# ARO Report 77-1

RIA-77-U982

# TRANSACTIONS OF THE TWENTY-SECOND CONFERENCE OF ARMY MATHEMATICIANS

TECHNICAL
LIBRARY




19971103 132

Approved for public release; distribution unlimited. The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

DTIC QUALITY INSPECTED 3

Sponsored by

The Army Mathematics Steering Committee

on behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT

AND ACQUISITION

US ARMY RESEARCH OFFICE

Report No. 77-1

February 1977

# TRANSACTIONS OF THE TWENTY-SECOND CONFERENCE OF ARMY MATHEMATICIANS

Sponsored by the Army Mathematics Steering Committee

Host

Watervliet Arsenal, Watervliet, New York

12-14 May 1976

Approved for public release; distribution unlimited.
The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

US Army Research Office

PO Box 12211

Research Triangle Park, North Carolina

# FOREWORD

On a continuing basis, the Army Mathematics Steering Committee (AMSC) sponsors three annual conferences. These meetings, in the areas of applied mathematics, numerical analysis and statistics, are designed to promote better communications among Army scientists. The oldest member of this series, the Conference of Army Mathematicians, held its twenty-second meeting at the Benet Weapons Laboratories, US Army Watervliet Arsenal, Watervliet, New York, on 13-14 May 1976. Dr. Moayyed A. Hussain, the Chairman on Local Arrangements, took this assignment seriously, and he, together with other members at Watervliet Arsenal, are due the thanks of all the attendees for an exceptionally well-planned meeting.

The ninth Conference of Army Mathematicians also had as its host Watervliet Arsenal. Statistics from these two meetings point out some of the changes taking place in these affairs. The ninth Conference had 65 attendees, while the present meeting entertained 94 persons. The 1963 meeting had one invited speaker and 24 contributed papers, while the 1976 Conference had 6 invited speakers and 44 contributed papers. The most encouraging statistic in these figures is the increase in the number of contributed articles. While 5 of the 44 papers in this class were given by University professors, this still leaves a sizable increase in the number of scientific papers being presented by Army scientists.

The Subcommittee on Applied Mathematics of the AMSC has charge of the planning of the Conference of Army Mathematicians. It selects invited speakers whose fields stress areas of applications of mathematics which meet the needs of the Army. It also selects some speakers that address fields which meet the special interests of the host installation. From the titles of the addresses of the invited speakers listed below, one may note that the requirements of the host in the area of fracture mechanics is stressed in several of these talks.

Nonlocal Elasticity and Fracture Mechanics
Professor A. C. Eringen, Princeton University

Unsteady Problems in Combustion Using Activation Energy Asymptotic
Professor John Buckmaster, University of Illinois

A Return to Input-Output Methods in Statistical Theory
Professor Thomas Kailath, Stanford University

Three-Dimensional Cracks and Weight Functions
Dr. Hans S. Bueckner, General Electric Company

Recent Developments in the Theory of Elasticity and Rupture of Fluid Infiltrated Solids
Professor James Rice, Brown University

In addition to the above speakers, Professor George H. Handelman of Rensselaer Polytechnic Institute gave an invited address at the banquet which was held on the first evening of the Conference.

Members of the AMSC were pleased that representatives of the Air Force, the Navy, and the Department of National Defence of Canada were in attendance at this symposium. They were also pleased to note the host installation had 22 of their staff members listening to the presented papers.

The last two articles appearing in these Transactions were not given at the Conference of Army Mathematicians. These papers, one by Dr. Achi Brandt and the other by Professor Gene H. Golub, resulted from invited addresses delivered at the 1976 Army Numerical Analysis and Computers Conference held 11-12 February 1976 at the US Army Research Office.

TABLE OF CONTENTS*


| TITLE | PAGE |
|---|---|
| Foreward | iii |
| Table of Contents | v |
| Program | ix |
| State of Stress in the Neighborhood of a Sharp Crack Tip<br>A. Cemal Eringen | 1 |
| Bending of a Cracked Strip Including Crack Surface Interference<br>O. L. Bowie and C. E. Freese | 19 |
| Singularity Analysis by the Finite Element Method<br>Dennis M. Tracey and Thomas S. Cook | 33 |
| Crack Tip Fields in Steady Crack Growth with Linear Strain Hardening<br>John C. Amazigo and John W. Huchinson | 51 |
| Finite-Difference Solution of Poisson's Equation in Rectangles of Arbitrary Proportions<br>J. Barkley Rosser | 53 |
| Solutions to Initial Value Problems Using Finite Elements - Unconstrained Variational Formulations<br>Julian J. Wu | 75 |
| The Numerical Solution of Free Boundary Problems by Mathematical Programming<br>Richard S. Sacher | 101 |
| A Numerical Integration Error Analysis Utilizing a Wronskian Technique<br>Larry A. Whatley and S. Bart Childs | 105 |
| Input Controllable Stochastic Model<br>Sheafen Frank Kuo | 115 |
| A Scanning Electron Microscope Investigation of Statically Loaded Foundation Materials<br>Raymond E. Aufmuth | 127 |
| Phase II Secure Voice Program - An Independent Army Analysis<br>Theodore S. Trybul | 139 |


*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.

Solving Control Problems Using Discrete Controls
Randy J. Schuetz and Bart Childs . . . . . 147

On Generalized Feller Equation
Siegfried H. Lehnigk . . . . . .157

A Pertubation Method for Free Boundary Problems of Elliptic Type
B. J. Fleishman and Thomas J. Mahar . . . . . 159

Determination of Propagation Constants in Scattering from Dielectric-Coated Wires
Leon Kotin . . . . . .169

Activation Energy Asymptotics and Unsteady Flames
J. Buckmaster and G. S. S. Ludford . . . . . .183

A Model for Shock Induced Structural Transformations
Paul Harris . . . . . 203

Some New Methods for Solving Linear Equations
Thomas Kailath . . . . . .211

An Exact Solution to an Elastic-Plastic Deformation Problem in a Radially Stressed Annular Plate
Peter C. T. Chen . . . . . .227

An Effective Stiffness Viscoelastic Composite Beam Theory
Charles R. Thomas . . . . . 239

Using Fast Transforms to Compute the Weight Distribution of a Linear Code
Bart F. Rice . . . . . 273

Factorial and Hadamard Series for Bessel Functions of Orders Zero and One
Alexander S. Elder . . . . . 277

Finite and Infinite Inhomogeneous Ladder Networks
C. C. Yang and T. N. Lee . . . . . 289

Automatic Numerical Integration Using VP-Splines
Royce W. Soanes, Jr. . . . . .313

Time Evolution of an Orthogonal Matrix
James M. Wilkes . . . . . 325

The Weight Functions of Mode I of the Penny-Shaped and of the Elliptic Crack
Hans F. Bueckner . . . . . 335

The Buckling Pressure of an Elastic Plate Floating on Water and Stressed Uniformly Along the Periphery of an Internal Hole
Shunsuke Takagi . . . . . 357

Nonlinear Theory of the Response of Pavements to Vibratory Loads
Richard A. Weiss . . . . . 419

Characterization of Behind Armor Effects for Long Rod Penetrators
Victor D. Maki . . . . . 467

Mathematical Models of Systems and Tactics in Land Combat
Roger F. Willis . . . . . 473

Evaluation of Several 'Best Fit' Methods as they Pertain to the Superposition of Solutions in a Multipoint Boundary Value Program
John H. Walker and S. Bart Childs . . . . . 485

A Statistical Study of Numerical Analysis Applied to the Regression of nth Order Differential Equations
Craig D. Hunter and S. Bart Childs . . . . . 497

Multi-Level Adaptive Solutions to Boundary-Value Problems
Achi Brandt . . . . . 509

Singular Value Decomposition: Applications and Computations
Gene H. Golub and Franklin T. Luk . . . . . 577

List of Attendees . . . . . 607

PROGRAM

THE 22nd CONFERENCE OF ARMY MATHEMATICIANS
Maggs Research Center, Watervliet Arsenal
Watervliet, New York

All general and technical sessions will be held in Rooms 240 and 215, on the second floor of Maggs Research Center, Bldg. 115, Watervliet Arsenal, Watervliet, New York

Wednesday, 12 May 1976

0745 BUS FROM HOLIDAY INN TO WATERVLIET ARSENAL

0800-0830 REGISTRATION - RECEPTION LOUNGE, 1st FLOOR, MAGGS RESEARCH CENTER

0830-0845 OPENING OF THE CONFERENCE, WELCOMING REMARKS - ROOM 240

0845-0945 GENERAL SESSION I - ROOM 240

SPEAKER: Professor A. Cemal Eringen
School of Engineering and Applied Science
Princeton University
Princeton, New Jersey

TITLE: Nonlocal Elasticity and Fracture Mechanics

CHAIRMAN: Dr. E. A. Saibel
US Army Research Office
P.O. Box 12211
Research Triangle Park, North Carolina

0945-1000 BREAK

Wednesday AM

1000-1200 TECHNICAL SESSION I - ROOM 240

CHAIRMAN: Dr. T. Davidson
Chief, Materials Engineering Division
Benet Weapons Laboratory
Watervliet Arsenal, Watervliet, New York

BENDING OF A CRACKED STRIP INCLUDING CRACK SURFACE INTERFERENCE
O. L. Bowie and C. E. Freese, Army Materials and Mechanics Research Center, Watertown, Massachusetts

DYNAMIC FRACTURE UNDER SHOCK LOADING CONDITIONS
John F. Mescall, Army Materials and Mechanics Research Center, Watertown, Massachusetts

SINGULARITY ANALYSIS BY THE FINITE ELEMENT METHOD
Dennis M. Tracey and Thomas S. Cook, Army Materials and Mechanics Research Center, Watertown, Massachusetts and Southwest Research Institute, San Antonio, Texas, respectively

SINGULAR BEHAVIOR AT THE TIP OF A GROWING CRACK IN A BILINEAR ELASTIC-PLASTIC MATERIAL
John C. Amazigo and John W. Hutchinson, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York and Division of Engineering and Applied Physics, Harvard University, Cambridge, Massachusetts, respectively

ASSESSMENT OF STRENGTH-PROBABILITY-TIME RELATIONSHIPS IN CERAMICS
Edward M. Lenoe and Donald M. Neal, Army Materials and Mechanics Research Center, Watertown, Massachusetts

1000-1200 TECHNICAL SESSION II - ROOM 215

CHAIRMAN: Dr. Aivars Celmins
Chief of Fluid Mechanics Branch
Applied Mathematics and Science Lab
Ballastic Research Laboratory
Aberdeen Proving Ground, Maryland

FINITE-DIFFERENCE SOLUTION OF POISSON'S EQUATION IN RECTANGLES OF ARBITRARY PROPORTIONS
J. Barkley Rosser, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

## Wednesday AM

1000-1200 TECHNICAL SESSION II - ROOM 215 (Continued)

ON A GENERAL METHOD FOR GENERAL PURPOSE HEAT DIFFUSION EQUATION
R. Yalamanchili, GEN Thomas J. Rodman Laboratory, Rock Island Arsenal, Rock Island, Illinois

SOLUTIONS TO INITIAL VALUE PROBLEMS USING FINITE ELEMENTS - UNCONSTRAINED VARIATIONAL FORMULATIONS
Julian J. Wu, Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

THE NUMERICAL SOLUTION OF FREE-BOUNDARY PROBLEMS BY MATHEMATICAL PROGRAMMING
R. S. Sacher, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York

A NUMERICAL INTEGRATION ERROR ANALYSIS UTILIZING A WRONSKIAN TECHNIQUE
Lawrence A. Whatley and S. Bart Childs, Intern Training Center, DARCOM, Alexandria, Virginia, and Texas A&M University, Texakana, Texas

## Wednesday PM

1200-1315 LUNCH (OFFICERS' CLUB)

1315-1515 TECHNICAL SESSION III - ROOM 240

CHAIRMAN: Roger F. Willis
US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico

AN INPUT CONTROLLABLE PROBABILITY MODEL
Frank Kuo, US Army Construction Engineering Research Laboratory, Champaign, Illinois

A SCANNING ELECTRON MICROSCOPE STUDY OF STATICALLY LOADED FOUNDATION MATERIALS
Raymond E. Aufmuth, US Army Construction Engineering Research Laboratory, Champaign, Illinois

Wednesday PM

1315-1515 TECHNICAL SESSION III - ROOM 240 (Continued)

PHASE II SECURE VOICE PROGRAM - AN INDEPENDENT ARMY ANALYSIS
Theodore S. Trybul, DARCOM, Alexandria, Virginia

SOLVING CONTROL PROBLEMS USING DISCRETE CONTROLS
Randy J. Schuetz and S. Bart Childs, Intern Training Center, DARCOM, Alexandria, Virginia, and Texas A&M University, Texakana, Texas

1315-1515 TECHNICAL SESSION IV - ROOM 215

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

ON THE GENERALIZED FELLER EQUATION
Siegfried H. Lehnigk, US Army Missile Command, Redstone Arsenal, Alabama

A PERTURBATION METHOD FOR FREE BOUNDARY PROBLEMS OF ELLIPTIC TYPE
B. A. Fleishman and Thomas J. Mahar, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York

CONSTITUTIVE EQUATIONS FOR TWO-PHASE FLOW
Donald A. Drew, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York

DETERMINATION OF PROPAGATION CONSTANTS IN SCATTERING FROM DIELECTRIC-COATED WIRES
Leon Kotin, US Army Electronics Command, Fort Monmouth, New Jersey

1515-1530 BREAK

1530-1630 GENERAL SESSION II - ROOM 240

SPEAKER: Professor John Buckmaster
Mathematics Department
University of Illinois
Urbana, Illinois

Wednesday PM

1530-1630 GENERAL SESSION II - ROOM 240 (Continued)

TITLE: Unsteady Problems in Combustion Using Activation Energy Asymptotic

CHAIRMAN: Dr. Donald Eccleshall
Chief, Applied Mathematics and Science Lab
Ballastic Research Laboratory
Aberdeen Proving Ground, Maryland

Wednesday Evening

1800 BANQUET - OFFICERS' CLUB

SPEAKER: Professor George H. Handelman
Dean, School of Science
Rensselaer Polytechnic Institute
Troy, New York

MASTER OF CEREMONY: Dr. F. W. SCHMIEDESHOFF
Director of Research, Watervliet Arsenal
Watervliet, New York

**********

Thursday, 13 May 1976

0800 BUS FROM HOLIDAY INN TO WATERVLIET ARSENAL

0830-1030 TECHNICAL SESSION V - ROOM 240

CHAIRMAN: Dr. Alma Gray
Physical Sciences Division
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York

Thursday AM

0830-1030 TECHNICAL SESSION V - ROOM 240 (Continued)

THE STRUCTURE OF GROUPS WITH INDEX-3 SUBGROUPS AND LANDAU'S SECOND THEOREM
L. V. Meisel*, D. M. Gray* and E. Brown**
*Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York
**Department of Physics, Rensselaer Polytechnic Institute, Troy, New York

PHASE-SPACE TRANSLATIONAL AND PERTURBATION METHODS IN NONRELATIVISTIC QUANTUM ELECTRODYNAMICS AND THEIR APPLICATION TO LASERS
R. A. Shatas, S. S. Mitra, and W. C. Henneberger, Quantum Physics, Physical Sciences Directorate, Redstone Arsenal, Alabama

A MODEL FOR SHOCK INDUCED PHASE TRANSFORMATIONS
Paul Harris, Concepts and Effectiveness Division, Picatinny Arsenal, Dover, New Jersey

BIFURCATION PROPERTIES OF LASER MODEL HAMILTONIANS
Charles M. Bowden and R. Gilmore, Quantum Physics, Physical Sciences Directorate, Redstone Arsenal, Alabama and Institut de Physique Théorique, Université de Louvain, B-1348 Louvain-La-Neuve, Belgium, respectively

0830-1030 TECHNICAL SESSION VI - ROOM 215

CHAIRMAN: Dr. Siegfried H. Lehnigk
Physical Sciences Directorate
US Army Missile Command
Redstone Arsenal, Alabama

A PULSATING REACTION FRONT IN SOLID FUEL COMBUSTION
B. J. Matkowsky and G. I. Sivashinsky, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York

STABILITY THEORY FOR SIMPLE FLUIDS
M. Slemrod, Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York

## Thursday AM

0830-1030 TECHNICAL SESSION VI - ROOM 215 (Continued)

EXACT METHODS IN HEAT TRANSFER PROBLEMS
John F. Polk, Detonation and Deflagration Dynamics Laboratory, US Army Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

EXTREMUM VARIATIONAL PRINCIPLES FOR LINEAR DIFFUSION-TYPE EQUATIONS
Ben Noble, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

1030-1045 BREAK

1045-1145 GENERAL SESSION III - ROOM 240

SPEAKER: Professor Thomas Kailath
Department of Electrical Engineering
Stanford University
Stanford, California

TITLE: A RETURN TO INPUT-OUTPUT METHODS IN STATISTICAL SYSTEM THEORY

CHAIRMAN: Dr. Merle M. Andrew
Head, Mathematical Sciences Division
Air Force Office of Scientific Research
Bolling Air Force Base
Washington, D. C.

1145-1300 LUNCH (OFFICERS' CLUB)

1300-1515 TECHNICAL SESSION VII - ROOM 240

CHAIRMAN: John F. Mescall
Army Materials and Mechanics Research Center
Watertown, Massachusetts

FINITE ORTHOTROPIC PLATE WITH CIRCULAR HOLE LOADED BY FRICTIONLESS RIGID INCLUSION
K. R. Gandhi, Army Materials and Mechanics Research Center, Watertown, Massachusetts

Thursday PM

1300-1515 TECHNICAL SESSION VII - ROOM 240 (Continued)

AN EXACT SOLUTION TO AN ELASTIC-PLASTIC DEFORMATION PROBLEM IN A RADIALLY STRESSED ANNULAR PLATE
Peter C. T. Chen, Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

A PROBABILISTIC THEORY OF THE INTRINSIC TIME TO FRACTURE
K. C. Valanis, Division of Materials Engineering, University of Iowa, Iowa City, Iowa

FINITE ELEMENTS FOR ELASTIC-PLASTIC ANALYSIS AND ITS APPLICABILITY TO DUCTILE FRACTURE
T. P. Rich, Army Materials and Mechanics Research Center, Watertown, Massachusetts

AN EFFECTIVE STIFFNESS VISCOELASTIC COMPOSITE BEAM THEORY
Charles R. Thomas, Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

1300-1515 TECHNICAL SESSION VIII - ROOM 215

CHAIRMAN: Dr. Leon Kotin
Communication/Automatic Data Processing Lab
US Army Electronics Command
Fort Monmouth, New Jersey

USING FAST TRANSFORMS TO COMPUTE THE WEIGHT DISTRIBUTION OF A LINEAR CODE
Bart F. Rice, National Security Agency, Fort Meade, Maryland

FACTORIAL AND HADAMARD SERIES FOR BESSEL FUNCTIONS OF ORDERS ZERO AND ONE
Alexander S. Elder and Emma M. Wineholt, US Army Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

ON A CLASS OF FINITE AND INFINITE NONUNIFORM CONTINUED FRACTIONS
T. N. Lee and C. C. Yang, Department of E.E. and C.S., The George Washington University, Washington, D. C. and Applied Mathematics Division, Naval Research Laboratory, Washington, D. C., respectively

## Thursday PM

1300-1515 TECHNICAL SESSION VIII - ROOM 215 (Continued)

AUTOMATIC NUMERICAL INTEGRATION USING VP-SPLINES
Royce W. Soanes, Jr., Benet Weapons Laboratory, Watervliet Arsenal, Watervliet, New York

TIME EVOLUTION OF AN ORTHOGONAL MATRIX
James M. Wilkes, Army Materiel Test and Evaluation Directorate, White Sands Missile Range, New Mexico

1515-1530 BREAK

1530-1630 GENERAL SESSION IV - ROOM 240

SPEAKER: Dr. Hans S. Bueckner
Turbine Department, General Electric Company
Schenectady, New York

TITLE: Three-Dimensional Cracks and Weight Functions

CHAIRMAN: Professor Ben Noble
Director, Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

**********

## Friday, 14 May 1976

0800 BUS FROM HOLIDAY INN TO WATERVLIET ARSENAL

0830-1030 TECHNICAL SESSION IX - ROOM 240

CHAIRMAN: Dr. San-Li Pu
Applied Mathematics and Mechanics Division
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York

0830-1030 TECHNICAL SESSION IX - ROOM 240 (Continued)

THE BUCKLING PRESSURE OF AN ELASTIC PLATE FLOATING ON WATER AND STRESSED UNIFORMLY ALONG THE PERIPHERY OF AN INTERNAL HOLE
Shunsuke Takagi, US Army Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire

NONLINEAR THEORY OF THE RESPONSE OF PAVEMENTS TO VIBRATORY LOADS
Richard A. Weiss, US Army Engineer Waterways Experiment Station, Vicksburg, Mississippi

STABILITY ANALYSIS OF A HIGH-SPEED SLIDER-CRANK MECHANISM WITH AN ELASTIC CONNECTING ROD
Shih-Chi Chu and K. C. Pan, GEN Thomas J. Rodman Laboratory, Rock Island Arsenal, Rock Island, Illinois

CHARACTERIZATION OF BEHIND ARMOR EFFECTS FOR LONG ROD PENETRATORS
Victor D. Maki, US Army Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland

0830-1030 TECHNICAL SESSION X - ROOM 215

CHAIRMAN: Dr. Badrig M. Kurkjian
US Army Material Development Readiness Command
DARCOM
Alexandria, Virginia

STABILITY OF SOLUTIONS OF THE LINEAR COMPLEMENTARITY PROBLEM
Stephen M. Robinson, Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

MODELS OF SYSTEMS AND TACTICS IN COMBAT
Roger F. Willis, US Army TRADOC Systems Analysis Activity, White Sands Missile Range, New Mexico

EVALUATION OF SEVERAL "BEST FIT" METHODS AS THEY PERTAIN TO THE SUPERPOSITION OF SOLUTIONS IN A MULTIPOINT BOUNDARY VALUE PROGRAM
John Walker and S. Bart Childs, Intern Training Center, DARCOM, Alexandria, Virginia and Texas A&M University, Texakana, Texas

## Friday AM

0830-1030 TECHNICAL SESSION X - ROOM 215 (Continued)

A STATISTICAL STUDY OF NUMERICAL ANALYSIS APPLIED TO THE REGRESSION OF N-th ORDER DIFFERENTIAL EQUATIONS
Craig D. Hunter and S. Bart Childs, Intern Training Center, DARCOM, Alexandria, Virginia and Texas A&M University, Texakana, Texas

1030-1045 BREAK

1045-1145 GENERAL SESSION V - ROOM 240

SPEAKER: Professor James Rice
Engineering Division
Brown University
Providence, Rhode Island

TITLE: RECENT DEVELOPMENTS IN THE THEORY OF ELASTICITY AND RUPTURE OF FLUID INFILTRATED SOLIDS

CHAIRMAN: Dr. Robert E. Weigle
Director, Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York

1200 ADJOURN

# STATE OF STRESS IN THE NEIGHBORHOOD OF A SHARP CRACK TIP*

A. Cemal Eringen

Princeton University

ABSTRACT

Field equations of nonlocal elasticity are solved to determine the state of stress in the neighborhood of a line crack in an elastic plate subject to uniform tension perpendicular to the line of crack at infinity. It is found that no stress singularity is present at the crack tip. When the maximum hoop stress is equated to the cohesive stress Griffith criterion of fracture is obtained with the Griffith constant fully determined. Cohesive stress necessary to break the atomic bonds are calculated for Aℓ, Ni, Fe, LiF, Diamond and Zn. The results are in excellent agreement with those known in the atomic theory of lattices and experiments.

## 1. INTRODUCTION

The determination of the state of stress near the tip of a sharp crack in an elastic plate subject to uniform tension perpendicular to the line of crack at infinity, Fig. 1, is one of the most fundamental problems in fracture mechanics. The solution of this problem was first given by Inglis [1913] and it was used by Griffith [1920] to establish his celebrated criterion for fracture of solids. The classical elasticity solution of this problem gives a hoop stress with a $\sqrt{r}$ singularity near the crack tip, where r is the distance from the crack tip. Thus, according to classical elasticity the stress is infinite at the crack tip for even a minute amount of applied tension. Since a plate with a sharp crack possesses a certain amount of resistance to fracture until the applied tension, $t_o$, reaches a critical value determined by the so-called Griffith criterion

$$t_o{}^2 \ell = C_G \tag{1.1}$$

where $\ell$ is the half crack length and $C_G$ is an experimental constant (Griffith constant), it must be concluded that classical elasticity solution fails to apply near the crack tip. This conclusion is responsible for the abandonment of maximum stress hypothesis for failure which has been prominent in structural mechanics. Consequently, for brittle solids, since the time of Griffith, two distinct fracture criteria have been in use, one for structural members with no cracks and one for those containing cracks. In fact, the state of the art is more involved, far beyond this dichotomy, and many other fracture criteria have been introduced by other authors to over-

---

*The present work was supported by The Army Research Office at Durham.

General lecture presented (under the title of "Nonlocal Elasticity and Fracture Mechanics") at the 22nd Conference of Army Mathematicians, Maggs Research Center, Watervliet Arsenal, Watervliet, NY, May 12-14, 1976.

come this stress singularity (e.g., J-integral, Barenblatt theory [1962] Khristianowich [1955] -Dugdale theory [1960], Goodier & Kanninen locally nonlinear theory [1966], etc.). Below we give a brief discussion of these theories. A thorough discussion of the status of the art is to be found in Goodier's [1968] article.

Griffith Criterion. Griffith assumes that the work done to extend a line crack of length $2\ell$ an amount of $2d\ell$ must be equal to the work of the surface tension. In this way he arrived at the formula (1.1) with

$$C_G = \frac{2E}{\pi(1-\nu^2)}\gamma \tag{1.2}$$

where E is the Young's modulus, $\nu$ is the Poisson's ratio for an isotropic elastic plate and $\gamma$ is the surface tension energy. The surface tension energy $\gamma$ he employed is that borrowed from fluid statics. In obtaining (1.1), the Inglis' solution for the elliptic hole was used with the provision that in the limit the minor axis of the ellipse approach to zero. This theory has been under criticism for over half a century nevertheless surviving all criticisms. Basic complaints may be summarized as:

(i) Crack tip stress is infinite no matter how small the applied load is.

(ii) The crack opens up into an ellipse, so that the shear strain at the tip is too large ($\pi/4$) for the linear theory to be applicable.

(iii) Ellipse shrinking to a crack may not be "uniform," mathematically, i.e., other shapes may give different limits.

(iv) The surface tension energy $\gamma$ borrowed from fluid statics may not be appropriate for solids.

Barenblatt Model. To overcome the objections (i) and (ii) Barenblatt [1962] assumed that the tip region of the crack is not free of tractions but there exists a "cohesive stress," $\sigma(x)$, distributed in such a way as to bring the crack tip to a cusp, Fig. 2. He then determined the shape of $\sigma(x)$ to achieve the cusp form.

Khristianowich-Dugdale Model. Khristianovich [1955] and Dugdale [1966] assumed that beyond the crack tip over a small length s there is a constant cohesive stress distribution to close up ends of the crack, Fig. 3.

Clearly both Barenblatt and Khristianwich-Dugdale theories are objectionable for their uses of heuristic assumptions not justifiable on the basis of any physical principles or experimental work.

Goodier-Kanninen Model. According to Goodier and Kanninen [1966] the atomic interactions are important at the tip of a crack. In order to overcome the objection (iii) they use nonlinear springs along the tip of the crack. The extent of nonlinear springs and their properties are left to our discretion. While the basic idea of inclusion of long range interatomic interactions are worthy of careful attention the model contains arbitrary factors and functions to be fixed to suit the purpose.

Remarkably common to all these models is the unequivocable realization that *near the crack tip interatomic cohesive forces must be important.*

There exist solutions of Inglis' problem by using polar theories, e.g., couple stress theory (Sternberg and Muki [1967]), micropolar theory (Kim and Eringen [1973]). These results also contain the same type of singularities and therefore no further progress is possible on these grounds.

Recently we have developed a continuum theory that takes into account the effect of long range interatomic attractions. According to this theory the stress at a point of an elastic solid is influenced by the strains at *all* points of the body. All known physical and thermodynamic principles were satisfied (cf., Eringen[1972a,b], Eringen & Edelen [1972]). When the nonlocal theory is employed for the solution of the crack tip problem one finds that the stress field at the crack tip is no longer singular and therefore it is possible to revert back to the maximum stress hypothesis for fracture criterion. Remarkably enough this theory not only gives Griffith's criterion without any new assumption but also determines the Griffith constant. In fact the cohesive stress calculated for various materials are in excellent agreement with those known from the atomic theory of lattices and experiments. The main purpose of the present paper is an exposition of these results.

## 2. BASIC EQUATIONS OF NONLOCAL ELASTICITY

Basic equations of linear, homogeneous, isotropic, nonlocal elastic solids with vanishing body and inertia forces are (cf., Eringen [1972b])

$$t_{k\ell,k} = 0 \tag{2.1}$$

$$t_{k\ell} = \int_V [\lambda'(|\underset{\sim}{x}'-\underset{\sim}{x}|)\, e_{rr}(\underset{\sim}{x}')\, \delta_{k\ell} + 2\mu'(|\underset{\sim}{x}'-\underset{\sim}{x}|)\, e_{k\ell}(\underset{\sim}{x}')]\, dv(\underset{\sim}{x}') \tag{2.2}$$

$$e_{k\ell} = \tfrac{1}{2}(u_{k,\ell} + u_{\ell,k}) \tag{2.3}$$

where the only difference from classical elasticity is in the stress constitutive equations (2.2) which states that the stress $t_{k\ell}(\underset{\sim}{x})$ at a point $\underset{\sim}{x}$ depends on strains, $e_{k\ell}(\underset{\sim}{x}')$, at *all* points of the body. For homogeneous and isotropic solids the material moduli $\lambda'(|\underset{\sim}{x}'-\underset{\sim}{x}|)$ and $\mu'(|\underset{\sim}{x}'-\underset{\sim}{x}|)$ are functions of the distance between the points $\underset{\sim}{x}'$ and $\underset{\sim}{x}$. The integral in (2.2) is over the volume $V$ of the body enclosed within the surface $\partial V$.

Here and throughout we employ Cartesian tensors with repeated indices that indicate summation over the range (1,2,3) and indices following a comma partial differentiation, with respect to space coordinates, e.g.

$$u_{k,\ell} \equiv \partial u_k / \partial x_\ell$$

In our previous work [1972b, 1974] we have obtained the form of $\lambda'(|\underset{\sim}{x}'-\underset{\sim}{x}|)$ and $\mu'(|\underset{\sim}{x}'-\underset{\sim}{x}|)$ for which the dispersion curves of plane waves coincide with those obtained in Born-von Kármán theory of lattice dynamics within the entire Brillouin zone. Accordingly

$$(\lambda',\ \mu') = (\lambda,\mu)\ \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)$$

$$\alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|) = \begin{cases} \alpha_o(a - |\underset{\sim}{x}'-\underset{\sim}{x}|) & , \quad |\underset{\sim}{x}'-\underset{\sim}{x}| \leq a \\ 0 & , \quad |\underset{\sim}{x}'-\underset{\sim}{x}| > a \end{cases} \tag{2.4}$$

where a is the lattice parameter, $\lambda$ and $\mu$ classical Lamé constants and $\alpha_o$ is a normalization constant to be determined from

$$\int_V \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)\ dv(\underset{\sim}{x}') = 1 \tag{2.5}$$

Since the nonlocal effects are most important along the edge of the crack we use (2.4) and (2.5) at $x_2 = 0$ to determine $\alpha_o$. This gives $\alpha_o = 6/\pi a^3$. Upon carrying (2.4) into (2.2) we will have

$$t_{k\ell} = \int_V \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)\ \sigma_{k\ell}(\underset{\sim}{x}')\ dv(\underset{\sim}{x}') \tag{2.6}$$

where

$$\sigma_{k\ell}(\underset{\sim}{x}') \equiv \lambda\ e_{rr}(\underset{\sim}{x}')\ \delta_{k\ell} + 2\mu\ e_{k\ell}(\underset{\sim}{x}') \tag{2.7}$$

is the classical Hooke's law.

Substituting (2.6) into (2.1) and using the identity

$$\alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)_{,k}\ \sigma_{k\ell}(\underset{\sim}{x}') = -\alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)_{,k'}\ \sigma_{k\ell}(\underset{\sim}{x}')$$

$$= -(\alpha\ \sigma_{k\ell})_{,k'} + \alpha\ \sigma_{k\ell,k'}$$

and Green-Gauss theorem we obtain

$$-\oint_{\partial V} \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)\ \sigma_{k\ell}(\underset{\sim}{x}')\ da_k(\underset{\sim}{x}') + \int_V \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)\ \sigma_{k\ell,k'}(\underset{\sim}{x}')\ dv(\underset{\sim}{x}') = 0 \tag{2.8}$$

Here the surface integral may be dropped if the effect of the surface tensions are negligible or the body extends to infinity in all directions. We assume this is the case so that

$$\int_V \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)\ \sigma_{k\ell,k'}(\underset{\sim}{x}')\ dv(\underset{\sim}{x}') = 0 \tag{2.9}$$

It is not difficult to prove that if $\alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)$ has a bounded support and $\sigma_{k\ell,k}$ is continuous in $V$ then the necessary and sufficient condition for (2.9) to be satisfied is, cf., Eringen [1976]

$$\sigma_{k\ell,k} = 0 \tag{2.10}$$

Equation (2.10) together with (2.7) are none other than Navier's equation for the displacement field $\underset{\sim}{u}(\underset{\sim}{x})$. From this result it follows that

*Theorem. The displacement field of the nonlocal elasticity (under the conditions stated above) satisfy Navier's equation.*

For the displacement boundary-value problem (1st boundary-value problem) this implies that:

*Corollary. The displacement field of the first boundary value problem of the nonlocal elasticity is identical to that of the classical elasticity.*

Note, however, that to obtain the stress field we must substitute $\sigma_{k\ell}$ obtained from the classical theory into (2.6) and carry out the volume integration. Thus, for boundary conditions on the tractions we must employ

$$t_{k\ell} n_k = t_\ell \qquad \text{on } \partial V_{\underset{\sim}{t}} \tag{2.11}$$

on that part of the surface $\partial V_{\underset{\sim}{t}}$ where the traction $t_\ell$ is prescribed.

## 3. CRACK PROBLEM

Consider a plate weakened by a sharp line crack of length $2\ell$. The plate is subject to a uniform compression $t_o$ at the crack surface and free of tractions at infinity. The displacement field $u_1 = u(x,y)$, $u_2 = v(x,y)$ in the upper half plane $y > 0$ are given by the classical elasticity solution (cf., Sneddon [1951, p. 404]).

$$\begin{aligned} u &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{i}{k} \left[ |k|\, A(k) + \left( |k| y - \frac{\lambda+3\mu}{\lambda+\mu} \right) B(k) \right] \exp(-|k|y - ikx)\, dk \\ v &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[ A(k) + y\, B(k) \right] \exp(-|k|y - ikx)\, dk \end{aligned} \tag{3.1}$$

where A(k) and B(k) are two functions to be determined from the boundary conditions at $y = 0$. These conditions are:

$$\begin{aligned} & t_{yx} = 0 && , && y = 0 && , && \forall\ x \\ & t_{yy} = -t_o && , && y = 0 && , && |x| < \ell \\ & v = 0 && , && y = 0 && , && |x| \geq \ell \end{aligned} \tag{3.2}$$

To obtain the solution of the crack problem with crack surface free of tractions and the plate is subject to a uniform tension $t_{yy} = t_o$ at $y = \infty$ (Fig. 1) to the solution of the above problem we superimpose a uniform stress field $t_{yy} = t_o$.

