justify in practical terms.

ACKNOWLEDGMENTS

This paper has benefited from discussions with many persons, but I wish to thank especially D. R. Schuette for invaluable help with the computations, and J. M. Hoem and W. F. Trench for their careful reading of the manuscript, which led to significant improvements. I am solely responsible for any errors that may be found.

REFERENCES

Aitken, A. C. (1926), "The Accurate Solution of the Difference Equation Involved in Whittaker's Method of Graduation, and its Practical Application," Transactions of the Faculty of Actuaries, 11, 31-9.

Andrews, George H., and Nesbitt, Cecil J. (1965), "Periodograms of Graduation Operators," Transactions of the Society of Actuaries, 17, 1-27.

Benjamin, B., and Haycocks, H. W. (1970), The Analysis of Mortality and Other Actuarial Statistics, Cambridge: Cambridge University Press.

De Forest, Erastus L. (1873), "On Some Methods of Interpolation Applicable to the Graduation of Irregular Series, Such as Tables of Mortality, &c., &c.," Smithsonian Report 1871, 275-339.

_____ (1875), "Additions to a Memoir on Methods of Interpolation Applicable to the Graduation of Irregular Series," Smithsonian Report 1873, 319-49.

_____ (1876), Additions to a Memoir on Methods of Interpolation Applicable to the Graduation or Adjustment of Irregular Series of Observed Numbers, New Haven: Tuttle, Morehouse, and Taylor Co.

_____ (1877), "On Adjustment Formulas," The Analyst, 4, 79-86 and 107-13.

Dresden, Arnold (1942), "On the Iteration of Linear Homogeneous Transformations," Bulletin of the American Mathematical Society, 48, 577-9.

Elphinstone, M. D. W. (1951), "Summation and Some Other Methods of Graduation -- the Foundations of Theory," Transactions of the Faculty of Actuaries, 20, 15-77.

Greville, T. N. E. (1947), "Actuarial Note: Adjusted Average Graduation Formulas of Maximum Smoothness," Record of the American Institute of Actuaries, 36, 249-64.

_____ (1948), "Actuarial Note: Tables of Coefficients in Adjusted Average Graduation Formulas of Maximum Smoothness," Record of the American Institute of Actuaries, 37, 11-30.

_____ (1957), "On Smoothing a Finite Table: a Matrix Approach," Journal of the Society for Industrial and Applied Mathematics, 5, 137-54.

_____ (1966), "On Stability of Linear Smoothing Formulas," SIAM Journal on Numerical Analysis, 3, 157-70.

_____ (1974a), Part 5 Study Notes, Graduation (1974 edition), Chicago: Education and Examination Committee of the Society of Actuaries.

_____ (1974b), "On a Problem of E. L. De Forest in Iterated Smoothing," SIAM Journal on Mathematical Analysis, 5, 376-98.

Hannan, E. J. (1970), Multiple Time Series, New York: John Wiley and Sons.

Hardy, George F. (1909), The Theory of the Construction of Tables of Mortality and of Similar Statistical Tables in Use by the Actuary, London: Institute of Actuaries.

Henderson, Robert (1916), "Graduation by Adjusted Average," Transactions of the Actuarial Society of America, 17, 43-48.

_____ (1924), "A New Method of Graduation," Transactions of the Actuarial Society of America, 25, 29-40.

_____ (1938), Mathematical Theory of Graduation, New York: Actuarial Society of America.

Macaulay, Frederick R. (1931), The Smoothing of Time Series, New York: National Bureau of Economic Research.

Maclean, Joseph B. (1913), "Graduation by the Summation Method. Some Elementary Notes," Transactions of the Actuarial Society of America, 14, 256-76.

Miller, Morton D. (1946), Elements of Graduation, New York and Chicago: Actuarial Society of America and American Institute of Actuaries.

Noble, Ben (1969), Applied Linear Algebra, Englewood Cliffs, N. J.: Prentice-Hall, Inc.

Oldenburger, Rufus (1940), "Infinite Powers of Matrices and Characteristic Roots," Duke Mathematical Journal, 6, 357-61.

Schiaparelli, G. V. (1866), "Sul Modo di Ricavare la Vera Espressione delle Legge della Natura dalle Curve Empiriche," appendix to Effemeredi Astronomiche di Milano per l'Anno 1867, Milan.

Schoenberg, I. J. (1946). "Contributions to the Problem of Approximation of Equidistant Data by Analytic Functions," Quarterly of Applied Mathematics, 4, 45-99 and 112-41.

_____ (1948), "Some Analytical Aspects of the Problem of Smoothing," in Studies and Essays Presented to R. Courant on his 60th Birthday, New York: Interscience Publishers.

_____ (1953), "On Smootning Operations and Their Generating Functions," Bulletin of the American Mathematical Society, 59, 199-230.

Sheppard, W. F. (1913), "Reduction of Errors by Means of Negligible Differences," Proceedings of the Fifth International Congress of Mathematicians, 2, 348-84.

Shiskin, Julius, Young, Allan H., and Musgrave, John C. (1967), The X-11 Variant of the Census Method II Seasonal Adjustment Program, Washington: U. S. Department of Commerce, Bureau of the Census.

Trench, William F. (1974), "Inversion of Toeplitz Band Matrices," Mathematics of Computation, 28, 1089-95.

Vaughan, Hubert (1933), "Summation Formulas of Graduation with a Special Type of Operator," Journal of the Institute of Actuaries, 64, 428-48.

Whittaker, E. T. (1923), "On a New Method of Graduation," Proceedings of the Edinburgh Mathematical Society, 41, 63-75.

Wiener, Norbert (1949), Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications, New York: John Wiley and Sons.

Wolfenden, Hugh H. (1925), "On the Development of Formulae for Graduation by Linear Compounding, with Special Reference to the Work of Erastus L. De Forest," Transactions of the Actuarial Society of America, 26, 81-121.

_____ (1942), The Fundamental Principles of Mathematical Statistics, Toronto: Macmillan Co. of Canada, Ltd.

NUMERICAL CALCULATION OF THE SOLUTION OF THE VISCOELASTIC DEFORMATION OF AN INFINITE FLOATING ICE PLATE UNDER A CIRCULAR LOAD

Shunsuke Takagi
U.S. Army Cold Regions Research and Engineering Laboratory
Hanover, New Hampshire 03755

The unresolved problem submitted to the 21st (1975) Conference of Army Mathematicians (Ref. 1) was completely answered with the aid of Professor Ben Noble, Director, Mathematics Research Center, University of Wisconsin. The theoretical insight gained in the above study enabled me to devise a simple numerical integral method as shown below.

PRINCIPLE OF OUR NUMERICAL INTEGRATION

Stated in general terms, our problem is to numerically integrate the integral containing a product of Bessel functions,

$$I = \int_0^\infty \Phi(\beta) J_1(\beta A) J_0(\beta R) d\beta, \tag{1}$$

where A and R are positive numbers; $\Phi(\beta)$ is finite in the range of integration, and asymptotically

$$\Phi(\beta) \sim a\beta^{-n}, \tag{2}$$

in which a is constant. The value of n in our formulas was ≥ 4. We shall change (1) to a finite integral. Let

$$[\text{Abs } I]_N^\infty = \int_N^\infty |\Phi(\beta) J_1(\beta A) J_0(\beta R)| d\beta.$$

Choosing N so large that $\Phi(\beta)$, $J_1(\beta A)$, and $J_0(\beta R)$ may be replaced with respective asymptotic expressions, the upper bound of the right-hand side can be estimated:

$$[\text{Abs } I]_N^\infty = \int_N^\infty |\Phi(\beta) J_1(\beta A) J_0(\beta R)| d\beta$$

$$\leq \int_N^\infty \frac{a}{\beta^n} \sqrt{\frac{2}{\pi\beta A}} \sqrt{\frac{2}{\pi\beta R}}\, d\beta$$

$$= \frac{2a}{\sqrt{RA}} \frac{1}{nN^n} \frac{1}{\pi}$$

$$< \epsilon = 10^{-5}.$$

Choose N to make

$$N^n > \frac{2a}{\sqrt{RA}} \frac{1}{n\varepsilon} \frac{1}{\pi} .$$

Then we can convert I to a finite integral,

$$I \fallingdotseq \int_0^N \Phi(\beta) J_1(\beta A) J_0(\beta R) d\beta. \tag{3}$$

We shall call $[\text{Abs } I]_N^\infty$ the absolute remainder.

This method worked well, because $n \geq 4$. If n is close to zero, this method does not work. A method for the general case is developed in Appendix B.

FIELD MEASUREMENT

Frankenstein [2] measured the deflection of a lake ice. He placed a tank on a frozen lake, filled it with water, and measured the deflection with the measuring rods at the distances as shown in Fig. 1. One of the results of his measurement is shown in Fig. 2.

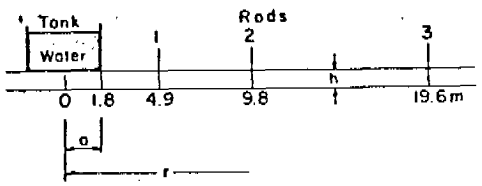


Fig. 1. Field Measurement

THEORETICAL CURVE

We assumed the lake ice to be a linear viscoelastic material of the Maxwell-Voight type four-element model, as shown in Fig. 3. This model is known to satisfactorily represent the creep of ice [Jellinek and Brill 3].

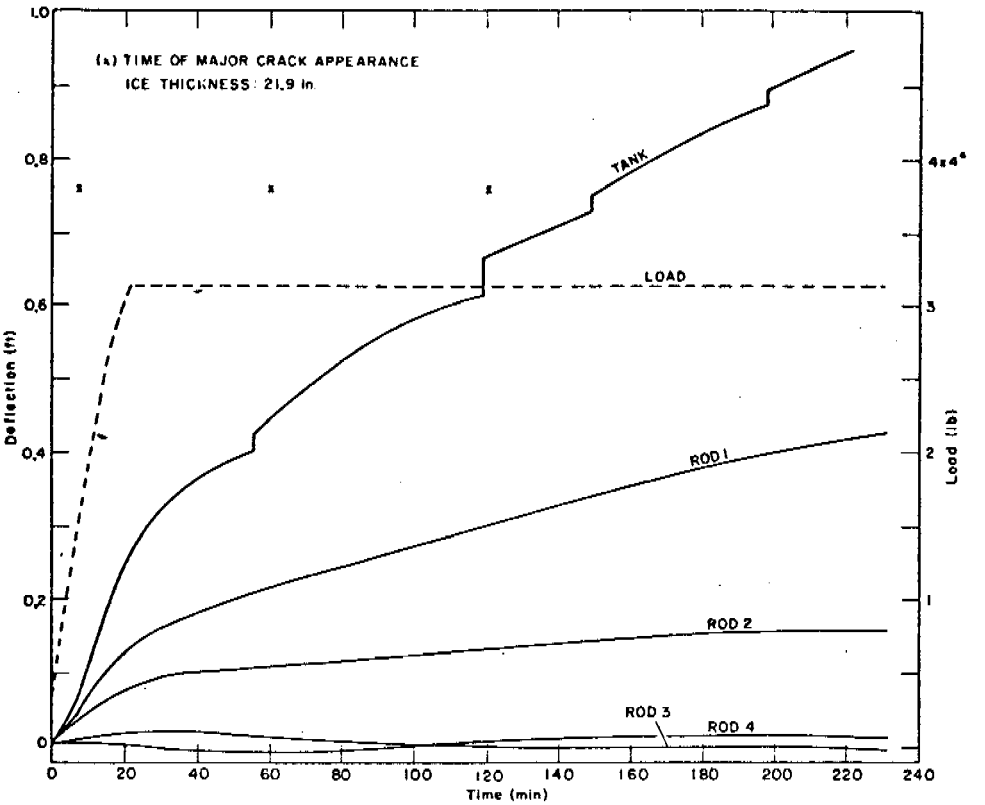


Fig. 2. A result of Frankenstein's measurement (Ref. 2, Test 8).

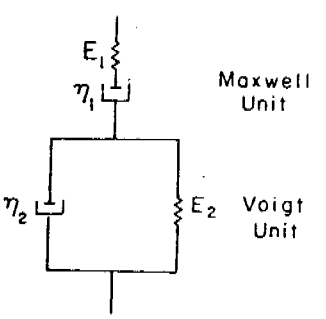


Fig. 3. Maxwell-Voight four-element model.

First we solved the case of step loading, i.e., under the assumption that the circular load P of radius a was applied at time t = 0 and kept constant for t>0. The process of solution is shown in Appendix A. For our curve-fitting we need only the deflection solution,

$$\omega = \frac{P}{\pi A \rho \ell^2} \int_0^\infty \left\{ 1 + \frac{\beta^4(\tau-\alpha_2)}{\sqrt{DESC}} e^{-\alpha_2 T} - \frac{\beta^4(\tau-\alpha_1)}{\sqrt{DESC}} e^{-\alpha_1 T} \right\} J_1(\beta A) J_0(\beta R) d\beta, \qquad (4)$$

in which the functions of the integral variable β are

$$DESC = (\tau\beta^4+1+\tau)^2 - 4\tau(\beta^4+E)$$

$$\left.\begin{matrix}\alpha_1 \\ \alpha_2\end{matrix}\right\} = \frac{\tau\beta^4+1+\tau \mp \sqrt{DESC}}{2(\beta^4+E)} .$$

The radius a of the loading circle and the radial coordinate r are nondimensionalized to A and R,

$$A = \frac{a}{\ell}$$

$$R = \frac{r}{\ell} ,$$

by use of the characteristic length

$$\ell^4 = \frac{E_0 h^3}{12\rho(1-\nu)} ,$$

where E_0 is defined by

$$\frac{1}{E_0} = \frac{1}{E_1} + \frac{1}{E_2} ,$$

and h, ρ, and ν are the thickness, density and the Poisson ratio, respectively, of the ice plate. Time t is nondimensionalized to T

$$T = \frac{E_0 t}{\eta_1} .$$

Nondimensional material constants

$$\tau = \frac{\eta_1 E}{\eta_2 E_0} \quad \text{and} \quad E = \frac{E_0}{E_1}$$

are used.