Substituting (3.1) into (2.3) and (2.7) we calculate:

$$\sigma_{yy}(x',y') = -\frac{2\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[|k|\ A(k) - \left(\frac{\mu}{\lambda+\mu} - |k|y'\right) B(k)\right] \exp(-|k|y' - ikx')\ dk$$

(3.3)

$$\sigma_{yx}(x',y') = -\frac{2i\mu}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[k\ A(k) + \left(k\ y' - \frac{|k|}{k}\frac{\lambda+2\mu}{\lambda+\mu}\right) B(k)\right] \exp(-|k|y' - ikx')\ dk$$

According to (2.6) then we have

$$t_{yy}(x,y) = \int_0^{\infty}\int_{-\infty}^{\infty} \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)\ \sigma_{yy}(x',y')\ dx'\ dy' \quad ,$$

(3.4)

$$t_{yx}(x,y) = \int_0^{\infty}\int_{-\infty}^{\infty} \alpha(|\underset{\sim}{x}'-\underset{\sim}{x}|)\ \sigma_{yx}(x',y')\ dx'\ dy'$$

Substituting (3.3) and (2.4) into (3.4) and after carrying out integrations on x' and y', we set y = 0 in these equations and in $(3.1)_2$ to form the boundary conditions (3.2). As in classical treatment $(3.2)_1$, can be used to determine B(k) in terms of A(k). The process is lengthy and tedious. We only give the resulting expression

$$\begin{aligned}(3.5)\qquad B(k) = k\Big[&\left(\frac{1}{3}k^2a^2 + \frac{4}{3}\right)\cos(ka) + \frac{1}{3}ka\sin(ka) \\ &+ \frac{1}{3}k^3a^3\,\mathrm{Si}(ka) - \frac{1}{6}\pi k^3a^3 - \frac{4}{3}\Big] A(k) \Big/ \Big[\Big(\frac{1}{3}k^2a^2\frac{\lambda+2\mu}{\lambda+\mu} \\ &+ \frac{4}{3}\frac{\lambda+2\mu}{\lambda+\mu} - \frac{1}{10}k^2a^2 - \frac{4}{5} + \frac{1}{20}k^4a^4\Big)\cos(ka) \\ &+\Big(\frac{1}{3}ka\frac{\lambda+2\mu}{\lambda+\mu} - \frac{3}{10}ka + \frac{1}{20}k^3a^3\Big)\sin(ka) + \frac{1}{3}k^3a^3\frac{\lambda+2\mu}{\lambda+\mu} \\ &+ \frac{1}{20}k^5a^5\Big)\mathrm{Si}(ka) - \frac{1}{6}\pi k^3a^3\frac{\lambda+2\mu}{\lambda+\mu} - \frac{4}{3}\frac{\lambda+2\mu}{\lambda+\mu} \\ &- \frac{1}{40}\pi k^5a^5 + \frac{4}{5}\Big]\end{aligned}$$

where Si(z) is the sine integral defined by

$$\mathrm{Si}(z) = \int_0^z \frac{\sin t}{t}\,dt$$

With B(k) given by (3.5), the boundary condition $(3.2)_1$ is satisfied and $(3.2)_2$ and $(3.2)_3$ lead to

$$(2/\pi)^{\frac{1}{2}} \int_0^\infty A(k)\cos(kx)\,dk = 0 \qquad , \qquad x \geq \ell$$

$$(2/\pi)^{\frac{1}{2}} \int_0^\infty k\,\bar{\alpha}(ka)\,A(k)\cos(kx)\,dk = T_o \qquad , \qquad x < \ell \tag{3.6}$$

where

$$T_o \equiv t_o(\lambda+2\mu)/2\mu(\lambda+\mu) \tag{3.7}$$

$$\begin{aligned}\bar{\alpha}(ka) = -\frac{6}{\pi}\Bigg\{&\Bigg[\left(\frac{1}{3}k^2a^2+\frac{4}{3}\right)\cos(ka)+\frac{1}{3}ka\sin(ka)\\&+\frac{1}{3}k^3a^3\,\mathrm{Si}(ka)-\frac{1}{6}\pi k^3a^3-\frac{4}{3}\Bigg]^2\Big/\Bigg[\left(\frac{1}{3}k^5a^5\right.\\&+\frac{4}{3}k^3a^3-\frac{1}{10}k^5a^5\,\frac{\lambda+\mu}{\lambda+2\mu}-\frac{4}{5}k^3a^3\,\frac{\lambda+\mu}{\lambda+2\mu}\\&\left.+\frac{1}{20}k^7a^7\,\frac{\lambda+\mu}{\lambda+2\mu}\right)\cos(ka)+\left(\frac{1}{3}k^4a^4-\frac{3}{10}k^4a^4\,\frac{\lambda+\mu}{\lambda+2\mu}\right)\\&+\frac{1}{20}k^6a^6\,\frac{\lambda+\mu}{\lambda+2\mu}\Big)\sin(ka)+\left(\frac{1}{3}k^6a^6+\frac{1}{20}k^8a^8\,\frac{\lambda+\mu}{\lambda+2\mu}\right)\mathrm{Si}(ka)\\&-\frac{1}{6}\pi k^6a^6-\frac{4}{3}k^3a^3-\frac{1}{40}\pi k^8a^8\,\frac{\lambda+\mu}{\lambda+2\mu}+\frac{4}{5}k^3a^3\,\frac{\lambda+\mu}{\lambda+2\mu}\Bigg]\Bigg\}\end{aligned}$$

The dual integral equations (3.6) must be solved to determine A(k). When this is done, we will have the problem solved.

It is interesting to note that in the continuum limit $a \to 0$ $\bar{\alpha} \to 1$ and (3.6) revert the dual integral equations obtained in classical elasticity for the same problem. With a complicated kernel function $\bar{\alpha}(ka)$ the solution of (3.6) cannot be affected in closed form. However, we can take advantage of the known classical solution to reduce the problem to a Fredholm integral equation which is more amenable to numerical treatment. To this end let $A_c(k)$ denote the solution of the dual integral equations of the classical theory

$$(2/\pi)^{\frac{1}{2}} \int_0^\infty A_c(k)\cos(kx)\,dk = 0 \qquad , \qquad x \geq \ell$$

$$(2/\pi)^{\frac{1}{2}} \int_0^\infty k\,A_c(k)\cos(kx)\,dk = T_o \qquad , \qquad x < \ell \tag{3.8}$$

Subtracting (3.8) from (3.6) we will have

$$\int_0^\infty [A(k) - A_c(k)]\cos(kx)\,dk = 0 \qquad , \qquad x \geq \ell$$

$$\int_0^\infty k[A(k) - A_c(k)]\cos(kx)\,dk = \int_0^\infty k\,[1-\bar{\alpha}(ka)]\,A(k)\cos(kx)\,dk \qquad , \qquad x < \ell$$

Treating the right-hand side of these equations as known, we copy the solution of these equations from Sneddon [1951, p. 70].

$$A(k) - A_c(k) = (2\ell^2/\pi)\left[J_o(k\ell)\int_0^1 (1-\eta^2)^{\frac{1}{2}}\left\{\int_0^\infty \zeta[1-\bar{\alpha}(\zeta a)]\right.\right.$$

$$\left.A(\zeta)\cos(\zeta\eta\ell)\,d\zeta\right\}d\eta + k\ell\int_0^1 (1-u^2)^{\frac{1}{2}}\,du$$

$$\left.\int_0^1\left\{\int_0^\infty \zeta[1-\bar{\alpha}(\zeta a)]A(\zeta)\cos(\zeta\ell\eta u)\,d\zeta\right\}\eta^2 J_1(\ell k\eta)d\eta\right]$$

where $J_o(z)$ and $J_1(z)$ are Bessel functions. After carrying out integrations in y and u we obtained the following integral equation of the second kind

$$(3.9)\qquad A(\kappa) - \int_0^\infty \eta(\eta^2-\kappa^2)^{-1}\left[\eta\, J_o(\kappa)\, J_1(\eta) - \kappa\, J_o(\eta)\, J_1(\kappa)\right]$$

$$\cdot\,[1-\bar{\alpha}(\eta\varepsilon)]\,A(\eta)\,d\eta = A_c(\kappa)$$

where

$$\kappa \equiv k\ell \quad , \quad \eta = \zeta\ell \quad , \quad \varepsilon \equiv a/\ell \quad ,$$

$$(3.10)\qquad A(\kappa) \equiv (2/\pi)^{\frac{1}{2}}\left[2\mu(\lambda+\mu)/\ell^2 t_o(\lambda+2\mu)\right]A(k) \quad ,$$

$$A_c(\kappa) \equiv (2/\pi)^{\frac{1}{2}}\left[2\mu(\lambda+\mu)/\ell^2 t_o(\lambda+2\mu)\right]A_c(k) = J_1(\kappa)/\kappa$$

in which the last equality follows from the classical solution for $A_c(k)$ in the case of $t_o$ = const.

When (3.9) is solved for $A(\kappa)$ then the displacement and stress fields follow from (3.1), (3.3) and (3.4). Along the crack line (y = 0) these are given by

$$v(\xi,0)\,[2\mu(\lambda+\mu)/(\lambda+2\mu)]/t_o\ell = \int_0^\infty A(\kappa)\cos(\kappa\xi)\,d\kappa$$

$$(3.11)\qquad t_{yy}(\xi,0)/t_o \equiv t(\xi,0)/t_o = -\int_0^\infty \kappa\,\bar{\alpha}(\kappa\varepsilon)\,A(\kappa)\cos(\kappa\xi)\,d\kappa$$

where $\xi \equiv x/\ell$ .

The integral equation (3.9) is non-singular for all $\varepsilon \neq 0$. For $\varepsilon = 0$ we have $A(\kappa) = A_c(\kappa) = J_1(\kappa)/\kappa$. It is also clear that $1-\bar{\alpha}(\kappa\varepsilon) = 0(\varepsilon^2)$ for small $\varepsilon<<1$. The smallest length crack may be constructed by one missing atom. In this case $a = \ell$ and $\varepsilon = 1$. Thus $0 \leq \varepsilon \leq 1$ for a micro-crack of 100 atomic length $\varepsilon = 1/50$. It is thus expected that the contribution of the integral in (3.9) will become appreciable only for submicroscopic cracks of few atomic distances. In fact, this turned out to be the case when (3.9) was solved by means of electronic computers.

The numerical calculations were carried out over a two Brillouin zone, $k = 2\pi/a$, by discretizing (3.9) over 150 grid points. The results will be reported elsewhere. Here, however, we give some typical cases. In fact, we have found that the classical solution $A_c$ is perfectly satisfactory for $2\ell/a \geq 40$ (still a submicroscopic crack).

The stress concentration for the case when the crack surface is free of traction but the plate is subject to uniform tension $t_{yy} = t_o$ at infinity is given by

$$P(x) = [t_{yy}(x,0)/t_o] + 1 \tag{3.12}$$

The fact that the classical solution $A_c$ of the dual integral equation (3.8) satisfies the boundary conditions extremely well for $2\ell/a \geq 40$ can be seen from Fig. 4. For other details and error estimates depending on $\varepsilon$ the reader is referred to Eringen, et al [1976].

## 4. COHESIVE STRESS-FRACTURE CRITERION

The stress concentration factor

$$C(\nu) \equiv (2\ell/a)^{-\frac{1}{2}} P(\ell) \tag{4.1}$$

is shown in Table 1 for various Poisson's ratio $\nu = \lambda/2(\lambda+\mu)$ valid for $2\ell/a \geq 100$. It is clear that $0.676 \leq C(\nu) \leq 0.845$. For $\nu = 0.25$, $C \simeq 0.713$ for $2\ell/a > 100$.

By means of (4.1) we make the following very significant observations:

(i) The stress field based on nonlocal theory has no singularity so long as $a \neq 0$. In the continuum limit $a \to 0$, and the classical square root singularity occurs.

(ii) A maximum stress hypothesis can now be used to predict the failure. In fact, we state that: *When* $t_{yy\ max} = t_c$ = *cohesive stress the fracture will occur.* From (3.1) it therefore follows that

$$t_o^2 \ell = [a/2\ C^2(\nu)]\ t_c^2 \equiv C_G \tag{4.2}$$

This is the Griffith criterion for brittle fracture, with extra benefit that the Griffith constant $C_G$ is now fully determined. Interestingly, no *ad hoc* constant (e.g., surface energy $\gamma$) occurs in (4.2) and from the value of $C_G$ it is clear that it is a material property, i.e., it is known once the cohesive stress $t_c$, lattice parameter a, and the Poisson's ratio $\nu$ are known.

(iii) The verification of the fact that fracture toughness, classically defined by $K_I \equiv \sqrt{\pi \ell}\, t_o$ is a material property led many experimentalists to carry out long and arduous experiments (cf., Freed et al. [1971]; Brown and Strawley [1966]). If (4.2) is used we see that

$$(4.3) \qquad K_I = (\pi C_G)^{\frac{1}{2}} = (\pi\, a/2)^{\frac{1}{2}}\, t_c/C(\nu)$$

is indeed a material property.

(iv) Cohesive stress $t_c$ may be calculated for a given solid by use of (4.2). Griffith surface energy $\gamma$ appearing in (1.2) has been the subject of a great deal of experimentation. If we equate (1.2) to (4.2) we obtain

$$(4.4) \qquad t_c^{\,2} a = K\gamma$$

where

$$(4.5) \qquad K = 8C^2(\nu)\mu/\pi(1-\nu)$$

Calculations may now be carried out for various materials. Employing the experimental values listed in Table 2, we have calculated $t_c/E$ based on the nonlocal theory. The results are recorded in the next to the last column of Table 2. The entries in the last column of this table are the estimates of $t_c/E$ based on atomic considerations, Lawn and Wilshaw [1975, p. 160].

The remarkably close values obtained should be considered to be indicative of the far reaching power of the nonlocal theory.

Acknowledgment

I wish to express my appreciation and thanks to Professor J. Rice, Dr. R. Chang and Dr. G. Sih for some valuable discussions regarding surface energy and to Mr. C. Speziale for computer work.

TABLE 1.

Stress Concentration factor at Crack Tip
vs. Poisson's Ratio ($\frac{2\ell}{a} = 100$)

| $\nu$ | C |
|---|---|
| 0 | .676 |
| .05 | .682 |
| .10 | .687 |
| .15 | .695 |
| .20 | .703 |
| .25 | .713 |
| .30 | .723 |
| .35 | .743 |
| .40 | .764 |
| .45 | .796 |
| .50 | .845 |

TABLE 2. Cohesive Stress. $t_c$:

$K \equiv 8C^2\mu/\pi(1-\nu)$

a = Atomic dist.

$\gamma$ = Surface energy

| Experimental[1] | | | | | | Present Theory | | | | | Atomistic Models[2] |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Type | Crystal | $\gamma$ CGS | $\mu x10^{-11}$ CGS | $\nu$ | a $A^o$ | $C(\nu)$ | $Kx10^{-11}$ CGS | $K\gamma$ x $10^{-14}$ | $t_c$ x $10^{-11}$ | $t_c/E$ | $t_c/E$ |
| Face C. | Al | 840 | 2.51 | 0.347 | 2.86 | 0.743 | 5.402 | 4.538 | 1.260 | 0.186 | |
| | Ni | 1725 | 7.48 | 0.276 | 2.49 | 0.718 | 13.57 | 23.41 | 3.066 | 0.16 | |
| Body C. | Fe | 1975 | 6.92 | 0.291 | 2.48 | 0.721 | 12.93 | 25.53 | 3.209 | 0.18 | 0.23 |
| Ionic | LiF | 480 | 4.40 | 0.068 | 2.014 | 0.684 | 5.621 | 2.698 | 1.157 | 0.123 | |
| Diam. | C | 5400 | 50.9 | 0.187 | 1.54 | 0.701 | 78.32 | 422.9 | 16.57 | 0.137 | 0.17 |
| Hex. | Zn | 575 | 3.83 | 0.333 | 2.66 | 0.736 | 7.925 | 4.557 | 1.309 | 0.128 | 0.11 |

[1] $\gamma$, $\mu$ and $\nu$ are taken from data collected by Rice and Thomson [1974, Table 1]. Lattice parameter a is from Kittel [1971, Table 5].

[2] Atomistic results are from Table 7.1, p. 160, Lawn and Wilshaw [1975].

REFERENCES


Inglis, C.E. [1913]: "Stresses in a Plate Due to the Presence of Cracks and Sharp Corners," Proc. Inst. Naval Architects.

Grifffith, A.A. [1920]: "The Phenomena of Rupture and Flow in Solids," Phil. Trans. Roy. Soc. (London) Ser. A221, 163-198.

Barenblatt, G.E. [1962]: "The Mathematical Theory of Equilibrium Cracks in Brittle Fracture," Advan. Appl. Mech., 7, 55-129.

Zheltov, Iu. P. and Khristianovich, S.A. [1955]: Izv. Acad. Nauk SSSR Otd Tekhn. Nauk. 5, 3-41.

Dugdale, D.S. [1960]: "Yielding of Steel Sheets Containing Slits," J. Mech. Phys. of Solids, 8, 100-104.

Goodier, J.N. and Kanninen, M. [1966]: "Crack Propagation in a Continuum Model with Nonlinear Atomic Separation Laws," Technical Report No. 165 (Contract under the Office of Naval Research) Division of Engineering Mechanics, Stanford University.

Goodier, J.N. [1968]: "Mathematical Theories of Brittle Fracture," Fracture, Vol. II, Ch. 1, edit. H. Liebowitz.

Sternberg, E. and Muki, R. [1967]: "The Effect of Couple Stresses on the Stress Concentration Around a Crack," Int. J. Solids & Structures 3, 69.

Kim, B.S. and Eringen, A.C. [1973]: "Stress Distribution Around an Elliptic Hole in an Infinite Micropolar Elastic Plate," Letters in Appl. & Engng. Sci. 1, 381-390.

Eringen, A.C. [1972a]: "Nonlocal Polar Elastic Continua," Int. J. Engng. Sci. 10, 1, 1-16.

Eringen, A.C. [1972b]: "Linear Theory of Nonlocal Elasticity and Dispersion of Plane Waves," Int. J. Engng. Sci. 10, 425-435.

Eringen, A.C. and Edelen, D.G.B. [1972]: "On Nonlocal Elasticity," Int. J. Engng. Sci. 10, 233-248.

Eringen, A.C. [1974]: "Nonlocal Elasticity and Waves," Continuum Mechanics Aspects of Geodynamics and Rock Fracture Mechanics, edit. by P. Thoft-Christensen, Dordrecht, Holland: D. Reidel Publishing Co., 81-105.

Eringen, A.C. [1976]: "Screw Dislocation in Nonlocal Elasticity," ONR Tech. Report #41, Princeton University Civil Engineering Research Report No. 76-SM-10.

Sneddon, I.N. [1951]: Fourier Transform, McGraw-Hill Book Co., New York.

Eringen, A.C., Speziale, C.G. and Kim, B.S. [1976]: "Crack Tip Problem in Nonlocal Elasticity," submitted for publication

Freed, C.N., Sullivan, A.M. and Stoop, J. [1971]: "Influence of Dimensions of Center-cracked Tension Specimen on $K_C$," ASTM Special Technical Publication No. 514, p. 109.

Brown, W.F. and Strawley, J.E. [1966]: "Plane Strain Crack Toughness Testing of High Strength Metallic Materials," ASTM Special Technical Publication No. 410, p. 20.

Lawn, B.R. and Wilshaw, T.R. [1975]: Fracture of Brittle Solids, Cambridge University Press (London).

Rice, J.R. and Thomson, R. [1974]: "Ductile Versus Brittle Behavior of Crystals," Philosophical Magazine, 29, p. 73.

Kittel, C. [1971]: Introduction to Solid State Physics, John Wiley & Sons (New York).

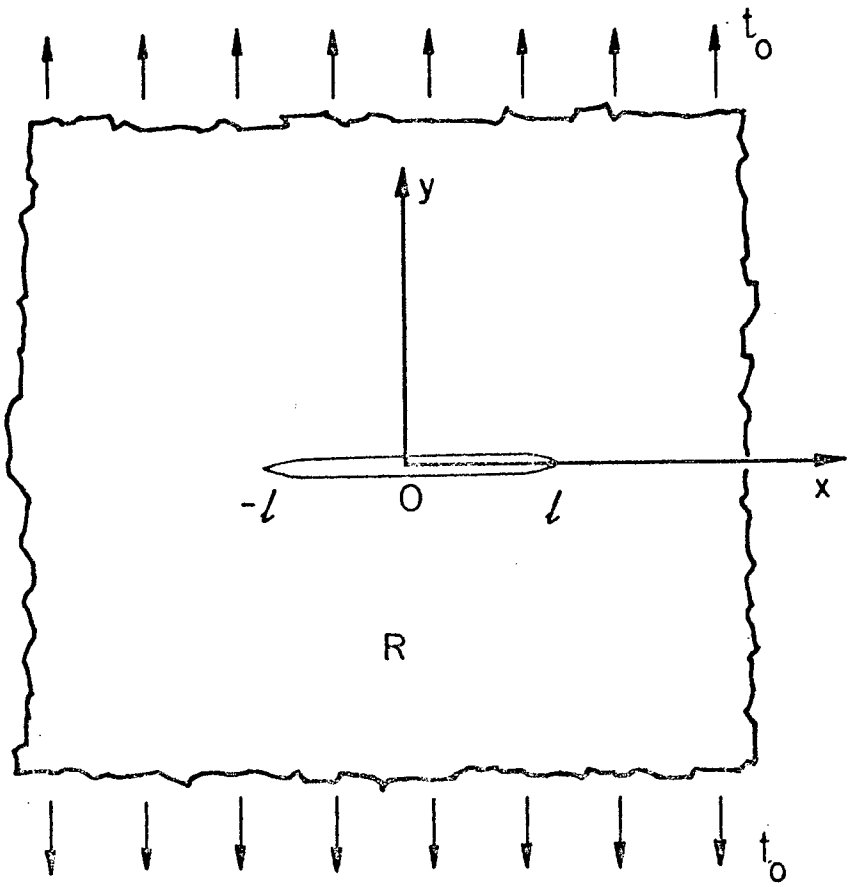



FIGURE 1

Elastic plate weakened by a crack

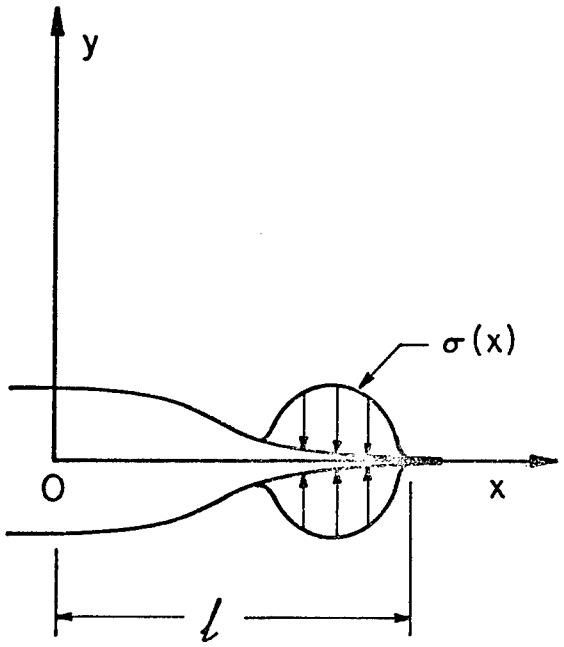



FIGURE 2

Barenblatt model assumes that cohesive normal stress $\sigma(x)$ act at the tip region of the crack surface. The form of $\sigma(x)$ is to be determined to give cusps at tips.

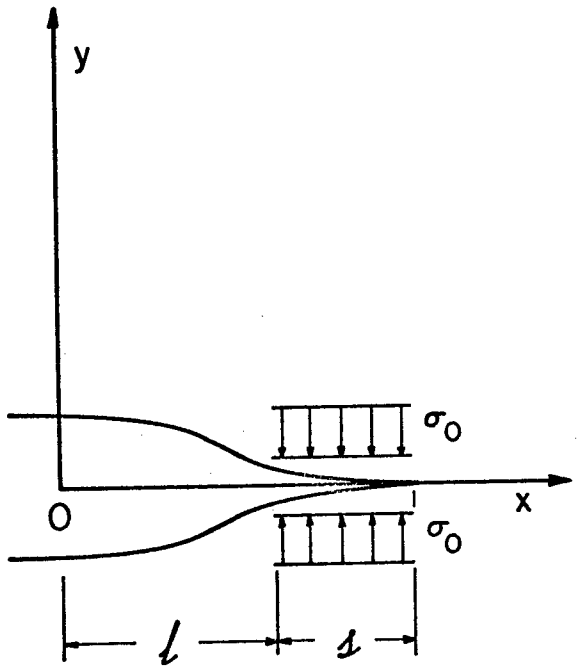



FIGURE 3

Khristianowich-Dugdale Model

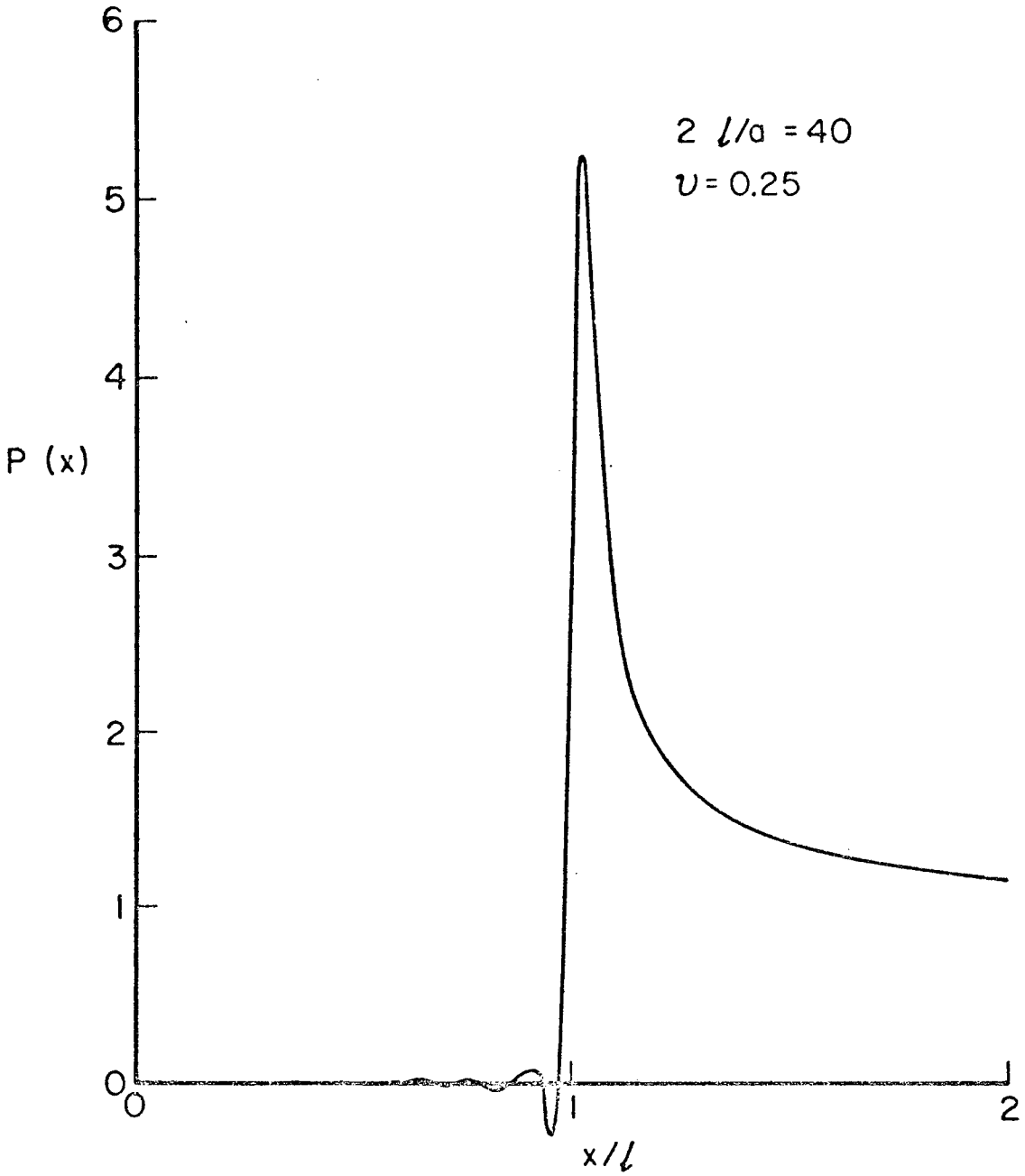



FIGURE 4

Stress concentration along the line of crack

# BENDING OF A CRACKED STRIP INCLUDING CRACK SURFACE INTERFERENCE

O. L. Bowie and C. E. Freese
Army Materials and Mechanics Research Center
Watertown, Massachusetts
Presented at the 22nd Conference of Army Mathematicians,
May 12-14
Watervliet Arsenal, Watervliet, New York, 1976

ABSTRACT. In the analysis of cracks lying in a compressive stress field, the classical solution of elasticity frequently yields unacceptable physical results - often predicting an overlapping of the crack faces. A first order correction to these solutions can be found by admitting crack surface interference and searching for a physically compatible displacement field.

The problem of a center (or edge) cracked strip under in-plane bending is solved from this viewpoint. A necessary condition for a physically compatible solution is shown to be the vanishing of the stress intensity factor at the crack tip in the otherwise compressive field. Numerical results indicate that the classical solution for the stress intensity factors at the crack tip in the tensile field underestimates the corrected solution by approximately ten percent.

1. INTRODUCTION. Every so often the simplifying assumptions of the classical linear theory of elasticity can lead to mathematical solutions which are physically unrealistic. We are familiar with the need for retaining the non-linear terms of the strain-displacement relations to account for the instability or buckling phenomena observed in the behavior of thin shells. Another type of deficiency arises in the analysis of configurations involving cracks lying in compressive stress fields.

A simple example illustrating the subject of this investigation is provided by a rectangular strip with a central crack loaded by a uniform uniaxial compression normal to the direction of the crack, Figure 1a. Assuming no friction across the crack surfaces, the obvious physically acceptable solution for this problem predicts the tangency of the crack surfaces AOB and AO'B with a stress state of uniform compression acting throughout the strip and across the crack surfaces. Compare this solution with that of reversing the signs for uniaxial tensile loading - an assumption consistent with the superposition argument of classical elasticity. Clearly the resulting infinite compressive stresses at the crack tips and the negative displacements predicting an overlapping of the crack surfaces (Figure 1b) arrived at by such an argument is a physically unacceptable solution of the problem.

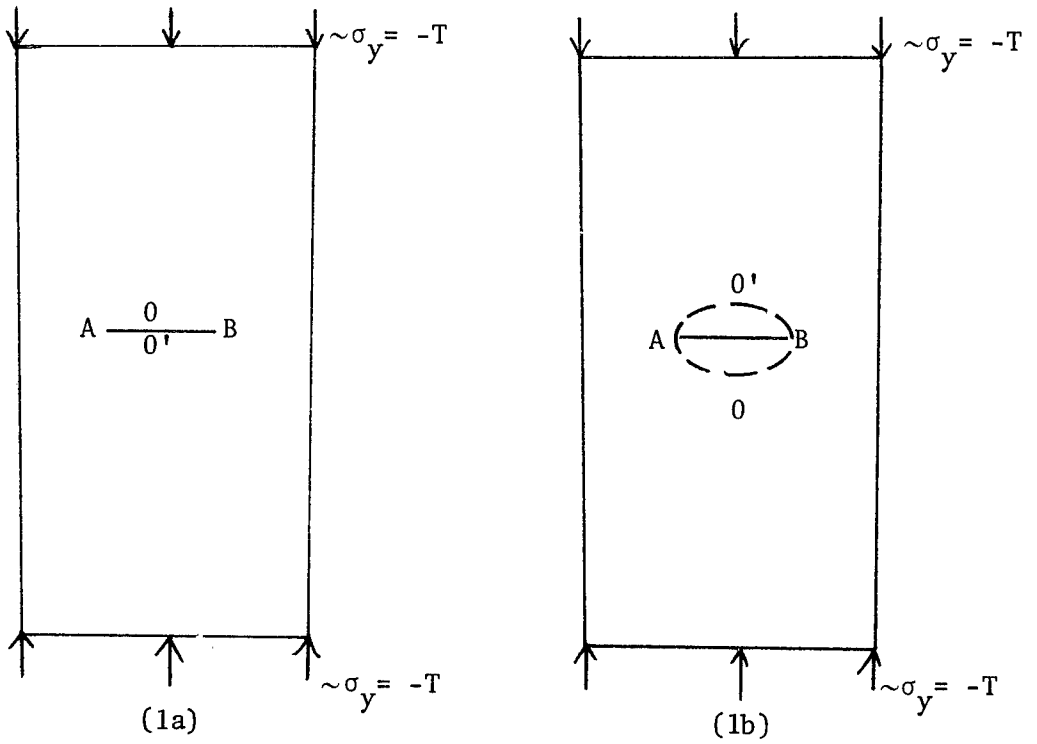



Figure 1. Central crack in rectangular strip under uniaxial compression, $\sigma_y = -T$.

The "overlapping" problem illustrated above carries over, usually more subtly, to a variety of crack solutions when a portion of the crack configuration lies in a stress field which is compressive. A positive symptom of overlapping in the vicinity of a crack tip can be inferred from the sign of the stress intensity factor of linear fracture mechanics. If, for example, $K_I$ (the conventional Mode I stress intensity component) is negative, then there exists local overlapping at the crack tip.

A plan of modifying the classical solution by tolerating crack closure but no overlapping is adopted in this paper. The problems corresponding to internal and edge cracks in a strip under in-plane bending are analyzed from this viewpoint and the "error" in the classical solutions is assessed.

2. CENTRAL CRACK IN AN INFINITE SHEET UNDER BENDING. First, we consider the problem of a crack of length 2L with center at $Z_o$ in an infinite sheet under in-plane bending (Figure 2). When $Z_o = 0$, the crack is centrally located with respect to the applied load and crack tips A and B obviously lie in compressive and tensile stress fields, respectively. We shall now show that both the classical and the modified solutions of this problem can be found by the Muskhelishvili [1] method of analysis.

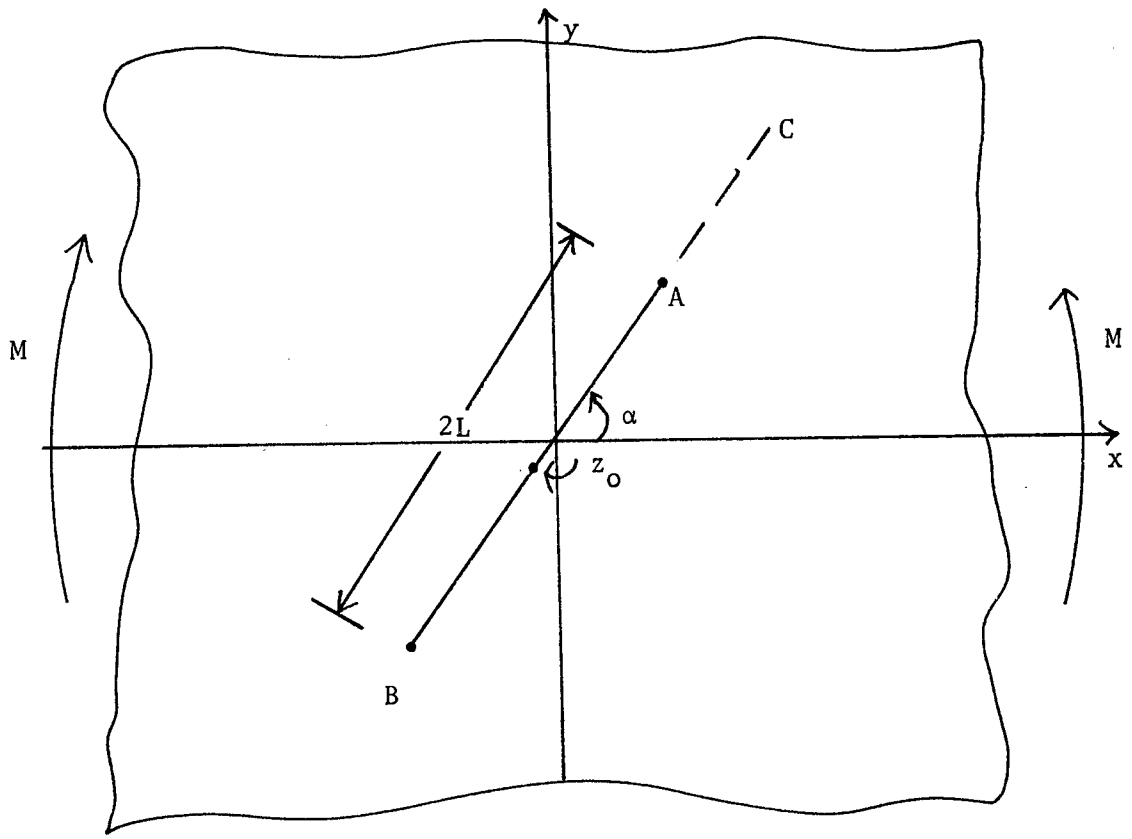



Figure 2. Crack in an infinite sheet under in-plane bending.

The Muskhelishvili analysis depends on the determination of two analytic stress functions $\phi(z)$ and $\psi(z)$ with the stresses and displacements defined as

$$\sigma_y + \sigma_x = 2[\phi'(z) + \overline{\phi'(z)}]$$

$$\sigma_y - \sigma_x + 2i\tau_{xy} = 2[\bar{z}\phi''(z) + \psi'(z)] \qquad (1)$$

$$2\mu(u + iv) = \kappa\phi(z) - z\overline{\phi'(z)} - \overline{\psi(z)}$$

where primes denote differentiation and bars complex conjugates. The constants $\mu$ and $\kappa$ are defined as $\mu = E/2(1+\nu)$ and $\kappa = 3-4\nu$ (plane strain) and $\kappa = (3-\nu)/(1+\nu)$ (plane stress) where E and $\nu$ are Young's modulus and Poisson's ratio, respectively.