We list in the following the asymptotic expansions of the non-Bessel factors, $\Phi(\beta)$, contained in the integrands of the deflection solution (4), and the stress solutions (A.37) and (A.38) in Appendix A:

$$
\begin{aligned}
&\alpha_1 \sim \beta^{-4} \\
&\alpha_2 \sim \tau(1+\beta^{-4}) \\
&e^{-\alpha_1 T} \sim 1-T\beta^{-4} \\
&e^{-\alpha_2 T} \sim e^{-\tau T} \\
&\frac{\tau-\alpha_1}{\sqrt{DESC}} \sim \beta^{-4}-\beta^{-8} \\
&\frac{\tau-\alpha_2}{\sqrt{DESC}} \sim -(1-E)\beta^{-8}.
\end{aligned}
\tag{5}
$$

The actual loading was the ramp/steady loading, as shown below:

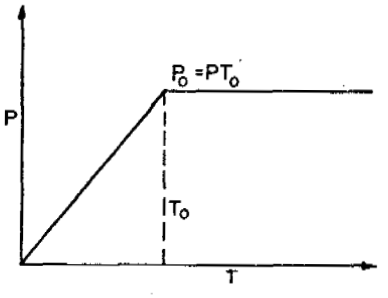


Fig. 4. Definition of the ramp/steady loading.

To calculate the deflection under the ramp/steady loading, define the influence function $\omega_0(T)$ by letting P=1 in (4),

$$
\omega_0(T) = \frac{1}{\pi A \rho \ell^2} \int_0^\infty \left[1 + \frac{\beta^4(\tau-\alpha_2)}{\sqrt{DESC}} e^{-\alpha_2 T} - \frac{\beta^4(\tau-\alpha_1)}{\sqrt{DESC}} e^{-\alpha_1 T}\right] J_1(\beta A) J_0(\beta R) d\beta. \tag{6}
$$

Then, deflection $\omega(T)$ for $0 \leq T \leq T_0$ is given by

$$\omega(T) = \int_0^T \omega_0(T-\lambda)\dot{P}d\lambda \tag{7}$$

and for $T_0 \leq T$ by

$$\omega(T) = \int_0^{T_0} \omega_0(T-\lambda)\dot{P}d\lambda \ , \tag{8}$$

where

$$\dot{P} = \frac{P_0}{T_0} \ .$$

Substituting (6) into (7) and integrating with regard to λ, we get the deflection $\omega(T)$ for $0 \leq T \leq T_0$:

$$\omega(T) = \frac{\dot{P}}{\pi A \rho \ell^2}(U_1 - U_2 + U_3), \tag{9}$$

where

$$U_1 = \int_0^\infty \frac{1}{\alpha_1}(e^{-\alpha_1 T} - 1 + \alpha_1 T) J_1(\beta A) J_0(\beta R) d\beta$$

$$U_2 = \int_0^\infty \left[\frac{\beta^4(\tau - \alpha_1)}{\sqrt{DESC}} - 1 \right] \frac{1}{\alpha_1}(1 - e^{-\alpha_1 T}) J_1(\beta A) J_0(\beta R) d\beta$$

$$U_3 = \int_0^\infty \frac{\beta^4(\tau - \alpha_2)}{\sqrt{DESC}} \frac{1}{\alpha_2}(1 - e^{-\alpha_2 T}) J_1(\beta A) J_0(\beta R) d\beta .$$

The absolute remainders are as follows:

$$[\text{Abs } U_1]_N^\infty < \frac{T^2}{4\pi\sqrt{AR}} N^{-4}$$

$$[\text{Abs } U_2]_N^{\infty} < \frac{T}{2\pi\sqrt{AR}} N^{-4}$$

$$[\text{Abs } U_3]_N^{\infty} < \frac{1-E}{2\pi\sqrt{AR}} \frac{1}{\tau}(1-e^{-\tau T})N^{-4}.$$

Substituting (6) into (8) and integrating with regard to λ, we get the deflection $\omega(T)$ for $T_0 \leq T$:

$$\omega(T) = \frac{P}{\pi A \rho \ell^2}(I_1+I_2+I_3), \tag{10}$$

where

$$I_1 = \int_0^{\infty} \left\{1 - \frac{\beta^4(\tau-\alpha_1)}{\sqrt{DESC}} \frac{1}{\alpha_1 T_0}(e^{\alpha_1 T_0}-1)\right\} J_1(\beta A)J_0(\beta R)d\beta$$

$$I_2 = \int_0^{\infty} \frac{\beta^4(\tau-\alpha_2)}{\sqrt{DESC}}(1-e^{-\alpha_1 T}) \frac{1}{\alpha_1 T_0}(e^{\alpha_1 T_0}-1)J_1(\beta A)J_0(\beta R)d\beta$$

$$I_3 = \int_0^{\infty} \frac{\beta^4(\tau-\alpha_2)}{\sqrt{DESC}}(1-e^{-\alpha_1 T})J_1(\beta A)J_0(\beta R)d\beta.$$

The absolute remainders are as follows:

$$[\text{Abs } I_1]_N^{\infty} < (2\pi\sqrt{AR})^{-1}N^{-4}$$

$$[\text{Abs } I_2]_N^{\infty} < \frac{T}{2\pi\sqrt{AR}} N^{-4}$$

$$[\text{Abs } I_3]_N^{\infty} < \frac{1-E}{2\pi\sqrt{AR}} \frac{e^{\tau T_0}-1}{\tau T_0} e^{-\tau T} N^{-4}$$

CURVE FITTING

The material constants found by the curve fitting are shown in Table 1. They vary with the location of the measurement.

TABLE 1

	Tank	Rod 1	Rod 2	Rod 3
Distance	1.8m	4.9m	9.8m	19.6m
λ	2	6.5	20	5
E	0.0005	0.007	0.05	0.1
$E_0(kg/_m2)$	1.766×10^6	9.813×10^7	6.869×10^8	9.813×10^{11}
η/E_0(sec)	2.815×10^6	4.896×10^5	1.101×10^6	2.448×10^6
TE(ramp)(m)	4.718×10^{-3}	5.812×10^{-3}	2.063×10^{-3}	
TE(flat)(m)	4.727×10^{-3}	2.730×10^{-3}	2.884×10^{-4}	2.750×10^{-3}

Material constants found by using the time-lapse curves of Frankenstein's concentrated load test (Ref. 2, Test 8)

To show the significance of the material-constant variation with the measurement locations, we chose the material constants determined at Rod 1 of Frankenstein's concentrated-load time-lapse curve, and computed the deflections at the other measurement locations. Fig. 5 shows the comparison of the computed curves and the measured data. The left and right columns show the ramp and steady portions of the deflection curves, respectively. They are designated by (r) and (s) respectively.

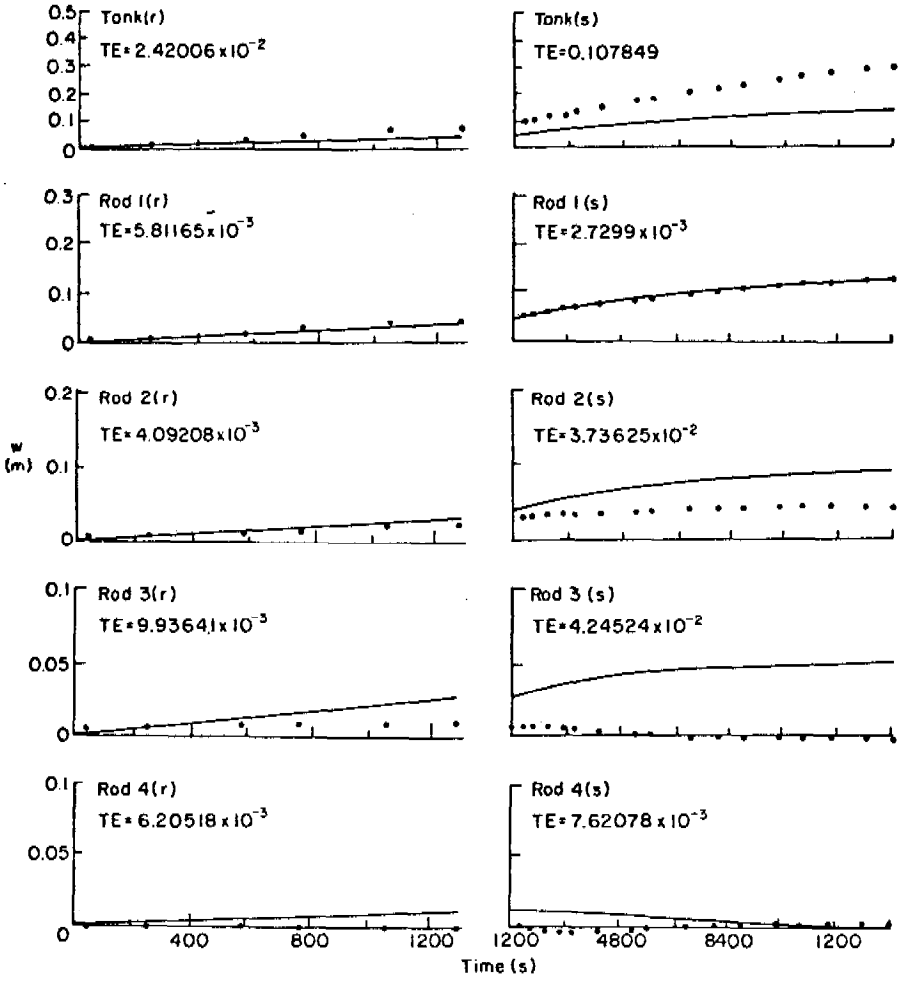


Fig. 5. Comparison of the calculated curves and measured points of Frankenstein's concentrated load test. Material constants are determined by use of the measurement at Rod 1.

To express the degree of curve fitting we devised the trapezoidal error (TE). In Fig. 6, let a, A and b, B be two pairs of measured and computed deflections at two consecutive times t_1 and t_2, respectively.

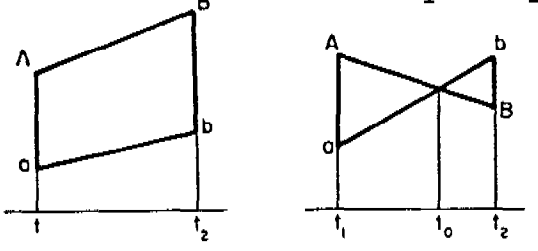


Fig. 6. Elements of TE

We squared A-a and B-b and, in case of the left figure where the errors are of the same sign, computed the area of the trapezoid whose bases are $(A-a)^2$ and $(B-b)^2$ and the height t_2-t_1. In case of the right figure where the errors change sign, we calculated the sum of the areas of the two triangles whose bases are $(A-a)^2$ and $(b-B)^2$ and heights t_0-t_1 and t_2-t_0 respectively, where t_0 is the abscissa of the intersection. Let S be the area thus computed. Then the TE is defined by

$$TE = \sqrt{\frac{\Sigma S}{T}} ,$$

where the summation is over all the intervals and T the sum of the abscissa intervals.

The TE indicates a sort of absolute maximum error. Its unit is m. If the deflections are of ordinary magnitude, the TE of order 10^{-3} and 10^{-2} means a good and tolerable fit, respectively. If the deflections are very small, as in the case of Rod 4, the smallness of the value of TE does not mean much. We did not list the computed values at Rod 4 in Table 1.

We evaluated the TE for all the possible cases. They are shown in Fig. 7. The abscissa is the distance from the center of the load. The measurement locations are noted on the abscissa axis. The circled points are those whose material constants are used to compute a set of TE. The sets of TE thus computed are connected with solid lines and labeled with the appelations of the circled measurement locations.

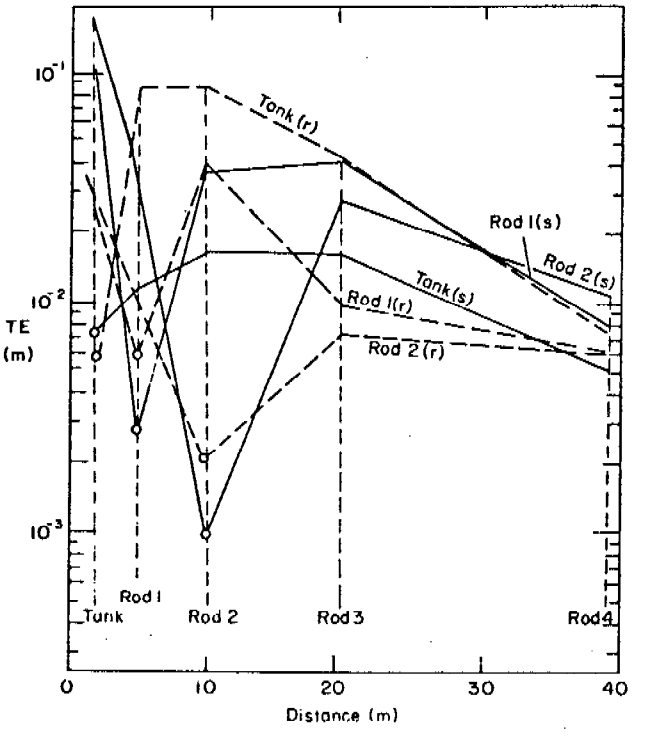


Fig. 7. The TE of Frankenstein's distributed load test No. 8, Ref. 2.