For a plate with no crack, the stress functions

$$\phi(z) = iTz^2/8 \ , \quad \psi(z) = -iTz^2/8 \qquad (2)$$

yield the stress distribution

$$\sigma_y = 0 \ , \quad \tau_{xy} = 0 \ , \ \sigma_x = -Ty \tag{3}$$

which is of course the desired loading for large $|z|$.

The physical region in Figure 2 can be described conveniently by the mapping

$$z = \omega(\zeta) = z_o + \ell^{i\alpha}(L/2)(\zeta + \zeta^{-1}) \tag{4}$$

The unit circle and its exterior in the $\zeta$-plane are mapped into the crack and its exterior in the z-plane. In particular, $\zeta$=1 maps into the crack tip A and $\zeta$ = -1 maps into the crack tip B.

The stress functions $\phi(z)$ and $\psi(z)$ can now be considered as $\phi(\zeta)$ and $\psi(\zeta)$ where $\phi'(z)$ now corresponds to $\phi'(\zeta)/\omega'(\zeta)$, etc. Using the well-known continuation arguments of Muskhelishvili, the crack is traction-free if we set

$$\psi(\zeta) = - \overline{\phi}(1/\zeta) - \overline{\omega}(1/\zeta)\phi'(\zeta)/\omega'(\zeta) \tag{5}$$

and the extended definition of $\phi(\zeta)$ leads to a function continuous across the unit circle. On the other hand, from (2) the loading conditions at infinity require

$$\phi(\zeta) \to iTz^2/8 \to iT(L/32)[L\ell^{2i\alpha}\zeta^2 + 4z_o\ell^{i\alpha}\zeta] \tag{6}$$

$$\psi(\zeta) \to - iTz^2/8 \to - iT(L/32)[L\ell^{2i\alpha}\zeta^2 + 4z_o\ell^{i\alpha}\zeta$$

for large $|\zeta|$.

Conditions (5) and (6) are satisfied by choosing

$$\phi(\zeta) = [iTL^2/32]\left\{\ell^{2i\alpha}\zeta^2 + 4(z_o/L)\ell^{i\alpha}\zeta - [\ell^{-2i\alpha} - 2]\zeta^{-2}\right. \tag{7}$$

$$\left. + [8i(\overline{z}_o/L)\sin\alpha + 4(z_o/L)\ell^{-i\alpha}]\zeta^{-1}\right\}$$

and this completes the formal solution.

3. THE CLASSICAL SOLUTION WHEN $z_o = 0$. The classical solution for the centrally located crack, $z_o = 0$, will first be considered. The crack tip A lies in apparently a compressive field and we can anticipate a physical incompatibility of the solution.

The stress intensity, $K_A^{(1)}$, at the crack tip A in general will be made up of Modes I and II and can be calculated from

$$K_{1A}^{(1)} - i\, K_{2A}^{(1)} = 2\phi'(1)[\ell^{i\alpha}\omega''(1)]^{-1/2} = -T(L^{3/2}/2)\sin^2\alpha(\sin\alpha + i\cos\alpha) \tag{8}$$

whence

$$K_{1A}^{(1)} = -(T/2)L^{3/2}\sin^3\alpha$$
$$K_{2A}^{(1)} = -(T/2)L^{3/2}\sin^2\alpha\cos\alpha \tag{9}$$

Similarly, at crack tip B (corresponding to $\zeta = -1$),

$$K_{1B}^{(1)} = (T/2)L^{3/2}\sin^3\alpha$$
$$K_{2B}^{(1)} = -(T/2)L^{3/2}\sin^2\alpha\cos\alpha \tag{10}$$

A clue to the unacceptability of the solution is negativeness of $K_{1A}^{(1)}$.

In order to examine the physical compatibility of the displacements of the crack surfaces, we introduce a $(\xi,\eta)$ coordinate system where $\xi$- and $\eta$ are along and normal to, respectively, the crack direction. Then

$$u_\xi + i\, u_\eta = \ell^{-i\alpha}(u + iv) = \ell^{-i\alpha}(\kappa + 1)\phi(\sigma)/2\mu \tag{11}$$

for the crack boundary where $\sigma = \ell^{i\theta}$ are points on the unit circle in the $\zeta$-plane. The condition for no "overlapping" of the crack boundaries can be written as

$$u_\eta(\theta) - u_\eta(-\theta) \geq 0 \qquad 0 \leq \theta \leq \pi \tag{12}$$

When $z_o = 0$,

$$u_\eta(\theta) - u_\eta(-\theta) = (\kappa + 1)TL^2\sin\alpha\sin 2\theta[\cos 2\alpha - 1]/16\mu \tag{13}$$

which (except for the trivial cases $\alpha = o, \pi$) clearly violates the no overlapping condition (12) in the interval $o < \theta < \pi/2$.

4. DETERMINATION OF A PHYSICALLY ACCEPTABLE SOLUTION. The plan for determining a physically acceptable solution depends on admitting crack closure over segments of the crack without overlapping. If the crack tip is involved in the region of overlapping, as is the case in the present problem, a necessary condition for an acceptable solution can be expressed in terms of the stress intensity factors from a consideration of the local stress and displacement fields.

Consider the crack tip A and the displacements $u_\xi$ and $u_\eta$ to the first order of the local crack tip expansion. A necessary condition for no local overlapping can easily be shown, $K_1 \geq 0$, from a consideration of $u_\eta$. If, in addition, we assume crack closure in the neighborhood of A, then $\sigma_\eta$ must be non-tensile across this interval. Therefore, a necessary condition for a physically acceptable solution is $K_1 = 0$ at A. No claim as to the sufficiency of this condition can be made as the stress intensity reflects only the dominant term of the local solution. A solution arrived at on this basis must still be tested for its overall consistency.

In the present case, we consider $z_o$ as undetermined and impose the vanishing of $K_1$ at A. Since

$$\phi'(1) = - iT(L^2/4) \left\{ \sin^2\alpha + (2/L)(\sin \alpha)\ \mathrm{Im}\ z_o \right\} \tag{14}$$

it follows that $K_1 = 0$ at A if we choose

$$\mathrm{Im}\ z_o = - (L/2)\sin \alpha \tag{15}$$

Although there are no restrictions on Re $z_o$, we choose $z_o$ so that the crack passes through the origin of coordinates, thus

$$z_o = - (L/2)\ell^{i\alpha} \tag{16}$$

With this choice of $z_o$, we reexamine the non-overlapping condition (12). On the crack,

$$u_\eta = (\kappa + 1)TL^2 \left\{ 4 \sin^3 \alpha \sin \theta(1 - \cos \theta) + \cos \alpha(1 + 2 \sin^2\alpha)(1 - 2 \sin^2\theta - 2 \cos \theta) \right\} /32\mu, \tag{17}$$

thus,

$$u_\eta(\theta) - u_\eta(-\theta) = (\kappa + 1)TL^2\sin^3\alpha \sin \theta(1 - \cos \theta)/4\mu \tag{18}$$

which clearly satisfies (12) for $0 \leq \theta \leq \pi$ and hence is a physically acceptable displacement field.

The stress intensity factors in the present case are

$$K_{1A}^{(2)} = K_{2A}^{(2)} = 0$$

$$K_{1B}^{(2)} = TL^{3/2}\sin^3\alpha \tag{19}$$

$$K_{2B}^{(2)} = - TL^{3/2}\sin^2\alpha \cos \alpha$$

Furthermore, it is easily verified that the forces normal to the segment AC in Figure 2 are compressive.

It is clear that the present solution can be considered as a physically acceptable solution for a central crack along the segment BC where closure occurs on the segment AC. We do assume, of course, that the frictional properties of the crack surfaces are consistent with a continuous displacement solution along AC, i.e. closure without slippage.

A comparison with the previously derived classical solution for a centrally located crack can now be made by observing the change in the stress intensity calculation at point B. The crack AC corresponds to a crack length of 2L if an effective half crack length of 2L/3 is used in the calculation of (19). Thus, the "corrected" stress intensity factors at B are

$$K_{1B} = T(2L/3)^{3/2}\sin^3\alpha$$

$$K_{2B} = - T(2L/3)^{3/2}\sin^2\alpha \cos \alpha \tag{20}$$

Since

$$K_{1B}/K_{1B}^{(1)} = K_{2B}/K_{2B}^{(1)} = 2(2/3)^{3/2} \tag{21}$$

the classical estimate of the stress intensity factor at B is in error on the non-conservative side by approximately nine percent.

5. CENTRAL CRACK IN A FINITE STRIP UNDER BENDING. We consider, now, the more difficult problem of a central crack in a strip of finite width under bending, Figure 3, where the solution cannot be found in closed form and the previous arguments must be carried out numerically. For the configuration in Figure 3, Benthem and Koiter [2] have estimated the crack tip stress intensity factors at B for the classical solution of the problem by using an effective asymptotic argument.

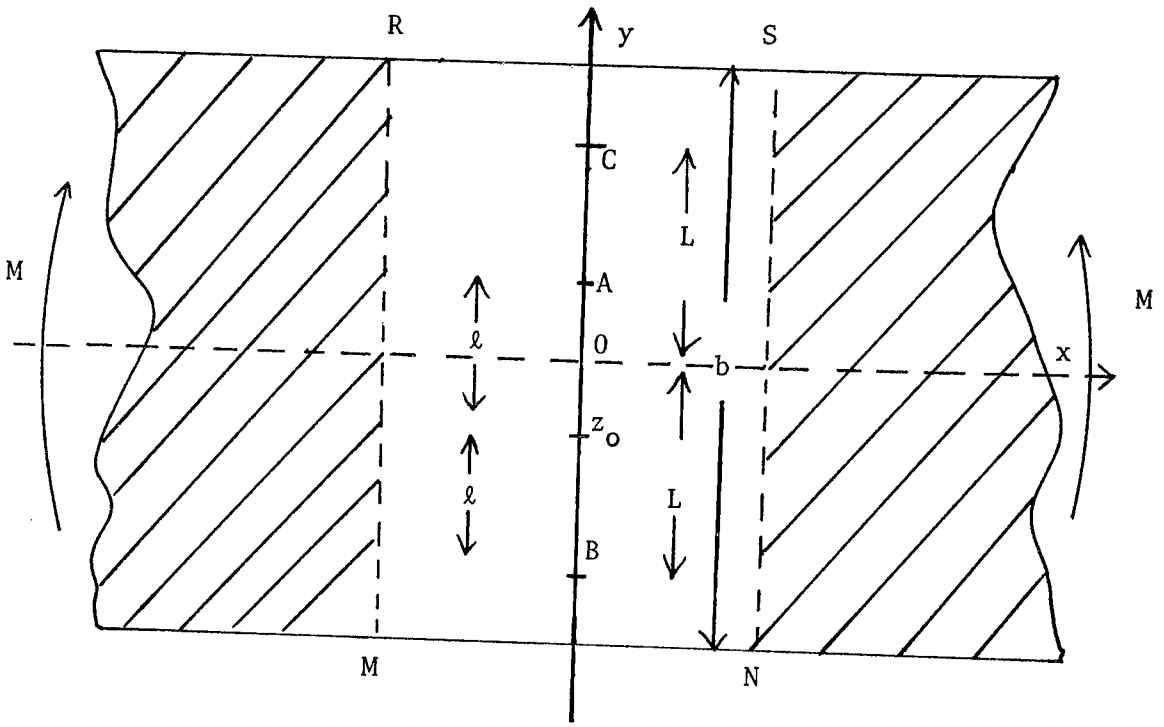



Figure 3. Central crack in a strip under bending.

The solution was carried out using the MMC (Modified Mapping Collocation) method combined with finite elements [3,4]. This plan is based on "partitioning" the region and using a representation of the solution appropriate to each sub-region. The boundary conditions along with appropriate "stitching" conditions between the representations must be satisfied by the solution. The details of this approach have been previously documented and will not be repeated here.

The partitioning plan is indicated in Figure 3. The region MRSN was described using the mapping function

$$z = z_o + i(\ell/2)(\zeta + 1/\zeta) \qquad (22)$$

which clearly maps the unit circle in the $\zeta$-plane into the crack AB. A series representation of the solution was chosen in the corresponding parameter region and traction-free conditions on the crack were enforced by the continuation argument, e.g., Equation (5). The boundary conditions on RS and MN and the stitching conditions on RM and SN were imposed by the collocation arguments outlined in [4]. In the complementary regions (the shaded

areas in Figure 3) a finite element representation of the solution was taken. Imposed on this representation were the appropriate stitching conditions, traction-free boundary conditions and end loading,

$$\sigma_x = - (3M/2b^3)y \tag{23}$$

Again, we seek a value of $z_o$ such that the stress intensity $K_{1A} = 0$ and the crack displacements and the forces on AC are physically compatible with our argument. It was found that $z_o$ can be determined quite readily by iteration. From the infinite sheet solution, it is evident that for small $\ell/b$ ratios, $z_o \simeq \ell/2$. With this as a guide for the first approximation, only a few trials were required to find the proper value of $z_o$ for successively increasing values of $\ell/b$.

The numerical results are presented in Table 1. Again the results are to be compared with the classical solution for the central crack, $z_o = 0$. The effective half crack length, L, is evidently

$$L = \ell + |z_o| \tag{24}$$

Table 1

"Corrected Stress Intensity Factors, $K_{1B}$, for Central Crack in Strip under Bending

| L/b | $\ell/b$ | $z_o/b$ | OA/b | $\frac{K_{1B}}{Mb^{-3/2}}$ | $\frac{K_{1B}^{**}}{Mb^{-3/2}}$ | $\frac{K_{1B}}{K_{1B}^{**}}$ |
|---|---|---|---|---|---|---|
| 0.1 | 0.067 | -0.033 | 0.033 | 0.0259 | 0.0237 | 1.09 |
| 0.2 | 0.133 | -0.067 | 0.067 | 0.0733 | 0.0672 | 1.09 |
| 0.3 | 0.200 | -0.100 | 0.100 | 0.136 | 0.124 | 1.10 |
| 0.4 | 0.270 | -0.130 | 0.140 | 0.213 | 0.193 | 1.10 |
| 0.5 | 0.340 | -0.160 | 0.180 | 0.304 | 0.276 | 1.10 |
| 0.6 | 0.414 | -0.186 | 0.228 | 0.417 | 0.379 | 1.10 |
| 0.7 | 0.492 | -0.208 | 0.284 | 0.567 | 0.516 | 1.10 |
| 0.8 | 0.574 | -0.226 | 0.350 | 0.796 | 0.727 | 1.10 |
| 0.9 | 0.668 | -0.232 | 0.432 | 1.280 | 1.163 | 1.10 |
| 1.0 | 0.763* | -0.237* | 0.526* | | | |

*Extrapolated
**Benthem and Koiter [2]

It is interesting to compare the present results with the classical results $K_{1B}^{**}$ of Benthem and Koiter. Within one percent, the classical solution underestimates the $K_{1B}$ values by nine percent for all values of L/b.

6. MODIFICATION OF THE ASYMPTOTIC APPROXIMATION. In [2], Benthem and Koiter introduced a non-dimensional factor K by writing

$$K_1 = K\ 3LM(aL/b)^{1/2}/2(b^3 - L^3) \tag{25}$$

where K is a polynomial in L/b. An approximate solution for K was found by order of magnitude considerations of the two limiting cases, $L/b \to 0$ and $a/b \to 0$.

The modifications of their arguments for the "corrected" solution for $L/b \to 0$ can now be carried out by using our solution for the central crack in an infinite sheet. In particular, if the order of magnitude considerations of [2] are modified by Equation (20), then, at the crack tip B,

$$K_{1B} \to (2/3)^{3/2}[3ML^{3/2}/2b^3][1 + 0(L^4/b^4] \quad \text{for } L/b \to 0 \tag{26}$$

From a comparison of Equations (25) and (26),

$$K \to (2/3)^{3/2}[1 + (1/2)(L/b) + (3/8)(L/b)^2 - (11/16)(L/b)^3 + 0(L^4/b^4] \quad \text{for } L/b \to 0 \tag{27}$$

For the second limiting case, $a/b \to 0$, by using the anti-symmetry of the classical problem and the "edge dam" solution, the authors of [2] found

$$K \to 2/(\pi^2 - 4)^{1/2} = 0.826 \quad \text{for } a/b \to 0 \tag{28}$$

Unfortunately, due to the non-linearity of our present solution no such limit can be rigorously argued. On the other hand, a reasonable estimate of this limit can be found by extrapolation of the data. From Table 1, the segment 0A can be extrapolated as $0A \to 0.52b$ as $a/b \to 0$. Furthermore, the stress distribution $\sigma_y$ along the centerline from A to the edge is very nearly linear. From equilibrium conditions, it can be argued that the local stress at B is nine percent higher than in the classical case. Thus,

$$K \to (1.09)(0.826) \quad \text{for } a/b \to 0 \tag{29}$$

(Although (29) is an extrapolated estimate, it was verified that reasonable variations in the approximation altered this result by no more than one percent.)

Therefore, the simplest polynomial interpolation between these asymptotic results yields

$$K = (2/3)^{3/2}[1 + (1/2)(L/b) + (3/8)(L/b)^2 - (11/16)(L/b)^3 + .464\ (L/b)^4] \tag{30}$$

Equation (30) is identical with the K in [2] if $(2/3)^{3/2}$ were replaced by 1/2.

7. EDGE CRACK IN A STRIP UNDER BENDING. At about the same time as the author's solution [5], Paris and Tada [6] considered the solution for an edge crack in a strip under bending again allowing for interference of segments of the crack surfaces, Figure 4.

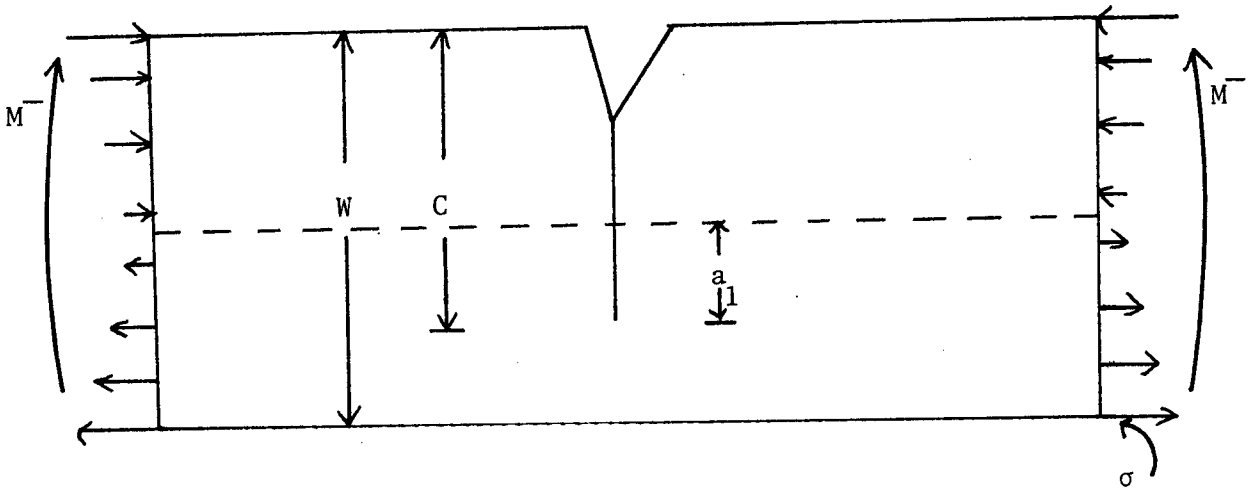



Figure 4. Edge crack in strip under bending.

It is obvious physically that for $C/W \leq 1/2$, assuming no friction between the crack surfaces, the admissible solution is one which predicts the strip is in uniform bending with the crack surfaces interferring and carrying a compressive load. For $C/W \geq 1/2$, it is also clear that the solution is identical to our results for the central crack with a modified interpretation of the parameters.

In Paris and Tada's analysis, the crack tip stress intensity, K, was approximated by

$$K = G(C/W)H(C/W)\sqrt{\pi C} \tag{31}$$

where

$$G(C/W) = (2/3)^{3/2}(2C/W)(1 - W/2C)^{3/2} \tag{32}$$

based on the solution for a central crack in an infinite sheet under bending. (Note that in Equation (31), the alternate introduction of $\sqrt{\pi}$ in the definition of stress intensity factors has been made.) The function H(C/W) was taken as the correction for the effect of the finite width of the strip. Paris and Tada did not calculate H(C/W) exactly, instead, they assumed an approximation based on the finite width correction for a center cracked finite width strip under tension. Their numerical results are listed in Table 2.

The results which we have derived can be applied with the following changes in notation.

$$L = a_1 = C - W/2$$

$$b = W/2 \qquad (33)$$

$$a = b - L$$

$$L/b = 2(C/W) - 1 = \lambda$$

$$M = (2/3)b^2\sigma$$

$$K_1 = \bar{K}$$

Then,

$$\bar{K}/\sigma\sqrt{\pi C} = R(\lambda)\lambda^{3/2}\sqrt{1 - \lambda}/(1 - \lambda^3)\sqrt{1 + \lambda}$$

where (34)

$$R(\lambda) = (2/3)^{3/2}[1 + \lambda/2 + 3\lambda^2/8 - 11\lambda^3/16 + .464\ ^4]$$

A comparison of the results is shown in Table 2.

Table 2. Values of $\bar{K}/\sigma\sqrt{\pi C}$

| C/W | $\lambda$ | Equation (31) | Equation (34) |
|---|---|---|---|
| 0.50 | 0.0 | 0.0000 | 0.0000 |
| 0.55 | 0.1 | 0.0165 | 0.0164 |
| 0.60 | 0.2 | 0.0453 | 0.0445 |
| 0.70 | 0.4 | 0.129 | 0.118 |
| 0.80 | 0.6 | 0.259 | 0.217 |

The approximation of H(C/W) used by Paris and Tada appears to exaggerate the stress intensity $\bar{K}$ for the deeper cracks. For cyclic bending of an edge cracked strip, the moment $\bar{M}$ contributes to the crack opening after the crack has reached the half width of the strip. The $\bar{K}$ contributes for further crack growth can then be determined from Equation (34).

8. OBSERVATIONS. The problem of "crack overlapping" occurs in several of the classical analyses found in the literature. The results of this investigation would appear to indicate that the errors so introduced are sufficient to warrant a more careful consideration of such solutions.

## References


[1] Muskhelishvili, N.I., "Some Basic Problems of Mathematical Theory of Elasticity," Nordhoff, Gromingen, Holland, (1953).

[2] Benthem, J.P. and Koiter, W.T., "Methods of Analysis and Solutions of Crack Problems," (Edited by G.C. Sih) Nordhoff, Leyden, Holland, pp. 131-178 (1973).

[3] Freese, C.E., "Collocation and Finite Elements - A Combined Method," AMMRC TR 73-28, Watertown, MA, June (1973).

[4] Freese, C.E. and Bowie, O.L., "Stress Analysis of Configurations Involving Small Fillets," Journ. of Strain Analysis, 10, pp. 53-58 (1975).

[5] Bowie, O.L. and Freese, C.E., "On the 'Overlapping' Problem in Crack Analysis," (to appear in Eng. Fracture Mechanics, (1976).

[6] Paris, P.C. and Tada, H., "The Stress Intensity Factors for Cyclic Reversed Bending of a Single Edge Cracked Strip Including Crack Surface Interference," Int. Journ. of Fracture, 11, pp. 1070-1072, (1975).

# SINGULARITY ANALYSIS BY THE FINITE ELEMENT METHOD

DENNIS M. TRACEY[1] and THOMAS S. COOK[2]
1. Army Materials & Mechanics Research Center, Watertown, Massachusetts
2. Southwest Research Institute, San Antonio, Texas

## SUMMARY

A finite element formulation is described for problems with solution functions known to have local $r^{\lambda}$ variation (s), $0<\lambda<1$, and thus singular gradients. Special 3-node triangular elements encircle the singularity and focus to share a common node at the singular point. The shape function of each triangle has the appropriate $r^{\lambda}$ mode and a smooth angular mode expressed in element natural coordinates. As with standard elements, the unknowns are the nodal values of the function. Even if the precise angular form of the asymptotic solution is known, the formulation makes no attempt to embed it, but instead piecewise approximates it. This allows assembly of the element coefficient matrix using standard procedures without nodeless variables and bandwidth complications.

The conditions of continuity, low order solution capability, and accurate numerical integration of the singularity element are discussed with a view towards establishing the general range of applicability of the formulation. Numerical applications to the elastic fracture mechanics problems of composite bondline cracking and crack branching are discussed.

## INTRODUCTION

We are considering here the problem of attaining accurate numerical representation of a function $\phi$ (x,y) when near discrete points in the domain $\phi$ varies as $r^{\lambda}$, $0<\lambda< 1$. Standard shape functions cannot properly model the singular gradient of $r^{\lambda}$ so our approach has been to design a special singularity element. Beyond embedding the proper singularity into the shape function, the usual questions of interelement continuity, constant state representation, and accurate numerical integration are addressed.

Interelement continuity should be maintained[1] for $\phi$ and its derivatives up to one order less than that occurring in the governing volume integral, denoted by I, of the problem. Subsequently, it will be shown that the singularity element has $\phi$ interelement continuity but no guaranteed continuity of $\phi$ gradients across edges. Strictly speaking then, it is limited to problems where $I = I(\phi, \partial\phi/\partial x_i)$. For example, this is the case in the potential energy formulation of elasticity where the governing

functional of displacement involves only first order derivatives. The virtual work formulation of plasticity is another such case with $\phi$ representing the displacement increment.

The other finite element convergence criterion[1] is that an element should be capable of representing fields with constant values of $\phi$, or derivatives of $\phi$ up to the order occurring in I. This is necessary because in the limit of vanishing element size, $\phi$ and its derivatives should, within the element, equal the pointwise constant values. From a practical standpoint the constancy conditions are important only when constant state conditions exist over the finite subdomain occupied by the element. The boundary conditions of a singularity problem can cause smooth as well as singular $\phi$ variations near the singular point. The constancy capability of the elements at the singular point is important only if the smooth terms are, on an element average basis, comparable in value to the singular terms. The element introduced below has $\phi$ modes of the constant and $r^{\lambda}$ type. It does not have the polynomial terms necessary to represent non-zero constant derivatives. Since the singular mode dominates the uniform mode as the singularity is approached, the lack of the latter mode is of diminishing consequence as element size is reduced, and thus convergence is achievable in this sense. However it is clear that the element is not suited for problems without an "active" singularity.

## FORMULATION

The element described here is a generalization of the singular element suggested[2] for analysis of the $r^{1/2}$ elastic crack tip singularity. The element is a 3 node triangle, and has one of its nodes at the singular point. The power form variation is chosen in the direction away from the singular point; low order smooth variation is chosen in the angular direction. Figure 1a illustrates the modeling with one of a necessary group of triangles at the singular point, node I. The shape function is developed in terms of the

oblique coordinates $\xi$, $\eta$ which vary over the range [0, 1] within the element. The radial edges correspond to $\eta = 0, 1$. The edge $\xi = 0$ is actually a point - the singular point-and the far transverse edge is $\xi = 1$. The transformation to cartesian coordinates follows

$$\underline{x} = \underline{x}^I (1 - \xi) + \underline{x}^J \xi(1 - \eta) + \underline{x}^K \xi\eta \tag{1}$$

It is straightforward to show that $\xi$ is always a linear function of r times a trigonometric function of angular orientation within the element and that $\eta$ is solely a trigonometric function of angle. As an example, the isosceles triangle of Fig. 1b has the transformation equations

$$\begin{aligned} \xi &= (r \cos\theta) / x_o \qquad\qquad = x/x_o \\ \eta &= (\tan\theta/\tan\alpha + 1)/2 \quad = (y/x \cdot x_o/y_o + 1)/2 \end{aligned} \tag{2}$$

With $\xi$ being a linear function of r, $\phi$ varies as $r^\lambda$ when $\xi^\lambda$ terms are chosen in the shape functions; such a choice yields the interpolation function

$$\phi = \phi^I (1 - \xi^\lambda) + \phi^J \xi^\lambda (1 - \eta) + \phi^K \xi^\lambda \eta \tag{3}$$

For the isosceles triangle this corresponds to

$$\begin{aligned} \phi = \phi^I (1 - (x/x_o)^\lambda) &+ 1/2\ \phi^J (1 - y/x \cdot x_o/y_o)(x/x_o)^\lambda \\ &+ 1/2\ \phi^K (1 + y/x \cdot x_o/y_o)(x/x_o)^\lambda \end{aligned} \tag{4}$$

By using a group of these elements about the singularity, the angular form of the asymptotic solution is approximated in a piecewise smooth fashion. The singular radial variation is embedded throughout the region occupied by the elements.

On the radial edges $\phi$ is a two parameter function, e.g. on IJ

$$\phi = \phi^I + (\phi^J - \phi^I)\ \xi^\lambda \qquad (5)$$

so that there is continuity of $\phi$ on these edges. On JK $\phi$ is a linear function of position which guarantees continuity with an element such as the bilinear isoparametric. Derivative continuity across element edges is not guaranteed so that, as previously discussed, the element strictly applies only to those problems whose governing integrals are independent of second and higher order $\phi$ derivatives.

The element is capable of representing a constant $\phi$ condition as can be seen by substituting a constant for the nodal values in the interpolation function and observing that $\phi$ then equals the constant. Without a linear term in the shape function the constant first derivative condition cannot be met. In analysis of deformable solids where $\phi$ would be the displacement function, situations such as rigid rotation and uniform thermal expansion correspond to a linear $\phi$ mode. The element cannot directly accommodate these cases, but by choosing a small enough element the singular mode will dominate the exact solution making the exclusion of the linear mode inconsequential.

The singular nature of the $\phi$ gradients does not preclude the possibility of accurate numerical integration in forming the coefficient matrix. It is assumed from the outset that the $r^\lambda$ variation gives rise to an integrable singularity. Standard methods of integration have been developed for polynomial variations so that these can be used only for the angular integration. In general the problem is to integrate terms of the form

$$\int_0^1 \left[ \int_0^1 f(\xi)\ \xi\ d\xi \right] g(\eta)\ d\eta \qquad (6)$$

The determinant of the Jacobian, $\partial(x, y)/\partial(\xi, \eta)$, accounts for the factor $\xi$ of the inner integrand. For the examples below a 2-point Gauss rule was used for the $\eta$ integration. The form of $f(\xi)$ must be scrutinized before

choosing a $\xi$ integration rule. For elasticity the governing integral is a quadratic function of the shape function first derivatives and this results in

$$f = \xi^{2\lambda - 2} \tag{7}$$

Hence

$$\int_0^1 f\ \xi\ d\xi = \int_0^1 \xi^{2\lambda - 1}\ d\xi = 1/2\lambda \tag{8}$$

For the elasticity examples below, the numerical technique employed to achieve precisely the result (8) was a specialized 1-point rule: one integration station was used at location $\xi = (2\lambda)^{1/(1 - 2\lambda)}$ and its weight was unity. It is easily appreciated that standard methods of integration can be very much in error for this problem, particularly for $\lambda < 0.5$. Hence, generally speaking, detailed investigation of $f(\xi)$ is required for design of an adequate integration procedure.

## EXAMPLES

The examples are problems of elastic fracture mechanics. The finite element approach employed was that based upon the principle of minimum potential energy, so that $\phi$ of the last section now stands for the displacement vector function. The first problem is the bimaterial elastic strip with a pressurized crack normal to and terminating at the bondline. The geometry is illustrated in Fig. 2. The material on the left is cracked and designated as material 1 with shear modulus $\mu_1$ and Poisson's ratio $\nu_1$; material 2 to the right has properties $\mu_2$, $\nu_2$. Crack length, plate width, and height are related by $a/b = a/h = 1/9$. The left end of the crack being surrounded completely by one material is a singular point with displacement varying as $r^{1/2}$. The bondline crack tip has a singularity dependent upon the bimaterial elastic properties. Displacement varies as $r^{\lambda}$ with $\lambda$ a function of $\mu_1/\mu_2$ and also the type of planar deformation, i.e. plane stress vs. plane strain.[3] The examples here are plane strain and the material combination is aluminum-epoxy. For aluminum $\mu = 3.846 \times 10^6$ psi,

$\nu = 0.3$; and for epoxy $\mu = 0.1667 \times 10^6$ psi, $\nu = 0.35$. With aluminum as the cracked material $m = \mu_2/\mu_1 = 0.043$ and $\lambda = 0.1752$. When epoxy is the cracked material $m = 23.08$ and $\lambda = 0.6619$.

Figure 3 shows the mesh used in the crack location. Symmetry allowed modeling just the upper half of the strip. Isosceles triangles with a radial dimension of $a/100$ and angular extent of 15° were used as singularity elements about each crack tip. Of course, about each tip the appropriate value of $\lambda$ was used to generate the element stiffnesses. The radial dimension of the singularity elements is a crucial aspect of the finite element model. The singularity elements should be entirely within the region where displacement is accurately represented by the $r^{\lambda}$ form. The crack opening displacement data from available singular integral equation solutions[3] were used to establish the suitability of the radial dimension $a/100$. When there is no basis for judgment of the range of dominance of the leading power term in the full solution, a convergence study must be conducted by successively decreasing element size to establish accuracy estimates of the singularity solution.

Bilinear isoparametric elements were used to model the plate away from the singularities. The total mesh involved 429 nodes and 433 elements. The forces specified to be acting on the crack face nodes were calculated, in terms of the uniform pressure p, consistent with the element shape functions. Thus, the singularity element node on the crack face had an applied normal force per unit thickness equal to $.01\ pa/(1 + \lambda)$.

Three features of the solutions to be discussed are the angular distribution of stress about the bondline crack tip, the crack opening behavior near the bondline, and the stress intensity factors. The angular variation of the normalized stress $\sigma_{yy}/p$ through the ring of bond tip singular elements is given in Figure 4. Data are given for both $\mu_2/\mu_1$ combinations. Along with the finite element data at the twelve discrete midpoint angles, singular integral equation (SIE) data are given at angles of 0, 90 and 180° and $r = 0.005a$. The first striking characteristic of the distribution is the discontinuity of stress across the bondline.

Independent of which material is cracked, at 90°- the bondline - the aluminum is stressed higher than the epoxy. Hence, when the epoxy is cracked $\sigma_{yy}(90^-)$ exceeds $\sigma_{yy}(90^+)$, and just the opposite when the aluminum is cracked. There is very good agreement between the SIE and finite element solutions with the exception of the $90^+$ values for $m = 0.043$. The finite element mesh is perhaps too coarse in the angular sense to accomodate the large gradient in the range 90-180° for $m = 0.043$, so that mesh refinement might improve this deviation.

In Figure 5 the normalized crack opening displacement $u_y/a$ is plotted as a function of distance from the bondline crack tip to $r/a = 0.16$ for the two $\mu_2/\mu_1$ cases. The data corresponds to a unit value of crack face pressure. The finite element data appear in discrete fashion in the plot and for comparison purposes the SIE solutions are presented and are represented by the solid curves. There is excellent agreement between the solutions for $m = 23.08$, and this is true over the entire crack face, $0 \leq r/a \leq 2$. While the SIE and finite element data agree at $r/a = 0.01$ for $m = 0.043$, the solutions differ by 5-10% over most of the crack face, including near the embedded end. There is a dramatic difference in the opening behavior local to the bondline for the two cracked cases. The SIE curves demonstrate the behavior which is expected from the $r^{0.175}$ and $r^{0.662}$ asymptotic displacement solutions. With epoxy bonded to cracked aluminum there is a rapid gradient in opening which is intuitively consistent with the stiffness mismatch. The intersection of the two curves is near $r/a = 0.01$, the location of the first finite element node, and the opening displacements $u_y/a$ there are $0.193 \times 10^{-6}$ for $m = 23.08$, and $0.222 \times 10^{-6}$ for $m = 0.043$.

The stress intensity factor, generalized for both the embedded and bondline crack tips is defined as

$$K = \lim_{r \to o} \sqrt{2}\, r^{1-\lambda}\, \sigma_{yy}(r, o) \qquad (9)$$

To deduce K from the displacement data, the following equation was used

$$K = 2\sqrt{2}\,\lambda\,\mu^{*}\,u_{y}(r,\pi)/r^{\lambda} \qquad (10)$$

The modulus $\mu^{*}$ is defined from the relationship

$$u_{y}(r,\pi) = r\,\sigma_{yy}(r,o)/2\lambda\mu^{*} \qquad (11)$$

$\mu^{*}$ is an algebraic function of the bimaterial constants and the eigenvalue $\lambda$. For the plane strain homogeneous material case, $m=1$, $\mu^{*}$ is equal to $\mu/2(1-\nu)$.