We carried out the curve fitting on the assumption of the linear viscoelasticity, i.e., that the material constants are kept absolutely constant during the increase of both the stress and the deformation. However, Dr. Andrew Assur, an ice mechanics expert, CRREL, notified me that the real material constants are nonlinear, i.e., that they change with the stress and the deformation. He told me that the variations shown in this paper are reasonable from the nonlinear viewpoint. At the writing of this paper we could not complete examination from the nonlinear viewpoint.

ASYMPTOTIC DEFLECTION

We shall show in the following that only one material constant is contained in the asymptotic formulas. Therefore, curve fitting must be performed for sufficiently small times.

Referring to the asymptotic relationships in (5). one finds that, when T is large, both the step-loading formulation (6) and the ramp/steady loading formulation (10) reduce to

$$\omega = \frac{P}{\pi A \rho \ell^2} \frac{K}{A} , \tag{11}$$

where

$$\frac{1}{A}K = \int_0^\infty (1-e^{-T\beta^{-4}}) J_1(\beta A) J_0(\beta R) d\beta . \tag{12}$$

It is considered in this derivation that $t>0$, and that only large values of β are effective in the integration. Letting

$$x = \beta A,$$

(12) becomes

$$K = \int_0^\infty (1-e^{-T_A x^{-4}}) J_1(x) J_0(\frac{R}{A}x) dx, \tag{13}$$

where

$$T_A = TA^4 \tag{14}$$

$$= \frac{ta^4}{h^3} \frac{12\rho(1-\nu)}{n_1} . \tag{15}$$

Thus all the material constants are lumped into the second factor of (15). The stress formulas, although not mentioned here, can be similarly transformed.

As shown in the Appendix B, (13) cannot be analytically integrated. It must be numerically integrated. To effect the numerical integration, we have chosen the non-Bessel factor in (13) to become zero at $x = \infty$. The absolute remainder is estimated:

$$[\text{Abs}_{-} K]_N^{\infty} < \frac{T}{2\pi\sqrt{AR}} N^{-4}.$$

Graphs of integral K for the values of R/A = 0.2 and 2.0 are shown in Fig. 8. When $T_A = \infty$, the non-Bessel factor becomes equal to one. At this limit, therefore, K = 1 when R < A, and K = 0 when R > A. As shown in the graphs, this limit is almost reached when $T_A > 1000$.

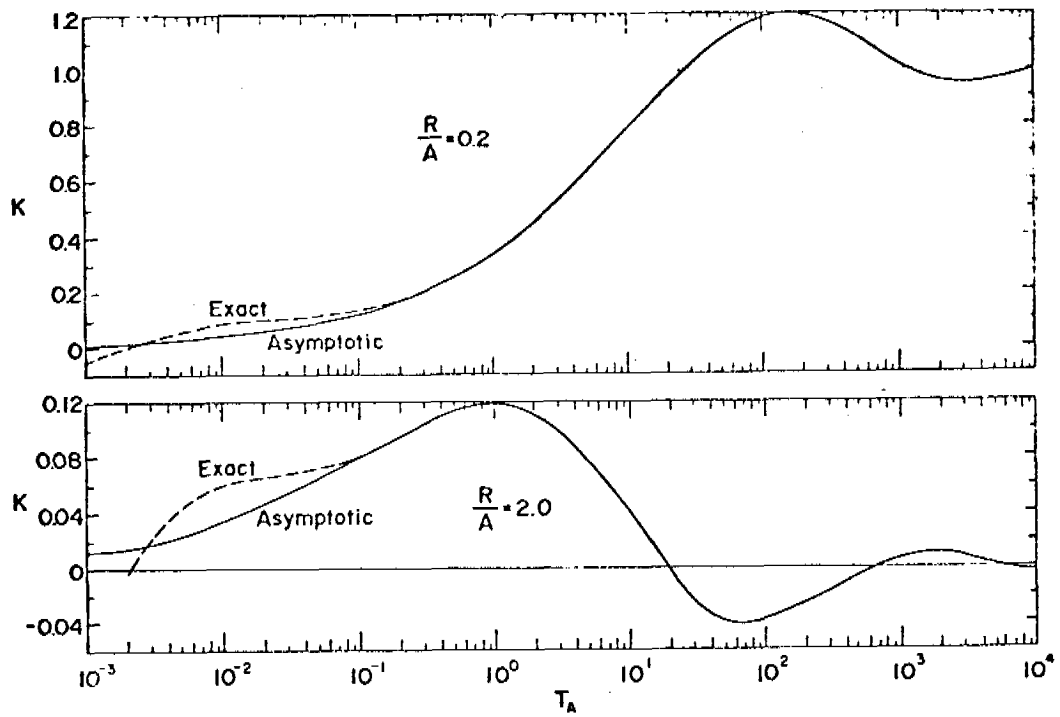


Fig. 8. Graphs of asymptotic integral K in (6.4).

Exact integral K was formulated for the ramp/steady loading, and evaluated by use of a set of constants: $T_0 = 6\text{x}10^3$ sec, $\tau = 10$, $E = 1/6$, $\eta_1/E_0 = 6.12\text{x}10^4$ sec = 17 hrs., $E_0 = 7\text{x}10^8 \text{kg/m}^2$, and $\nu = 0.5$. These constants give $\ell = 29.31\text{m}$ and $T_A = t(2.48 \times 10^{-3}\text{day}^{-1})$. As shown in Fig. 8, the asymptotoc integral K is very close to the exact integral in the range $T_A > 0.1$. The above constants were the values we used at the outset of the numerical computation for the rough estimate. We did not use other sets of constants to evaluate the integral K.

APPENDIX A

Analytical solution for the step loading

The Problem

We shall consider the viscoelastic ice plate floating on water extending horizontally to infinity. (Historical background of the analytical study is mentioned in Appendix C.) We shall use the Maxwell-Voigt type four-element model (Fig. 3) to describe the viscoelastic deformation of ice.

Using the notation of Fig. 3, one can show that this model gives the stress-strain relationship which we show in an operator form,

$$\varepsilon = \left[\frac{1}{E_1} + \frac{1}{\eta_1 \frac{\partial}{\partial t}} + \frac{1}{E_2 + \eta_2 \frac{\partial}{\partial t}} \right] \sigma \quad , \tag{A.1}$$

where t is time. To extend the one-dimensional relationship (A.1) to the three-dimensional relationship, we assume, as explained by Flügge [4] that ε and σ are deviatoric, and relate them by

$$\sigma = 2G\varepsilon \quad ,$$

where G is the rigidity modulus relative to the three-dimensional deformation. Using (A.1), 2G is given as an operator,

$$\frac{1}{2G} = \frac{1}{E_1} + \frac{1}{\eta_1 \frac{\partial}{\partial t}} + \frac{1}{E_2 + \eta_2 \frac{\partial}{\partial t}} \quad . \tag{A.2}$$

The differential equation describing the deflection w of an elastic plate on water is

$$D\nabla^4 w + \rho w = q \quad , \tag{A.3}$$

where ∇^4 is the biharmonic operator

$$\nabla^4 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^2 \quad ,$$

ρ the density of water, q the load per unit area, D the flexural rigidity defined by

$$D = 2Gh^3/[\,12(1-\nu)\,]\ , \tag{A.4}$$

in which h is the thickness of the ice plate, and ν Poisson's ratio. Substitute 2G from (A.2) into (A.4); substitute D thus found into (A.3); then one finds the differential equation governing the viscoelastic deformation of a floating ice plate. We shall show this equation later in the nondimensional form.

We assume the load q to be step loading at t = 0 distributed uniformly over the circle of radius a with the center at origin. Let r be the radial distance from origin. Then

$$\begin{aligned} q &= q_o U(t) \qquad \text{for} \quad 0 \leqq r < a \\ &= 0 \qquad\quad \text{for} \quad a < r\ , \end{aligned} \tag{A.5}$$

where U(t) is the step function. Our problem is axisymmetric, and the biharmonic operator ∇^4 reduces to

$$\nabla^4 = \left(\frac{\partial}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} \right)^2 .$$

We shall nondimensionalize our differential equation. We define the characteristic length ℓ by

$$\ell^4 = E_o h^3/[12\rho(1-\nu)]\ , \tag{A.6}$$

where

$$\frac{1}{E_o} = \frac{1}{E_1} + \frac{1}{E_2}\ . \tag{A.7}$$

We have chosen E_o, rather than E_1 or E_2, to define ℓ because E_o is related to the secondary creep [Nevel 5], which is the main interest in our field observation.

Let D_1 be defined by

$$D_1 = D/(\rho \ell^4)\ . \tag{A.8}$$

Use of (A.4) and (A.8) to

$$D_1 = 2D/E_o \quad . \tag{A.9}$$

Substituting G from (A.2) into (A.9) yields

$$D_1 = 1/\{\frac{E_o}{E_1} + \frac{E_o}{\eta_1 \frac{\partial}{\partial t}} + \frac{E_o}{E_1 + \eta_2 \frac{\partial}{\partial t}}\} \quad . \tag{A.10}$$

Choose nondimensional time T,

$$T = E_o t/\eta_1 \quad , \tag{A.11}$$

and a parameter τ,

$$\tau = \eta_1 E_2/(\eta_2 E_o) \quad . \tag{A.12}$$

Then (A.10) becomes

$$D_1 = 1/\{E + \frac{1}{\frac{\partial}{\partial T}} + \frac{\eta_1/\eta_2}{\tau + \frac{\partial}{\partial T}}\} \quad , \tag{A.13}$$

where

$$E = E_o/E_1 \quad . \tag{A.14}$$

It is noted that

$$0 \leq E \leq 1 \quad . \tag{A.15}$$

Clearing the denominator, (A.13) becomes

$$D_1 = \frac{\partial}{\partial T}(\frac{\partial}{\partial T} + \tau)/\{E\frac{\partial^2}{\partial T^2} + (1 + \tau)\frac{\partial}{\partial T} + \tau\} \quad , \tag{A.16}$$

where use is made of the relation

$$E\tau + \eta_1/\eta_2 = \tau \quad ,$$

which one can prove by use of (A.12), (A.14), and (A.7).

Define the nondimensional length R by

$$R = r/\ell \quad . \tag{A.17}$$

Replace D in (A.3) with D in (A.8) and (A.3) becomes

$$D_1 \nabla_R^4 w + w = q/\rho \quad , \tag{A.18}$$

where

$$\nabla_R^4 = \left(\frac{d^2}{dR^2} + \frac{1}{R}\frac{d}{dR}\right)^2 \quad .$$

With D_1 given by (A.16), (A.18) is the differential equation to be solved.

The Solution

Denote the Hankel transform of f(R) by $\tilde{f}(\beta)$,

$$\tilde{f}(\beta) = \int_0^\infty f(R) J_o(\beta R) R dR \quad , \tag{A.19}$$

and the two-sided Laplace transform [Van del Pol and Bremers 6] of g(T) by $\bar{g}(s)$,

$$\bar{g}(s) = s\int_{-\infty}^{\infty} g(T) e^{-sT} dT \quad . \tag{A.20}$$

Denote the inverse of (A.20) by

$$g(T) = L^{-1}(\bar{g}(s)) \quad . \tag{A.21}$$

Applying these two transforms, (A.18) becomes

$$\bar{D}_1 \beta^4 \tilde{\bar{w}} + \tilde{\bar{w}} = \tilde{\bar{q}}/\rho \quad ,$$

where

$$\bar{D}_1 = \frac{s(s + \tau)}{Es^2 + (1 + \tau)s + \tau} \quad . \tag{A.22}$$

Applying the two transforms to q in (A.5), one gets

$$(1/\rho)\tilde{\bar{q}} = (P/(\pi A\rho\ell^2))(1/\beta)J_1(\beta A) \quad , \tag{A.23}$$

where

$$P = \pi a^2 q \tag{A.24}$$

and

$$A = a/\ell \quad . \tag{A.25}$$

Thus the transformed solution is given by

$$\tilde{\bar{w}} = \frac{P}{\pi A\rho\ell^2}\frac{1}{\beta(1 + \bar{D}_1\beta^4)}J_1(\beta A) \quad .$$

Performing the Hankel inverse,

$$\bar{w} = \frac{P}{\pi A\rho\ell^2}\int_0^\infty \frac{1}{1 + \bar{D}_1\beta^4}J_1(\beta A)J_o(\beta R)d\beta \quad . \tag{A.26}$$

Performing the Laplace inverse,

$$w = \frac{1}{\pi A\rho\ell^2}\int_0^\infty L^{-1}\left(\frac{1}{1 + \bar{D}_1\beta^4}\right)J_1(\beta A)J_0(\beta R)d\beta \quad . \tag{A.27}$$

To find $L^{-1}(1/(1 + \bar{D}_1\beta^4))$, compute the partial fraction,

$$\frac{1}{s}\frac{1}{1 + \bar{D}_1\beta^4} = \frac{1}{s} + \frac{\beta^4(\tau - \alpha_2)}{\sqrt{DESC}}\frac{1}{s + \alpha_2} - \frac{\beta^4(\tau - \alpha_1)}{\sqrt{DESC}}\frac{1}{s + \alpha_1} \quad ,$$

where $-\alpha_1$ and $-\alpha_2$ are the roots of the quadratic equation

$$(E + \beta^4)s^2 + (\tau\beta^4 + 1 + \tau)s + \tau = 0 \quad . \tag{A.28}$$

They are given by

$$\left.\begin{matrix}\alpha_1\\ \alpha_2\end{matrix}\right\} = \frac{\tau\beta^4 + 1 + \tau \mp \sqrt{DESC}}{2(\beta^4 + E)} \quad , \tag{A.29}$$

where

$$\text{DESC} = (\tau\beta^4 + 1 + \tau)^2 - 4\tau(\beta^4 + E) \quad , \tag{A.30}$$

which transform to

$$= [\tau(\beta^4 + 1) - 1]^2 + 4\tau(1 - E) \quad . \tag{A.31}$$

From (A.31), it is clear that

$$\text{DESC} > 0 \quad . \tag{A.32}$$

The roots α_1 and α_2 are therefore always real. Moreover, inspection of (A.29) and (A.30) shows that both α_1 and α_2 are always positive.