From eqn. (10), the stress intensity factor at the embedded tip, $K/p\sqrt{a}$, computed from the finite element data at $r/a=0.01$ was found to equal 0.89 when $m=23.08$, and 1.52 when $m=0.043$. For a homogeneous plate the result is 1.00, and this shows the degree to which the aluminum reduces the severity of the singularity in the cracked epoxy, and how much more severe the singularity is in aluminum when epoxy is bonded to it. The values for $K/pa^{1-\lambda}$ at the bondline crack ends are 2.85 for $m=23.08$, and 0.112 for $m=0.043$. The SIE displacement data predicts essentially the same K values with the exception of the embedded tip $m=0.043$ value which is 10% lower, consistent with the displacement deviation mentioned above. A detailed discussion of the results of the bimaterial crack problem will be reserved for a future specialized paper[4].

The second example is the branch crack in an elastic tension strip, Figure 6a. The main crack emanates from the free edge at 45° and its projected length normal to the tension is $W/4$. $W$ is the strip width, and $3W$ is the strip length. The branch normal to the tension has length $W/80$. There are two singularities in this problem each with local $r^{\lambda}$ displacement distributions. The right end of the branch has the usual crack tip singularity with $\lambda = 1/2$, while the angle on the upper face of the crack is a reentrant corner with $\lambda = 0.674$. These conclusions are drawn from the asymptotic analysis of reference[5]. The finite element mesh at the branch is shown in Figure 6b. The singularity elements were chosen to have a radial extent 5% of the branch length and an angular dimension of 22.5°

The angular variation of the normalized polar stress $\sigma_{\theta\theta}/\sigma_{\infty}$ about the bend singularity is given in Figure 7. The data are from the singularity element midpoints. The stress state is essentially entirely compressive with peak compression equal to 3.1 $\sigma_{\infty}$ at $\theta = 125°$. This suggests that forking would not occur from this point. At the right end of the branch the stress intensity factors $K_I$, $K_{II}$ were deduced from the singularity element crack face nodal displacements. If $\delta$ represents the relative opening displacement of the nodes on the two crack faces and $\Delta$ the relative sliding displacement, the equations used to determine $K_I$ and $K_{II}$ for this plane stress example were

$$8K_I = \delta E\sqrt{2/r}$$

$$8K_{II} = \Delta E\sqrt{2/r}$$

Notice that the factor $\sqrt{\pi}$ is not used in these definitions. The value of $K_I$ was found to be 4% lower than the value for a normal to the tension unbranched crack with length (1.05) W/4,

$$K_I = 1.49\,\sigma_{\infty}\sqrt{(1.05)\,W/4}$$

$K_{II}$ was determined to be negligible in relation to $K_I$, $K_{II}/K_I < 10^{-2}$. An additional problem was considered which had the above geometry altered by extending the branch length to W/40. $K_I$ again was 4% lower than that of the projected length crack

$$K_I = 1.52\,\sigma_{\infty}\sqrt{(1.10)\,W/4}$$

and $K_{II}/K_I < 10^{-4}$.

## CONCLUSIONS

The solutions to the crack problems are judged to be very accurate. The agreement between the singular integral equation and finite element results for the bimaterial problems supports this conclusion. Certainly no standard finite element formulation can be expected to provide reasonable solutions to problems such as these. The formulation proposed here allows routine analysis of a class of singularity problems which heretofore has been approached only with elaborate analytical methods. The singular element proposed is simple to implement since it is easily programmed using techniques which today are commonplace.

## REFERENCES


1. O. C. Zienkiewicz, The Finite Element Method in Engineering Science, McGraw-Hill, London, 1971

2. D. M. Tracey, "Finite Elements for Determination of Crack Tip Elastic Stress Intensity Factors," Engr. Fracture Mechs., 1971, 3, pp. 255-265.

3. T. S. Cook and F. Erdogan, "Stresses in Bonded Materials with a Crack Perpendicular to the Interface," Int. J. Engr. Sci., 1972, 10, pp. 677-697.

4. T. S. Cook and D. M. Tracey, "Stress Distribution in a Bimaterial Plate Containing a Crack Normal to the Bond," to be published.

5. M. L. Williams, "Stress Singularities Resulting From Various Boundary Conditions in Angular Corners of Plates in Extension," J. Appl. Mechs., 1952, 19, Trans. ASME, 74, Series E, pp. 526-528.

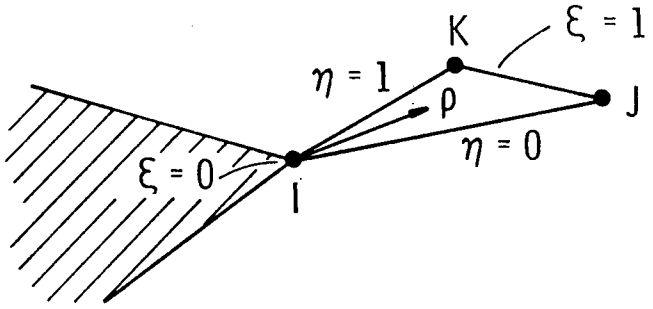



FIGURE 1a

General Triangle Terminating at Singular Point 1

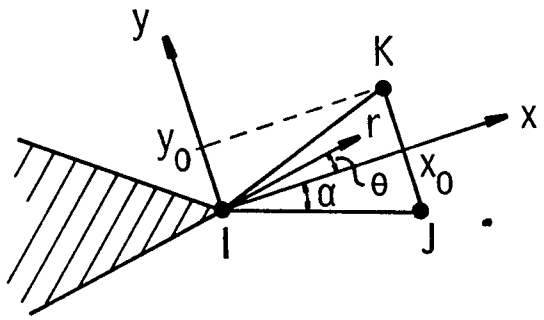



FIGURE 1b

Isosceles Triangle at Singularity

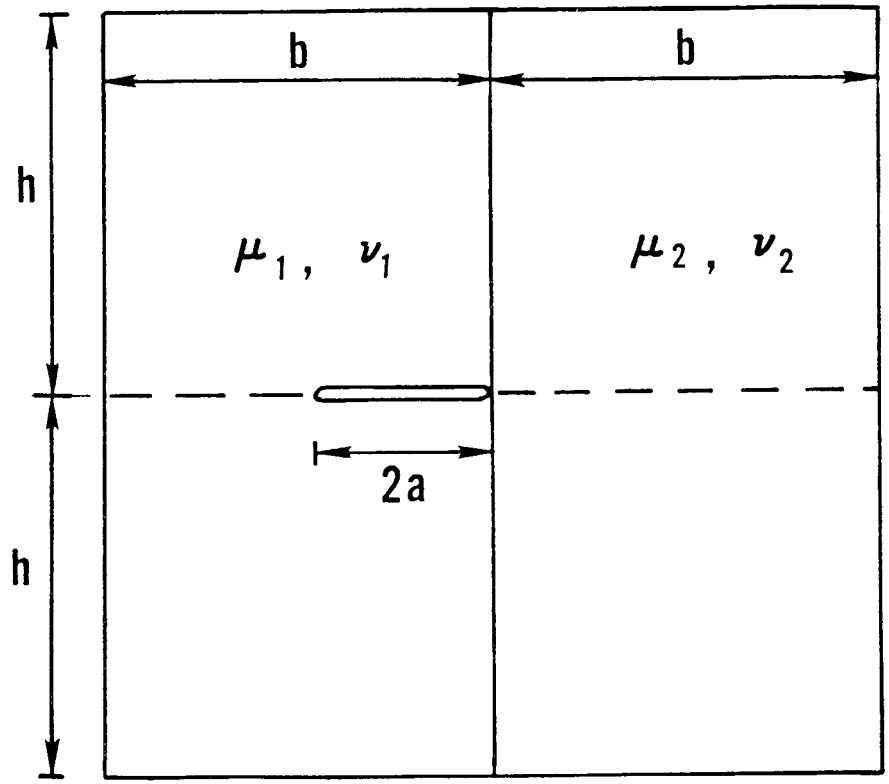



FIGURE 2
Cracked Bimaterial Strip

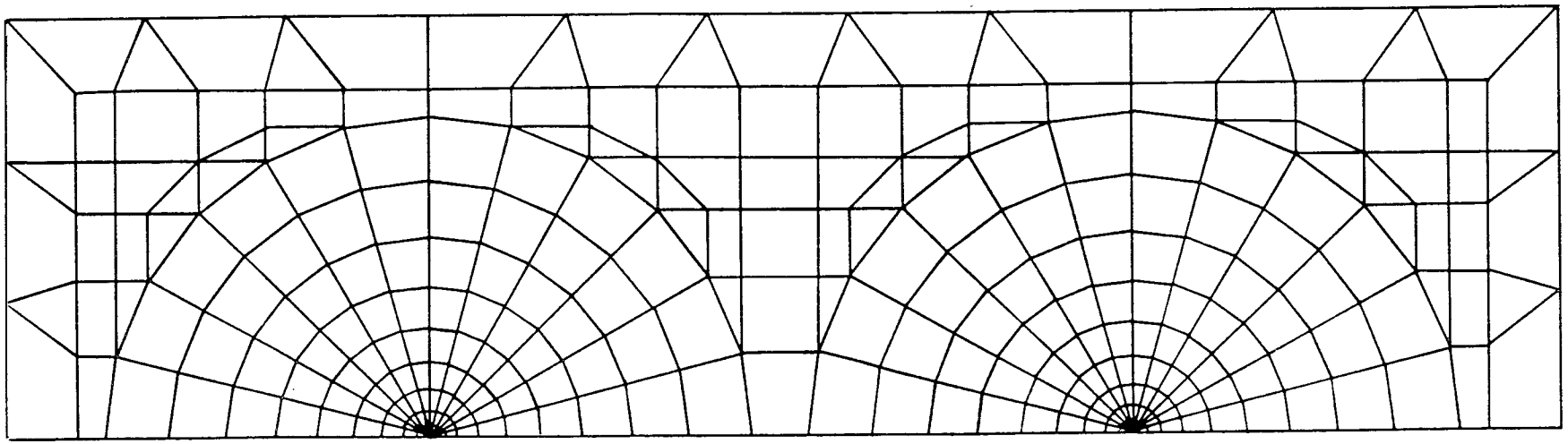


FIGURE 3

Mesh in Crack Location, Bimaterial Problem

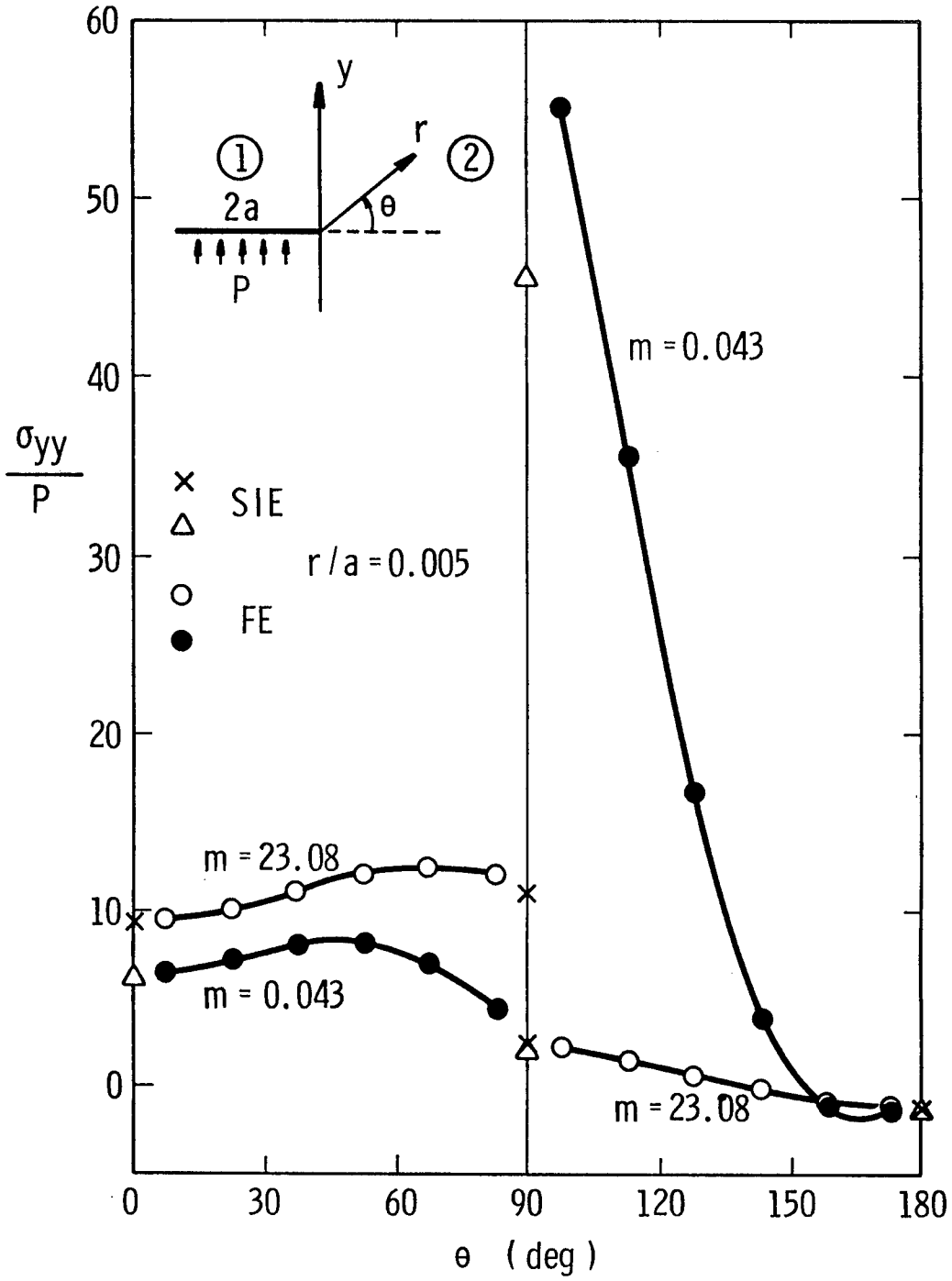



FIGURE 4

Bondline Cracktip Angular Stress Variation

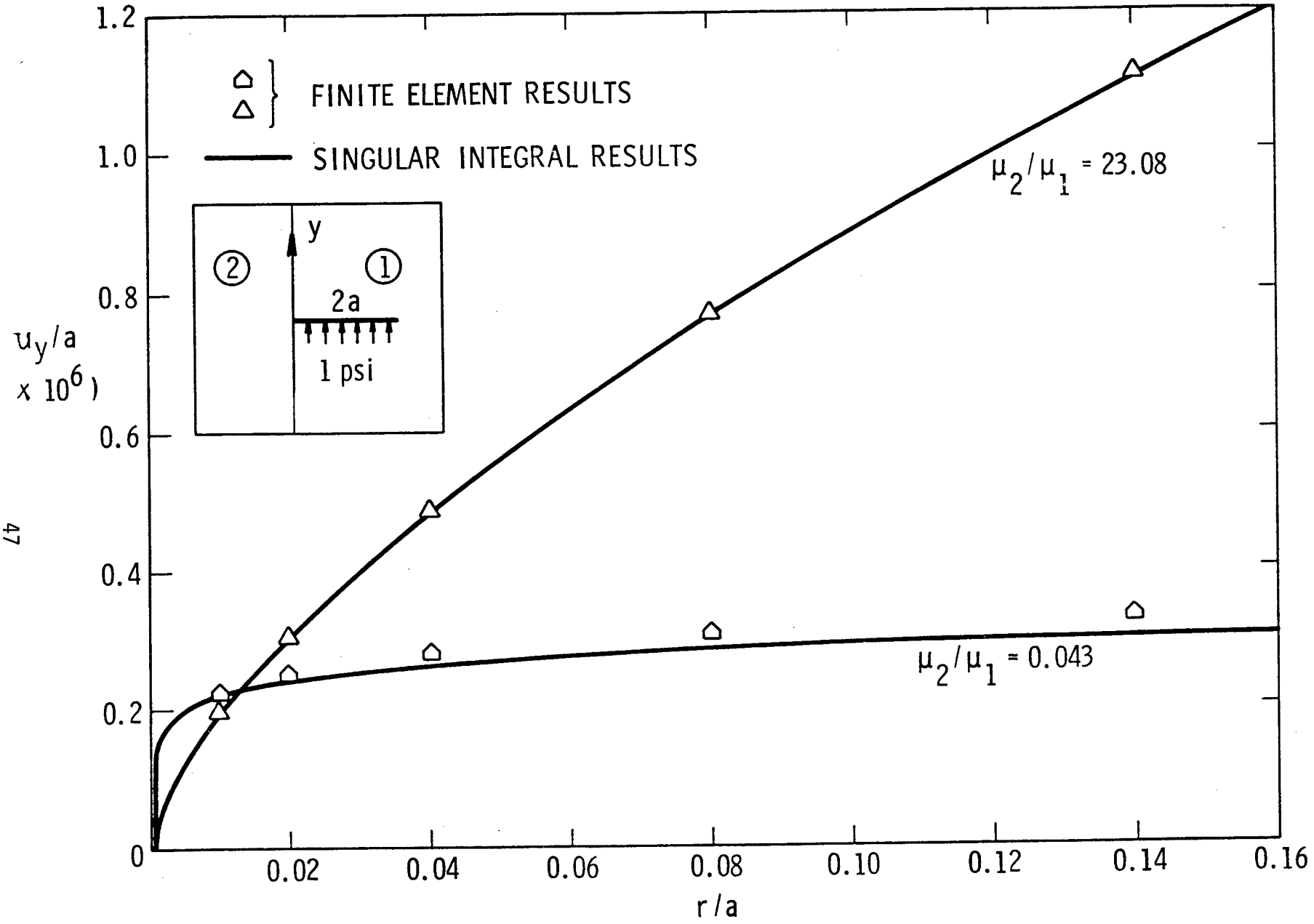



FIGURE 5

Crack Opening Displacement vs. Distance From Aluminum-Epoxy Bondline

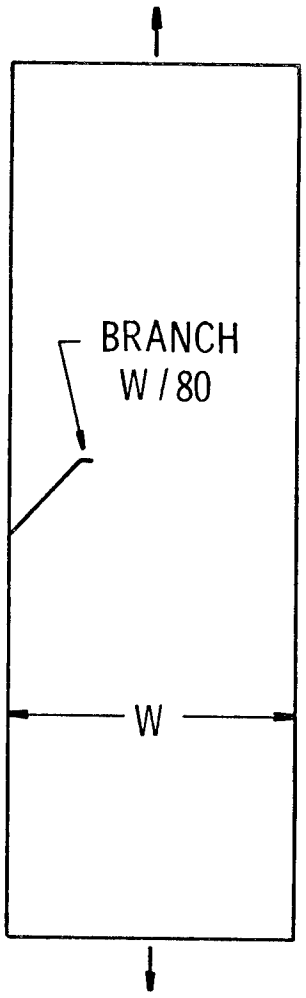



FIGURE 6a

Strip With Branch Crack

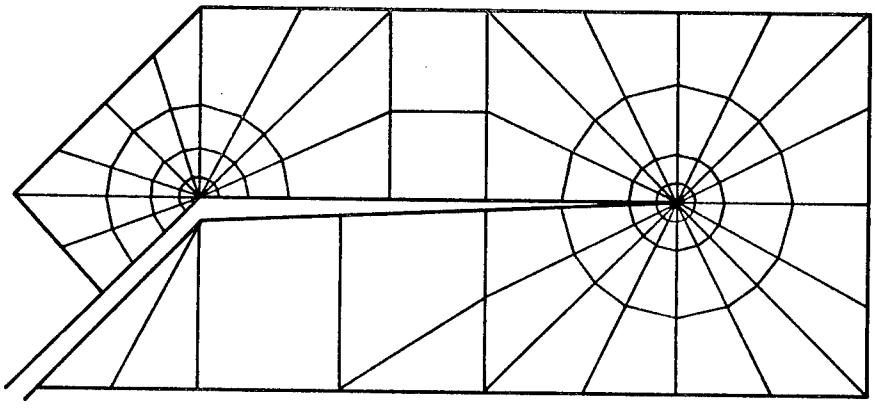


FIGURE 6b

Mesh in Location of Branch

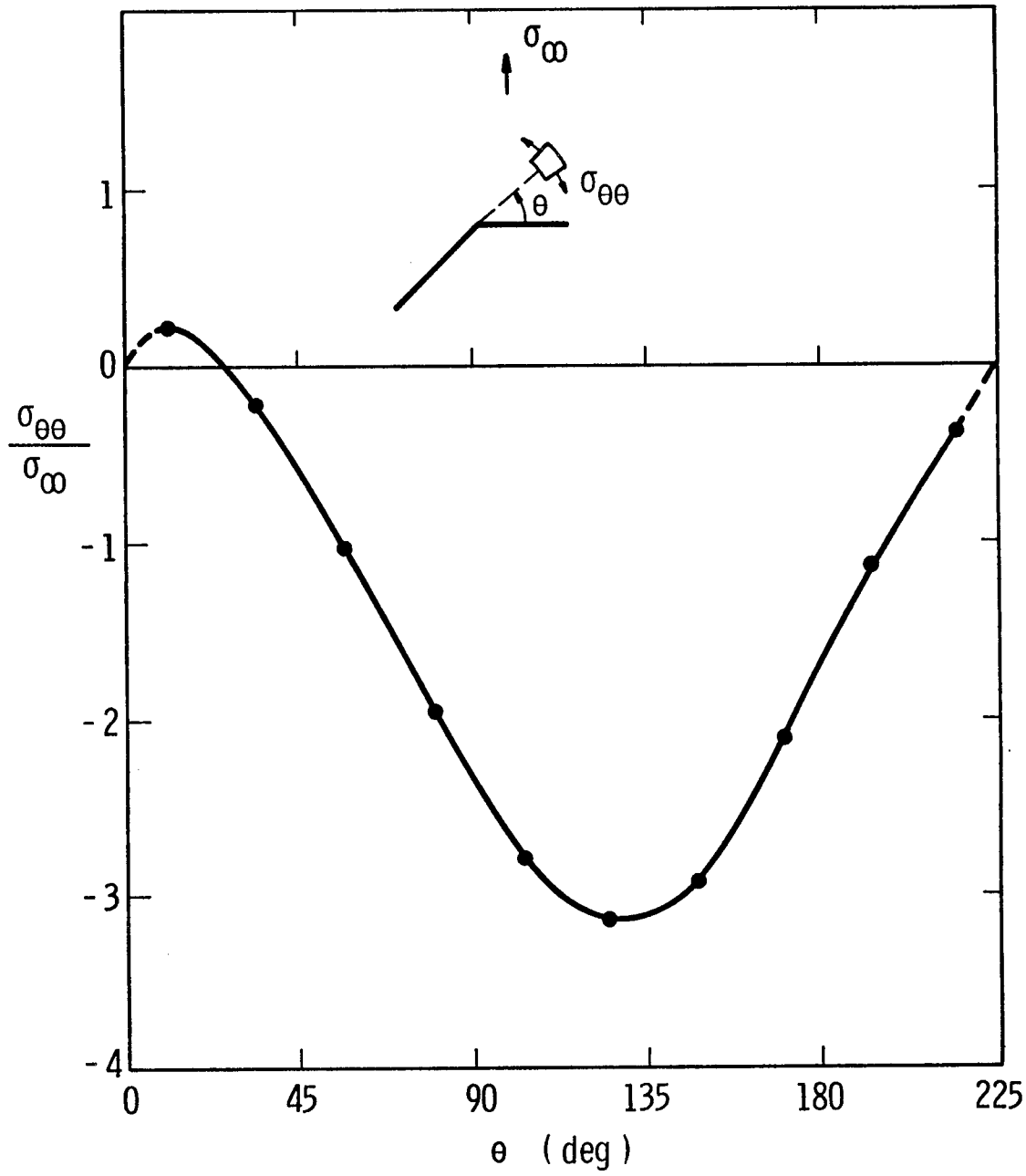



FIGURE 7

Polar Stress Variation About Bend Singularity, Branch Crack Problem

# CRACK TIP FIELDS IN STEADY CRACK GROWTH WITH LINEAR STRAIN HARDENING

John C. Amazigo
Department of Mathematical Sciences
Rensselaer Polytechnic Institute, Troy, New York 12181

and

John W. Hutchinson
Division of Engineering and Applied Physics
Harvard University, Cambridge, Massachusetts 02138

## SUMMARY

Singular stress and strain fields are found at the tip of a crack growing steadily and quasi-statically into an elastic-plastic strain hardening material. The material is characterized by $J_2$ flow theory together with a bilinear effective stress-strain curve. Anti-plane shear, plane stress and plane strain are each considered. Numerical results are given for the order of the singularity, details of the stress and strain-rate fields, and the near-tip regions of plastic loading and elastic unloading.

This paper is to be published in the Journal of Mechanics and Physics of Solids.

# FINITE-DIFFERENCE SOLUTION OF POISSON'S EQUATION IN RECTANGLES OF ARBITRARY PROPORTIONS

J. Barkley Rosser

Mathematics Research Center, University of Wisconsin, Madison, Wisconsin

## 1. Introduction.

We consider the problem of getting an approximation of reasonably good accuracy by finite-difference methods for the function $u(x, y)$ which satisfies Poisson's equation

$$\nabla^2 u(x, y) = f(x, y) \tag{1.1}$$

inside a rectangle $R$, and satisfies various boundary conditions on the boundary of $R$. When $f(x, y) \equiv 0$, (1.1) reduces to Laplace's equation, and the problem is appreciably simpler.

This problem has been much studied. A common approach is to cover $R$ exactly with a mesh or grid of small rectangles, after which one can replace (1.1) by a finite-difference approximation involving values of $u(x, y)$ at the grid points. One then tries to solve this finite-difference analogue of (1.1) to a suitable degree of accuracy. In order to employ this approach when high accuracy is required, it has been necessary to require that the ratio of the sides of $R$ must be rational since use of high order methods usually requires that one cover $R$ exactly with a grid of squares. However, the conformal transformation method of Papamichael and Whiteman [2] will lead more often than not

The author wishes to acknowledge the sponsorship of the United States Army under Contract No. DAAG29-75-C-0024 and of the Science Research Council under grant B/RG 4121 at Brunel University.

to a rectangle in which the ratio is not rational, and covering with a grid of squares is not possible. Even when the ratio is rational, there may be difficulties. Suppose, from some engineering problem, one is confronted with a rectangle $R$ of base six and five-eighths and height five and seven-eighths. If this is to be covered exactly with squares, there must be $53N$ squares along the base and $47N$ squares along a vertical side, where $N$ is a positive integer. With such a covering, many popular methods would operate at less than maximum efficiency.

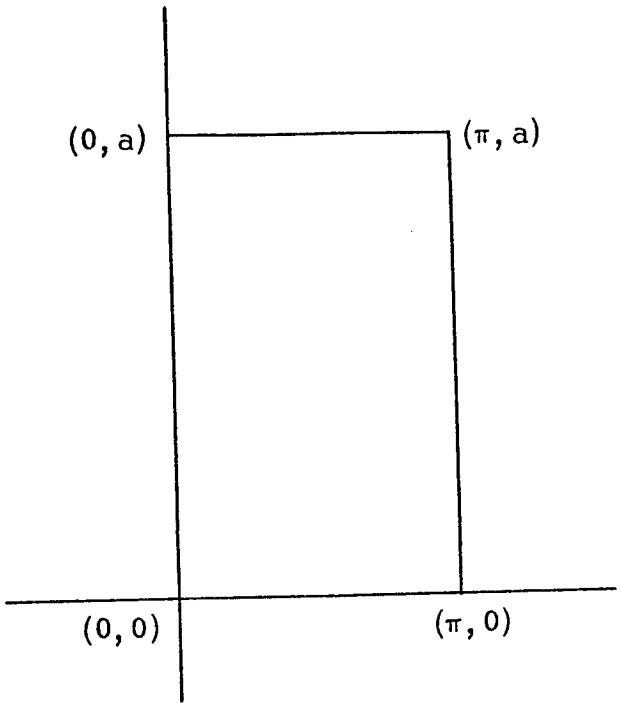
Accordingly, we will propose a method of getting good accuracy with moderate labor for rectangles of arbitrary proportions.

## 2. Formulation of the problem.

By rotation, translation, and scaling, as needed, we can take the rectangle $R$ to be that shown in Figure 1. By rotating through another $90^\circ$ and translating and scaling again, if need be, we can assure that $a \geq \pi$. If $a = \pi$, we have a square, and familiar approaches suffice. So we assume $a > \pi$.

We consider first the case of Dirichlet boundary conditions. That is, we wish to approximate the function $u(x, y)$ which is continuous on and inside $R$, satisfies

$$\nabla^2 u(x, y) = f(x, y) \tag{2.1}$$




The rectangle R

Figure 1

inside R, and on the sides of R satisfies the Dirichlet boundary conditions

(2.2) $u(0, y) = g_o(y) \qquad 0 < y < a$

(2.3) $u(\pi, y) = g_\pi(y) \qquad 0 < y < a$

(2.4) $u(x, 0) = h_o(x) \qquad 0 < x < \pi$

(2.5) $u(x, a) = h_a(x) \qquad 0 < x < \pi$ .

Because we seek a $u(x, y)$ which is continuous on R, as well as inside, we are thereby assuming that $g_o(y)$ and $g_\pi(y)$ are continuous for $0 \leq y \leq a$, that $h_o(x)$ and $h_a(x)$ are continuous

for $0 \leq x \leq \pi$, and that

$$g_o(0) = h_o(0) , \tag{2.6}$$

$$g_o(a) = h_a(0) , \tag{2.7}$$

$$g_\pi(0) = h_o(\pi) , \tag{2.8}$$

$$g_\pi(a) = h_a(\pi) . \tag{2.9}$$

If there should be discontinuities in the boundary conditions, or their derivatives, this would induce still another source of errors in the solutions, besides those due to truncation and round off. See Rosser [3]. "Jump" discontinuities can be "removed" by the methods on pp. 221-222 of Milne [4]. More complicated discontinuities can sometimes be "removed", but one cannot count on doing this. For the present treatment, we assume that the boundary conditions and their low order derivatives are continuous. This includes continuity at the corners, as exemplified by (2.6) through (2.9). Or, if we replace (2.2) by

$$u_x(0, y) = j_o(y) \qquad 0 < y < a ,$$

then continuity of the first derivatives at the corners would require

$$j_o(0) = h'_o(0)$$

$$j_o(a) = h'_a(0) .$$

3. Finite-difference approximations.

There are finite-difference approximations of various orders. The higher order methods of solution, involving the higher order approximations,

can be used effectively only when the function $f(x, y)$ which appears in (2.1) has suitable high order smoothness; that is, when it is continuous and has continuous derivatives of suitable orders. Thus the reader must exercise discrimination in choosing which order method to use. When they can be used, the high order methods permit the use of coarse meshes. This can greatly reduce the labor of computation.

For difference approximations of order 2, one can use mesh elements which are rectangles, rather than squares. See Hockney [1]. In this case, there would be no trouble if the ratio of the sides of $R$ were irrational. For difference approximations of order 4, one can also use mesh elements which are rectangles. See Rosser [5]. For difference approximations of order 6, it appears that the mesh elements have to be squares. Details are presented in Rosser [5]. If $f(x, y)$ in (2.1) is sufficiently smooth, this permits one to use quite a coarse mesh, greatly reducing the computational labor. However, this raises the question how to proceed if the ratio of the sides of $R$ is irrational.

4. Ill-proportioned rectangles.

We take $h$ to be the side of the square mesh element. We arrange that the squares can be fitted along the base of $R$. That is, we take $M$ to be a positive integer, and define

$$h = \frac{\pi}{M} . \tag{4.1}$$

We take $N$ to be the integer part of $aM/\pi$; in symbols

$$N = \left[\frac{aM}{\pi}\right] . \tag{4.2}$$

Then

$$Nh \leq a , \tag{4.3}$$

$$(N + 1)\, h > a . \tag{4.4}$$

If

$$Nh = a , \tag{4.5}$$

then we can fill up the rectangle $R$ exactly with $MN$ squares of side $h$, and the methods of Rosser [5] are applicable. So we are interested only in the case $Nh < a$. We could assume this, but it is not required for the analysis which follows. If we should have (4.5) holding, then some of the steps of the subsequent analysis would be quite trivial but not incorrect in any way.

We begin by defining

$$b = Nh \tag{4.6}$$

$$c = a - b = a - Nh . \tag{4.7}$$

We take $R_b$ to be the rectangle with corners $(0,0)$, $(0,b)$, $(\pi,0)$, and $(\pi,b)$, and take $R_c$ to be the rectangle with corners $(0,c)$, $(0,a)$, $(\pi,c)$ and $(\pi,a)$.

We choose $h_b(x)$ to be a smooth function such that

$$h_b(0) = g_o(b)$$

$$h_b(\pi) = g_\pi(b) .$$

The better we can choose $h_b(x)$ to approximate $u(x, b)$; the more we can curtail certain computations later. With the limited information available at this stage, we content ourselves with taking

$$h_b(x) = h_a(x) + (1 - \frac{x}{\pi})(g_o(b) - h_a(0)) + \frac{x}{\pi}(g_\pi(b) - h_a(\pi)) .$$

We take $u_b(x, y)$ to be the function which is continuous on and inside $R_b$, satisfies (2.1) inside $R_b$, and on the sides of $R_b$ satisfies the boundary conditions

$$u_b(0, y) = g_o(y) \qquad 0 \le y \le b \tag{4.8}$$

$$u_b(\pi, y) = g_\pi(y) \qquad 0 \le y \le b \tag{4.9}$$

$$u_b(x, 0) = h_o(x) \qquad 0 \le x \le \pi \tag{4.10}$$

$$u_b(x, b) = h_b(x) \qquad 0 \le x \le \pi . \tag{4.11}$$

We take $u_c(x, y)$ to be the function which is continuous on and inside $R_c$, satisfies (2.1) inside $R_c$, and on the sides of $R_c$ satisfies the boundary conditions

$$u_c(0, y) = g_o(y) \qquad c \le y \le a \tag{4.12}$$

$$u_c(\pi, y) = g_\pi(y) \qquad c \le y \le a \tag{4.13}$$

$$u_c(x, c) = u_b(x, c) \qquad 0 \le x \le \pi \tag{4.14}$$

$$u_c(x, a) = h_a(x) \qquad 0 \le x \le \pi . \tag{4.15}$$

By our definition of $h_b(x)$, we see that $u_b(x, y)$ has continuous boundary conditions around the rectangle $R_b$. Then it follows by (4.14) that the same holds for $u_c(x, y)$ relative to the rectangle $R_c$. This is why in (4.8) through (4.15) we can use $\le$ rather than $<$.

By (4.1) and (4.6) we can fill up the rectangle $R_b$ exactly with $MN$ squares of side $h$. Thus we can use the 9-point difference approximation of Rosser [5] to get accurate approximations for $u_b(x, y)$ inside $R_b$ at the grid points $(mh, nh)$. From these, we can get accurate approximations for $u_b(mh, c)$. By (4.14) these are part of the boundary values for $u_c(x, y)$. Thus it is necessary to determine them to order $h^6$. By the principle of the maximum, it is also sufficient. For a given $m$, the point $(mh, c)$ is on a vertical grid line. Thus one can determine $u_b(mh, c)$ to order $h^6$ by using a high order interpolation formula in one dimension on the values at the six grid points $(mh, 0)$, $(mh, h)$, $(mh, 2h)$, $(mh, 3h)$, $(mh, 4h)$, and $(mh, 5h)$.

By (4.14), this gives us good approximations to $u_c(x, c)$ at $x = h, 2h, \ldots, (M - 1)h$. By (4.1) and (4.7) we can fill up the rectangle $R_c$ exactly with $MN$ squares of side $h$. Thus we can use the 9-point difference approximation of Rosser [5] to get accurate approximations for $u_c(x, y)$ inside $R_c$ at the grid points $(mh, c + nh)$. Then we can get accurate approximations for $u_c(mh, b)$ by the method mentioned earlier.

We define $R_{bc}$ to be the rectangle which is the intersection of the rectangles $R_b$ and $R_c$. In $R_{bc}$, the function $u_c(x, y) - u_b(x, y)$ is harmonic. Also, it is zero along the bottom and along the two vertical sides. So on and inside $R_{bc}$ we have

$$u_c(x, y) - u_b(x, y) = \sum_{r=1}^{\infty} a_r \frac{\sinh r(y - c)}{\sinh r(b - c)} \sin rx \tag{4.16}$$

where

$$a_r = \frac{2}{\pi} \int_0^{\pi} \{u_c(x, b) - u_b(x, b)\} \sin rx \, dx \, . \tag{4.17}$$

Clearly the $|a_r|$ are bounded by

$$2 \max_{0 \le x \le \pi} |u_c(x, b) - u_b(x, b)| \, . \tag{4.18}$$

We recall (see (4.11)) that

$$u_b(x, b) = h_b(x) \, .$$

Presumably $u_c(x, b)$ is fairly close to $u(x, b)$. If also we were lucky enough to choose $h_b(x)$ fairly close to $u(x, b)$, then by (4.18) the $a_r$ will be fairly small. This will save computational effort later.