Thus one finds that

$$L^{-1}\left(\frac{1}{1 + \tilde{D}_1\beta^4}\right) = 1 + \frac{\beta^4(\tau - \alpha_2)}{\sqrt{\text{DESC}}} e^{-\alpha_2 T} - \frac{\beta^4(\tau - \alpha_1)}{\sqrt{\text{DESC}}} e^{-\alpha_1 T} \quad . \tag{A.33}$$

Substituting (A.33) into (A.27), solution w is found:

$$w = \frac{P}{\pi A \rho \ell^2} \int_0^\infty \{1 + \frac{\beta^4(\tau - \alpha_2)}{\sqrt{\text{DESC}}} e^{-\alpha_2 T} - \frac{\beta^4(\tau - \alpha_1)}{\sqrt{\text{DESC}}} e^{-\alpha_1 T}\} J_1(\beta A) J_0(\beta R) d\beta . \tag{A.34}$$

The radial and hoop stresses are given by

$$\sigma_r = -\frac{6D}{h^2}\left(\frac{\partial^2 w}{\partial r^2} + \frac{\nu}{r}\frac{\partial w}{\partial r}\right)$$

and

$$\sigma_\theta = -\frac{6D}{h^2}\left(\frac{1}{r}\frac{\partial w}{\partial r} + \nu\frac{\partial^2 w}{\partial r^2}\right).$$

Changing D to D_1 by use of (A.8) and r to nondimensional R by use of (A.17), they become

$$\sigma_r = -\frac{6\rho\ell^2}{h^2} D_1\left(\frac{\partial^2 w}{\partial R^2} + \frac{\nu}{R}\frac{\partial w}{\partial R}\right) ,$$

and

$$\sigma_\theta = -\frac{6\rho\ell^2}{h^2} D_1\left(\frac{1}{R}\frac{\partial w}{\partial R} + \nu\frac{\partial^2 w}{\partial R^2}\right) ,$$

where D_1 is the operator on T given by (A.16). The two-sided Laplace transform yields

$$\bar{\sigma}_r = -\frac{6\rho\ell^2}{h^2}\left(\frac{\partial^2}{\partial R^2} + \frac{\nu}{R}\frac{\partial}{\partial R}\right)\overline{D_1 w} \tag{A.35}$$

and

$$\bar{\sigma}_\theta = -\frac{6\rho\ell^2}{h^2}\left(\frac{1}{R}\frac{\partial}{\partial R} + \nu\frac{\partial^2}{\partial R^2}\right)\overline{D_1 w} , \tag{A.36}$$

where $\overline{D_1 w}$ is the Laplace transform of $\overline{D_1 w}$.

Using (A.26) one gets

$$\overline{D_1 w} = \frac{P}{\pi A\rho\ell^2}\int_0^\infty \frac{\bar{D}_1}{1 + \bar{D}_1\beta^4} J_1(\beta A)J_0(\beta R)d\beta .$$

The Laplace inverse of this is

$$D_1 w = \frac{P}{\pi A\rho\ell^2}\int_0^\infty L^{-1}\left(\frac{\bar{D}_1}{1 + \bar{D}_1\beta^4}\right)J_1(\beta A)J_0(\beta R)d\beta .$$

To find $L^{-1}(\bar{D}_1/(1 + \bar{D}_1\beta^4))$, compute the partial fraction,

$$\frac{1}{s}\frac{\bar{D}_1}{1 + \bar{D}_1\beta^4} = \frac{\tau - \alpha_1}{\sqrt{DESC}}\frac{1}{s + \alpha_1} - \frac{\tau - \alpha_2}{\sqrt{DESC}}\frac{1}{s + \alpha_2} .$$

Thus one finds

$$L^{-1}\left(\frac{\bar{D}_1}{1 + \bar{D}_1\beta^4}\right) = \frac{\tau - \alpha_1}{\sqrt{DESC}} e^{-\alpha_1 T} - \frac{\tau - \alpha_2}{\sqrt{DESC}} e^{-\alpha_2 T} .$$

Thus the inverseof (A.35) is

$$\sigma_r = \frac{6P}{\pi Ah^2}\int_0^\infty J_1(\beta A)\{J_0(\beta R) - \frac{1 - \nu}{\beta R}J_1(\beta R)\}\frac{(\tau - \alpha_1)e^{-\alpha_1 T} - (\tau - \alpha_2)e^{-\alpha_2 T}}{\sqrt{DESC}}\beta^2 d\beta . \tag{A.37}$$

The inverse of (A.36) is

$$\sigma_\theta = \frac{6P}{\pi Ah^2}\int_0^\infty J_1(\beta A)\{\nu J_0(\beta R) + \frac{1-\nu}{\beta R}J_1(\beta R)\}\frac{(\tau-\alpha_1)e^{-\alpha_1 T} - (\tau-\alpha_2)e^{-\alpha_2 T}}{\sqrt{DESC}}\beta^2 d\beta \ . \tag{A.38}$$

Tabulation of σ_r and σ_θ becomes easier if linear combinations of (A.37) and (A.38) that do not contain ν are computed.

APPENDIX B

ANALYTICAL BACKGROUND OF THE NUMERICAL INTEGRATION

B1. The follwing theorem shows the condition under which the integral (1) becomes either discontinuous or continuous at R = A.

Theorem 1. The integral (1) is discontinuous or continuous at R = A when n in (2) is equal or larger than zero, respectively.

Proof. We can rewrite (1) to a one-parameter integral

$$I(\alpha) = \int_0^\infty f(x,\alpha)dx \tag{B.1}$$

by letting $x = \beta A$, i.e. $\alpha = R/A$, where $f(x,\alpha)$ is continuous with regards to x and α. The condition that $I(\alpha)$ is a continuous function of α is that the integral (B.1) converges uniformly with respect to α (c.f. Titchmarch [7] p. 25). The integral (B.1) uniformly converges when $\underline{n} > 0$, but does not when $\underline{n} = 0$.

B2. We shall consider in the following the integral (1) whose non-Bessel factor $\phi(\beta)$ is finite in the range of integration but asymptotically becomes zero on a more general form rather than in the specific form (2).

Let an asymptotic expansion of $\phi(\beta)$ be

$$\phi(\beta) \sim \sum_{n=0}^{m} \phi_n(\beta) \ . \tag{B.2}$$

Rewrite (1) to

$$I = I_0 + \sum_{n=0}^{m} K_n \ , \tag{B.3}$$

where

$$I_o = \int_o^\infty \{\phi(\beta) - \sum_{n=o}^{n_1} \phi_n(\beta)\} J_I(\beta A) J_o(\beta R) d\beta \tag{B.4}$$

and

$$K_n = \int_o^\infty \phi_n(\beta) J_1(\beta A) J_o(\beta R) d\beta. \tag{B.5}$$

We choose such an integer m that makes I_o fast convergent. We choose such a function $\phi_n(\beta)$ that makes (B.5) analytically integrable. The following Theorem is useful for the choice of $\phi_n(\beta)$.

Theorem 2. Let $F(z)$ be an even function of complex variable $z = x + iy$ that becomes zero at $z = \infty$ and possesses only algebraic singularities (pole or branch points) on the upper half plane but no poles on the real axis. Then the integral

$$\int_o^\infty F(x) J_1(ax) J_o(bx) dx,$$

where a and b are positive, transform to the following contour integrals,

$$= \frac{1}{a} F(o) + \frac{1}{2} \int_{-\infty}^{+\infty} F(z) H_1^{(1)}(az) J_o(\beta z) dz, \tag{B.6}$$

when $o<b<a$

$$= \frac{1}{2} \int_{-\infty}^{+\infty} F(z) J_1(az) H_o^{(1)}(bz) dz, \tag{B.7}$$

when $o<a<b$

where $\int_{-\infty}^{+\infty} dz$ means the integral along the contour in Fig. 9, where radius ε is infinitesimal, and the cut is along the negative real axis.

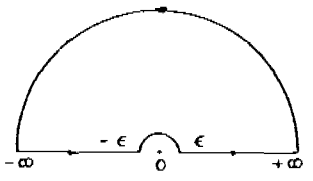


Fig. 9. Contour of integrations in (B.6) and (B.7)

Proof. The proof in ref. [1] must be revised to the following:

Consider the contour integrals

$$I(a>b) = \frac{1}{2}\oint_{-\infty}^{+\infty} F(z)H_1^{(1)}(az)J_0(bz)dz \qquad (a)$$

when $o<b<a$, and

$$I(a<b) = \frac{1}{2}\oint_{-\infty}^{+\infty} F(z)J_1(az)H_0^{(1)}(bz)dz \qquad (b)$$

when $o<a<b$. Use of the asymptotic formulas show that $H_1^{(1)}(az)J_0(bz)$ and $J_1(az)H_0^{(1)}(bz)$ are zero on the infinitely large circle when $o<b<a$ and $o<a<b$, respectively. Therefore we may consider only the contour along the real axis.

$$I(a>b) = \frac{1}{2}\int_{-\infty}^{+\infty} F(z)H_1^{(1)}(az)J_0(bz)dz \qquad (c)$$

and

$$I(a<b) = \frac{1}{2}\int_{-\infty}^{+\infty} F(z)J_1(az)H_0^{(1)}(bz)dz. \qquad (d)$$

Divide the real axis in three parts, $-\infty\sim-\varepsilon$, $-\varepsilon\sim\varepsilon$, and $\varepsilon\sim\infty$. Let $z = -x$ in the region $-\infty\sim-\varepsilon$, and $z = x$ in the region $+\varepsilon\sim\infty$. Neglecting the infinitesimal terms, one gets

$$F(z) = F(o)$$

$$H_1^{(1)}(az) = -2i/(\pi az)$$

and

$$H_0^{(1)}(bz) = [2i/\pi]\,\ell og(bz/2).$$

Then (c) and (d) become

$$I(a>b) = \frac{1}{2}F(o) + \int_0^{\infty} F(x)J_1(ax)J_0(bx)dx \qquad (e)$$

and

$$I(a<b) = \int_0^\infty F(x)J_1(ax)J_0(bx)dx. \qquad (f)$$

Eq (e) and (f) prove (B.6) and (B.7), respectively.

B3. The need of Theorem 2 appears frequently in the mathematical study of the problems of floating ice plate and also the problems of elastic plate on an elastic foundation. A similar integral including only one Bessel function was proved by Dougal as early as in 1903 ([8] p. 138 and 147).

When $t = 0$, our solutions of the viscoelastic plate reduces to the solution of the elastic plate. The elastic solution thus found is composed of the following integrals:

$$M_0 = \int_0^\infty \frac{1}{1+x^4} J_1(ax)J_0(bx)dx \qquad (B.8)$$

$$M_1 = \int_0^\infty \frac{x}{1+x^4} J_1(ax)J_0(bx)dx \qquad (B.9)$$

$$M_2 = \int_0^\infty \frac{x^2}{1+x^4} J_1(ax)J_0(bx)dx, \qquad (B.10)$$

where $s = AE^{1/4}$ and $b = RE^{1/4}$ we can carry out these integrals by direct or indirect application of Theorem 2:

$$M_0 = \mathrm{ber}(b)\mathrm{ker}'(a) - \mathrm{bei}(b)\mathrm{kei}'(a) + a^{-1} \quad \text{when } b \leq a$$

$$= \mathrm{ber}'(a)\mathrm{ker}(b) - \mathrm{bei}'(a)\mathrm{kei}(b) \quad \text{when } a \leq b$$

$$M_1 = \mathrm{ber}'(b)\mathrm{ker}'(a) + \mathrm{bei}'(b)\mathrm{kei}'(a) \quad \text{when } b \leq a$$

$$= \mathrm{ber}'(a)\mathrm{ker}'(b) + \mathrm{bei}'(a)\mathrm{kei}'(b) \quad \text{when } a \leq b$$

$$M_2 = \text{bei}(b)\text{ker}'(a) + \text{ber}(b)\text{kei}'(a) \qquad \text{when } b \leq a$$

$$= \text{bei}'(a)\text{ker}(b) + \text{ber}'(a)\text{kei}(b). \qquad \text{when } a \leq b$$

M_0 and M_2 are found by directly applying the theorem. M_1 is found by differentiating M_0 with regard to b. Wyman [9] derived M_0 by integrating a concentrated-load elastic-plate solution over the loading circle.