On and inside $R$ define

$$v(x, y) = \sum_{r=1}^{\infty} a_r b_r \frac{\sinh r(a - y)}{\sinh ra} \sin rx \, , \tag{4.19}$$

where

$$b_r = \frac{\sinh rc}{\sinh r(b - c)} \, . \tag{4.20}$$

On and inside $R_b$ define

$$u(x, y) = u_b(x, y) + v(x, y) + \sum_{r=1}^{\infty} a_r \frac{\sinh r(y - c)}{\sinh r(b - c)} \sin rx \, . \tag{4.21}$$

We see that $u(x, y)$ is continuous on and inside the rectangle $R_b$, satisfies (2.1) inside $R_b$, and on three sides satisfies the boundary conditions (4.8), (4.9), and (4.10). By (4.16), we see that on and inside $R_{bc}$ we have

$$u(x, y) = u_c(x, y) + v(x, y) . \tag{4.22}$$

We use (4.22) to define $u(x, y)$ for the rest of the rectangle $R_c$. Then $u(x, y)$ is continuous on and inside the rectangle $R_c$, satisfies (2.1) inside $R_c$, and on three sides satisfies the boundary conditions (4.12), (4.13), and (4.15).

Thus we see that $u(x, y)$ is exactly the function $u(x, y)$ that we were seeking to obtain.

We have obtained accurate approximations for $u_b(x, y)$ and $u_c(x, y)$ at various grid points. If $M$ is of reasonable size, then $c$ is small, since $0 \leq c < h$ by (4.7), (4.3), and (4.4). As $a$ is greater than $\pi$, and $b = a - c$ by (4.7), we see that the series on the right of (4.19) is rapidly convergent for $0 \leq y \leq a$. Also, the series appearing on the right of (4.21) is rapidly convergent for small $y$, certainly for $0 \leq y \leq h$. If in addition the $a_r$ are all quite small (see (4.18)), then very few terms of the series are needed to get high accuracy. So, using the known approximations for $u_b(mh, nh)$, we can get approximate values for $u(x, y)$ for small $y$ by (4.21). For all other values of $y$, we can use the known approximations for $u_c(mh, c + nh)$ to get approximate values for $u(x, y)$ by (4.22).

The calculation of the $a_r$ presents no problem. Not more than four or five will be required; fewer if the $a_r$ are all small. Observe that the values of $u_b(x, b)$ are given by (4.11). Also, we had got accurate approximations for $u_c(mh, b)$. So we can use a numerical quadrature formula to calculate the $a_r$ by (4.17).

CAUTION. If $r$ is not fairly small compared to $N$, then there will be fairly few abscissa points in each cycle of $\sin rx$ in (4.17); in such case the usual quadrature formulas are not trustworthy. One can get twice, or four times, or eight times, as many abscissa points by interpolating to get approximations for $u_c(x, b)$ at the additional abscissa points (recall that $u_b(x, b)$ is given by (4.11)). For this interpolation one can use a high order one dimensional interpolation formula on the values $u_c(0, b)$, $u_c(h, b)$, $u_c(2h, b), \ldots$ .

We need high accuracy for only the first one or two of the $a_r$, because of the very rapid convergence of the series appearing on the right of (4.19) and (4.21). In any case, one should increase the number of abscissa points, as needed, to the point where one can use a quadrature formula with assurance. Also, by a little foresight in the choice of $M$, one can arrange that, after increasing the number of abscissa points if needed, one can use a high order quadrature formula, like Bode's Rule, for example.

5. Tests for accuracy.

One advantage of using the 9-point difference approximation when one can exactly fill up the rectangle with squares is that one can make a first calculation, for less than a quarter of the calculating effort, with squares twice as large on a side, and then repeat with the smaller squares. Because the error is of the order of $h^6$, one can get an estimate of the error.

This can be done with the present procedure by choosing $M$ divisible by 2. If $N$ is not divisible by 2, the values of $b$ and $c$ which are used with the squares of side $2h$ will not be the same as those which are used with the squares of side $h$. However, this does not matter.

One dividend that will accrue from making an initial calculation with squares of side $2h$ is that from this calculation one can derive a very good approximation to take for $h_b(x)$. Then, for the calculation with squares of side $h$, the $a_r$ will be very small, so that not more than two or three of them will be needed.

6. Neumann boundary conditions.

Suppose we have the same rectangle $R$, and impose on $u(x, y)$ the same conditions as before, except that on top of the rectangle $R$ we specify values to be taken by $u_y(x, a)$. That is we replace (2.5) by the Neumann condition

$$u_y(x, a) = k_a(x) \qquad 0 < x < \pi . \tag{6.1}$$

We postpone to the latter part of the section a discussion of how one would handle this in the case in which $a/\pi$ is rational, so that one can fill up $R$ exactly with squares of side $h$. For the moment, let us assume that this can be done, and explain how to generalize to the case in which $a/\pi$ is irrational.

We proceed very nearly as in Section 4. Instead of the definition given there of $h_b(x)$, we use

$$h_b(x) = (1 - \frac{x}{\pi})\, g_o(b) + \frac{x}{\pi}\, g_\pi(b) \,. \tag{6.2}$$

We take $u_b(x, y)$ as before, but for $u_c(x, y)$ we replace (4.15) by the analogue of (6.1), namely

$$\frac{\partial}{\partial y}\, u_c(x, a) = k_a(x) \qquad 0 < x < \pi \,. \tag{6.3}$$

Everything now goes the same, down to the definition of $v(x, y)$. Let us pause a moment, and think what we require of $v(x, y)$. Clearly it should be harmonic, so that $u(x, y)$, as defined in part by (4.21) and in part by (4.22), will satisfy (2.1) inside $R$. Also, we wish $v(x, y)$ to be zero on the vertical sides of $R$, so that there $u(x, y)$ will satisfy the proper boundary conditions. Also, on the bottom of $R$, we must have

$$v(x, 0) = \sum_{r=1}^{\infty} a_r \frac{\sinh rc}{\sinh r(b - c)} \sin rx \qquad 0 < x < \pi \tag{6.4}$$

so that by (4.21) $u(x, y)$ will satisfy the right boundary conditions on the bottom of $R$. Finally, looking at (4.22), we see that if $u(x, y)$

is to satisfy the right boundary conditions on the top of R, we must have

(6.5) $$v_y(x, a) = 0 \qquad 0 < x < \pi .$$

All these conditions can be met by simply replacing the factor

$$\frac{\sinh r(a - y)}{\sinh ra}$$

in the definition of $v(x, y)$ by

$$\frac{\cosh r(a - y)}{\cosh ra} .$$

In this case, since it is unlikely that (6.2) makes $h_b(x)$ come out very close to $u(x, b)$, we cannot count on the $a_r$ being particularly small, so that two or three more of them might have to be calculated. It might be better to turn the rectangle R upside down and proceed as follows.

Consider next the case in which the Neumann condition is at the bottom of R. That is, $u(x, y)$ satisfies (2.2), (2.3), and (2.5), but (2.4) is replaced by

(6.6) $$u_y(x, 0) = k_o(x) \qquad 0 < x < \pi .$$

Again, we proceed nearly as in Section 4. We can now take $h_b(x)$ the same as in Section 4, which should lead to smaller values of the $a_r$, so that we can get by with calculating fewer of them. For the definition of $u_b(x, y)$, we replace (4.10) by the analogue of (6.6), namely

(6.7) $$\frac{\partial}{\partial y} u_b(x, 0) = k_o(x) \qquad 0 < x < \pi .$$

We take $u_c(x, y)$ as in Section 4, and continue the same down to the definition of $v(x, y)$. A key requirement is that $u(x, y)$, as defined by (4.21), shall satisfy the proper boundary conditions at the bottom of R. In Section 4, this required that

$$(6.8) \qquad v(x, y) + \sum_{r=1}^{\infty}{}' a_r \frac{\sinh r(y - c)}{\sinh r(b - c)} \sin rx$$

should be zero when $y = 0$. This was accomplished by the proper choice of the $b_r$. Now we must assure that the partial derivative of (6.8) with respect to $y$ shall be zero when $y = 0$. Again, this is accomplished by the proper choice of the $b_r$; specifically we now take

$$(6.9) \qquad b_r = \frac{-\sinh ra}{\sinh r(b - c)} \frac{\cosh rc}{\cosh ra} .$$

All else remains the same.

Next consider the case in which there are Neumann conditions both at the top and the bottom of R. That is, $u(x, y)$ satisfies (2.2) and (2.3), but (2.4) is replaced by (6.6) and (2.5) is replaced by (6.1). We proceed much as in Section 4. In the definition of $u_b(x, y)$ we replace (4.10) by (6.7), and in the definition of $u_c(x, y)$ we replace (4.15) by (6.3). We define $h_b(x)$ by (6.2). It is then easily verified that we should replace

$$\frac{\sinh r(a - y)}{\sinh ra}$$

in the definition of $v(x, y)$ by

$$\frac{\cosh r(a - y)}{\cosh ra}$$

and define

$$b_r = \frac{\cosh ra}{\sinh r(b-c)} \frac{\cosh rc}{\sinh ra} . \tag{6.10}$$

One can of course have Neumann conditions on one or both of the vertical sides. Let us consider first the case in which there are Neumann conditions on both vertical sides, but Dirichlet conditions at the top and bottom. Rotation by $90^\circ$ would reduce this to the case just considered. However, this is not desirable, since we would then lose the qualification that the height is greater than the base. It was this that assured the rapid convergence of the Fourier series in (4.19) and (4.21).

So we assume that (2.4) and (2.5) hold, but that (2.2) and (2.3) are replaced by

$$u_x(0, y) = j_o(y) \qquad 0 < y < a \tag{6.11}$$

$$u_x(\pi, y) = j_\pi(y) \qquad 0 < y < a . \tag{6.12}$$

We proceed analogously to Section 4, except that we use cosines instead of sines throughout. Because it is desirable to have $u_x(x, y)$ continuous around the boundary we define

$$h_b(x) = h_a(x) + \frac{1}{2\pi}(x - \pi)^2(h_a'(0) - j_o(b)) + \frac{x^2}{2\pi}(j_\pi(b) - h_a'(\pi)) . \tag{6.13}$$

We define $u_b(x, y)$ and $u_c(x, y)$ as in Section 4, except that they now have Neumann conditions on their vertical sides. We replace (4.16) and (4.17) by

$$u_c(x, y) - u_b(x, y) = \sum_{r=0}^{\infty} a_r \frac{\sinh r(y - c)}{\sinh r(b - c)} \cos rx \tag{6.14}$$

where

$$a_0 = \frac{1}{\pi} \int_0^{\pi} \{u_c(x, b) - u_b(x, b)\}dx \tag{6.15}$$

$$a_r = \frac{2}{\pi} \int_0^{\pi} \{u_c(x, b) - u_b(x, b)\}\cos rx\, dx\ . \tag{6.16}$$

When $r = 0$, we define

$$\frac{\sinh r(y - c)}{\sinh r(b - c)} = \frac{y - c}{b - c}\ .$$

Exactly analogous changes are made in (4.19) and (4.21).

If, in addition to the Neumann conditions on the vertical sides, we replace one or both of the Dirichlet conditions on the top or bottom by Neumann conditions, we can modify the procedure just outlined quite analogously to the way in which we modified the procedure of Section 4 earlier in this section.

It will be noted that we are allowing the possibility of Neumann conditions on all four sides. For this, there will be a solution only if the boundary conditions satisfy a certain criterion. If they do, the solution is not unique, but any two solutions differ by a constant. The procedure outlined will produce one of this infinity of solutions if and only if there is a solution.

To handle the case of a Dirichlet condition on the left side and a Neumann condition on the right side, we replace $\sin rx$ by

$$\sin(r - \tfrac{1}{2})x\ ,$$

with suitable related changes. To handle the case of a Dirichlet condition on the right side and a Neumann condition on the left side, we replace $\sin rx$ by

$$\cos(r - \tfrac{1}{2})x .$$

We consider finally how to handle the case in which the rectangle has a rational ratio of the sides, and we have filled it exactly with squares of side $h$, and wish to approximate $u(x, y)$ at the grid points. At interior grid points, we can use one of the formulas of Rosser [5]. On boundaries where there are Dirichlet boundary conditions, we assign $\bar{u}_{m,n}$ the specified value. This leaves only the boundary points where there is a Neumann condition to be dealt with. Suppose, for example, that the condition (6.11) holds on the left side of $R$. We note that

$$(6.17) \qquad hf_x(x, y) \cong -\frac{137}{60} f(x, y) + 5f(x + h, y) - 5f(x + 2h, y) + \frac{10}{3} f(x + 3h, y) - \frac{5}{4} f(x + 4h, y) + \frac{1}{5} f(x + 5h, y)$$

holds to within terms of order $h^6$. If we take $x = 0$ and $y = nh$, we get by (6.11)

$$(6.18) \qquad hj_o(nh) \cong -\frac{137}{60} \bar{u}_{0,n} + 5\bar{u}_{1,n} - 5\bar{u}_{2,n} + \frac{10}{3} \bar{u}_{3,n} - \frac{5}{4} \bar{u}_{4,n} + \frac{1}{5} \bar{u}_{5,n} .$$

One could use a higher order formula than (6.17), but it probably suffices. A heuristic argument for this is as follows. By the principle of the maximum, if we wish to determine interior points to order $h^6$,

it is sufficient to determine the boundary points to order $h^6$. However, if the interior points are given to order $h^6$, (6.18) will determine $\bar{u}_{0,n}$ to order $h^6$.

Use of (6.18) with the formulas of Rosser [5] results in a rather messy matrix of coefficients of the $\bar{u}_{m,n}$. However, one is probably using such a coarse mesh that this matrix would be less than $100 \times 100$, perhaps even less than $50 \times 50$. If so, probably the quickest method of solution is to use the standard computer routine for solving simultaneous linear equations. If this is done, it does not much matter if the matrix is messy or not.

If it happens that one is solving the Laplace equation, with $f(x, y) \equiv 0$, and has a zero normal derivative along one side, say $j_o(y) \equiv 0$, one can use the reflection principle to replace (6.18) by something which seems conceptually simpler. However, it involves three boundary grid points and three interior points, and so is probably about as much bother on a computer as (6.18), which also involves six grid points.

If one has Neumann conditions on one or more sides, and so is using (6.18), one might consider the following procedure, which would bypass the treatment in Section 4 altogether. Almost always, there is at least one side with Dirichlet conditions. By rotating and relinquishing the qualification $a > \pi$, if need be, we can arrange to have Dirichlet

conditions on top. If, in the notation of Section 4, we have $0 < c < h$, the difficulty is that we have no good way to write down an equivalent of (3.7) of Rosser [5] for the values of $u(x, y)$ at the row of grid points $(mh, Nh)$, $1 \leq m \leq M - 1$. As a substitute, write down (3.7) of Rosser [5] for the 9-point formula centered at $(mh, a - h)$. It involves values of $u(x, y)$ at $((m - 1)h, a - h)$, $((m - 1)h, a - 2h)$, $(mh, a - h)$, $(mh, a - 2h)$, $((m + 1)h, a - h)$, $((m + 1)h, a - 2h)$, as well as at the boundary points $((m - 1)h, a)$, $(mh, a)$, and $((m + 1)h, a)$, at which latter points $u(x, y)$ is known. Now, by a high order one dimensional interpolation formula, we can write each of $u(rh, a - h)$ and $u(rh, a - 2h)$, approximately as a linear combination of $u(rh, nh)$ for $n \leq N$; we do this for $r = m - 1$, $r = m$, and $r = m + 1$. So we get a formula involving $u(rh, Nh)$, $u(rh, (N - 1)h)$, etc., for $r = m - 1$, $m$, $m + 1$, which we can use in place of (3.7) of Rosser [5]. Probably interpolation of order eight should be used. This makes the matrix still messier, but if we are having to deal with a messy matrix anyhow, because of the Neumann conditions, the idea might be worth considering.

## REFERENCES


[1] R. W. Hockney, "The potential calculation and some applications," Methods in Computational Physics, vol. 9 (1970), pp. 135-211.

[2] N. Papamichael and John R. Whiteman, "A numerical conformal transformation method for harmonic mixed boundary value problems in polygonal domains," Zeit. für Angew. Math. Phys., vol. 24 (1973), pp. 304-316.

[3] J. Barkley Rosser, "Effect of discontinuous boundary conditions on finite-difference solutions," Technical Report TR/30, Brunel University, 1973, and MRC Technical Summary Report #1383, 1973. To appear in Zeit. für Angew. Math. Phys.

[4] William E. Milne, "Numerical solution of differential equations," John Wiley and Sons, Inc., New York, 1960.

[5] J. Barkley Rosser, "Nine-point difference equations for Poisson's equation," MRC Technical Summary Report #1523, 1975. To appear in Computers & Mathematics with Applications.

# SOLUTIONS TO INITIAL VALUE PROBLEMS USING FINITE ELEMENTS - UNCONSTRAINED VARIATIONAL FORMULATIONS

Julian J. Wu
Research Directorate
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. This paper presents a variational formulation which treats initial value problems and boundary problems in a unified manner. The basic ingredients of this theory are (1) adjoint variable and (2) unconstrained variations. It is an extension of the finite element-unconstrained variational formulation used previously in solving several nonconservative stability problems. The technique which makes this extension possible is described. This formulation thus enables one to adapt such numerical technique as the finite element method, which has had great success and popularity for solution of boundary value problems, for solutions of initial value problems as well. These formulations are given here for a forced vibration problem, a heat (mass) transfer problem and a wave propagation problem. Numerical calculations in conjunction with finite elements for two specific examples are obtained and compared with known exact solutions.

1. INTRODUCTION. In its application to the solutions of engineering problems, the finite element discretization has been implemented almost exclusively to the spatial dimensions. For dynamic or time-dependent problems whose solutions as functions of time are of interest, a step-by-step procedure of finite difference, i.e., the quasi-static approach is usually employed. The answer to the question why the time dimension has not been treated equally with the spatial variables in the finite element discretization must be related, in part at least, to the development of variational methods, since the finite element procedure can be viewed most readily as an extremizing sequence associated with a variational statement. While there are numerous variational principles for boundary value problems, few exist for initial value problems. Like many problems involving nonconservative forces, the difficulty appears to be that initial value problems are nonself-adjoint and thus they do not possess variational principles in the classical sense. In conjunction with problems involving nonconservative forces, certain constrained variational principles (sometimes called extended Hamilton's principles —See, for example, ref. [1]) were used for finite element solution formulations [2, 3]. Shortly afterwards, using the combined notion of the Lagrange multipliers and the adjoint variable, some unconstrained variational statements were established and used as bases for finite element solutions [4, 5]. This approach has been shown to be more

advantageous in terms of simplicity, versatility and the rate of convergence compared with the constrained variational approach [5, 6].

Fried was first to treat the time-dimension identically with the space dimensions in using the finite elements [7]. His solution formulations, however, emanate from constrained variational principles. In contrast, this paper presents a generalization of the unconstrained variational approach to time-dependent problems.

At this point, the variational principles of integrals of convolution developed by Gurtin [8, 9] should be mentioned. The applications of these principles in conjunction with finite elements in the time-dimension [10, 11, 12, 13] have so far failed to show any advantage over the procedure described by Fried. In fact, all these analyses had to resort to either the Fried's or some other similar step-by-step procedure to complete the solutions in the time-dimension.

In this paper, the use of unconstrained variational principles - finite elements for usual boundary value problems is first illustrated and the advantages over the constrained formulations are pointed out. The unconstrained variational principles can always be constructed through the use of the Lagrange multipliers. The unconstrained variations are then shown to lead naturally to (nonself-) adjoint variational statements. Thus, nonconservative problems can be formulated easily using finite elements. The application to a control problem is given [14]. With the introduction of a cross-product term involving two-point boundary (initial) values, the unconstrained variational - finite element formulation is again easily extended to include time-dependent problems. This formulation is obviously simpler compared with those derived from Gurtin's variational principles because no convolutional integrals are needed. It is also easier to use and more versatile than the Fried's procedure due to the fact that no boundary or initial conditions are involved in the solution formulation and because of the nature of the Lagrange multipliers. As further examples of application, finite element matrix equations are derived for several transient problems including a force vibration, a heat transfer and a wave propagation problem. Detailed formulations and numerical results of two examples are given and comparisons with some known exact solutions are made.

2. LAGRANGE MULTIPLIER AND FINITE ELEMENT FORMULATIONS. One of the advantages of the finite element method is its capability of solving large complicated problems in a routine manner. However, the same concepts used in a program for large systems may be understood using relatively simple problems.

Let us consider the stability of a Euler's column. The governing equations are as follows:

D.E. $$E\,I\,u'''' + P\,u'' + \omega^2\,\rho A u = 0 \tag{1a}$$

B.C. $$u(0) = u'(0) = 0 \tag{1b),(1c}$$

$$u''(\ell) = 0 \tag{1d}$$

$$E\,I\,u'''(\ell) + P\,u'(\ell) = 0 \tag{1e}$$

where u is the lateral displacement, a prime (') denotes differentiation with respect to the coordinate x; E is the Young's modulus, $\rho$, density of the material; I is the second moment, A, area of the cross-section, $\ell$, length of the beam and $\omega$ is the eigenvalue. For eqs.(1), a usual variational principle can be written:

$$\delta J_1\,(u) = 0 \tag{2a}$$

where

$$J_1(u) = \frac{1}{2}\int_0^{\ell} [E\,I\,(u'')^2 - P(u')^2 + \omega^2\rho A u^2]\,dx \tag{2b}$$

To establish the equivalence between eqs. (1) and (2), one simply carries out the variation of $J_1$ in eq. (2a):

$$\delta J_1 = \int_0^{\ell} [E\,I\,u''\delta u'' - P\,u'\delta u' + \omega^2\,\rho A\,u\,\delta u]\,dx \tag{3a}$$

$$= \int_0^{\ell} [E\,I\,u'''' + P\,u'' + \omega^2\,\rho A u]\,\delta u\,dx$$

$$+ [E\,I\,u''\,\delta u' - (E\,I\,u''' + P\,u')\,\delta u]_{x=\ell}$$

$$- [E\,I\,u''\,\delta u' - (E\,I\,u''' + P\,u')\,\delta u]_{x=0} \tag{3b}$$

From eq. (3b) one observes that for the coordinate functions and their variations satisfying the boundary conditions in eqs. (1b - 1e), eq. (1a) implies eq. (2a) and vice versa. The finite element formulation for this problem begins with eq. (3a).

Let

$$u(x) = \underset{\sim}{a}^T(x)\,\underset{\sim}{U} \tag{4}$$

where $\underset{\sim}{a}(x)$ is the displacement-function vector and $\underset{\sim}{U}$, the generalized displacement vector. Upon the substitution of eq. (4) into eq. (3a), one immediately obtains

$$\delta\underset{\sim}{U}^T \left\{ \underset{\sim}{K}_1 + \omega^2\underset{\sim}{M} \right\} \underset{\sim}{U} = 0 \tag{5}$$

where

$$\underset{\sim}{K}_1 = \int_0^{\ell} [E\,I\,\underset{\sim}{a}''\,\underset{\sim}{a}''^T - P\underset{\sim}{a}'\,\underset{\sim}{a}'^T]\,dx \tag{6a}$$

$$\underset{\sim}{M} = \int_0^{\ell} \rho A\,\underset{\sim}{a}\,\underset{\sim}{a}^T\,dx \tag{6b}$$

Eq. (5) is not yet ready to be solved since neither $\tilde{U}$ nor $\delta\tilde{U}$ consists of independent elements due to the boundary conditions requirements placed on u(x).

Let us now consider a slightly different variational principle:

$$\delta J_2 = 0 \tag{7a}$$

with

$$J_2 = \frac{1}{2}\int_0^{\ell} [E\,I\,(u'')^2 - P\,(u')^2 + \omega^2\,\rho A u^2]\,dx + \frac{1}{2}\,\alpha_1\,[u(0)]^2 + \frac{1}{2}\,\alpha_2\,[u'(0)]^2 \tag{7b}$$

where $\alpha_1$ and $\alpha_2$ are the Lagrange multipliers.

Carrying out the variation of eqs. (7), we have

$$\delta J_2 = \int_0^{\ell} [E\,I\,u''\,\delta u'' - Pu'\,\delta u + \omega^2\rho A u^2]\,dx + \alpha_1\,u(0)\,\delta u(0) + \alpha_2\,u'(0)\,\delta u'(0) \tag{8a}$$

$$= \int_0^{\ell} [E\,I\,u'''' + Pu'' + \omega^2\,\rho A u]\,\delta u\,dx + [E\,I\,u''\,\delta u' - (E\,I\,u''' + Pu')\,\delta u]_{x=\ell} - [(E\,I\,u'' - \alpha_2 u' - (E\,I\,u''' + Pu' + \alpha_1 u)\,\delta u]_{x\,=\,0} \tag{8b}$$

Eq. (8b) states that a necessary and sufficient condition for $\delta J_2 = 0$ is the problem defined by the following set of equations:

$$E\,I\,u'''' + Pu'' + \omega^2\rho A u = 0 \tag{9a}$$

$$E\,I\,u''(0) - \alpha_2\,u'(0) = 0 \tag{9b}$$

$$E\,I\,u'''(0) + Pu'(0) + \alpha_1\,u(0) = 0 \tag{9c}$$

$$E\,I\,u''(\ell) = 0 \tag{9d}$$

$$E\,I\,u'''(\ell) + Pu'(\ell) = 0 \tag{9e}$$

provided that the variation $\delta u$ is completely arbitrary, comparing eqs. (9) and (1), it is seen that eqs. (1) is a special case of (9) as $\alpha_1$, $\alpha_2$ approach to infinity. From eq. (8a), we can see that the finite element matrix equation now becomes

$$\delta \underset{\sim}{U}^T \left\{ \underset{\sim}{K}_2 + \omega^2 \underset{\sim}{M} \right\} \underset{\sim}{U} = 0 \tag{10}$$

where

$$\underset{\sim}{K}_2 = \underset{\sim}{K}_1 + \alpha_1 \underset{\sim}{a}(0) \underset{\sim}{a}^T(0) + \alpha_2 \underset{\sim}{a}'(0) \underset{\sim}{a}'^T(0) \tag{11}$$

The matrix $\underset{\sim}{K}$ in eq. (11) has been defined in eq. (5) and the superscript T denotes the transpose of a matrix (a vector). Since $\delta u$ is arbitrary, $\delta \underset{\sim}{U}$ in eq. (10) is arbitrary, eq. (10) leads directly to the final matrix equation to be solved.

$$\left\{ \underset{\sim}{K}_2 + \omega^2 \underset{\sim}{M} \right\} \underset{\sim}{U} = 0 \tag{12}$$

It is then clear that the method of Lagrange multipliers, used in conjunction with the finite element method, will not only facilitate the solution formulations but also encompass a larger class of problems to be solved compared with the use of constrained variational statements. The applications of the same general concept can be extended further.

## 3. FROM UNCONSTRAINED VARIATIONS TO ADJOINT VARIATIONAL STATEMENTS.

We have noted that the variation $\delta u$ in eq. (8) is quite independent of the function $u$ itself and nothing will be changed if we simply replace $\delta u$ with $\delta v$ to emphasize this independence. This substitution, however, has suggested the adjoint variational principles. Let us consider

$$\delta J_3 = 0 \tag{13a}$$

$$J_3 = \int_0^{\ell} (E\,I\,u''v'' - Pu'v' + \omega^2 \rho A u v)\, dx$$
$$+ \alpha_1 u(0)v(0) + \alpha_2 u'(0)v(0) + \alpha_3 Pu'(\ell)v(\ell) \tag{13b}$$

Carrying out the variations, we have:

$$\delta J_3 = (\delta J_3)_u + (\delta J_3)_v \tag{14}$$

where

$$(\delta J_3)_u = \int_0^{\ell} (E\,I\,u''\delta v'' - Pu'\delta v' + \omega^2 \rho A u \delta v)\, dx$$
$$+ \alpha_1 u(0)\delta v(0) + \alpha_2 u'(0)\delta v'(0) + \alpha_3 u'(\ell)\, \delta v(\ell) \tag{15a}$$
$$= \int_0^{\ell} (E\,I\,u'''' + Pu'' + \omega^2 \rho A u)\, \delta v\, dx$$
$$+ [E\,I\,u''\delta v' - (E\,I\,u''' + Pu' - \alpha_3 u')\, \delta v]_{x=\ell}$$

$$- \left[ (E\,I\,u'' - \alpha_2 u')\,\delta v' - (E\,I\,u''' + Pu' + \alpha_1 u)\,\delta v \right]_{x=0} \qquad (15b)$$

and

$$(\delta J_3)_v = \int_0^{\ell} (E\,I\,v''\,\delta u'' - Pv'\delta u' + \omega^2 \rho A v \delta u)\,dx$$

$$+ \alpha_1 v(0)\,\delta u(0) + \alpha_2 v'(0)\,\delta u'(0) + \alpha_3 v(\ell)\,\delta u'(\ell) \qquad (16a)$$

$$= \int_0^{\ell} (E\,I\,v'''' + Pv'' + \omega^2 \rho A v)\,\delta u\,dx$$

$$+ \left[ (E\,I\,v'' + \alpha_3 v)\,\delta u' - (E\,I\,v''' + P\,v')\,\delta u \right]_{x=\ell}$$

$$- \left[ (E\,I\,v'' - \alpha_2 v')\,\delta u' - (E\,I\,v''' + Pv' + \alpha_1 v)\,\delta u \right]_{x=0} \qquad (16b)$$

From eq. (15a), it is clear that a necessary and sufficient condition for $(\delta J_3)_u = 0$ is the problem defined by the following set of equations:

D.E. $$E\,I\,u'''' + Pu'' + \omega^2 \rho A u = 0 \qquad (17a)$$

B.C. $$E\,I\,u''(\ell) = 0 \qquad (17b)$$

$$E\,I\,u'''(\ell) + (P - \alpha_3)\,u'(\ell) = 0 \qquad (17c)$$

$$E\,I\,u''(0) - \alpha_2 u'(0) = 0 \qquad (17d)$$

$$E\,I\,u'''(0) + P\,u'(0) + \alpha_1 u(0) = 0 \qquad (17e)$$

Now eqs. (9) has become a special case of eqs. (17) when $\alpha_3 = 0$. In addition, the problem defined by $(\delta J_3)_v = 0$ of eqs. (16) is called the adjoint problem to eqs. (17). For $\alpha_3 = 0$, the adjoint problem is identical to the problem itself — hence, the self-adjoint system. Now, considering

$$\alpha_3 = k\,P \qquad (18)$$

in eq. (17c), we have

$$E\,I\,u'''(\ell) - K\,P\,u'(\ell) = 0 \qquad (19)$$

$$K = k - 1 \qquad (20)$$

Eq. (19) defines the boundary condition of a general non-conservative load. It is also clear from eq. (19) that $K$ is a dimensionless design constant which defines the small angle between the direction of the applied load $P$ and the tangent of the deflected column at the end. Since $(\delta J_3)_u = 0$ alone defines the boundary value problem of eq. (17) and vice versa, we need not at all to be concerned with the adjoint problem. Now it is a simple matter to modify the finite element matrix equation as

$$\left\{ \underset{\sim}{K}_3 + \omega^2 \underset{\sim}{M} \right\} \underset{\sim}{U} = 0 \tag{22}$$

where

$$\underset{\sim}{K}_3 = \underset{\sim}{K}_2 + \alpha_3 \, \underset{\sim}{a}'(\ell) \, \underset{\sim}{a}^T(\ell) \tag{23}$$

## 4. FINITE ELEMENTS FOR INITIAL AND INITIAL-BOUNDARY VALUE PROBLEMS.

(1) A Forced Vibration Problem. Let us first consider a problem of "one" degree of freedom, i.e., a mass-spring system. The differential equation and initial conditions are

$$m \ddot{u} + k u = f(t), \quad 0<t<T \tag{24a}$$

$$u(0) = u_0 \tag{24b}$$

$$\dot{u}(0) = u_1 \tag{24c}$$

where $u(t)$ is the displacement of the mass centre from its equilibrium position, $m$, the amount of mass and $k$, the spring constant. The function $f(t)$ is given, so are the constants $u_0$ and $u_1$. The constant $T$ appeared in the bounds of eq. (24a) is any given positive number other than infinity. In order to formulate approximate solutions for eqs. (24) the way we did in the previous section, let us consider a more general case

$$m \ddot{u} + k u = f(t) \tag{25a}$$

$$\dot{u}(T) - \alpha \, [\, u(0) - u_0 \,] = 0 \tag{25b}$$

$$\dot{u}(0) = u_1 \tag{25c}$$

where $\alpha$ is a parameter, obviously eqs. (25) reduce to (24) when $\alpha$ approaches to $\infty$. Now, with eqs. (25), we are able to write an unconstrained variational statement as follows:

$$\delta J_4 = 0 \tag{26a}$$

where

$$J_4 = \int_0^T [ - m\dot{u}\dot{v} + kuv - f(t)v ] \, dt$$

$$+ m\alpha [u(0) - u_0] v(T) - mu_1 v(0) \qquad (26b)$$

Since

$$(\delta J_4)_u = \int_0^T [ - m\dot{u}\delta\dot{v} + ku\delta v - f(t) \delta v ] \, dt$$

$$+ m\alpha [ u(0) - u_0 ] \delta v(T) - mu_1 \delta v(0) \qquad (27a)$$

$$= \int_0^T [ m\ddot{u} + ku - f(t) ] \delta v \, dt$$

$$- m \left\{ \dot{u}(T) - \alpha [u(0) - u_0] \right\} \delta v(T)$$

$$+ m [\dot{u}(0) - u_1 ] \delta v(0) \qquad (27b)$$

The already familiar form of eqs. (27) state that (a), $(\delta J)_u = 0$ is a necessary and sufficient condition for eqs. (25), and (b), eq. (27a) provides us the finite element matrix equation. Thus, if we assume as before that

$$u(t) = \underset{\sim}{a}^T(t) \, \underset{\sim}{U}$$

$$v(t) = \underset{\sim}{a}^T(t) \, \underset{\sim}{V}$$

Eq. (27a) yields

$$\delta \underset{\sim}{V}^T \, \underset{\sim}{K}_4 \, \underset{\sim}{U} = \delta \underset{\sim}{V}^T \, \underset{\sim}{F} \qquad (28)$$

where

$$\underset{\sim}{K}_4 = \int_0^T ( -m \, \dot{\underset{\sim}{a}} \, \dot{\underset{\sim}{a}}^T + k \, \underset{\sim}{a} \, \underset{\sim}{a}^T) \, dt$$

$$+ m\alpha \, \underset{\sim}{a}(T) \, \underset{\sim}{a}^T(0) \qquad (29)$$

and

$$\underset{\sim}{F} = \int_0^T f(t) \, \underset{\sim}{a} \, dt + m\alpha u_0 \, \underset{\sim}{a}(t) + m \, \dot{u}_0 \, \underset{\sim}{a}(0) \qquad (30)$$

Again, since $\delta \underset{\sim}{V}$ is unconstrained eq. (28) leads directly to

$$\underset{\sim}{K}_4 \, \underset{\sim}{U} = \underset{\sim}{F} \qquad (31)$$

which is the final equation to be solved.