The continuity of M_0 and M_2 at $a = b$ is obvious on the strength of Theorem 1. We will show, however, a direct proof in the following. We shall prove that

$$\text{ber}(x)\text{ker}'(x) - \text{bei}(x)\text{kei}'(x) + x^{-1}$$

$$= \text{ber}'(x)\text{ker}(x) - \text{bei}'(x)\text{kei}(x) \tag{B.11}$$

and

$$\text{bei}(x)\text{ker}'(x) + \text{ber}(x)\text{kei}'(x) = \text{bei}'(x)\text{ker}(x) + \text{ber}'(x)\text{kei}(x). \tag{B.12}$$

To prove this, note that

$$w_1(x) = \text{ber}(x) + i\,\text{bei}(x) \tag{B.13}$$

and

$$w_2(x) = \text{ber}(x) + i\,\text{kei}(x) \tag{B.14}$$

are the solutions of the differential equation

$$\frac{d^2w}{dx^2} + \frac{1}{x}\frac{dw}{dx} - iw = 0. \tag{B.15}$$

This can be proved by decomposing the equation

$$\left(\frac{d^2}{dx^2} + \frac{1}{x}\frac{d}{dx}\right)^2 w + w = 0$$

of which (B.13) and (B.14) are the solutions.

One can find that the Wronskian

$$\begin{vmatrix} w_1(x) & w_2(x) \\ w_1'(x) & w_2'(x) \end{vmatrix}$$

is equal to

$$= -x^{-1}.$$

Thus one has the identity

$$\begin{vmatrix} \mathrm{ber}(x) + i\,\mathrm{bei}(x) & \ker(x) + i\,\mathrm{kei}(x) \\ \mathrm{ber}'(x) + i\,\mathrm{bei}'(x) & \ker'(x) + i\,\mathrm{kei}'(x) \end{vmatrix} = -\frac{1}{x}$$

of which the real part gives (B.11) and the imaginary part gives (B.12).

Theorem 2 can be extended in many ways. Nevel [10] found that

$$\int_0^\infty F(x)dx = \frac{i}{\pi}\int_{-\infty}^{+\infty} F(z)\log z\,dz \qquad (B.16)$$

for an odd function $F(z)$ that does not have any pole on the real axis and vanishes at $z = \infty$.

B4. It is impossible to apply Theorem 2 to the integrals of w at (4), σ_r at (A.37), and σ_θ at (A.38) by the following reason.

The function $\exp(-\alpha_2 T)$ has essential singularities at the roots of

$$\beta^4 + E = 0,$$

because

$$\lim_{\beta^4 \to -E} \alpha_2 = \infty.$$

The function $\exp(-\alpha_1 T)$ does not possess any essential singularities because the limit of

$$\alpha_1 = 2\tau/\{\tau\beta^4 + 1 + \tau + \sqrt{(\tau\beta^4 + 1 + \tau)^2 - 4\tau(\beta^4 + E)}\}$$

is finite. However, the real part of α_1 becomes negative, and exp $(-\alpha_1 T)$ diverges, as $|\beta| \to \infty$ in a certain range of directions.

Theorem 2 does not apply to integral K at (13) because the point $x = 0$ is an essential singularity.

Only the other alternative for the integration is the use of Barnes' integral method. It consists in substituting the integrals

$$J_\nu(x) = \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \frac{\Gamma(-s)(\frac{1}{2}x)^{\nu+2s}}{\Gamma(\nu+s+1)} \, ds \tag{B.17}$$

and

$$\pi e^{\frac{1}{2}(\nu+1)\pi i} H_\nu^{(1)}(z) = \frac{1}{2\pi i} \int_{-c-\infty i}^{-c+\infty i} \Gamma(-\nu-s)\Gamma(-s)(-\tfrac{i}{2}z)^{\nu+2s} ds \tag{B.18}$$

for $J_\nu(x)$ and $H_\nu^{(1)}(z)$, where c is a real number satisfying $c > R(\nu)$, z is complex, and x is real. One can usually exchange the order of integration to carry out the integration with regard to x or z. Then, one can carry out the rest of the integration in most cases by use of the theorem of residue. Only the forms (B.17) and (B.18) serve this purpose. The other Barnes' representations of $J_\nu(x)$ and $H_\nu^{(1)}(x)$ do not enable one to carry out the above two procedures.

However, as mentioned by Watson ([11] p. 192), (B.17) does not hold true for $\nu = 0$, and (B.18) does not hold true when $\nu = 0$ and z is real. In these two cases, the integrands of (B.17) and (B.18) become proportional to s^{-1} as s approaches $i\infty$ as the limite of the imaginary axis. Therefore we cannot use Barnes' integral method to carry out our integrals.

APPENDIX C

Historical Background of the Analytical Study

Since ancient times floating ice plates have been used to cross rivers and lakes. Recently traffic load has increased. Vehicles have become heavier; aircraft landing and parking also add to weights. In these several years oil companies started to use ice plates as drilling platforms.

Formulation of the creep of a floating ice plate began after World War II with the intense development of the linear viscoelasticity theory. In 1947 Golushkevich (referred to by Kheysin [1]) presented an analysis assuming that ice behaves elastically for volumetric deformations and viscoelastically for deviatoric deformations. Kheysin [1] used a general viscoelastic thin-plate theory to analyze the infinitely-wide floating ice plate. He used the Maxwell model (Fig. 1), and considered only a concentrated load. Nevel [12] also used the Maxwell model, but considered a distributed load. He limited his numerical computation only to the center of the load.

William L. Ko, as reported by Garbaccio [14,15], used the Maxwell-Voigt type four-element model (Fig. 1), which is known to represent the creep of ice satisfactorily well [Jellinek and Brill 3]. In addition to thin-plate theory, Ko used Reissner's plate theory, which includes the deformation due to vertical shear forces. Garbaccio [15] numerically evaluated Ko's solution for specific values of material constants rather than for nondimensional parameters. Garbaccio's numerical answers show a strong effect of the discontinuity of the load distribution on the values of deflection. It is reasonable to suspect that his numerical evaluation may contain some errors.

Yakunin [16, 17] has solved the same problem as Ko, but Yakunin used only thin-plate theory. Unfortunately, only an abstract of Yakunin's work is available to the western researchers.

Katona [18] and Vaudrey and Katona [19] solved the same problem with a finite-element viscoelastic computer program.

We solved this problem analytically, and also developed an effective method of numerical integration. However, the theoretical curves did not satisfactorily fit the field-test curves. It is now evident that a large scale laboratory test eliminating the variation due to natural conditions must be carried out and the applicability of the theoretical assumptions must be tested.

APPENDIX D.

References

1. S. Takagi. Integration of $\int_0^\infty F(x)J_0(ax)J_1(bx)dx$. Transactions of the 21st Conference or Army Mathematicians, U.S. Army Research Office. p. 511. 1976.

2. G. E. Frankenstein, Load test data for lake ice sheets, Technical Report 89, U.S. Army Cold Regions Research and Engineering Laboratory (1963).

3. H. H. G. Jellinek and R. Brill, Viscoelastic properties of ice, J. Appl. Phys. 27, 1198-1209 (1956).

4. W. Flügge, Viscoelasticity, Blaisdell Pub. Co. (1967).

5. D. E. Nevel, Creep theory for a floating ice sheet, Special Report 76-4, U.S. Army Cold Regions Research and Engineering Laboratory, June 1976.

6. B. Van der Pol and H. Bremmer, Operational Calculus based on the Two-sided Laplace Integral, Cambridge at the University Press (1959).

7. E. C. Titchmarsh, The Theory of Functions, Oxford University Press (1939).

8. J. Dougall, An analytical theory of the equilibrium of an isotropic elastic plate, Transaction of the Royal Society of Edinburgh, vol. 41, Part I, no 8, 1903-04, 129-228.

9. M. Wyman, Deflections of an infinite plate, Canadian Jl of Research, vol. A28 (1950), 293-302.

10. D. E. Nevel, Private communication.

11. G. N. Watson. A Treatise on the Theory of Bessel Functions. Cambridge at the University Press. 1962.

12. D. Ye Kheysin, K zadache uprugo-plasticheskogo izgiba ledyanogo pokrova (On the problem of the elastic-plastic bending of an ice cover) Trudy Arkticheskogo i Antarkticheskogo Nauchno-Issledovatel'-skgo Instituta. Tom 267, p. 143-49 (1964.

13. D. E. Nevel, Time-dependent deflection of a floating ice sheet, U.S. Army Cold Regions Research and Engineering Laboratory Research Report 196. Hanover, N.H. (1966).

14. D. H. Garbaccio, Creep of floating ice sheet, Naval Civil Engineering Laboratory Report No. CR-67.025, Port Heuneme, Calif. (1967).

15. D. H. Garbaccio, Creep of floating ice sheets computer calculations, Naval Civil Engineering Laboratory Report No. CR-69014, Port Hueneme, Calif. (1968).

16. A. Ye. Yakunin, K voprosu ob izgibe ledianogo prokrova s uchetom viazkokh svolstv l'da (Calculation of ice-cover bending allowing for viscous properties of ice), Novosibirskiy Institut Inzhenerov Zheleznodoroznogo Transporta. Transactions vol. 79, pp. 79-82. (1968). (Available as CRREL Draft translation 425).

17. A. Ye. Yakunin, Issledovaniye vliyaniya vremeni deystviya nagruzki na nesushchuyu sposobnost' ledyanogo pokrova (The investigation of the effect of the loading time on the bearing capacity of an ice cover), Novosibirskiy Institut Inzhenerov Zheleznodoroznogo Transporta, Speciality 01.022, Resistance of Materials and Structural Mechanics, p. 1-22. (1970).

18. M. G. Katona, Ice Engineering: Viscoelastic finite element formulation, Naval Civil Engineering Laboratory Technical Report R803, Port Hueneme, Calif. (1974).

19. K. D. Vaudry and M. G. Katona, Viscoelastic finite element analysis of sea ice sheets, International Association of Hydraulic Research, Third International Symposium on Ice Problems. U. S. Army Cold Regions Research and Engineering Laboratory, Haniver, N.H. (1975).

USE OF ALGEBRAIC METHODS IN THE DESIGN OF CONTROLLERS AND OBSERVERS FOR SYSTEMS WITH TIME DELAYS

Edward W. Kamen
School of Electrical Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

ABSTRACT. Many systems contain time delays that have a significant effect on the overall operation of the system. For example, time delays can result from telemetry between ground-based guidance systems and drones, and from reaction or decision times of human operators in radar-tracking systems. Nyquist-type results and functional analytical techniques are available for designing controllers for such systems, but it is difficult to implement these methods due to the infinite dimensionality of the underlying vector spaces. In contrast to these approaches, the presentation given here deals with an approach based on the algebraic structure of the system model. In this algebraic setting, designs can be constructed using computations based on matrices and vectors defined over rings of operators.

1. INTRODUCTION. In the last decade there has been a good deal of effort (see the survey [1]) devoted to the study of dynamical systems described by functional differential equations in n-dimensional space. Much of this mathematical theory centers on the class of linear systems with time delays given by a set of delay differential equations. Systems with time delays have appeared in many engineering problems for the last several decades [2,3]. In many of these applications the delays resulted from the flow of fluids or gases in various types of industrial processes (e.g. see [4]). In many new applications, time delays result from telemetry or communication links between components of a system

located large distances from each other. For instance, sizable time delays occur in the ground-based guidance of drones at White Sands Missile Range [5]. Time delays can also result from data processing in the execution of on-line control and estimation algorithms.

Systems with time delays arise in many military applications. For example, in addition to the drone control problem at White Sands, delays can result from reaction or decision times of human operators in radar-tracking systems and manned-vehicle systems [6]. In ground warfare, delays can result from the relocation of men or weapons.

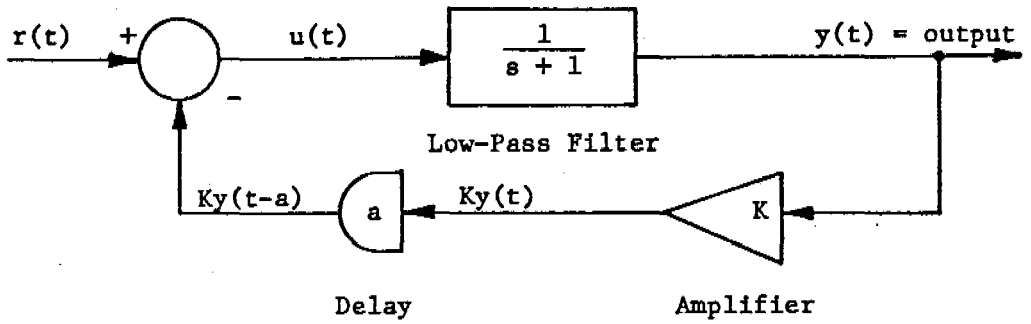
Despite recent efforts to develop an extensive mathematical theory for time-delay systems, it appears that few viable new techniques have been made available for the applications mentioned above. This is due in part to the "gap" between new mathematical theories and current engineering practices. However, the primary difficulty is due to the fact that time-delay systems are infinite-dimensional systems. Thus the implementation of mathematical results in general requires the use of finite-dimensional approximations. Further, if equations are to be implemented on a digital computer, it is necessary to consider "discrete-time" approximations with quantized magnitudes.

Although it is clear that approximations must be used in the study of time-delay systems, it is not at all clear as to when approximations should be brought into the theory. For example, given a system model with time delays, one could immediately approximate all delays by finite-dimensional elements, and then proceed by applying the theory of finite-dimensional systems. On the other hand, one could attempt to work with the time-delay model as much as possible, say by using symbolic computations, and then implement the resulting equations on a digital computer. In general, we would expect the latter procedure to yield a higher degree of accuracy since the time delays are not approximated until the last step. In fact, as illustrated in the next section, the first procedure can result in unstable designs.