(2) A Heat Conduction Problem. The one dimensional transient heat conduct problem can be described by the equation

$$\frac{\partial}{\partial x}\left(K \frac{\partial u}{\partial x}\right) - \rho c \frac{\partial u}{\partial t} - f(x,t) = 0 \tag{32a}$$

Boundary and initial conditions are

$$u(0,t) = g_0(t) \tag{32b}$$

$$u(L,t) = g(t) \tag{32c}$$

$$u(x,0) = h(x) \tag{32d}$$

where $K$ = thermal conductivity

$\rho$ = material density

$c$ = specific heat

$f(x,t)$ = heat source function

and

$g_0(t)$, $g_1(t)$ and $h(x)$ are prescribed functions

Let us consider

$$\delta J_5 = 0 \tag{33a}$$

$$J_5 = -\int_0^L \int_0^T \left[K \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} + \rho c \frac{\partial u}{\partial t} v + f(x,t)\, v\right] dt\, dx$$

$$+ \int_0 \alpha K \left[u(L,t) - g_1(t)\right] v(L,t)\, dt$$

$$- \int_0^T \alpha K \left[u(0,t) - g_0(t)\right] v(0,t)\, dt$$

$$- \int_0^L \rho c \left[u(x,0) - h(x)\right] v(x,0)\, dx \tag{33b}$$

since

$$(\delta J_5)_u = \int_0^L \int_0^T \left[K \frac{\partial u}{\partial x} \delta\left(\frac{\partial v}{\partial x}\right) + \rho c \frac{\partial u}{\partial t} \delta v + f(x,t)\, \delta v\right] dx\, dt$$

$$+ \int_0^T \alpha K \left[u(L,t) - g_1(t)\right] \delta v(L,t)\, dt$$

$$\int_0^T \alpha K \left[u(0,t) - g_0(t)\right] \delta v(0,t)\, dt$$

$$\int_0^L \rho c \left[u(x,0) - h(x)\right] \delta v(x,0)\, dx \tag{34a}$$

$$= \int_0^L \int_0^T \left[ \frac{\partial}{\partial x} \left( K \frac{\partial u}{\partial x} \right) - \rho c \frac{\partial u}{\partial t} - f(x,t) \right] \delta v \, dx dt$$

$$- \int_0^T K \left\{ \frac{\partial u(L,t)}{\partial x} - \alpha [u(0,t) - g_1(t)] \right\} \delta v(L,t) \, dt$$

$$+ \int_0^T K \left\{ \frac{\partial u(0,t)}{\partial x} - \alpha [u(0,t) - g_0(t)] \right\} \delta v(0,t) \, dt$$

$$+ \int_0^L \rho c \, [u(x,0) - h(x)] \, \delta v(x,0) \, dx \qquad (34b)$$

it is clear that $(\delta J_5)_{u = 0}$ is a necessary and sufficient condition for eqs (32) as $\alpha \to \infty$ and eq. (34a) provides the finite element matrix equation. We can write from eq. (34a),

$$- \int_0^L \int_0^T \left[ K \frac{\partial u}{\partial x} \delta \left( \frac{\partial v}{\partial x} \right) + \rho c \frac{\partial u}{\partial t} \delta v \right] dx dt$$

$$+ \int_0^T \alpha K \, [u(L,t) \, \delta v(L,t) - u(0,t) \, \delta v(0,t)] \, dt$$

$$+ \int_0^L \rho \, c \, u(x,0) \, \delta v(x,0) \, dt$$

$$= \int_0^L \int_0^T f(x,t) \, \delta v \, dx dt$$

$$+ \int_0^T \alpha K \, [g_1(t) \, \delta v(L,t) \, g_0(t) \, \delta v(0,t)] \, dt$$

$$+ \int_0^L \rho \, c \, h(x) \, \delta v(x,0) \, dx \qquad (35)$$

Now, let

$$u(x,t) = \underset{\sim}{a}^T(x,t) \, \underset{\sim}{U} \qquad (36a)$$

$$v(x,t) = \underset{\sim}{a}^T(x,t) \, \underset{\sim}{V} \qquad (36b)$$

in the usual manner, we have

$$\delta \underset{\sim}{V}^T \, \underset{\sim}{K} \, \underset{\sim}{U} = \delta \underset{\sim}{V}^T \, \underset{\sim}{F} \qquad (37)$$

$$\underset{\sim}{K} = - \int_0^L \int_0^T (K \, \underset{\sim}{a}_{,x} \, \underset{\sim}{a}^T_{,x} + \rho \, c \, \underset{\sim}{a} \, \underset{\sim}{a}^T_{,t}) \, dx \, dt$$

$$+ \int_0^T \alpha K \, [\underset{\sim}{a}(L,t) \, \underset{\sim}{a}(L,t) - \underset{\sim}{a}(0,t) \, \underset{\sim}{a}^T(0,t)] \, dt$$

$$+ \int_0^L \rho \, c \, \underset{\sim}{a}(x,0) \, \underset{\sim}{a}^T(x,0) \, dx \qquad (38)$$

and

$$F = \int_0^L \int_0^T f(x,t)\ a(x,t)\ dxdt + \int_0^T \alpha K\ [g_1(t)\ a(L,t) - g_0(t)\ a(0,t)]\ dt + \int_0^L \rho\ c\ h(x)\ \underset{\sim}{a}(x,0)\ dx \tag{39}$$

Again, since $\delta\underset{\sim}{V}$ in eq. (37) is completely arbitrary, we arrive at the final matrix equation to be solved.

$$\underset{\sim}{K}\ \underset{\sim}{U} = \underset{\sim}{F} \tag{40}$$

(3) A Wave Propagation Problem. For a quite general wave propagation problem, the following system can be written.

$$\frac{\partial^2 u}{\partial x^2} - c^2\ \frac{\partial^2 u}{dt^2} = f(x,t). \tag{41a}$$

$$u(0,t) = g_0(t) \tag{41b}$$

$$u(L,t) = g_1(t) \tag{41c}$$

$$u(x,0) = h_0(x) \tag{41d}$$

$$\dot{u}(x,0) = h_1(x) \tag{41e}$$

The extension of the previous formulation to this problem is straight forward. Let us consider

$$\delta J_6 = 0 \tag{42a}$$

where

$$J_6 = \int_0^L \int_0^T \left[ - \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + c^2\ \frac{\partial u}{\partial t}\frac{\partial v}{\partial t} - f(x,t)\ v\right] dxdt - \alpha \int_0^T [u(L,t) - g_1(t)]\ v(L,t)\ dt + \alpha \int_0^T [u(0,t) - g_0(t)]\ v(0,t)\ dt - \alpha \int_0^L [u(x,0) - h_0(x)]\ v(x,T)\ dx + \int_0^L [u(x,0) - h_1(x)]\ v(x,0)\ dx \tag{42b}$$

Again,

$$(\delta J_6)_u = \int_0^L \int_0^T \left[ - \frac{\partial u}{\partial x} \delta\left( \frac{\partial v}{\partial x} \right) + c^2 \frac{\partial u}{\partial t} \delta\left( \frac{\partial v}{\partial t} \right) - f(x,t)\delta v \right] dx dt$$

$$- \alpha \int_0^T [u(L,t) - g_1(t)] \delta v(L,t) \, dt$$

$$+ \alpha \int_0^T [u(0,t) - g_0(t)] \delta v(0,t) \, dt$$

$$- \alpha \int_0^L [u(x,0) - h_0(x)] \delta v(x,T) \, dx$$

$$+ \int_0^L [u(x,0) - h_1(x)] \delta v(x,0) \, dx \qquad (43a)$$

$$= \int_0^L \int_0^T \left[ \frac{\partial^2 u}{\partial x^2} - c^2 \frac{\partial^2 u}{\partial t^2} - f(x,t) \right] \delta v \, dx dt$$

$$+ \int_t \left\{ \frac{\partial u}{\partial x}(L,t) - \alpha [u(L,t) - g_1(t)] \right\} \delta v(L,t) \, dt$$

$$- \int_t \left\{ \frac{\partial u}{\partial x}(0,t) - \alpha [u(0,t) - g_0(t)] \right\} \delta v(0,t) \, dt$$

$$+ \int_V \left\{ \frac{\partial u}{\partial t}(x,t) - \alpha [u(x,0) - h_0(x)] \right\} \delta v(x,T) \, dx$$

$$- \int_V \left[ \frac{\partial u}{\partial t}(x,0) - h_1(x) \right] \delta v(x,0) \, dx \qquad (43b)$$

From eqs. (43), it is again clear that $(\delta J_6)_u = 0$ is a necessary and sufficient condition for eqs. (41) as $\alpha \to \infty$ and that eq. (43a) will yield the finite element matrix equation. From(43a) one has:

$$\int_0^L \int_0^T \left\{ - \frac{\partial u}{\partial x} \delta\left(\frac{\partial v}{\partial x}\right) + c^2 \frac{\partial u}{\partial t} \delta\left(\frac{\partial v}{\partial t}\right) \right\} dx \, dt$$

$$- \alpha \int_t u(L,t) \, \delta v(L,t) \, dt + \alpha \int_t u(o,t) \, \delta v(0,t) \, dt$$

$$- \alpha \int_V u(x,0) \, \delta v(x,T) \, dx$$

$$= \iint f(x,t) \, \delta v(x,t) \, dx \, dt$$

$$\alpha \int_t g_1(t)\ \delta\, v(L,t)\ dt + \alpha \int_t g_0(t)\ \delta v(0,t)\ dt$$

$$- \alpha \int_v h_0(x)\ \delta v(x,T)\ dx + \int_x h_1(x)\ \delta v(x,0)\ dx \qquad (44)$$

Again, let

$$u(x,t) = \underset{\sim}{a}^T(x,t)\ \underset{\sim}{U} \qquad (45a)$$

$$v(x,t) = \underset{\sim}{a}^T(x,t)\ \underset{\sim}{V} \qquad (45b)$$

Eq. (44) becomes, in matrix form,

$$\delta\, \underset{\sim}{V}^T\ \underset{\sim}{K}\ \underset{\sim}{U} = \delta\, \underset{\sim}{V}^T\ \underset{\sim}{F} \qquad (46)$$

where

$$\underset{\sim}{K} = \int_0^T \int_0^L (-\ \underset{\sim}{a}'\ \underset{\sim}{a}'^T + c^2\ \dot{\underset{\sim}{a}}\ \dot{\underset{\sim}{a}}^T)\ dxdt$$

$$- \alpha \int_0^T \underset{\sim}{a}(L,t)\ \underset{\sim}{a}^T(L,t)\ dt + \alpha \int_0^T \underset{\sim}{a}(0,t)\ \underset{\sim}{a}^T(0,t)\ dt$$

$$- \alpha \int_0^L \underset{\sim}{a}(x,t)\ \underset{\sim}{a}^T(x,0)\ dt \qquad (47)$$

$$\underset{\sim}{F} = \int_0^T \int_0^L f(x,t)\ \underset{\sim}{a}(x,t)\ dxdt$$

$$- \alpha \int_0^T g_1(t)\ \underset{\sim}{a}(L,t)\ dt + \alpha \int_0^T g_0(t)\ \underset{\sim}{a}(0,t)\ dt$$

$$+ \alpha \int_0^L h_0(x)\ \underset{\sim}{a}(x,T)\ dt + \int_0^L h_1(x)\ \underset{\sim}{a}(x,0)\ dt \qquad (48)$$

Due to the arbitrariness of $\delta\underset{\sim}{V}$ , eq. (46) leads directly to the final matrix equation

$$\underset{\sim}{K}\ \underset{\sim}{U} = \underset{\sim}{F} \qquad (49)$$

5. NUMERICAL DEMONSTRATIONS. Several numerical examples will be given in this section to demonstrate the application of the formulation described so far.

(1) Forced Vibration. We shall consider a special case of the forced vibration problem formulated earlier. The forcing function in eqs. (24) is taken to be a cosine function thus, rewrite eqs. (24),

$$m\ \ddot{u} + k\ u = f_0 \cos \omega_f t \qquad (50a)$$

$$u(0) = u_0 \tag{50b}$$

$$\dot{u}(0) = u_1 \tag{50c}$$

where $u_0$, $u_1$, $f_0$ and $\omega_f$ are given constants. In the finite element formulation, we shall replace eqs. (50) with the following set

$$m\,\ddot{u} + k\,u = f_0 \cos \omega_f t \tag{51a}$$

$$\dot{u}(t) - \alpha\,[u(0) - u_0] = 0 \tag{51b}$$

$$\dot{u}(0) - u_1 = 0$$

thus, eqs. (50) becomes a special case of (51) as $\alpha \to \infty$. It is convenient to nondimensionalize the independent variable $t$ and let

$$\tau = t/T \tag{52}$$

In terms of $\tau$, eqs. (51) become

$$\ddot{u} + T^2\,\omega^2\,u = f_1 \cos\,(T\,\omega_f\,\tau) \tag{53a}$$

$$\dot{u}(1) - T\,\alpha\,[u(0) - u_0] = 0 \tag{53b}$$

$$\dot{u}(0) - T\,u_1 = 0 \tag{53c}$$

where

$$f_1 = T^2\,f_0/m \quad \omega^2 = k/m \tag{54}$$

The exact solution for eqs. (53) can be easily written as

$$u(\tau) = A \cos\,(T\,\omega \cdot \tau) + B \sin\,(T\,\omega \cdot \tau)$$

$$+ \eta \cos\,(T\,\omega_f \cdot \tau) \tag{55}$$

with

$$\eta = \frac{f_0}{m(\omega^2 - \omega_f^2)}\,, \quad \beta = \frac{u_1}{\omega}$$

$$A = \frac{\alpha\,u_0 + T\,u_1 \cos\,(T\omega) - \eta\,[\alpha + T\omega_f \sin\,(T\omega_f)]}{\alpha + T\,\omega \sin\,(T\,\omega)} \tag{56}$$

To solve eqs. (53) using finite elements, one begins with the variational statement:

$$\delta\,J = 0 \tag{57a}$$

$$J = \int_0^1 [-\,\dot{u}\,\dot{v} + T^2\,\omega^2 uv - f(\tau)\,v]\,d\tau$$

$$+ T\,\alpha\,[u(0) - u_0]\,v(1) - T\,u_1\,v(0) \tag{57b}$$

Now that

$$(\delta J)_u = 0 \tag{58a}$$

$$= \int_0^1 [-\dot{u}\,\delta\dot{v} + T^2\,\omega^2\,u\delta v - f(\tau)\delta v]$$

$$+ T\,\alpha\,[u(0) - u_0]\,v(1) - Tu_1\delta v(0) \tag{58b}$$

$$= \int_0^1 [\ddot{u} + T^2\,\omega^2\,u - f(\tau)]\,\delta v\,dt$$

$$- \{u(0) - \alpha\,T[u(0) - u_0]\}\,\delta v(1)$$

$$+ \{u(0) - T\,u_1\}\,\delta v(0) \tag{58c}$$

From eq. (58b), one has

$$\int_0^1 [-\dot{u}\,\delta\dot{v} + T^2\omega^2\,u\,\delta v]\,dt + \alpha T\,u(0)\,\delta v(1)$$

$$= \int_0^1 f(\tau)\,\delta v\,d\tau + \alpha\,T\,u_0\,\delta v(1) + T\,u_1\,\delta v(0) \tag{59}$$

with

$$u(\tau) = \underset{\sim}{a}^T(\tau)\,\underset{\sim}{U}$$
$$v(\tau) = \underset{\sim}{a}^T(\tau)\,\underset{\sim}{V} \tag{60}$$

eq. (59) leads to

$$\delta\,\underset{\sim}{V}^T\,\underset{\sim}{K}\,\underset{\sim}{U} = \delta\,\underset{\sim}{V}^T\,\underset{\sim}{F}$$

or

$$\underset{\sim}{K}\,\underset{\sim}{U} = \underset{\sim}{F} \tag{61}$$

where

$$K = \int (-\dot{\underset{\sim}{a}}\,\dot{\underset{\sim}{a}}^T + T^2\,\omega^2\,\underset{\sim}{a}\,\underset{\sim}{a}^T)\,dt$$

$$+ \alpha\,T\,\underset{\sim}{a}(1)\,\underset{\sim}{a}(0) \tag{62}$$

$$\underset{\sim}{F} = \int_0^1 f(\tau)\,\underset{\sim}{a}\,d\tau + \alpha T u_0\,\underset{\sim}{a}(1) + T\,u_1\,\underset{\sim}{a}(0) \tag{63}$$

The results obtained from this finite element formulation are compared with the exact solutions as shown in Tables 1 - 3. The values of the parameters chosen for these data are $k = 1.0$, $m = 1.0$, $f_0 = 1.0$, $\omega_f = 0.5$, $u_0 = 1.0$, $\dot{u}_0 = 1.0$ the number of elements used is ten. The calculated $u$ and $\dot{u}$ for $T = 2.0$, 10.0 and 20.0 are given in Table 1, 2, 3 and 4 respectively. The forcing function $\cos \omega_f t$ and the solution $u(t)$ are also plotted in the range $0 \lesssim t \lesssim 20$ as shown in Figure 1.

(2) Solutions to a Transient Heat Conduction Problem. As another numerical example, we shall take the nondimensional heat transfer problem defined by the following set:

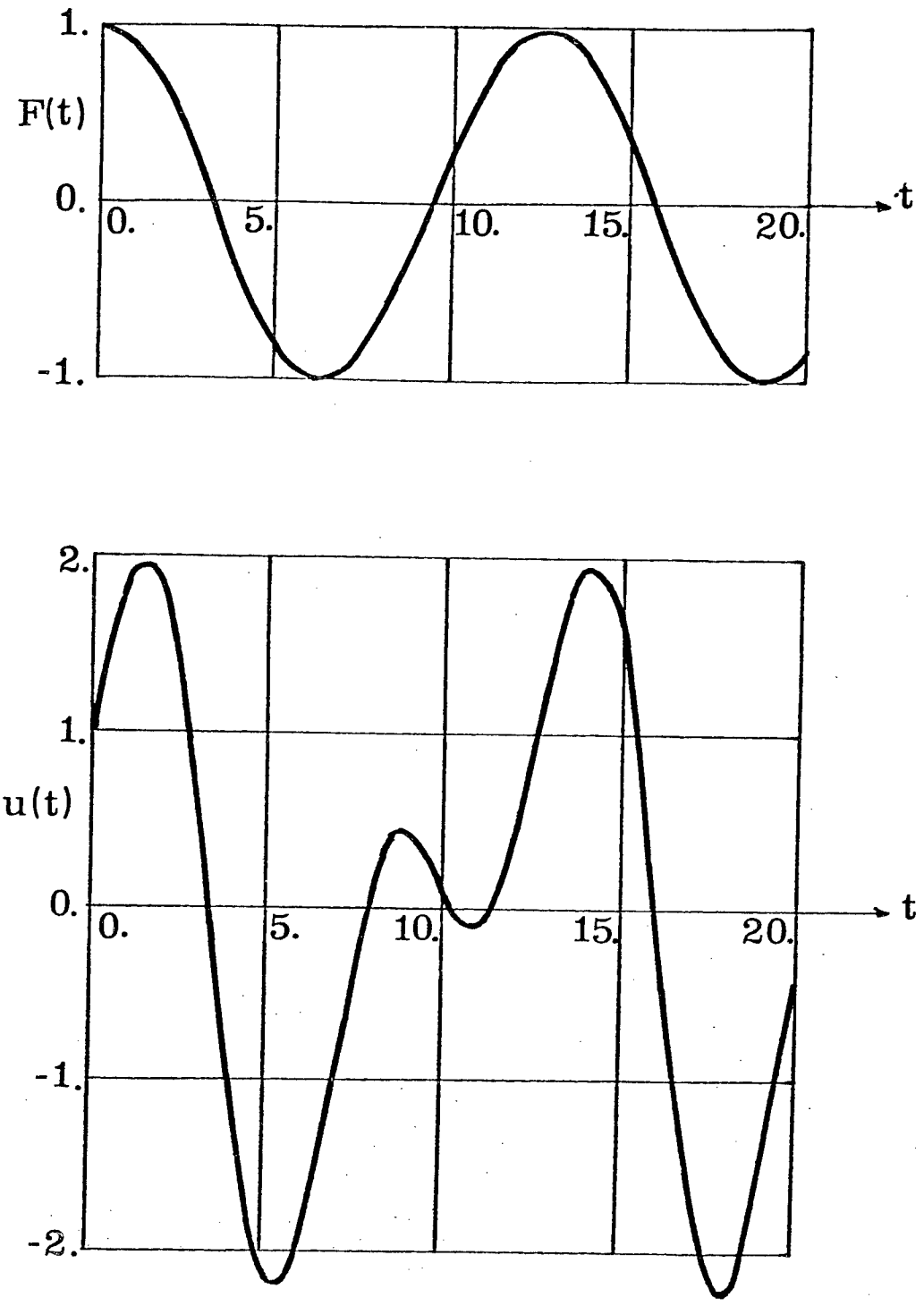



Figure 1. Forcing Function F(t) and Solution u(t) for the Vibration Problem

TABLE 1

Solutions to the Forced Vibration Problem Using FE-UVF
Compared with Exact Solutions (in Parentheses)

$0 \leq t \leq 2.0$

| t | u(t) | Exact | u(t) | Exact |
|---|---|---|---|---|
| 0 | 1.000 000 0 | (1.000 000 0) | 1.000 00 | (1.000 00) |
| 0.2 | 1.198 652 6 | (1.198 652 7) | 0.979 74 | (0.979 73) |
| 0.4 | 1.389 153 7 | (1.389 153 4) | 0.918 43 | (0.918 42) |
| 0.6 | 1.563 313 2 | (1.563 312 6) | 0.816 55 | (0.816 54) |
| 0.8 | 1.713 202 9 | (1.713 201 8) | 0.676 22 | (0.676 21) |
| 1.0 | 1.831 481 7 | (1.831 480 3) | 0.501 18 | (0.501 18) |
| 1.2 | 1.911 702 4 | (1.911 700 6) | 0.296 62 | (0.296 61) |
| 1.4 | 1.948 585 6 | (1.948 583 6) | 0.068 98 | (0.068 97) |
| 1.6 | 1.938 251 2 | (1.938 249 1) | - 0.174 24 | (-0.174 25) |
| 1.8 | 1.878 396 9 | (1.878 395 0) | - 0.424 82 | (-0.424 80) |
| 2.0 | 1.768 416 1 | (1.768 416 1) | - 0.674 13 | (-0.674 03) |

TABLE 2

Solutions to the Forced Vibration Problem Using FE-UVF
Compared with Exact Solutions (in Parentheses)

$0 \leq t \leq 10.0$

| t | u(t) | | $\dot{u}(t)$ | |
|---|---|---|---|---|
| 0 | 1.000 | (1.000) | 1.004 | (1.000) |
| 1.0 | 1.832 | (1.831) | 0.505 | (0.501) |
| 2.0 | 1.770 | (1.768) | - 0.675 | (-0.674) |
| 3.0 | 0.566 | (0.565 | - 1.614 | (-1.608) |
| 4.0 | - 1.094 | (-1.094) | - 1.518 | (-1.512) |
| 5.0 | - 2.123 | (-2.122) | - 0.435 | (-0.435) |
| 6.0 | - 1.920 | (-1.919) | 0.778 | (0.773) |
| 7.0 | - 0.843 | (-0.843) | 1.213 | (1.207) |
| 8.0 | 0.167 | (0.166) | 0.690 | (0.689) |
| 9.0 | 0.436 | (0.435) | - 0.126 | (-0.122) |
| 10.0 | 0.114 | (0.114) | - 0.385 | (-0.381) |

TABLE 3

Solution to the Forced Vibration Problem Using FE-UVF
Compared with Exact Solutions (in Parentheses)

$0 \lesssim t \lesssim 20.0$

| t | u(t) | | $\dot{u}(t)$ | |
|---|---|---|---|---|
| 0 | 1.000 | (1.000) | 1.05 | (1.00) |
| 2.0 | 1.778 | (1.768) | - 0.68 | (-0.67) |
| 4.0 | - 1.097 | (-1.094) | - 1.57 | (-1.51) |
| 6.0 | - 1.928 | (-1.919) | 0.82 | (0.77) |
| 8.0 | 0.173 | (0.166) | 0.71 | (0.69) |
| 10.0 | 0.116 | (0.114) | - 0.44 | (-0.38) |
| 12.0 | 0.453 | (0.462) | 0.88 | (0.85) |
| 14.0 | 1.956 | (1.950) | 0.06 | (0.03) |
| 16.0 | - 0.156 | (-0.162) | - 1.76 | (-1.71) |
| 18.0 | - 2.199 | (-2.186) | 0.15 | (0.14) |
| 20.0 | - 0.348 | (-0.342) | 1.10 | (1.08) |

D.E.: $$\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} = 0 \quad , \quad 0<x<1 \; ; \quad 0<t<T \tag{64a}$$

B.C.: $$u(0),t) = 1, \; \frac{\partial u}{\partial x}(1,t) = 0 \tag{64b,c}$$

I.C.: $$u(x,0) = 0 \tag{64d}$$

where T is any given finite real positive number. To facilitate computation, it is desirable to change the independent variable t into $\tau$ such that

$$\tau = t/T \tag{65}$$

thus, the system of eqs. (64) becomes

D.E.: $$\frac{\partial^2 u}{\partial x^2} - \frac{1}{T}\frac{\partial u}{\partial \tau} = 0 \; , \; 0<x<1 \; ; \; 0<\tau<1 \tag{66a}$$

B.C.: $$u(0,\tau) = 1 \; ; \; \frac{\partial u}{\partial x}(1,\tau) = 0 \tag{66b,c}$$

I.C.: $$u(x,0) = 0 \tag{66d}$$

According to our unconstrained variational formulation, this system is again replaced by the following:

D.E.: $$\frac{\partial^2 u}{\partial x^2} - \frac{1}{T}\frac{\partial u}{\partial \tau} = 0 \; , \; 0<x<1 \; ; \; 0<\tau<1 \tag{67a}$$

B.C.: $$\frac{\partial u}{\partial x}(0,\tau) + \alpha\,[u(0,\tau) - 1] = 0 \tag{67b}$$

$$\frac{\partial u}{\partial x}(1,\tau) = 0 \tag{67c}$$

I.C.: $$u(x,0) = 0 \tag{67d}$$

Clearly, eqs. (67) reduces to (66) as $\alpha \to \infty$. The variational statement can be written as

$$\delta J = 0 \tag{68a}$$

where

$$J = -\int_0^1\int_0^1 \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{1}{T}\frac{\partial u}{\partial \tau} v\right) dx dt + \alpha\int_0^1 [u(0,\tau) - 1]\, v(0,\tau)\, d\tau + \int_0^1 u(x,0)\, v(x,0)\, dx \tag{68b}$$

Due to the fact that $v(x,\tau)$ is unconstrained, it is a simple matter to show that

$$(\delta J)_u = 0 \tag{69}$$

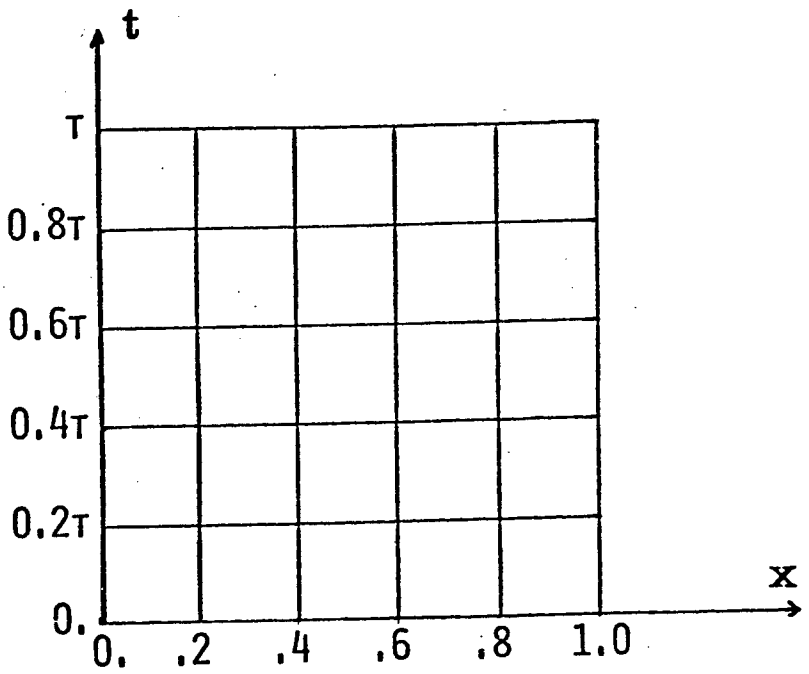



Figure 2. Finite Element Grid Scheme Used for a Transient Heat Conduction Problem

TABLE 4

Transient Heat Transfer Solutions u(x,t) Using FE-UVF
Compared with Exact Series Solutions (in Parentheses)

$0 < t < T = 1.00$

| t \ x | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|
| 0.2 | 1.000 (1.000) | 0.754 (0.757) | 0.583 (0.496) | 0.370 (0.405) | 0.264 (0.284) | 0.228 (0.179) |
| 0.4 | 1.000 (1.000) | 0.855 (0.853) | 0.713 (0.721) | 0.622 (0.616) | 0.552 (0.549) | 0.516 (0.526 |
| 0.6 | 1.000 (1.000) | 0.910 (0.910) | 0.828 (0.830) | 0.767 (0.767) | 0.725 (0.724) | 0.708 (0.710) |
| 0.8 | 1.000 (1.000) | 0.945 (0.945) | 0.896 (0.896) | 0.857 (0.857) | 0.832 (0.832) | 0.823 (0.823) |
| 1.0 | 1.000 (1.000) | 0.967 (0.967) | 0.937 (0.937) | 0.913 (0.913) | 0.897 (0.897) | 0.892 (0.892) |

TABLE 5

Transient Heat Transfer Solutions u(x,t) Using FE-UVT
Compared with Exact Series Solutions (in Parentheses)

$0 < t < T = 0.05$

| t \ x | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
|---|---|---|---|---|---|---|
| 0.01 | 1.000<br>(1.000) | 0.144<br>(0.157) | 0.014<br>(0.005) | 0.002<br>(0.000) | 0.000<br>(0.000) | 0.000<br>(0.000) |
| 0.02 | 1.000<br>(1.000) | 0.315<br>(0.317) | 0.047<br>(0.046) | (0.003)<br>(0.003) | (0.000)<br>(0.000) | (0.000<br>(0.000) |
| 0.03 | 1.000<br>(1.000) | 0.413<br>(0.414) | 0.103<br>(0.102) | 0.015<br>(0.014) | 0.001<br>(0.001) | 0.000<br>(0.000) |
| 0.04 | 1.000<br>(1.000) | 0.479<br>(0.480) | 0.157<br>(0.157) | 0.034<br>(0.034) | 0.005<br>(0.005) | 0.001<br>(0.001) |
| 0.05 | 1.000<br>(1.000) | 0.527<br>(0.527) | 0.206<br>(0.206) | 0.058<br>(0.058) | 0.012<br>(0.012) | 0.003<br>(0.003) |

is a necessary and sufficient condition of eqs. (67). Now the finite element matrix equations can be obtained from eq. (69).

$$(\delta J)_u = -\int_0^1\int_0^1 \left[\frac{\partial u}{\partial x}\,\delta\left(\frac{\partial v}{\partial x}\right) + \frac{1}{T}\frac{\partial u}{\partial \tau}\,\delta v\right] dx\,dt$$

$$+ \int_0^1 [u(0,\tau) - 1]\,\delta v(0,\tau)\,dt$$

$$+ \int_0^1 u(x,0)\,\delta v(x,0)\,dx = 0 \qquad (70)$$

or,

$$-\int_0^1\int_0^1 \left[\frac{\partial u}{\partial x}\,\delta\left(\frac{\partial v}{\partial x}\right) + \frac{1}{T}\frac{\partial u}{\partial \tau}\,\delta v\right] dx\,dt$$

$$+ \alpha\int_0^1 u(0,\tau)\,\delta v(0,\tau)\,dt + \int_0^1 u(x,0)\,\delta v(x,0)\,dx$$

$$= \alpha\int_0^1 \delta v(0,\tau)\,d\tau \qquad (71)$$

Using the usual procedure of discretization and the assumption of displacement functions, the final finite element matrix equation evidently can be derived from eq. (71). We shall omit the details here. The computational results are presented in Tables 4 and 5. The finite element grid scheme used is shown in Figure 2. As clearly shown in those tables, excellent agreement exists between the FE-UVF approach and the series solution. It is noted that the approximate solutions are less accurate invariably as they approach the initial time $t = 0$. This is probably due to the discontinuity of the initial boundary data at $x = 0$, $t = 0$.

## REFERENCES


1. Levinson, M., "Application of the Galerkin and Ritz Methods to Non-conservative Problems of Elastic Stability," Zeitschrift fur Angewandte Mathematic und Physik, Vol. 17, pp. 431-442 (1966).

2. Barsoum, R. S., "Finite Element Method Applied to the Problem of Stability of a Nonconservative System," International Journal for Numerical Methods in Engineering, Vol. 3, pp. 63-87 (1971).

3. Mote, C. D., "Nonconservative Stability by Finite Element," Journal of the Engineering Mechanics Division, Proceedings of the American Society of Civil Engineers, Vol. EM3, pp. 645-656 (June 1971).

4. Wu, J. J., "Column Instability under Nonconservative Forces, with Internal and External Damping - Finite Element Using Adjoint Variational Principles," Development in Mechanics, Vol. 7, pp. 501-514 (1973).

5. Wu, J. J., "A Unified Finite Element Approach to Column Stability Problems," Development in Mechanics, Vol. 8, pp. 279-294 (1975).

6. Wu, J. J., "On the Numerical Convergence of Matrix Eigenvalue Problems Due to Constraint Conditions," Journal of Sound and Vibrations, Vol. 37, pp. 349-358 (1974).

7. Fried, I., "Finite Element Analysis of Time Dependent Phenomena," Journal of the American Institute of Aeronautics and Astronautics, Vol. 7, No. 6, pp. 1170-1172 (June 1970).

8. Gurtin, M. E., "Variational Principles for Linear Elastodynamics," Archive for Rational Mechanics and Analysis, Vol. 16, No. 1, pp. 36-50 (1964).

9. Gurtin, M. E., "Variational Principles for Linear Initial-Value Problems," Quarterly of Applied Mathematics, Vol. 22, pp. 252-256 (1964).

10. Wilson, E. L. and Nickel, R. E., "Application of the Finite Element Method to Heat Conduction Analysis," Nuclear Engineering and Design, Vol. 4, pp. 276-286 (1966).

11. Dunham, R. S., Nickel, R. E. and Strickler, D. C., "Integration Operators for Transient Structural Response," Computers and Structures, Vol. 2, pp. 1-15 (1972).

12. Ghaboussi, J. and Wilson, E. L., " Variational Formulation of Dynamics of Fluid-Saturated Porous Elastic Solids,: Journal of the Engineering Mechanics Division, Proceedings of the American Society of Civil Engineers, Vol. EM4, pp. 947-963 (1972).

13. Atluri, S., "An Assumed Stress Hybrid Finite Element Model for Linear Elastodynamic Analysis," Journal of the American Institute of Aeronautics and Astronautics, Vol. 11, No. 7, pp. 1028-1031 (1973).

14. Wu, J. J., "On the Stability of a Free-Free Beam under Axial Thrust Subjected to Directional Control," Journal of Sound and Vibration, Vol. 43, pp. 45-52 (1975). Also see a correction note on this paper, ibid, Vol. 44, p. 309 (1976).

# THE NUMERICAL SOLUTION OF FREE BOUNDARY PROBLEMS BY MATHEMATICAL PROGRAMMING

Richard S. Sacher
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, New York 12181

## 1. INTRODUCTION

This paper is concerned with the numerical solution of free boundary problems by mathematical programming. In such problems, one seeks the solution of a partial differential equation (usually Laplace's or Reynolds' equation) satisfying prescribed conditions on the boundary of a region when a portion of the boundary is unknown and must be determined as part of the problem. The unknown boundary is called the free boundary.

Many of these boundary value problems have not yielded to analytical methods of solution. Recently, however, a novel transformational approach has met with more success. Specifically, the free boundary problem is reformulated as a variational inequality which, in turn, is equivalent to a certain constrained minimization problem in a Sobolev (function) space. Although this latter problem is still computationally intractable, finite difference or finite element approximations yield a difficult, but solvable, sparse, specially-structured quadratic programming problem of potentially very large size. It is the solution of this last problem with which we are concerned and for which an algorithm will be stated.

## 2. APPLICATIONS

Free boundary problems arise in a variety of situations. Rohde and McAllister [8] have developed the variational inequalities for the finite-length journal bearing problem, in which one is concerned with a cylindrical rod (the journal) rotating within a tube (the bearing). The inner surface of the bearing is coated with a thin film of lubricant and we wish to know the pressure distribution on the film. At a certain point, the pressure becomes so low that the lubricant vaporizes, thus creating the free boundary interface.

In the area of fluid dynamics, Baiocchi et al. [1] have reformulated certain problems dealing with stationary fluid flow through porous media as variational inequalities. These include porous dams in which the free boundary is the interface between the wet and dry part of the dam. Brézis and Stampacchia [2], [3] have studied the determination of steady subsonic flows for nonviscous compressible and incompressible fluids past a two-dimensional convex body by using a hodograph transformation to obtain an equivalent free boundary problem for which a variational inequality problem can be stated.

## 3. THE QUADRATIC PROGRAMMING PROBLEM

The common denominator of these and several other free boundary problems is that their associated quadratic programming problem

$$\text{Minimize } f(x) = \frac{1}{2}\langle x, Mx\rangle + \langle q, x\rangle$$

$$\text{subject to } x \geq 0$$

has certain special attributes which can be exploited in the development of efficient algorithms. The matrix M is a block-tridiagonal Stieltjes matrix (ie., symmetric, diagonally dominant with nonpositive off-diagonal entries). Furthermore, the diagonal blocks are themselves tridiagonal matrices and the off-diagonal blocks are diagonal matrices.