The main purpose of this paper is to show that various problems involving time-delay systems can be studied via the second procedure men-

tioned above. The theory is based on a module framework for linear functional differential equations [7,8,9]. Although this has not yet been attempted, it should be possible to apply the results of the module approach to the engineering applications mentioned above.

2. TRANSFER-FUNCTION TECHNIQUES. Classical techniques for the study of time-delay systems are based for the most part on transfer-function representations. Although few general results are available, in some cases it is possible to apply standard methods such as the Nyquist stability test. To illustrate this, consider the following control system:



In this example, 1/(s+1) is the transfer function of the open-loop system (a low-pass filter in this case). The input u(t) of the open-loop system is given by

$$u(t) = -Ky(t-a) + r(t)$$

where K is the gain of the amplifier and r(t) is an external signal. The a-second time delay in the feedback loop could be a result of the controller (in this case the amplifier) being located a large distance from the open-loop system. The transfer function T(s) of the closed-loop system

is given by

$$(1) \qquad T(s) = \frac{1}{s + 1 + Ke^{-as}}$$

where e^{-as} is the transfer function of the time delay.

It is well known that the closed-loop system is exponentially stable if and only if the zeros of the characteristic function $s + 1 + Ke^{-as}$ are in the left-half of the complex plane. Unfortunately, in general it is difficult to compute zeros of characteristic functions containing exponentials e^{-as}. However, in some cases, including the above example, stability can be determined by using the Nyquist encirclement criterion: Given a fixed value of a, the range of positive values of K for which the closed-loop system is stable can be computed by plotting $e^{-as}/(s + 1)$ for $s = j\omega$, $-\infty < \omega < \infty$.

As an example, let's set a = .1. Since in this case the time delay is 1/10 th of the time constant of the open-loop system, it may appear that the delay can be neglected. If we neglect the .1 second delay, we find that the closed-loop system is stable for any positive value of K. If we consider the delay and apply the Nyquist test, we find that the closed-loop system is stable when $0 < K < 16.3$ and unstable when $K > 16.3$. Hence the delay has a very significant effect on stability. This phenomenon is well known in engineering, for example see [10, pages 346-350].

Now let's consider a rational approximation to the delay so that the above system can be studied using results for rational transfer functions. Writing

$$(2) \qquad e^{-as} = \frac{1}{1 + as + (as)^2/2 + (as)^3/6 + \ldots} \quad ,$$

we get a rational approximation by truncating the series in (2). Again let a = .1, and consider the following first-order approximation

$$(3) \qquad e^{-.1s} \approx \frac{1}{1 + .1s} = \frac{10}{s + 10} \quad .$$

The approximation (3) may seem reasonable in this case since the bandwidth of $10/(s + 10)$ is ten times the bandwidth of the open-loop system (given by $1/(s + 1)$). Using (3), we have that the transfer function $\hat{T}(s)$ of the resulting system approximation is given by

$$\hat{T}(s) = \frac{s + 10}{s^2 + 11s + 10(K+1)} \quad . \tag{4}$$

From (4), we see that the system approximation is stable for any $K > 0$. Thus with the first-order approximation of the delay, we completely miss the fact that the given system is unstable for $K > 16.3$.

Now consider the second-order approximation

$$e^{-.1s} \approx \frac{1}{1 + .1s + .005s^2} \quad .$$

The transfer function of the resulting system approximation is given by

$$\hat{T}(s) = \frac{s^2 + 20s + 200}{s^3 + 21s^2 + 220s + 200(K+1)} \quad .$$

Using the Routh-Hurwitz test, we find that the system approximation is stable when $0 < K < 22.1$ and unstable when $K > 22.1$. Thus with this approximation, we do see that the system is unstable for large values of K, although the critical value of K (22.1) is off a good deal from the actual critical value (16.3).

These results show that neglecting or approximating time delays can lead to serious problems such as instability. Moreover, in many cases the order of the rational approximations must be rather large, resulting in system models that are difficult to work with because of the high dimensionality. Given these problems, there is much interest in developing techniques (e. g. design procedures) that do not require the approximation of time delays. As will be discussed later, such techniques are already available. They are based primarily on state-space

models which are defined in the next section.

3. STATE-SPACE MODELS. Again consider the time-delay system with transfer function given by (1). It follows from (1) that the system has an internal (or state) representation given by the following delay differential equation

$$\frac{dy(t)}{dt} = -y(t) - Ky(t-a) + r(t) \quad . \tag{5}$$

As a generalization of (5), we can consider the class of m-input p-output time-delay systems given by the following state model

$$\frac{dx(t)}{dt} = A_0x(t) + A_1x(t-a) + Bu(t) \tag{6a}$$

$$y(t) = Cx(t). \tag{6b}$$

In (6a,b), A_0, A_1 (resp. B,C) are $n \times n$ (resp. $n \times m$, $p \times n$) matrices over the field R of real numbers, $u(t) \in R^m$ is the input or control function, $y(t) \in R^p$ is the output function, and $x(t) \in R^n$ is the state.

As is well known, to solve (6a) for $t > 0$ we need initial data consisting of the values of $x(t)$ for $-a \le t \le 0$. Although $x(t)$ is usually referred to as the state at time t, as a result of the delay term in (6a) the actual state at time t is the function segment $x(\sigma)$, $t - a \le \sigma \le t$. Thus the space of (actual) states for the system (6a,b) is some infinite-dimensional vector space of function segments. There are several candidates for this function space. Examples include the space $C([-a,0];R^n)$ of R^n-valued continuous functions defined on the interval $[-a,0]$ and the space $R^n \times L^p([-a,0];R^n)$ where $L^p([-a,0];R^n)$ is the space of R^n-valued p-integrable functions defined a.e. on $[-a,0]$.

Much of the mathematical theory of (6a), or generalizations of (6a), is based on a characterization of (6a) in terms of an ordinary differential equation in a Banach or Hilbert space such as $C([-a,0];R^n)$. For details, see the book [11] by Hale and the papers (e.g. [12]) of Delfour

and Mitter. However, this approach does not fully exploit the finiteness of (6a) resulting from the assumption here that x(t) belongs to n-dimensional space. This finiteness can be retained in an operator setting by expressing (6a) as a vector differential equation with operator coefficients. The constructions are as follows.

Let L_+^{loc} denote the space of real-valued Lebesgue-measurable locally-integrable functions defined a.e. on R with supports bounded on the left. Let d denote the delay operator given by

$$d: L_+^{loc} \to L_+^{loc}: f(t) \mapsto f(t-a)$$

where a is some fixed positive number.
With $(L_+^{loc})^n$ = space of n-element column vectors over L_+^{loc}, we can extend the delay operator d to $(L_+^{loc})^n$ by defining

$$d: (f_1(t),\ldots,f_n(t))^T \mapsto (f_1(t-a),\ldots,f_n(t-a))^T$$

where T denotes the transpose operation.

Then viewing x(t) as an element of $(L_+^{loc})^n$, we can write (6a) in the form

$$\frac{dx(t)}{dt} = (A_0 + A_1 d)x(t) + Bu(t) \tag{7}$$

where $A_0 + A_1 d$ is a n × n matrix operator acting on elements in $(L_+^{loc})^n$.

As a generalization of (7), we can consider vector differential equations with coefficient matrices defined over a ring of delay operators: Let R[d] denote the set of all finite sums of the form $\sum_i a_i d^i$ where $a_i \in R$ and $d^i: f(t) \mapsto f(t-ia)$. With the usual addition and multiplication operations, R[d] is a ring of delay operators and with the scalar multiplication

$$R[d] \times (L_+^{loc})^n \to (L_+^{loc})^n: (\sum_i a_i d^i, x(t)) \mapsto \sum_i a_i x(t-ia) ,$$

$(L_+^{loc})^n$ is a module over the ring R[d].

Now given matrices F(d), G(d), and H(d) defined over R[d], consider the following system equations

$$\frac{dx(t)}{dt} = F(d)x(t) + G(d)u(t) \tag{8a}$$

$$y(t) = H(d)x(t). \tag{8b}$$

By definition of F(d), G(d), H(d), we can write

$$F(d) = \sum_i F_i d^i,\quad G(d) = \sum_i G_i d^i,\quad H(d) = \sum_i H_i d^i$$

where the F_i, G_i, H_i are matrices over R. Thus (8a,b) can be written in the form

$$\frac{dx(t)}{dt} = \sum_i F_i x(t-ia) + \sum_i G_i u(t-ia) \tag{9a}$$

$$y(t) = \sum_i H_i x(t-ia). \tag{9b}$$

Therefore, the class of delay differential equations given by (9a,b) can be studied in terms of vector differential equations (8a,b) defined over a ring of operators. This observation, along with results based on the operator ring structure, was first made in [7]. It is now known [8,9] that the operator ring framework applies to a very large class of time-delay systems, including systems with noncommensurate delays and distributed delays.

4. APPLICATIONS OF THE OPERATOR STRUCTURE. In this section we briefly consider several topics that can be studied via the operator ring framework. We restrict our attention to systems over R[d] given by (8a,b), although many of the results discussed below apply to more general rings of operators.

a. Computation of Solutions. As shown in [8], initial data for solving (8a) can be incorporated into the operator structure, so that complete solu-

tions can be expressed in terms of algebraic operations. In fact the complete solution can be written directly in terms of the initial data and the coefficient matrices F(d), G(d) of the system equation (8a). In other words, solutions can be computed by employing matrix operations defined over rings of operators. It should be possible to implement portions of this operational calculus by using symbolic computations.

b. Realization. Given a system specified by its transfer function matrix T(s), the problem of realization is concerned with the construction of a state model from T(s). For systems with time delays, T(s) is often a matrix of rational functions in s and e^{-as}, in which case we want to construct a state model of the form (8a,b).

The computation of a realization from T(s) is not difficult; however, the computation of a realization with the number of coordinates of x(t) minimal among all possible realizations is a nontrivial problem in the multiterminal case. Such realizations are said to be minimal. In [7] a constructive procedure is given for computing minimal realizations defined over R[d]. The construction is based on a method for computing a basis of module $R^n[d]$ from a set of generators. These results can also be used to reduce overdetermined systems of delay differential equations.

c. System Properties. As before, consider a time-delay system given by the state equation

$$\frac{dx(t)}{dt} = F(d)x(t) + G(d)u(t) \tag{10}$$

where F(d) (resp. G(d)) is a $n \times n$ ($n \times m$) matrix over R[d]. Let N(d) denote the $n \times mn$ matrix over R[d] given by

$$N(d) = [G(d), F(d)G(d), \ldots, F^{n-1}(d)G(d)] .$$

In the special case when F(d) and G(d) are over R, so that the system (10) is finite dimensional, the matrix N(d) is called the controllability matrix

of the system. This term arises from the well-known result [13] that a finite-dimensional system is reachable (or controllable) if and only if the rank of N(d) is equal to n (in this special case N(d) is over the field R). Motivated by this result, we give the following

Definition. The system (10) is reachable in the strong sense (resp. weak sense) if the columns of N(d) generate the module $R^n[d]$ (resp. N(d) has rank n viewed as a matrix over the quotient field of R[d]).

This definition was first given by Morse [14]. Although strong and weak reachability are not equivalent to the dynamical properties of Euclidean and functional controllability, they (or variants of these concepts) are related to certain dynamical properties such as stabilizability (see [15]). The concepts of strong and weak reachability are particularly interesting because it is possible to determine, in an algorithmic fashion, whether or not a given system has these properties.

d. State-Feedback Control. For the system (10) we can consider state feedback by setting

$$u(t) = -K(d)x(t) + r(t)$$

where K(d) is a m × n matrix over R[d] and r(t) is an external input. Note that since K(d) is over the ring R[d], we are allowing time delays in the feedback loop. This is reasonable since the given system (10) is defined over R[d]. However, the constraint that K(d) be over R[d] is too severe. In addition to "pure" delays given by elements of R[d], we need to consider distributed delays. That is, let

$$u(t) = -K(d)x(t) - \int_{t-h}^{t} L(t-\sigma)x(\sigma)d\sigma + r(t) \tag{11}$$

where L(t) is a m × n matrix of integrable functions with bounded supports contained in [0,h], h > 0. The last term on the right side of (11) is referred to as a distributed delay operation. As an example, suppose that L(t) is a scalar function equal to 1 for $0 \le t \le h$ and equal to 0 otherwise.

Then

$$\int_{t-h}^{t} L(t-\sigma)x(\sigma)d\sigma = \int_{t-h}^{t} x(\sigma)d\sigma$$

In this case the distributed delay operation is the finite-interval integrator. It is interesting that this particular operation often appears in the control theory of time-delay systems.

For feedback controls of the form (11), or variants of this form, algebraic procedures have been developed for eigenvalue (or pole) assignment and stabilizability (see [14,15,9]). These results rely heavily on the algebraic properties of the operator ring framework.

A very interesting and important problem is the computation of feedback controls of the form (11) that minimize the following cost functional

$$J = \int_{0}^{\infty} [x^{T}(t)Qx(t) + u^{T}(t)Pu(t)]dt$$

where P and Q are symmetric positive definite matrices over R. It is known [1] that optimal controls can be computed by solving a set of coupled ordinary and partial differential equations, referred to as the Riccati equations. But there is some evidence that optimal controls can be expressed in terms of the operator ring structure, which could result in simplified procedures for computing controls. This problem is currently under investigation.

e. Duality and Observer Theory. Another nice consequence of the operator framework is that there is a natural concept of dual system arising from notion of the dual of a homomorphism on finite free modules. The dual system is defined as follows.