One computationally successful approach to this problem is a modification of the block- (or line-) successive overrelaxation method. This algorithm requires that we partition the vector $x = (x_1, x_2, \ldots, x_m)$ where $x_i \in R^{n_i}$ and conformably partition M and q. For this special class of problems, we may state the algorithm as follows:

Algorithm

Step 0. Let $x^0 = (x_1^0, x_2^0, \ldots, x_m^0)$ be any nonnegative vector, eg., $x^0 = 0$. Let $\omega \in (0,2)$ be given. Set k=0 and i=1.

Step 1. Determine $\bar{x}_i^{k+1} \geq 0$ which minimizes (over the nonnegative orthant $R_+^{n_i}$)

$$f(x_1^{k+1}, x_2^{k+1}, \ldots, x_{i-1}^{k+1}, v, x_{i+1}^{k}, \ldots, x_m^{k})$$

$$= \frac{1}{2} \langle v, M_{ii} v \rangle + \langle (q_i + M_{i,i-1} x_{i-1}^{k+1} + M_{i,i+1} x_{i+1}^{k}), v \rangle + c_i$$

where $c_i$ may be taken to be zero.

Step 2. Define

$$\omega_i^{k+1} = \max \{\bar{\omega} : \bar{\omega} \leq \omega, \; x_i^k + \bar{\omega}(\bar{x}_i^{k+1} - x_i^k) \geq 0$$

$$x_i^{k+1} = x_i^k + \omega_i^{k+1}(\bar{x}_i^{k+1} - x_i^k) .$$

Step 3. If i=m, go to Step 4. Otherwise, return to Step 1 with i replaced by i+1.

Step 4. Define

$$S = \{(i,j) : (x_i^{k+1})_j > 0\} \cup \{(i,j) : (x_i^{k+1})_j = 0, \; (q_i + \sum_{\ell=1}^{m} M_{i\ell} x_\ell^{k+1})_j < 0\}.$$

If $\max_{(i,j) \in S} |(q_i + \sum_{\ell=1}^{m} M_{i\ell} x_\ell^{k+1})_j| \leq \varepsilon$, stop. An approximate solution is at hand. If not, return to Step 1 with k replaced by k+1 and i=1.

Step 1 requires that we solve a smaller quadratic programming problem whose quadratic form contains a tridiagonal Stieltjes matrix. For a discussion of some fast methods to do this, we refer the reader to [5]. For more details on the development of and computational experience with the algorithm given above, see [4]. From a consideration of storage requirements and speed, one may conclude that this algorithm is competitive, if not superior, to other methods described in the literature.

## 4. REFERENCES


[1] C. BAIOCCHI, V. COMINCIOLI, L. GUERRI and G. VOLPI, "Free boundary problems in the theory of fluid flow through porous media : A numerical approach", Calcolo 10 (1973), 1-86.

[2] H. BRÉZIS, "A new method in the study of subsonic flows", Lecture Notes in Mathematics No. 446 (J. Goldstein, ed.) Springer-Verlag, New York, 1975.

[3] H. BRÉZIS and G. STAMPACCHIA, "The hodograph method in fluid-dynamics in the light of variational inequalities", to appear in J. Arch. Rat. Mech. Anal., 1976.

[4] R. W. COTTLE, G.H. GOLUB and R.S. SACHER, "On the solution of large, structured linear complementarity problems : The block-tridiagonal case", To appear.

[5] R.W. COTTLE and R.S. SACHER, "On the solution of large, structured linear complementarity problems : The tridiagonal case", To appear.

[6] C.W. CRYER, "The method of Christopherson for solving free boundary problems for infinite journal bearings by means of finite differences", Math. of Comp. 25 (1971), 435-443.

[7] C.W. CRYER, "The solution of a quadratic programming problem using systematic overrelaxation", J. SIAM Control 9 (1971), 385-392.

[8] S.M. ROHDE and G.T. McALLISTER, "A variational formulation for a class of free boundary problems arising in hydrodynamic lubrication", Int. J. Eng. Sci 13 (1975), 841-850.

# A NUMERICAL INTEGRATION ERROR ANALYSIS UTILIZING A WRONSKIAN TECHNIQUE

Larry A. Whatley

Quality and Reliability Division
USAMERADCOM
ATTN:DRXFD-TQ (Whatley)
Ft. Belvoir, Virginia 22060
Formerly, Intern Training Center, DARCOM

S. Bart Childs, Ph.D.

Department of Industrial Engineering
Texas A&M University
Texarkana, Texas

ABSTRACT. An error analysis is performed upon the superposition and numerical integration procedures of a method of solution of multipoint boundary value problems utilizing power series expansions. The procedure involves the evaluation of the relative error of the Wronskian, which provides a scalar function characterization of the error of integrators of a matrix of solutions. The error behavior is investigated by using different integration step sizes and orders (terms of the power series).

Evaluations are performed with numerical solutions of specified accuracy or order. Example applications are included.

1. INTRODUCTION. Numerical integration is commonly used by engineers and scientists as a tool for solving ordinary differential equations. Those equations which cannot be solved exactly or in closed form can often be solved using numerical integration techniques. There are drawbacks to each particular integration scheme. The most important considerations are: the origin of the problem, guidelines from the theory of the algorithm, the computer being used, and the class or problems to be considered, Shampine and Allen (1973).

Many techniques, in the form of "canned" routines or pre-programmed methods, and their variations, are available to the user. It is now possible to obtain numerical solutions using techniques which require lengthy operations. The more popular integration techniques (i.e. Adams methods, Runge-Kutta, etc.) provide reasonable results for a wide range of applications. They are subject to some disadvantages, the most common being their susceptability to round-off error, Ralston and Wilf (1960).

An alternate method of numerical integration has been investigated which is based upon the expansion of power series. The method is relatively free of the disadvantages of the more popular techniques and significantly more efficient for certain classes of problems, Doiron (1967). Research done previously by Fehlberg (1964) has shown the power series technique to be five to six times faster than a Runge-Kutta method for the same specified accuracy in certain selected problems. The power series methods generally require more user effort.

The purpose of this study is to investigate the error in the integration via power series expansions. It has been shown that the Wronskian can be used as a meaningful check on the solvability and superposition procedures in the solution of boundary value problems. It has been proposed that the relative error of the Wronskian can provide some insight into the errors arising from this particular integration scheme, Childs et al. (1971).

2. DEVELOPMENT. The problems to be considered are presented as an ordinary differential equation written in the general linear form as

$$\dot{y} = Ly + f \tag{2-1}$$

where $L$ is a linear operator in the form of an $n$ x $n$ coefficient matrix (expressed as a constant or function of an independent variable). The letter $y$ represents the state variable vector and $\dot{y}$ denotes the derivative of $y$ with respect to the independent variable (in this case $t$). The vector $f$ is a vector of forcing functions. The above equation is subject to a set of specified boundary conditions.

$$q_i(y(t_i)) = b_i \qquad 0 \leq t_i \leq T \qquad\qquad i = 1,2,\ldots,m \tag{2-2}$$

where $m \geq n$. The operator $q_i$ is a linear combination of the elements of the vectors at $t = t_i$, that is equal to the boundary value $b_i$.

To meet the above boundary condition it is necessary to superimpose independent solutions of equation (2-1). The technique used is to superimpose the appropriate number of solutions of the homogeneous equation

$$\dot{H} = LH \tag{2-3}$$

upon a particular solution

$$\dot{p} = Lp + f \quad . \tag{2-4}$$

This can be written as

$$y = p + H\beta = p + \sum_{k=1}^{r} h^{(k)} \beta_k \qquad\qquad r \leq n \tag{2-5}$$

where $H$ is a matrix whose columns are homogeneous solutions. The superscript in parentheses indicates that vector is the $(\cdot)^{th}$ column of a matrix denoted by the capital letter and $\beta$ denotes the superposition constants. The letter $r$ denotes the number of homogeneous equations which is equal to the number of unknown elements of $y(o)$.

It is known from elementary differential equation techniques that the sum of a particular solution of a linear differential equation and a solution of its homogeneous differential equation is merely another particular solution of that differential equation.

Utilizing this fact, it can be established that $y$ can be expressed as a combination of particular solutions (Childs, 1971)

$$y = Pa = \sum_{k=0}^{r} p^{(k)} a_k \tag{2-6}$$

where $P$ is a matrix whose columns are solutions of equation (2-1), thus:

$$\dot{p}^{(k)} = Lp^{(k)} + f \quad . \tag{2-7}$$

We multiply each side of equation (2-7) by $a_k$ and sum these products

$$\sum_{k=0}^{r} \dot{p}^{(k)} a_k = L \sum_{k=0}^{r} p^{(k)} a_k + f \sum_{k=0}^{r} a_k \quad . \tag{2-8}$$

By comparing equation (2-1) with equation (2-8), it is obvious that the left hand side of equation (2-8) is the quantity $\dot{y}$ and the first term of the right hand side is the state vector $y$. Therefore, it is elementary that the superposition constants must obey

$$\sum_{k=0}^{r} a_k = 1 \quad . \tag{2-9}$$

After determining the superposition constants, subject to the above restriction, the solution becomes trivial and is generated utilizing the initial conditions

$$y(o) = P(o) \quad a \quad . \tag{2-10}$$

The reason for superposition of solutions is to satisfy the boundary conditions. It is necessary that the superimposed solutions be independent to be able to meet boundary conditions.

The requirement of independence is satisfied using a determinant of homogeneous solutions, which is usually known as the Wronskian. The independence of homogeneous solutions is satisfied when the matrix whose columns are these vectors of *rank r*.

$$rank\ (H) = r \tag{2-11}$$

which must contain at least one $r$ x $r$ submatrix of $H$ and has a *non-zero* determinant for the range of values of the independent variable.

This can be applied to superposition of particular solutions. Define $P$ as an $(n+1)$ x $(r+1)$ matrix in which the first row elements are *one (unity)* and remaining submatrix is $P$ (shown in equation 2-6).

$$\grave{P} = \left[\begin{array}{c} \underline{1}\ \underline{1} \cdots \underline{1} \\ P \end{array}\right] \tag{2-12}$$

The Wronskian of *rank n* has been shown to obey the following equation, Petrovski (1966).

$$det\ (H(t)) = det\ (H(o))\ exp\ (\int_o^t tr\ (L(\Phi))\ d\Phi) \tag{2-13}$$

where $tr\ (L(\Phi))$ is the *trace* (the summation of the principle diagonal of the matrix) of the coefficient matrix in equation (2-1). When the Wronskian is *non-zero* at the initial value of the independent variable, then it is theoretically *non-zero* for all values of the independent variable over any finite interval. The following theorem adapts (2-13) to particular solutions.

Theorem:

$$det\ (\grave{P}(t)) = det\ (\grave{P}(o))\ exp\ (\int_o^t tr\ (L(\Phi))\ d\Phi) \tag{2-14}$$

Proof:

The columns of $\grave{P}$ are $p^{(o)}$ and

$$p^{(k)} = p^{(o)} + h^{(k)} \quad . \tag{2-15}$$

The subtraction of one column of a matrix from all other columns does not affect the value of the determinant of that matrix. Therefore, subtracting the $0^{th}$ column of $\grave{P}$ from all other columns and comparing with equation (2-13) completes the proof:

$$det\ (\grave{P}) = det \left[\begin{array}{c|ccc} \underline{1} & \underline{1} & \cdots & \underline{1} \\ o & & & \\ \vdots & & H & \\ o & & & \end{array}\right] = det\ (H) \quad . \tag{2-16}$$

The Wronskian shows that solutions are (not) linearly independent and that a fundamental set of solutions (doesn't) exist.

The relative error of the Wronskian is defined as follows:

$$R(t) = \left| \frac{|W(t)| - |W_n(t)|}{|W(t)|} \right| \tag{2-17}$$

where $W_n\ (t)$ is evaluated utilizing particular solutions which come from numerical integration procedures and $W\ (t)$ is evaluated from (2-14).

3. AN EXAMPLE. The power series integration method was programmed in FORTRAN utilizing an Amdahl 470 digital computer. The program, subroutines and function routines used in the study were provided from unpublished studies, Childs (1975).

The results are for damped, forced harmonic oscillators described by the following equations:

$$\dot{y}_1 = y_2$$

$$\dot{y}_2 = -\lambda y_1 - \rho y_2 + \sin(t) \quad .$$

A set of independent particular solutions are created using arbitrarily chosen initial conditions:

$$y_1 = 1.$$

$$y_2 = 0.$$

Power series evaluations are then generated using these initial conditions and the recursive relationships. The set of particular solutions are then solved over the range of the independent variable, $t$.

By calculating the Wronskian using both numerical procedures and the analytical method, the relative error may be examined.

The results of the relative error of the power series integration procedure are compared to results obtained from previous studies by Childs et al. (1971) concerning the same problem using two different numerical integration procedures with $\lambda = 1.0$ and $\rho = 0.2$. The two integration procedures used to compare with the power series method are modified Euler and Runge-Kutta methods. They are order $h^2$ and $h^4$ respectively, where $h$ is the integration step size. The two plots in Figure 4-1 are log-log plots of $R(t)$ versus $h$ for the Euler and Runge-Kutta procedures. For these results it has been observed that the following relationship is true:

$$\gamma(R(t)) = R(\gamma t)$$

where $\gamma$ is a positive scalar quantity. From these results it has also been suggested that the relative error is dominated by the following proportionality for "reasonable" integration step sizes

$$|R(t)| \propto h^j \, t$$

where $j$ is the order of the integration formula used.

By comparing both cases (Figure 4-1.a and 4-1.b) it has been determined that they have slopes of 2 and 4 respectively. It has also been observed that for "large" step sizes the points tend away from the straight line due to approximation error and also for using "small" step sizes due to round off error.

Results for the power series integration procedure are presented in Figure 4-2 in the form of a log-log plot of $R(t)$ versus $h$ for different orders (terms in the power series). It is seen that a family of curves exist for different orders. It was observed that for constant step sizes the error decreases as the number of terms increase. As the step size increases, the number of terms must also increase in order to retain a specified accuracy. Like the Euler and Runge-Kutta procedure, the relative error tends toward linearity as it increases with step size. As the step size decreases for each "order curve" the error function tends toward the error specification. This observation can be explained by the evaluation subroutine used on accuracy specification of $1 \times 10^{-7}$. Thus, more accuracy was not attempted.

Since the curves are all similar, results will be explained for only one curve. For the "order curve" evaluated at 4 terms, the step size begins at .001. It is observed to be within accuracy specifications due to the fact that the power series integration is performed with such small step sizes. Since such small steps are used, all the terms (in this case, 4) are not required to meet the accuracy. As the step size increases, more terms are required to meet the accuracy specification. At the step size, .005, it is observed that the curve "dips". This occurs because at this step size more terms (in this case, 1) are required to meet the accuracy. From this point on the routine is utilizing all the terms of the power series in order to meet the accuracy requirement. However, as the step size increases, it is seen that the accuracy is not being met due to the larger steps being taken. It is also seen that the error function is linear while all terms of the power series are being used and would continue to be linear (within machine limitations) if it were not for round-off.

Results were also calculated for several values of $(\lambda,\rho)$. All tendencies held as shown in Figure 4-2.

4. CONCLUSIONS. The relative error of the Wronskian can apparently be used to determine if the step size used by an integration procedure is appropriate. The error would grow approximately linearly in a log-log plot. Further investigations should involve different systems of equations.

BIBLIOGRAPHY

Boyce, W.E. and R.C. DiPrima. Elementary Differential Equations and Boundary Value Problems. New York: John Wiley and Sons, Inc., 1969.

Childs, S.B. Unpublished Class Notes. "Computer Methods in Applied Sciences", Texas A&M University, Summer, 1975.

Childs, B., Luckinbill, D., Bryan, J., and Boyd, J.H., Jr., "Numerical Solutions of Multipoint Boundary Value Problems in Linear Systems", Int. J. Systems Science, 2, 49, 1971.

Childs, S.B. and H.P. Porter. Numerical Solutions of Nonlinear Multipoint Boundary Value Problems, in preparation for Applied Math and Computation Series for Addison Wesley, 1977.

Doiron, H.H., "Numerical Integration Via Power Series Expansions", Masters Thesis, The University of Houston, August, 1967.

Doiron, H.H., "An Indirect Optimization Method with Improved Convergence Characteristics", Ph.D. Dissertation, The University of Houston, August, 1968. Also available as NASA TM X-58088, May 1970.

Fehlberg. E. "Numerical Integration of Differential Equations by Power Series Expansions, Illustrated by Physical Examples". National Aeronautics and Space Administration Technical Note No. TND-2356, October, 1964.

Hull, T.E., Enright, W.H., Fellen, B.M., and Sedgwick, A.E. "Comparing Numerical Methods for Ordinary Differential Equations". SIAM J. Numerical Anal. Vol. 9, No. 4, December, 1972.

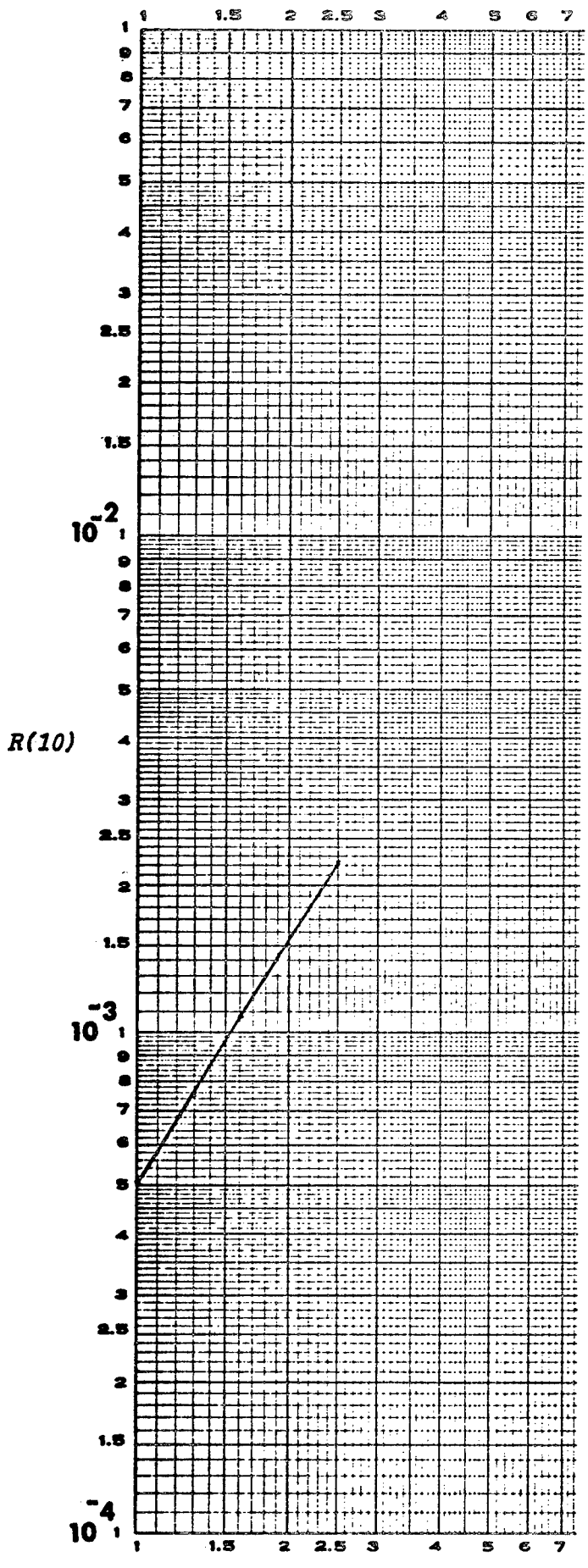
Levy, H. and E.A. Baggot. Numerical Solutions of Differential Equations. New York: Dover Publications, Inc., 1950.

Norman, A.C. "Computing with Formal Power Series". ACM Transactions on Mathematical Software. Vol. 1, No. 4, December, 1975.

Petrovski, G. Ordinary Differential Equations. Englewood Cliffs: Prentice Hall, Inc., 1966.

Ralston, A. and H.S. Wilf. Mathematical Methods for Digital Computers. New York: John Wiley and Sons, Inc., 1960.

Shampine, L.F. and R.C. Allen, Jr. Numerical Computing: An Introduction. Philadelphia: W.B. Saunders Company, 1973.



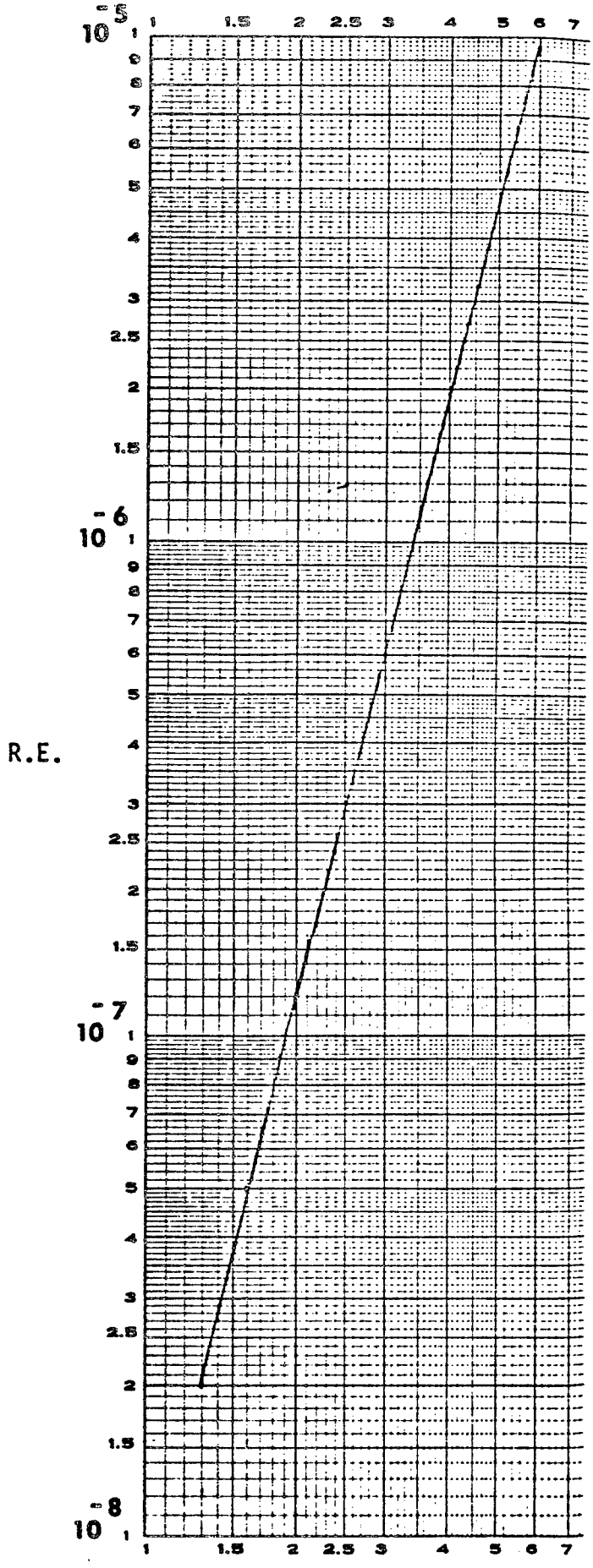




FIGURE 4-1

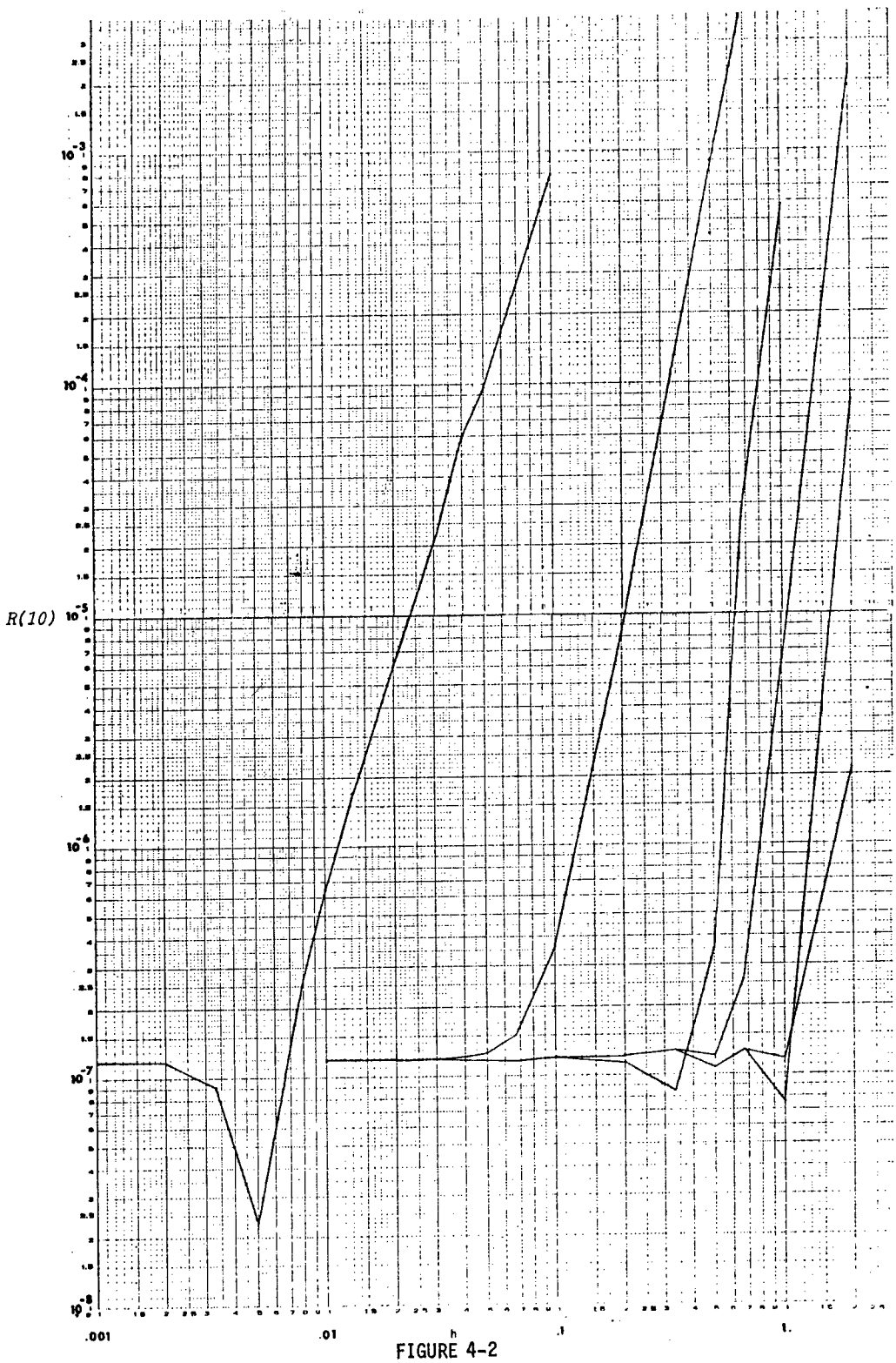



FIGURE 4-2

# INPUT CONTROLLABLE STOCHASTIC MODEL

Sheafen Frank Kuo
U. S. Army Construction Engineering Research Laboratory
P. O. Box 4005
Champaign, Illinois 61820

1. INTRODUCTION. This paper introduces a model which incorporates the principles of both Markov chains and finite state machines. Markov chains possess stochastic behavior in the transition between states but are not input controllable. Finite state machines, on the other hand, are input controllable between states, but do not have stochastic behavior. Basic concepts of an input controllable stochastic model and analysis of its short- and long-term behaviors are presented. Forecast accuracy (FA) of a model is defined and relations between strings and models are described. The first order derivative (FOD) of a model is introduced. A sufficient condition for a model and its FOD to have equal FA is proved. In addition, some applications are briefly discussed.

2. INPUT CONTROLLABLE STOCHASTIC MODEL.

A. Definition. An input Controllable Stochastic Model (ICSM) is a quadruple $H = \{I, O, S, \mu\}$ where I is the input set, S is the state set, O is the output set, and $\mu$ is a probabilistic function, such that

$$\mu: I \times S^t \times O \times S^{t+1} \to P$$

where

$$S^t, S^{t+1} \subset S$$

P = the set of real numbers between 0 and 1.

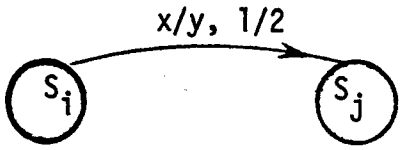
In other words, given input $x_i$ and present state $S_j$, $\mu$ assigns a probability $P_{ijmn}$ to each output $y_m$ and next state $S_n$. Using the property of the probability function gives

$$\sum_{x_m \in O} \sum_{S_n \in S} P_{ijmn} = 1, \text{ for all } x_i \in I,\ S_j \in S$$

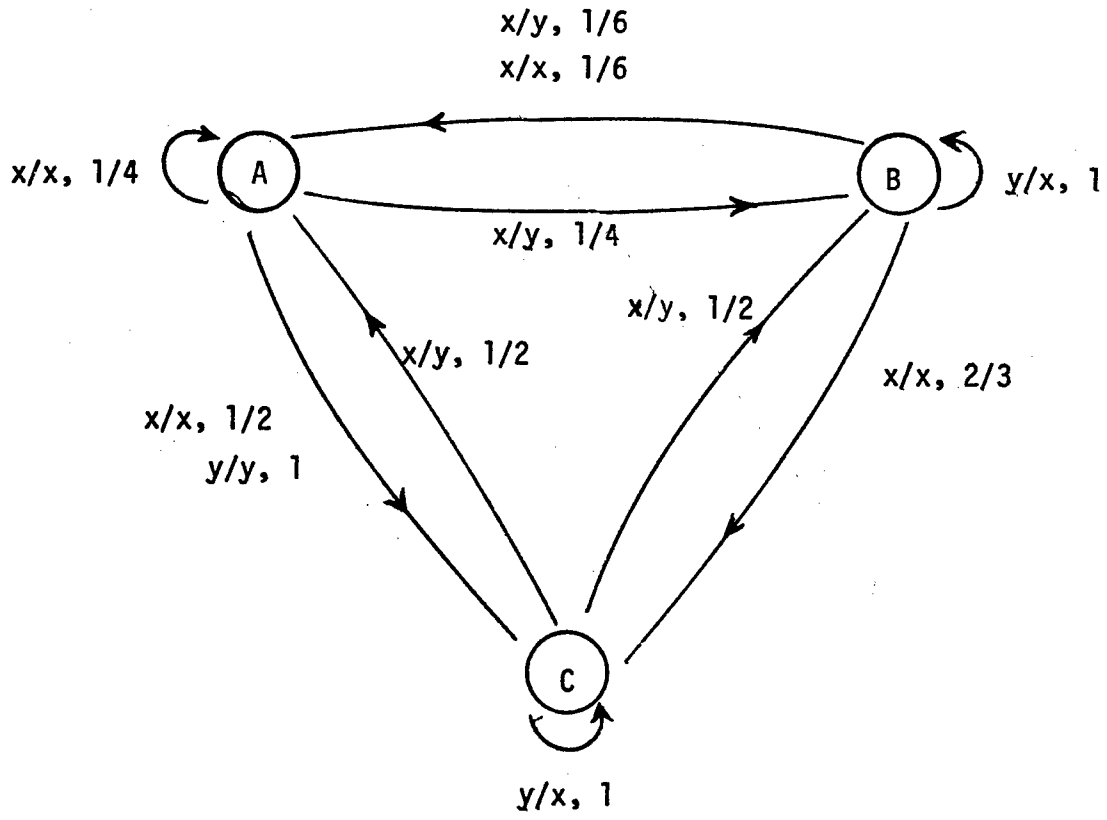
B. Example. Let $I = O = \{x, y\}$, $S = \{A, B, C\}$. $\mu$ is defined as follows:

$\mu(x, A, x, A) = 1/4$
$\mu(x, A, y, B) = 1/4$
$\mu(x, A, x, C) = 1/2$
$\mu(y, A, y, C) = 1$
$\mu(x, B, x, A) = 1/6$
$\mu(x, B, y, A) = 1/6$
$\mu(x, B, x, C) = 2/3$
$\mu(y, B, x, B) = 1$
$\mu(x, C, y, A) = 1/2$
$\mu(x, C, y, B) = 1/2$
$\mu(y, C, x, C) = 1$
$\mu = 0$ otherwise

C. Graphic Notation. Noting the input, output, and probability on an arc path between two states $S_i$ and $S_j$ gives the following:




This notation means that given input x and current state $S_i$, the probability of getting the next state $S_j$ and output y is 1/2. Making an arc between each communicable state would give a flow graph for that model. The flow graph of the last example is shown as follows:



3. INPUTTABLE MARKOV CHAIN (IMC). A special case of ICSM of interest in this paper is the model with empty output set O. This kind of model is called the Inputtable Markov Chain (IMC).

A. Definition. An IMC is a triple G = {I, S, k} where I is the input set, S is the state set, and k is a probability function which satisfies:

$$k(S_i, x_j, S_k) = \text{prob}\{S^{t+1} = S_k \mid S^t = S_i, x_j \text{ is input}\} \text{ for all}$$

$$S_i, S_k \epsilon S \text{ and } x_j \epsilon I$$

Hence

$$\sum_{S_k \epsilon S} k(S_i, x_j, S_k) = 1 \text{ for all } S_i \epsilon S \text{ and } x_j \epsilon I$$

B. Transition Matrix. Let K = {I, S, k} be an IMC, where I = $\{x_1, x_2 \ldots x_n\}$ and S = $\{S_1, S_2, \ldots S_n\}$. Using $P_{ijk} = k(S_i, x_j, S_k)$ gives:

$$Px_j = \begin{array}{c} \\ 1 \\ 2 \\ \cdot \\ \cdot \\ \cdot \\ m \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & m \end{array} \\ \begin{pmatrix} P_{1j1} & P_{1j2} & \cdots & P_{1jm} \\ P2j1 & P2j2 & \cdots & P2jm \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ P_{mj1} & P_{mj2} & \cdots & P_{mjm} \end{pmatrix} \end{array} \qquad (1)$$

as a conditional transition matrix of input $x_j$. Notice that the summation of each row is 1.

Suppose at each state $S_i$ the probability of getting input $x_j$ is $q_{ij}$. Let

$$Q_j = \begin{pmatrix} q_{1j} & & & \\ & q_{2j} & & 0 \\ & & \cdot & \\ & & & \cdot \\ 0 & & & \cdot \\ & & & & q_{mj} \end{pmatrix} \text{ be a diagonal matrix}$$

Let

$$P_j = Q_j \cdot Px_j \tag{2}$$

$$P = \sum_{j=1}^{n} P_j \tag{3}$$

Thus, P is a transition matrix without knowing an input variable.

## 4. SHORT- AND LONG-TERM BEHAVIOR.

A. Short-Term Behavior. From a given model one can explore the k-step state distribution; i.e., after the k-step, the model will go to a certain state with a certain probability. Two cases can be considered:

(1) Input string is given. If $x_1x_2 \ldots x_k$ is the input string, then the k-step conditional transition matrix is

$$Px_1x_2 \ldots x_k = Px_1 \cdot Px_2 \ldots Px_k \tag{4}$$

where $Px_j$ is the conditional transition matrix defined by equation 1.

(2) Input string is not given, but the input distribution matrix $Q_j$ is given. Equations 2 and 3 can then be used to find P, and the k-step transition matrix is as shown in equation 5.

$$P^k = \underbrace{P \cdot P \ldots P}_{k \text{ times}} \tag{5}$$

B. Long-Term Behavior. For long-term analysis, only the case without input is considered here. If k is large, calculating $P^k$ is somewhat cumbersome, but applying the z-transformation, which is a common way of calculating the power of a stochastic matrix, simplifies it. Let

$$q(k) = P^k$$

The z-transformation Q(z) of q(k) is defined as:

$$Q(z) = \sum_{k=0}^{\infty} q(k)z^{-k} \tag{6}$$

or $$Q(z)[I - z^{-1}P] = q(0) \tag{7}$$

since $$q(0) = 1$$

Hence $$Q(z) = [1 - z^{-1}P]^{-1} \tag{8}$$

Let the inverse transform of $Q(z) = y(k)$.