Definition. Given the system

$$\frac{dx(t)}{dt} = F(d)x(t) + G(d)u(t)$$

$$y(t) = H(d)x(t)$$

where $u(t) \in R^m$, $x(t) \in R^n$, $y(t) \in R^p$ and $F(d)$, $G(d)$, $H(d)$ are $n \times n$, $n \times m$, $p \times n$ matrices over $R[d]$, the dual system is given by

$$\frac{d\xi(t)}{dt} = F^T(d)\xi(t) + H^T(d)v(t)$$

$$\gamma(t) = G^T(d)\xi(t)$$

where $v(t) \in R^p$, $\xi(t) \in R^n$, $\gamma(t) \in R^m$ and T denotes matrix transposition.

As in the Kalman duality theory for finite-dimensional systems, the state-feedback control problem for the dual system corresponds to the state-observation problem for the given system. Hence results on state-feedback control in the dual system yield results on state observers for the given system.

ACKNOWLEDGEMENT

Much of the author's work discussed above has been sponsored by the Mathematics Division of the U. S. Army Research Office. This support is gratefully acknowledged.

REFERENCES

1. A. Manitius, "Optimal control of hereditary systems," in Control Theory and Topics in Functional Analysis, vol. III, International Atomic Energy Agency, Vienna, 1976, pp. 43-178.

2. R. Weiss, "Transportation lag-an annotated bibliography," IRE Trans. Automatic Control, AC-4, 1959, pp. 56-58.

3. N. Choksy, "Time-lag systems-a bibliography," IRE Trans. Automatic Control, AC-5, 1960, pp. 66-70.

4. W. Ray and M. Soliman, "The optimal control of processes containing pure time delays-I, Necessary conditions for an optimum," Chem. Eng. Sci., 25, 1970, pp. 1911-1925.

5. Private communication with A. Gilbert and B. Green of White Sands Missile Range, April 1977.

6. D. Kleinman, S. Baron, and W. Levison, "A control theoretic approach to manned-vehicle systems analysis," IEEE Trans. Automatic Control, AC-16, pp. 824-832.

7. E. Kamen, "On an algebraic theory of systems defined by convolution operators," Math. Systems Theory, 9, 1975, pp. 57-74.

8. E. Kamen, "An operator theory of linear functional differential equations," accepted for publication in J. Differential Equations.

9. E. Kamen, "Representation and realization of operational differential equations with time-varying coefficients," J. Franklin Institute, 301, 1976, pp. 559-571.

10. K. Ogata, Modern Control Engineering, Prentice-Hall, Inc., N. J., 1970.

11. J. Hale, Functional Differential Equations, Springer-Verlag, New York, 1971.

12. M. Delfour and S. Mitter, "Hereditary differential systems with constant delays. I. General case," J. Differential Equations, 12, 1972, pp. 213-235.

13. R. Kalman, "Lectures on controllability and observability," CIME Summer Course 1968, Cremonese, Rome, 1969.

14. A. Morse,"Ring models for delay differential systems," Automatica, 12, 1976, pp. 529-531.

15. E. Sontag,"Linear systems over commutative rings: A survey," Ricerche di Automatica, 7, 1976, pp. 1-33.

The Structure of Groups with Index-3 Subgroups

L.V. Meisel, D.M. Gray, and E. Brown

Physical Science Division, Watervliet Arsenal,

Watervliet, New York 12189

For any group G_o which contains an index-3 subgroup G, it is shown that either: (a) G is an invariant subgroup or (b) G contains an index-2 subgroup G_A where G_A is an invariant subgroup of G_o. For case (a), G and its cosets give rise to three operators which span a stable 3-dimensional subspace of the group algebra which further reduces to three 1-dimensional stable subspaces. For case (b), G_A and its cosets give rise to six operators which span a 6-dimensional stable subspace of the group algebra which reduces to two 1-dimensional and two 2-dimensional irreducible stable subspaces of the group algebra. The irreducible representations and the corresponding basis elements of the group algebra are given for both cases.

This paper was presented at the 22nd Conference of Army Mathematicians.

I. INTRODUCTION

In this paper we discuss some features of groups containing a subgroup consisting of one third of the group elements. We derive the structure of such groups and give explicit irreducible representations which are characteristic of them.

Interest in such groups springs from their relevance to the theory of second order phase transitions. In their classic text on Statistical Physics, Landau and Lifshitz[1] make the statement: "It appears that the following theorem is also true: No second-order phase transition can exist for transitions involving the decrease by a factor three of the number of symmetry elements (owing to the existence of third-order terms in the expansion of the thermodynamic potential)." In a recent review article Cracknell[2] conjectured that this theorem probably could not be proven in the general case.

Thus, we were motivated to a general study of such groups and to a proof of Landau's Theorem. The general study is presented here; the proof of Landau's Theorem will appear separately.[3]

II. STRUCTURE AND REPRESENTATIONS

We shall denote sets of group elements by capital letters and

members of sets by corresponding lower case letters, e.g. $h\varepsilon H$ or $h_A\varepsilon H_A$. The order of any such set will be denoted as an absolute value, e.g., the order of set H is denoted $|H|$. We shall be discussing a group G_o such that $|G_o|=3N$ containing a subgroup G with $|G|=N$ and the distinct left cosets $H=h_1G$ and $K=k_1G$. In an obvious notation, $G_o=G+H+K$.

We present theorems in Appendix I indicating that either G is an invariant subgroup of G_o or that it contains a subgroup G_A, an invariant subgroup of G_o, where $|G_A|=|G|/2$. In the latter case the cosets H and K divide into the cosets (H_A and H_B) and (K_A and K_B) with respect to G_A and the remaining elements of G are members of the coset G_B. Table I gives a multiplication table for the various sets of elements; its derivation is described in Appendix I. The sets G_A, G_B, H_A, H_B, K_A and K_B form the factor group G_o/G_A. In the event that G is an invariant subgroup G, H, and K form the factor group G_o/G; the multiplication table for these sets is the same as that in Table I for G_A, H_A, and K_A.

We may furthermore define the set operators $\tilde{G}_A\equiv|G_A|^{-1}\sum_{g_A\varepsilon G_A} g_A$, $\tilde{G}_B\equiv|G_A|^{-1}\sum_{g_B\varepsilon G_B} g_B$, $\tilde{H}_A\equiv|G_A|^{-1}\sum_{h_A\varepsilon H_A} h_A$ etc. These operators will have the multiplication table given in Table I where we now read each entry in the table as the corresponding set operator. The operators

span a 6-dimensional reducible stable subspace of the group algebra[4] and also form a six element group. The six element group contains three classes: $\tilde{G}_A$, $(\tilde{H}_A, \tilde{K}_A)$ and $(\tilde{G}_B, \tilde{H}_B, \tilde{K}_B)$. Thus, one can find two 1-dimensional and one 2-dimensional irreducible representation of the six element group. This corresponds to a reduction of the 6-dimensional subspace of the group algebra into two 1-dimensional stable subspaces and two 2-dimensional irreducible stable subspaces. (Each 2-dimensional subspace gives rise to the same irreducible representation.)

When G is an invariant subgroup, the operators $\tilde{G}_A \equiv |G|^{-1} \sum_{g \in G} g$, $\tilde{H}_A \equiv |G|^{-1} \sum_{h \in H} h$, and $\tilde{K}_A \equiv |G|^{-1} \sum_{k \in K} k$ form a three element group each element of which is in a class by itself yielding three 1-dimensional stable subspaces of the group algebra and three 1-dimensional representations.

The characters of the representations are given in Table II. It is straight forward to show that when G is not invariant the irreducible stable subspaces will be spanned by the following members of the group algebra:

$$\Gamma_{1+} \quad : \quad \tilde{A}_+ = \tilde{G}_+ + \tilde{H}_+ + \tilde{K}_+$$

$$\Gamma_{1-} \quad : \quad \tilde{A}_- = \tilde{G}_- + \tilde{H}_- + \tilde{K}_-$$

$$\Gamma_2 : \quad \tilde{E}_1 = \sqrt{1/6}\ (2\tilde{G}_+ - \tilde{H}_+ - \tilde{K}_+)$$

$$\tilde{E}_2 = \sqrt{1/2}\ (\tilde{H}_+ - \tilde{K}_+)$$

$$\Gamma_2' : \quad \tilde{E}_1' = \sqrt{1/2}\ (\tilde{H}_- - \tilde{K}_-)$$

$$\tilde{E}_2' = -\sqrt{1/6}\ (2\tilde{G}_- - \tilde{H}_- - \tilde{K}_-)$$

where $\tilde{G}_\pm \equiv \tilde{G}_A \pm \tilde{G}_B$, $\tilde{H}_\pm \equiv \tilde{H}_A \pm \tilde{H}_B$, and $\tilde{K}_\pm \equiv \tilde{K}_A \pm \tilde{K}_B$.

The normalization factors in the basis elements of Γ_2 and Γ_2' have been chosen to produce the following unitary representation:

$$D(\tilde{G}_A) \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad D(\tilde{H}_A) \equiv \begin{pmatrix} -a & -b \\ b & -a \end{pmatrix} \qquad D(\tilde{K}_A) \equiv \begin{pmatrix} -a & b \\ -b & -a \end{pmatrix}$$

$$D(\tilde{G}_B) \equiv \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad D(\tilde{H}_B) \equiv \begin{pmatrix} -a & b \\ b & a \end{pmatrix} \qquad D(\tilde{K}_B) \equiv \begin{pmatrix} -a & -b \\ -b & a \end{pmatrix}$$

where $a = 1/2$ and $b = \sqrt{3}/2$.

It is interesting to note how the 2-dimensional representation reduces when G is an invariant subgroup. Here $\tilde{G}_A = \tilde{G}_+ = \tilde{G}_-$, etc. and the complex conjugate pair of basis elements:

$$\tilde{E}_1 + i\tilde{E}_2 = \sqrt{2/3}\ (\tilde{G}_A + \omega\, \tilde{H}_A + \omega^* \tilde{K}_A)$$

and $$\tilde{E}_1 - i\tilde{E}_2 = \sqrt{2/3}\ (\tilde{G}_A + \omega^* \tilde{H}_A + \omega\, \tilde{K}_A)$$

where $\omega = \exp(2\pi i/3)$ are seen to span 1-dimensional stable subspaces of the group algebra with characters given by Γ and Γ^* of Table IIa.

(The third stable subspace corresponding to the representation Γ will be spanned by $\tilde{A}=\tilde{A}_+-\tilde{A}_-=\tilde{G}_A+\tilde{H}_A+\tilde{K}_A$.) In this case $\tilde{E}_1$ and $\tilde{E}_2$ are basis elements for the appropriate real pseudo-2-dimensional irreducible representation obtained upon invoking time reversal invariance.

A simple application of these ideas to the theory of continuous phase transitions is presented in Appendix II.

REFERENCES

1. L. D. Landau and E. M. Lifshitz, Statistical Physics, London, Pergamon Press (1958)
2. A. P. Cracknell, Advances in Physics 23, 673 (1974).
3. E. Brown, D. M. Gray, and L. V. Meisel submitted to Phys. Rev.
4. A review of group algebra concepts may be found in Per-Olov Löwdin, Rev. Mod. Phys. 39, 259 (1967).
5. See for example, theorems 1 and 2 on page 60 of E. P. Wigner, Group Theory and Its Application to the Quantum Mechanics Of Atomic Spectra, Academic Press (1959).
6. L. P. Bouckaert, R. Smoluchowski, and E. Wigner, Phys. Rev 50, 58 (1936)
7. J. Perel, B. W. Batterman, and E. I. Blount, Phys. Rev. 166, 616 (1968)

Table I. Multiplication table for the set operators. The table gives the product operation of an element (set) on the left and an element (set) at the top, e.g. $K_A H_B = G_B$.

	G_A	G_B	H_A	H_B	K_A	K_B
G_A	G_A	G_B	H_A	H_B	K_A	K_B
G_B	G_B	G_A	K_B	K_A	H_B	H_A
H_A	H_A	H_B	K_A	K_B	G_A	G_B
H_B	H_B	H_A	G_B	G_A	K_B	K_A
K_A	K_A	K_B	G_A	G_B	H_A	H_B
K_B	K_B	K_A	H_B	H_A	G_B	G_A

Table IIa. Character table for the irreducible representations of the set operator group for the case that $|G_B|=0$. ($\omega=\exp(2\pi i/3)$)

	G_A	H_A	K_A
Γ_o	1	1	1
Γ	1	ω^*	ω
Γ^*	1	ω	ω^*

Table IIb. Character table for the irreducible representations of the set operator group for the case that $|G_B|=|G_A|$. Read C_{H_A} as the class of H_A, etc.

	C_{G_A}	$2C_{H_A}$	$3C_{G_B}$
Γ_{1+}	1	1	1
Γ_{1-}	1	1	-1
or Γ'_2	2	-1	0

APPENDIX I. THEOREMS

All of the theorems presented here pertain to the group G_o which contains the subgroup G and the distinct left cosets H and K as defined in the main text. We give proofs only for a select few of the theorems; the omitted proofs are similar to those given. (N.B. In these theorems the subscripts i, j denote specific group elements.)

Theorem 1. If $g_i h_j \in H$ then $g_i H = H$ and $g_i K = K$. (N.B. Not every member of G will satisfy the hypothesis of theorem 1 unless G is an invariant subgroup.)

Proof: $g_i H = g_i h_j G = h_s G = H$. The first and third equalities follow from the definition of left cosets. The second equality follows from the hypothesis of the theorem. $g_i G = G$ since G is a group. Thus $g_i K = K$ by the rearrangement theorem[5]. QED

Theorem 2. If $g_i h_j \in K$ then $g_i H = K$ and $g_i K = H$.

Define G_A and G_B as the sets of elements g_i satisfying theorems 1 and 2 respectively.

Theorem 3. $G = G_A + G_B$.

Theorem 4. If $h_i h_j \in K$ then $h_i H = K$ and $h_i K = G$.

Theorem 5. If $h_i h_j \in G$ then $h_i H = G$ and $h_i K = K$.

Define H_A and H_B as the sets of elements h_i satisfying the hypotheses of theorems 4 and 5 respectively.

Theorem 6. $H=H_A+H_B$.

Theorem 7. If $k_i k_j \in H$ then $k_i K=H$ and $k_i H=G$.

Theorem 8. If $k_i k_j \in G$ then $k_i K=G$ and $k_i H=H$.

Define K_A and K_B as the sets of elements k_i satisfying the hypotheses of theorems 7 and 8, respectively.

Theorem 9. $K=K_A+K_B$.

Theorem 10. The multiplication table for the sets defined by theorems 1 to 9 is given by Table I.

Theorem 10 can be proven as follows:

The sets G_A, G_B, H_A, etc., have been defined in terms of the result of operating from the left on arbitrary elements of the sets G, H, and K. If h_{Ai} is an element of H_A then $h_{Ai}\begin{bmatrix}G\\H\\K\end{bmatrix}=\begin{bmatrix}H\\K\\G\end{bmatrix}$ etc. Using these properties we can find the set membership of any product. For example,

$$h_{Ai}\, k_{Bj}\begin{bmatrix}G\\H\\K\end{bmatrix} = h_{Ai}\begin{bmatrix}K\\H\\G\end{bmatrix} = \begin{bmatrix}G\\K\\H\end{bmatrix} = g_{Bs}\begin{bmatrix}G\\H\\K\end{bmatrix}$$

i.e., the product of an arbitrary element of H_A and an arbitrary element of K_B (in the order given) yields an element of G_B. By repeated application of these techniques the entire multiplication table is generated.

Theorem 11. Either $|G_A|=|G|$ or $|G_A|=|G|/2$.

Proof: Either $|G_B|=0$ or $|G_B|\neq 0$.

If $|G_B| = 0$ then $G = G_A$ and $|G_A| = |G|$.

If $|G_B| \neq 0$ then $g_{Bs}\, g_{Bj} = g_A \in G_A$ for all j. (Table I). Hence, $|G_A| \geq |G_B|$. (The rearrangement theorem implies that no element of G_A will be repeated as g_{Bj} runs over G_B.) Also $g_{Bs}\, g_{Aj} = g_{Bm} \in G_B$ for all j. (Table I). Hence $|G_B| \geq |G_A|$. Therefore $|G_B| = |G_A|$ and, since $G = G_A + G_B$ (from theorem 3), $|G_A| = |G|/2$. QED

Theorem 12. $|G_A| = |H_A| = |K_A|$.

Theorem 13. G_A is an invariant subgroup.

Left and right cosets with respect to G_A can be seen to be identical by examination of Table I.

Theorem 14. The operators $\tilde{G}_A, \tilde{G}_B, \tilde{H}_A, \tilde{H}_B, \tilde{K}_A$, and $\tilde{K}_B$ have the multiplication table of Table I.

The proof of this theorem goes as follows: Consider, for example, the operators $\tilde{H}_A$ and $\tilde{K}_B$. $\tilde{H}_A \tilde{K}_B = |G_A|^{-2} \sum_{j,s} h_{Aj}\, k_{Bs}$ $= |G_A|^{-2} \sum_{j,t} g_{Bt}$. The last equality follows from Table I as applied to sets of group operators G_A, G_B, H_A etc. (theorem 10), the rearrangement theorem and theorems 11 and 12. Thus, $\tilde{H}_A \tilde{K}_B = |G_A|^{-1} \sum_j \tilde{G}_B = \tilde{G}_B$. The entire table for operators now can be generated in this way.

Theorem 15. If $|G_B| = 0$, then $\tilde{G}_A$, $\tilde{H}_A$, and $\tilde{K}_A$ form a three element group with each element in a class by itself.

Theorem 16. If $|G_B| \neq 0$, then $\tilde{G}_A, \tilde{G}_B, \tilde{H}_A, \tilde{H}_B, \tilde{K}_A$, and $\tilde{K}_B$ form a six element group with classes: $\tilde{G}_A$, $(\tilde{H}_A, \tilde{K}_A)$, and $(\tilde{G}_B, \tilde{H}_B, \tilde{K}_B)$.

APPENDIX II. A SIMPLE APPLICATION

An example of the application of these ideas is found in the theory of continuous phase transitions. When a solid transforms from a structure of 3N symmetry operations to one of N operations one is interested in a function invariant with respect to the smaller group of symmetry operators which may also serve as a basis function for an irreducible representation (other than the identical representation) of the larger group. One sees that

$$\psi = \tilde{E}_1 \phi = \sqrt{1/6}\,(2\tilde{G}_+ - \tilde{H}_+ - \tilde{K}_+)\phi$$

(or $\psi = \tilde{E}_1' \phi$) will be an acceptable function by examination of the matrix representation of Γ_2 (or Γ_2'). In the particular case of a transition from a crystal having the symmetry of the full cubic group to one having the symmetry of the tetragonal group which singles out the z-axis one finds by taking

$\phi = z^2$,

$$\psi = \sqrt{1/6}\,(2z^2 - x^2 - y^2)$$

which has the partner function

$$\psi' = \sqrt{1/2}\,(\tilde{H}_+ - \tilde{K}_+)\, z^2 = \sqrt{1/2}\,(x^2 - y^2)$$

i.e., the representation Γ_2 corresponds to the representation of the full cubic group denoted Γ_{12} in BSW[6] notation. An extensive discussion of the transformation from cubic to tetragonal symmetry may be found in Perel, Batterman and Blount[7].

D. M. Gray: There is a definite connection between this and one aspect of our work on superconductivity. The A-15's (Cr_3Si structure) are an important class of high-T_c superconductors; many of the A-15's undergo a cubic to tetragonal phase transition at temperatures somewhat above T_c. If one considers only the point group of the lattice this is a factor of three reduction. Thus, when such transition is second-order, there must be a further lowering of the symmetry (by relative movement of the basis atoms) to be consistent with Landau's theory. Symmetry considerations can be used to limit the types of internal movement. See Perel, Batterman, and Blount, Phys. Rev. 166, 616 (1968).

ATTENDEES LIST

(By Organization)

23RD CONFERENCE OF ARMY MATHEMATICIANS

11-13 May 1977

Hosted By: Langley Directorate, USAAMRDL
NASA-Langley Research Center
Hampton, Virginia

BOLLING AIR FORCE BASE, NC

Robert W. Buchal

DAVID W. TAYLOR NAVAL SHIP R&D CENTER - Bethesda, MD

Dr. Elizabeth Cuthill
Kent Meals

GEORGIA INSTITUTE OF TECHNOLOGY - Atlanta, GA

Edward W. Kamen

HARRY DIAMOND LABORATORY - Adelphi, MD

Joseph M. Kirshner

MATHEMATICS RESEARCH CENTER - Madison, WI

M. G. Crandall
Walter Gautschi
Thomas N. E. Greville
Karl E. Lonngren
Ben Noble
John A. Nohel
Peter D. Robinson
J. Barkley Rosser
Alwyn C. Scott

NASA-LANGLEY RESEARCH CENTER - Hampton, VA

Dr. J. E. Duberg
Dr. Wayne D. Erickson
Jay Lambiotte
Dr. Stephen K. Park
D. H. Rudy
John N. Shoosmith
Dariene D. Stevens
Dr. Dean J. Weidman

OLD DOMINION UNIVERSITY SCHOOL OF ENGINEERING - NASA-LRC

V. M. Vatsa

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE & ENGINEERING (ICASE) - NASA-LRC

Alvin Bayliss
Jim Daywitt
Lois Mansfield
John Strikwerda
Tom Zang

NASA-LEWIS RESEARCH CENTER - Cleveland, OH

John P. Gyekenyesi

NEW YORK UNIVERSITY - New York, NY

Prof. Heinz O. Kreiss

DEPARTMENT OF NATIONAL DEFENSE
OPERATIONAL RESEARCH & ANALYSIS ESTABLISHMENT - Ottawa, Ontario Canada

Dr. M. A. Weinberger

PRINCETON UNIVERSITY - Princeton, NJ

Martin Kruskal

STATE UNIVERSITY OF NEW YORK - Stony Brook, NY

Yung Ming Chen
Ram P. Srivastav

USA AIR MOBILITY R&D LABORATORY, HEADQUARTERS - Moffett Field, CA

Dr. James T. Wong

USA AIR MOBILITY R&D LABORATORY, AMES DIRECTORATE - Moffett Field, CA

Don Adams

USA AIR MOBILITY R&D LABORATORY, EUSTIS DIRECTORATE - Fort Eustis, VA

Lou Bartek

USA AIR MOBILITY R&D LABORATORY, LANGLEY DIRECTORATE - Hampton, VA

Thomas L. Coleman
C. E. Hammond
Robert L. Tomaine

USA AIR MOBILITY R&D LABORATORY, LEWIS DIRECTORATE - Cleveland, OH

Albert Kascak
Jon Kring

USA BALLISTIC RESEARCH LABORATORY - Aberdeen Proving Ground, MD

Jad H. Batteh
Vitalius Benokraitis
Aivars Celmins
Alexander S. Elder
John D. Powell
Dr. J. M. Santiago

USA COLD REGIONS R&D LABORATORY - Hanover, NH

Yoshisuke Nakano
Shunsuke Takagi

USA COMMUNICATIONS COMMAND - Fort Huachuca, AZ

John L. Lazaruk

USA CONCEPTS ANALYSIS AGENCY - Bethesda, MD

Dr. Richard A. Robinson

USA ELECTRONICS COMMAND - Fort Monmouth, NJ

Leon Kotin
Walter Pressman

USA ELECTRONIC PROVING GROUND - Fort Huachuca, AZ

Bart Wilburn

USA INTELLIGENCE & SECURITY COMMAND - Warrenton, VA

Michael W. White

USA MATERIALS & MECHANICS RESEARCH CENTER - Watertown, MA

Donald Neal
Tien-Yu Tsui

USA MISSILE COMMAND - Redstone Arsenal, AL

Siegfried H. Lehnigk
Romas A. Shatas

USA NATICK R&D COMMAND - Natick, MA

Edward W. Ross

USA NATIONAL RANGE OPERATIONS DIRECTORATE - White Sands Missile Range, NM

William D. Johnston
Ernest J. Sanchez

USA RESEARCH OFFICE - Durham, NC

Dr. Jagdish Chandra
Dr. Francis Dressel
Robert T. Launer
Edward Saibel

USA TANK-AUTOMOTIVE COMMAND - Warren, MI

Dr. James L. Thompson

USA TRADOC SYSTEMS ANALYSIS ACTIVITY - White Sands Missile Range, NM

Dr. H. M. Sassenfeld

USA WATERVLIET ARSENAL - Watervliet, NY (Benet Weapons Laboratory)

Peter C. T. Chen
Dr. Moayyed A. Hussain
Dr. John D. Vasilakis
Dr. Julian J. Wu

UNIVERSITY OF MINNESOTA - Minneapolis, MN

Prof. David H. Sattinger

UNIVERSITY OF TEXAS - Arlington, TX

Stephen R. Bernfeld
Dr. V. Lakshmikantham
S. Leela

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARO Report Number 78-1
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): TRANSACTIONS OF THE TWENTY-THIRD CONFERENCE OF ARMY MATHEMATICIANS

5. TYPE OF REPORT & PERIOD COVERED: Interim Technical Report

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s):

8. CONTRACT OR GRANT NUMBER(s):

9. PERFORMING ORGANIZATION NAME AND ADDRESS:

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS: Army Mathematics Steering Committee on Behalf of the Chief of Research Development and Acquisition

12. REPORT DATE: February 1978

13. NUMBER OF PAGES: 654

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): US Army Research Office DRXRO
PO Box 12211
Research Triangle Park, NC 27709

15. SECURITY CLASS. (of this report): Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited. The findings in this report are not to be considered as official Department of the Army position; unless so designated by other authorized documents.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This is a technical report resulting from the Twenty-Third Conference of Army Mathematicians. It contains most of the papers on the agenda of this meeting. These treat various Army applied mathematical problems.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Nonneutral plasmas
Moving boundary
Tube flow equations
Finite element analysis
Elastic stress
Plastic Deformation
Shock propagation
Navier-Stokes equations
Banach space problems
2nd Order Linear Differential Equations
Crack problems
Bivariational bounds
Computer graphics
Managerial control
Data reduction
Volterra Integrodifferential equations
Helicopter rotor blades
Rotating beams
Laplace Transforms
Irregular surfaces
Group theoretic methods
Accretive operators
Harmonic functions
Large Linear Systems
Boundary value problems
Moving-Weighted-Averages
Floating Ice Plates
Algebraic Methods

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

☆U.S. GOVERNMENT PRINTING OFFICE: 1978-761-695/7