Therefore $$y(k) = z^{-1}[(I - z^{-1}P)^{-1}] \tag{9}$$

$$q(k) = y(k) \tag{10}$$

or $$y(k) = p^k \tag{11}$$

From equation 11,

$$\lim_{k \to \infty} p^k = \lim_{k \to \infty} y(k)$$

## 5. STATE PROBABILITY AND FORECAST ACCURACY OF A MODEL.

A. Definition. State probability (or state frequency) $P_i$ of state $S_i$ is defined as

$$P_i = \sum_j P_j P_{ji} \tag{12}$$

where $$P_{ji} = \sum_k q_{jk} P_{jki} \tag{13}$$

$q_{jk}$ = the probability of input $x_k$ at state $S_j$

$P_{jki}$ = the conditional probability of $S_j$ transferring to $S_i$, given input $x_k$

Obviously $$\sum_i P_i = 1 \tag{14}$$

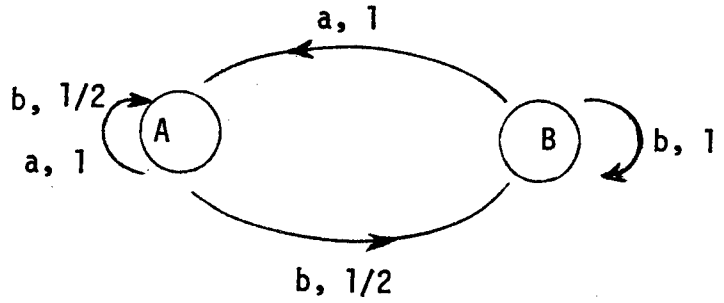
Using equations 12 through 14, $P_i$ for each i can be found.

The forecast accuracy [FA(R)] of a model R is defined as:

$$FA(R) = \sum_i \sum_k [P_i q_{ik} \cdot \max_j P_{ikj}]$$

Intuitively, FA(R) is the maximum average probability of forecasting the next state correctly, given the current input and state.

B. Example 1.




The figure shown above is a simple model R with two states and two input variables. Assume input a and b are equally probably at each state. Simple calculation usings equations 10, 11, and 12 gives:

$$\text{state frequency: } P_A = \frac{2}{3} \quad P_B = \frac{1}{3}$$

Therefore, state A is visited twice as frequently as state B is visited.

$$FA(R) = P_A \cdot q_{Aa} \cdot \max\{P_{AaA}, P_{AaB}\} + P_A \cdot q_{Ab} \cdot \max\{P_{AbA}, P_{AbB}\}$$
$$+ P_B \cdot q_{Ba} \cdot \max\{P_{BaA}, P_{BaB}\} + P_B \cdot q_{Bb} \cdot \max\{P_{BbA}, P_{BbB}\}$$

Since $$q_{Aa} = q_{Ab} = q_{Ba} = q_{Bb} = \frac{1}{2} \quad P_A = \frac{2}{3} \quad P_B = \frac{1}{3}$$

Hence $$FA(R) = \frac{1}{2}\left[\frac{2}{3} \cdot \max\{1, 0\} + \frac{1}{3} \cdot \max\{\frac{1}{2}, \frac{1}{2}\} + \frac{1}{3} \cdot \max\{1, 0\}\right.$$
$$\left. + \frac{2}{3} \cdot \max\{1, 0\}\right]$$

Therefore $$FA(R) = \frac{11}{12}$$

The average chance of forecasting the next state correctly is 11/12, given the current state and input.

It is trivial to see that a deterministic model, like a finite state machine, has a forecast accuracy of 1.

The following are some of the trivial properties of FA(R):

(1) $$FA(R) \geq \max_j \sum_i \sum_k P_i q_{ik} P_{ikj}$$

(2) $FA(R) = 1 - \sum_i \sum_k P_i q_{ik} \min_j P_{ikj}$ if number of state is 2

(3) Define $FA(R|k) = \sum_i P_i \max_j P_{ikj}$

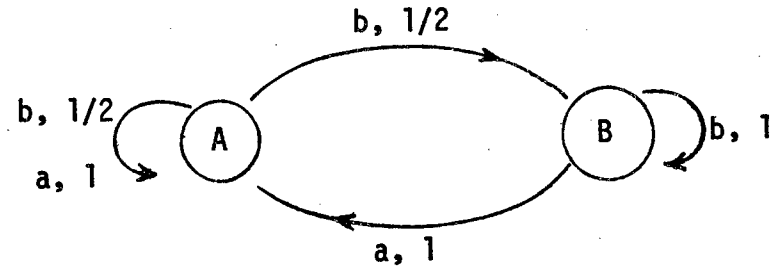
Then $\min_k FA(R|k) \leq FA(R) \leq \max_k FA(R|k)$

6. THE STRING AND MODEL. Consider the following string:

aAaAaAbAbBbBaAbAbBaAbBaAbB . . . (15)

where A and B are state variables and a and b are input variables.

After sufficient observation, a model like that shown below can be developed.




Combining the last state with the current state, or putting the current state to the left upper corner of the next state gives

A A A A B B A A B A B A B
aAaAaAbAbBbBaAbAbBaAbBaAbB . . .

Putting the upper characters down gives

A A A . . .
aAaAAaAAbAAbABbBBaBAbAAbABaBAbABaBAaABbB . . . (16)

String 16 is said to be a First Order Derivative (FOD) of string 15. FOD's are developed to increase the number of states so that the system is better described. For example, if string 15 is an observer's weather record with A and B meaning sunny and rainy, respectively, and a and b meaning decreasing temperature and increasing humidity, respectively. An FOD of String 15 can be derived to String 16 with AA as sunny, AB as cloudy, BA as partly cloudy, and BB as rainy. Therefore, String 16 is more descriptive than String 15.

## 7. THE DERIVATIVE OF THE MODEL.

A. Definition. Like a string, the model also has derivatives. Let R be an IMC.

$$R = \{I, S, H\},\ I = \{x_1, x_2, \ldots, x_n\},$$

$$S = \{S_1, S_2 \ldots s_m\},\ H = \{P_{x_j} \quad j = 1, \ldots n\}$$

where $P_{x_j}$ is a transition m-matrix under input $x_j$.

Define $S' = S \times S = \{S'_k \mid S'_k = S_i S_j;\ S_i, S_j \epsilon S\}$

Let $P_{ikj}$ be an element of $P_{x_k}$, $P'_{i_o k j_o}$ be an element of $P'_{x_k}$

where $P'_{x_k}$ is a transition matrix of dim $m^2 \times m^2$ under $x_k$

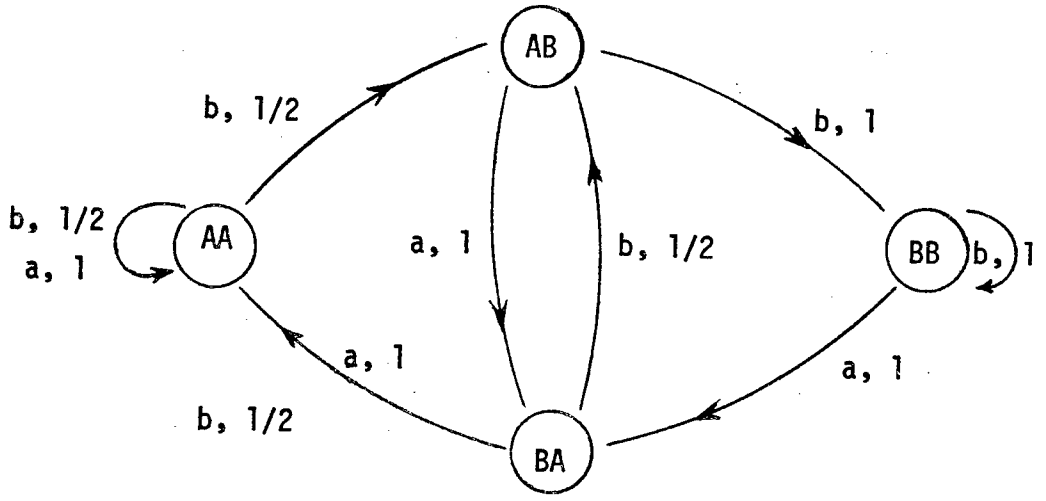
such that $P'_{i_o k j_o} = P_{ikj}$ if $S'_{i_o} = S_t S_i$, $S'_{j_o} = S_r S_j$

and if $i = r$, otherwise $P'_{i_o k j_o} = 0$

Define $H' = \{P'_{x_k} \mid k = 1, \ldots, n\}$ and $I' = I$

Then $R' = \{I', S', H'\}$ is an FOD of $R = \{I, S, H\}$

B. Example 2. The FOD R′ of the model R in Example 1 is as shown below.



Similar to Example 1, assume that input a and b are equally probable at each state. Then it can be shown that

$$FA(R) = FA(R') \tag{17}$$

In general, equation 17 is not true. However, the sufficient condition of it can be found.

Define index sets of R, R′ as follows:

$$T = \{r \mid S_r \in S\}$$

$$T' = \{r \mid S'_r \in S'\}$$

$$T'_i = \{r \mid S'_r \in S \times S_i\} \quad \text{where } S \times S_i = \{S_j S_i \mid j \in T\}$$

LEMMA 1:

(1) For all $i_o \in T'_i$ $j \in T$ there exists $j_o \in T'$ s.t. $P_{ikj} = P'_{i_o k j_o}$

(2) For all $i_o \in T'_i$ $j_o \in T'$ there exists $j \in T$ s.t. $P_{ikj} \geq P'_{i_o k j_o}$

Proof: Trivial

LEMMA 2:

For all $i_o \in T'_i$ $\max_{j \in T} P_{ikj} = \max_{j \in T'} P'_{i_o kj}$

Proof:

Let $$\max_{j \in T} P_{ikj} = P_{ikj_1} \quad \text{where } j_1 \in T \tag{18}$$

By Lemma 1

$$P_{ikj_1} = P'_{i_o k j_2} \leq \max_{j_o \in T} P'_{i_o k j_o} \quad \text{for all } i_o \in T'_i \tag{19}$$

Conversely, let $$\max_{j \in T'} P'_{i_o kj} = P'_{i_o k j_1} \tag{20}$$

By Lemma 1 $$P'_{i_o k j_1} \leq P_{ikj_3} \leq \max_{j \in T} P_{ikj}$$

Therefore $\max_{j \in T} P_{ikj} = \max_{j \in T'} P'_{i_o kj}$ for all $i_o \in T'_i$ QED

THEOREM:

If $q_{ik} = q'_{i_o k}$ for all $i_o \varepsilon T'_i$ and if $\sum_{i_o \varepsilon T'_i} P'_{i_o} = P_i$ for all $i$

Then $$FA(R) = FA(R')$$

Proof:

Since $$FA(R) = \sum_i \sum_k P_i q_{ik} \max_j P_{ikj}$$

And $$FA(R') = \sum_{i_o} \sum_k P'_{i_o} q'_{i_o k} \max_{j_o} P'_{i_o k j_o}$$

$$= \sum_i \sum_k \sum_{i_o \varepsilon T'_i} P'_{i_o} q'_{i_o k} \max_{j_o} P'_{i_o k j_o}$$

$$= \sum_i \sum_k \sum_{i_o \varepsilon T'_i} P'_{i_o} q_{ik} \max_j P_{ikj} \quad \text{(by Lemma 2)}$$

$$= \sum_i \sum_k [(q_{ik} \max_{j_o} P_{ikj}) P_i]$$

Hence $$FA(R) = FA(R')$$ QED

It is not surprising that for most models R and its FOD R', conditions

$$q_{ik} = q'_{i_o k} \quad \text{for all } i_o \varepsilon T'_i$$

$$\text{and } \sum_{i_o \varepsilon T'_i} P'_{i_o} = P_i \quad \text{for all } i$$

are easily satisfied; thus, the forecast accuracy of the FOD R' is not lost.

8. APPLICATION. The Markov Chain has been applied to many management science or systems analysis fields. IMC improves the Markov Chain because it has more features to adapt the real work of physical, economic, biological, or engineering systems [1], [4], [6]. The most important feature, the input controllability, allows one to understand a system by controlling input to find the subsequent changes in state (and output). Because the processing is stochastic, the finite state machine (or automata) cannot describe the procedure properly. If a model can be

build corresponding to a string of data, the model can then be tested and evaluated by calculating its forecast accuracy. The FOD is a useful tool in understanding the model, as illustrated by the weather forecasting example.

Some stochastic automata have already been applied to the reliability problem and decision process [8]. It is hoped that the discussion in this paper will create a new interest in the research in a discrete system.

## REFERENCES


Ackoff, R., *Progress in Operations Research* (Wiley, New York, 1961). [1]

Brzozowski, J. A., "Derivatives of Regular Expressions," *Journal of Association for Computing Machinery*, Vol. 11, pp. 481-494 (1964). [2]

Graver, D. P. and G. L. Thompson, *Programming and Probability Models in Operations Research* (Brooks/Cole, California, 1973). [3]

Harling, J., "Simulation Techniques in Operations Research," *Journal of Operations Research Society of America* (May - June, 1958). [4]

Howard, R. A., *Dynamic Programming and the Markov Process* (Wiley, New York, 1960). [5]

Leibowitz, M. L., "The Role of Computer Simulation in Military Operations Research," *Bulletin of Operations Research Society of America* (Fall, 1960). [6]

Levin, V. I., "One Method of Analyzing the Reliability of Finite Automata," *Automat. i Telemeh.*, Vol. 27, No. 4, pp. 114-118 (1966). [7]

Tou, J., *Applied Automata Theory* (Academic Press, New York and London, 1968). [8]

# A SCANNING ELECTRON MICROSCOPE INVESTIGATION OF STATICALLY LOADED FOUNDATION MATERIALS

RAYMOND E. AUFMUTH
Department of the Army
CONSTRUCTION ENGINEERING RESEARCH LABORATORY
P.O. Box 4005
Champaign, Illinois 61820

ABSTRACT. Selected rock samples were tested to failure in bending tension and compression test modes within the vacuum stage of a scanning electron microscope (SEM). The load was applied slowly such that crack initiation and growth could be observed and recorded by photography and video tape. The failure surfaces were further evaluated by standard methods to determine failure mechanisms involved for each test mode and rock type.

1. INTRODUCTION. An understanding of the physical properties and behavior of rock materials (rock engineering) is necessary to implement a systems approach for designing a structure. Structural design considerations may include rock removal, tunneling, use of rock as a foundation material, or any combination of these factors. Information about the fundamental mechanisms of the fatigue and failure properties of rock is essential and should be available to the design engineer. Since construction of underground structures such as tunnels for defense facilities, underground power plants, and hydraulic structures has increased, and since idealized construction sites are not always available, it is essential that rock failure mechanisms be controlled by proper design practice.

There have been few investigations concerning the failure modes of rock materials in simulated field tests, primarily because of the experimental problems associated with controlling rock failure. Wawersik, Brace, and Fairhurst (AROD Proposal 11278 EN) have investigated the post-failure behavior of selected materials. Brace and Sprunt have investigated the microcavities in crystalline rocks; and Brace (ARO Contract: DAHCO 4-73-C-0017) is presently investigating the microstructure in crystalline rocks with a scanning electron microscope. The study herein complements these and other investigations by advancing the state-of-the-art of failure mechanisms.

2. EQUIPMENT. The AMR 900 Scanning Electron Microscope (SEM) is a high-resolution instrument providing surface resolution of 100 to 200 Å and useful magnification of up to 50,000X. The depth of focus is accurate to tens of microns. This means that a fairly rough surface, such as a rock fracture surface, will remain in focus at high magnifications. The micrograph obtained appears similar to that obtained from the reflection light microscope, but it has much better resolution and depth of field.

The AMR 1300 stage and chamber door assembly is inserted in the AMR 900 Chamber to replace the standard door and stage. Depending on the type of test to be performed, the bending or tension-compression device is mounted on this assembly during its operation. A platform mounted on the base plate provides the X motion (right and left), the Y motion (backward and forward) and the Z motion (up and down). The Z motion is actuated by articulated shafts to the door and a single knob on the front face of the door. One revolution of the exterior knob represents a change in Z of 1 mm; the readout is such that one digit corresponds to a change in specimen neight of 0.1 mm. A counter-clockwise rotation of the stage raises the bending stage and closes the compression-tension heads. The same revolution and motion changes apply to the X and Y directions.

The *bending stage* (Figure 1) is custom-designed to load a rectangular specimen having maximum dimensions of 1 x 1 x 6 in. in simple three-point bending to a maximum load of 2000 lb. This stage is essentially a platform having a knife edge on its top surface that supports the specimen at the center of its bottom surface. A load bar connected to the platform by a ball screw and gear system is connected to the edge which bears down on the top of the specimen. The points of the specimen's tension links continually vary to accommodate specimens of 3 to 5 in. in length. The maximum bar deflection is 0.375 in., it is applied via the hand crank on the outside of the chamber door. Each digit of the readout corresponds to a specimen deflection of 0.004 mm at no load.

The bending device is loaded into the chamber parallel to the Y axis at an angle of 45 degrees to the horizontal. Two positions 180 degrees apart are possible, allowing observation of the tension face or a side face of the specimen.

The *tension-compression stage* (Figure 2) consists of two heads mounted on a pair of right- and left-hand ball screws. When the screws are rotated, the heads move either together or apart, but remain parallel. Compression specimens are placed between the flat surfaces of the heads for testing. Tension specimens may be held in place by various techniques. In this study, square steel heads with a centered slot and a pin hole normal to the slot were epoxied to the specimen ends. These in turn were connected by the pin to threaded rods, flattened at one end, which fed through the holes in the stage heads (Figure 3).

The gear train used for specimen deflection is the same used for the bending state; however, one digit of readout corresponds to 0.005 mm change in distance between the heads. Minimum distance between the heads is 0.25 in., and maximum distance is 4.0 in. The maximum load which may be applied to either failure mode (tension-compression) is 2000 lb. The stage itself is tilted at an angle of 15 degrees to the horizontal; however, since a specimen may be placed in any orientation between the heads, any desired tilt may be obtained.

3. MONITORING DEVICES. A secondary (backscatter) electron image for direct observation of the specimen is developed and displayed on a signal modulation unit. (This is the primary visual means of specimen observation.)

In addition, the secondary image may be displayed on a 9-in. square TV rate monitor display unit. This unit displays the same field as the previous module, but has a limited magnification range of from 100X to 10,000X. It has a built-in zoom capability that allows closeup display of a small area in the center of the TV and is operable at all magnifications.

Photomicrographs are obtained through a record oscilloscope 4 x 5 in. square and Polaroid 52-P/55-P/N, 4 x 5 in. film. An alphanumeric generator is integrated into the signal modular display unit in order to facilitate identification and description of the photomicrographs.

4. SPECIMEN PREPARATION. Table 1 lists the representative suite of rock samples chosen for evaluation in this study and summarizes their physical characteristics. One set of specimens was prepared for each of three test modes: bending, tension, and compression. In addition, three cross-sectional dimensions were prepared to determine any specimen size effects.

Bending (flexure) specimens were sawed into beams and ground square in lengths from 4 to 5 in. long and cross sections of 1/8, 1/4, and 1/2 in. square. A fine notch was filed into the top (tension) surface to control crack origin during scanning at high magnifications. This notch was approximately 1/16 in. deep for all specimens.

Tensile specimens were prepared in the same manner, but were cut 2-1/8 in. in length. Notches were ground into opposite sides of the specimens to minimize extraneous stress concentrations at other points in the specimen test mode. The intact cross-section varied from 3/16 to 1/4 in., depending on the specimen size. Only the tensile specimens were modified for testing; a steel head was epoxied to each end to facilitate application of pure tensile stresses.

Compression specimens were prepared similarly to the tensile specimens in lengths of 2 to 3 in., with no notching or other preparations made after grinding.

a. specimens were strain-gaged, coated under vacuum with gold-platinum to facilitate conductivity, and wrapped in aluminum foil. The purpose of the foil wrapping was to prevent spalling during testing or at a failure which could harm internal portions of the vacuum system.

5. SPECIMEN FAILURE EVALUATION. The selected rock specimens were evaluated for bending, tension and compression failure. Prior to the SEM evaluation, representative strain-gaged specimens were tested to failure in each mode outside the vacuum drawer to determine: (1) if failure at each size in each mode was feasible; (2) the extent of spalling, if any; and (3) where and how failure would occur on the specimen.

Problems encountered were associated with the compressive failure mode, which proved to spall excessively and to fail unpredictably along the entire length of the specimen. For the Westerly Blue granite and Traprock shale specimens, the test size had to be reduced to 1/4 in. square (in compression) to facilitate the 2000 lb maximum load.

6. SEM FAILURE METHODS. The bending (flexure) failure mode was first evaluated in the vacuum drawer by applying load to a specimen up to a strain level approaching failure. At the point approaching failure, the load application was slowed to approximately 0.3 mm/min. The notch area was scanned during this load application. Slow load application was continued until crack initiation, when the load was stopped and the crack scanned. If the crack was partial, loading was applied again while the crack tip was followed with a scan. Loading was halted periodically for a side to side scan. For the bending failure mode, there were no significant changes indicated on either side of the failure plane.

For the tensile failure mode evaluation, a slight seating load was applied manually to the specimen before placing it in a vacuum, so that the specimen would not rotate during load application. Since this test mode builds up stress prior to failure, most specimens failed rapidly, even at a very small load rate. In some cases, a scan was possible before complete separation. When side scans were performed in this failure mode, such secondary phenomena as grain separation were present.

7. FAILURE SURFACE EVALUATION. After completion of the SEM failure evaluation, one surface of each failed specimen was mounted on studs and coated with gold-platinum. These surfaces were then evaluated by standard SEM evaluation procedures and a standard stub stage. This evaluation, together with the SEM failure evaluation, was the basis of the failure analysis.

8. FAILURE ANALYSIS. When the beams failed in a bending mode, both intergranular and transgranular failure mechanisms were present, usually in approximately equal distribution; however, different rock types exhibited each failure mechanism to different degrees. The Bonne Terre limestone and Westerly Blue granite exhibit approximately equal distribution of the inter- and transgranular failure mechanisms. The Traprock shale and Murphy marble beams primarily displayed transgranular failure and intergranular failure. The Danby marble primarily showed intergranular failure and some transgranular failure, while the Berea sandstone exhibited 100 percent intergranular failure mechanism.

The tensile failure test mode showed no preferences to either rock type or crystal/grain size relative to the failure mechanisms. Both inter- and transgranular failure mechanisms were approximately equally distributed for each rock type evaluated. These figures also indicate the variation in crystal/grain size for the six rock types evaluated. The Westerly Blue granite, Murphy marble, and Bonne Terre limestone display good crystal cleavage planes. The Berea sandstone exhibits surface wear on the individual sand grains.

As anticipated in the compression mode, transgranular failure mechanisms were present due to the nature of the test. However, the Berea sandstone and Bonne Terre limestone exhibited an unexpectedly excellent intergranular failure mechanism. The Westerly Blue granite, Murphy marble, and Traprock shale exhibited predominantly (95 percent) transgranular failure; the Danby marble displayed both failure mechanisms, with intergranular failure predominating.

Table 2 summarizes the failure mechanisms relative to test mode and rock type.

9. SUMMARY AND CONCLUSIONS. Selected rock samples were prepared and tested to failure by bending, tension, and compression within the vacuum stage of a scanning electron microscope. Load was applied very slowly in order to observe crack initiation and growth. Crack growth was observed visually and recorded by both photography and video tape. The crack surfaces of the failed specimens were evaluated by standard methods, and two evaluation technques were used to determine the failure mechanisms for each test mode and rock type studied.

Conclusions. Based on the techniques of stub evaluation and failure in the vacuum stage, the following statements apply only to those test modes and rock materials studied herein:

a. Cross-section size differences had no effect on the failure mode. The only benefit derived from studying several sizes were facilitation of compression testing of granite and shale specimens.

b. The rock types evaluated in this study had no apparent effect on the failure mode or the failure mechanisms.

c. Crystal/grain size directly and significantly influences the failure mechanisms as follows:

(1) Large crystals/grains - failure was primarily transgranular for each test mode.

(2) Small crystals/grains - failure was primarily intergranular for each test mode.

d. Cementing agents have little or no effect on the gross failure mechanisms; however, failure in the cementing agent was exclusively transgranular.

10. RECOMMENDATIONS. This study has proved the feasibility and usefulness of applying a metallurgical research tool to geologic materials. The present study, in conjunction with studies by Brace of specimen preparation techniques, could yield valuable information in the area of geophysics and earthquake analysis. Studies relative to slickenside development in clay shales and other shear phenomena of soil and rock could be advanced by this approach.

11. REFERENCES.

a. W. F. Brace and W. R. Wawersik, "Post-Failure Behavior of Granite and Diabase," Rock Mechanics and Mining Sciences and Geomechanic Abstracts, Vol 3, No. 2 (1971), pp 61-85.

b. E. S. Sprunt and W. F. Brace, "Direct Observations of Microcavities in Crystalline Rocks," Rock Mechanics and Mining Sciences and Geomechanic Abstracts, Vol 11, No. 4 (1974), pp 139-150.

c. W. F. Brace, Rock Mechanics and Mining Sciences and Geomechanic Abstracts, Vol 3, No. 2 (1971); and E. S. Sprunt and W. F. Brace, "Direct Observation of Microcavities in Crystalline Rocks," Rock Mechanics and Mining Sciences and Geomechanic Abstracts, Vol 11, No. 4 (1974).

# TABLE 1

## PHYSICAL PROPERTY DATA

| Identification Code | Rock Type & Location | | Schmidt Rebound Hardness No. (R) Range | Avg | Point Tensile Strength psi Avg | L/D | Slake Durability (% Lost) Cycles 1 | 2 | Unit Weight gm/cc | Unconfined Compressive Strength psi x $10^3$ | psi x $10^6$ | Sonic Velocity @ 50% 6a ft/sec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MODO | Limestone Missouri (Bonne Terre) | No. Tests | 24-41 | 31 | 743 | 2.04 | 0.74 | 0.95 | 2.663 | 15.1 | 8.75 | 13944 |
| | | V% | | 48 | 4 | | 1 | 1 | 2 | 3 | | |
| | | | | 20.5 | | | | | | 5.6 | | |
| VTMA | Marble Vermont (Danby) | No. Tests | 18-30 | 21 | 299 | 2.03 | 1.54 | 2.12 | 2.708 | 7.95 | 5.85 | 20222 |
| | | V% | | 48 | 4 | | 1 | 1 | 2 | | | |
| | | | | 21.9 | 6.2 | | | | | 0.7 | | |
| NYTRSH | Shale New York (Traprock) | No. Tests | 27-45 | 30 | 563 | 2.02 | 0.57 | 0.98 | 2.695 | 9.15 | 8.25 | 17395 |
| | | V% | | 48 | 4 | | | | | 3 | | |
| | | | | 21.5 | 11.5 | | | | | 2.5 | | |
| GAMA | Marble Georgia (Murphy) | No. Tests | 19-36 | 28 | 375 | 2.07 | 1.41 | 2.09 | 2.708 | 9.7 | 7.5 | 19799 |
| | | V% | | 36 | 4 | | 1 | 1 | | 3 | | |
| | | | | 17.1 | 18.4 | | | | | 0.9 | | |
| VTWBGR | Granite Vermont (Blue Westerly) | No. Tests | 28-41 | 35 | 1171 | 2.00 | 0.57 | 0.70 | 2.643 | 26.0 | 7.636 | 16910 |
| | | V% | | 48 | 4 | | 1 | 1 | | 4 | | |
| | | | | 15.2 | 6.0 | | | | | 5.0 | | |
| CQSS | Sandstone Ohio (Berea) | No. Tests | 15-22 | 17 | 118 | 2.04 | 4.36 | 5.80 | 2.173 | 8.80 | 2.444 | 12002 |
| | | V% | | 45 | 4 | | 1 | 1 | | 4 | | |
| | | | | 19.4 | 18.2 | | | | | 11 | | |

TABLE 2

SUMMARY OF FAILURE MECHANISMS

| Rock | FAILURE MODE | | | Crystal/Grain Size |
|---|---|---|---|---|
| | Bending | Tension | Compression | |
| Bonne Terre Limestone | T - I | T - I | I | Macro Crystalline |
| Danby Marble | t - I | T - I | t - I | Fine - Macro-Crystalline |
| Traprock Shale | T - i | T - I | I | Micro Granular |
| Murphy Marble | T - i | T - I | T | Medium - Macro Crystalline |
| Westerly Blue Granite | T - I | T - I | T | Fine - Macro Crystalline |
| Berea Sandstone | I | T - I | T | Marco-Granular |

t, T* = Transgranular

i, I* = Intergranular

* Capital letter indicates predominance of failure mechanism.

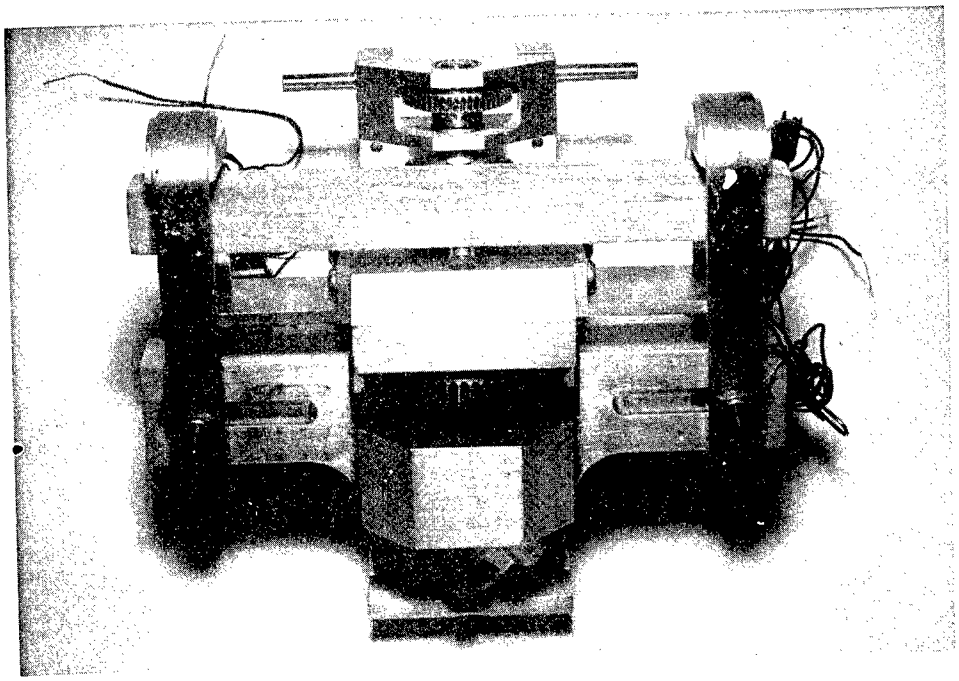


Figure 1: Bending Stage.

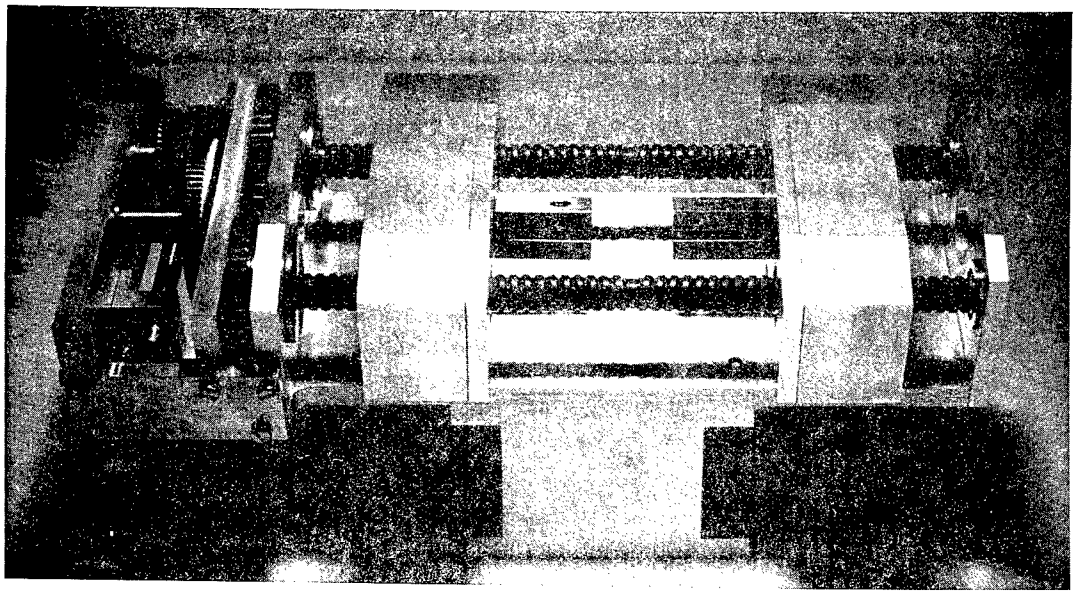


Figure 2: Tension-Compression Stage.

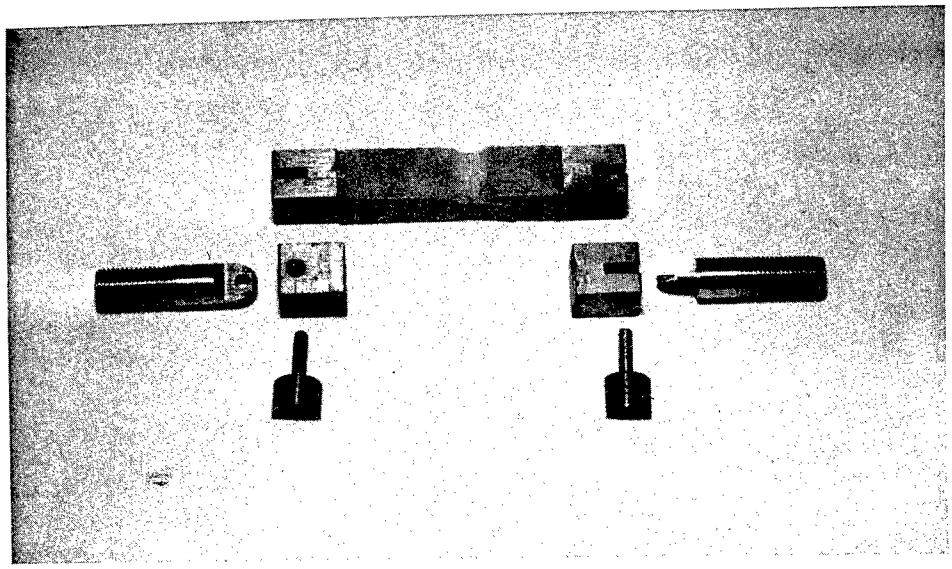


Figure 3: Tension Heads on Specimen.

# PHASE II SECURE VOICE PROGRAM - AN INDEPENDENT ARMY ANALYSIS

Theodore S. Trybul
Comptroller Directorate, Cost Analysis Division
HQ, US Army Materiel Development & Readiness Command
Alexandria, VA 22333

ABSTRACT. The Phase II Secure Voice Program (P2SVP) will develop, acquire and install a high quality, effective, long haul DOD secure voice system that will serve up to 10,000 subscribers in the 1985 time frame to provide requisite interoperability with strategic and tactical systems. It replaces the Phase I Automatic Secure Voice Communications (AUTOSEVOCOM) Network and Interim Conferencing for the National Military Command System (NMCS).

The independent army analysis was a unique effort because this was the first time the Army was asked by the Secretary of Defense to evaluate another agency's program.

The Director, Telecommunications and Command and Control Systems, Office of the Secretary of Defense requested the Army to prepare Independent Cost Estimates (ICE's) of the P2SVP alternatives developed by the DCA in support of Development Concept Paper (DCP) #153. These Independent Cost Estimates were to be prepared for the Defense Systems Acquisition Review Council, Office of the Secretary of Defense, Cost Analysis Improvement Group (DSARC, OSD, CAIG). This analysis provided input for the full-scale engineering development decision point.

HQ, DARCOM established a Systems Study Group (SSG), Chaired by myself, consisting of representatives (multi-and inter-disciplinary) from COA, CSA, ACC, ECOM, DCA, NSA, DCEC, DDR&E, DTACCS, and OSD. This SSG generated an ICE by analyzing the Phase II computer printouts at the Defense Communications Engineering Center (DCEC), supported by engineering judgement, mathematical analysis, expert opinion and historical data. These estimates were prepared in accordance with the Army Materiel Guide for Organizing and Presenting Cost Studies, and the HQ, Department of the Army Investment and O&S Cost Guides for Army Materiel Systems.

1. INTRODUCTION. An analysis of P2SV and alternatives was made previously by the Defense Communications Agency (DCA) in the form of an Economic Analysis Estimate (EAE). The ICE described in this paper provides an independent evaluation of the costs generated in that EAE. Such an evaluation is a normal procedure in the acquisition of Army materiel systems. Together with the benefits (effectiveness) calculations made in the EAE, it allows a ranking of the candidate systems to be made and gives visibility to the decision maker of the trade-offs involved.

2. DESCRIPTION OF ALTERNATIVES. Independent Cost Estimates were made on 4 alternatives: Worldwide Tenley, Narrowband, Wideband and Hybrid Systems for the Phase II Secure Voice Program. Summary descriptions are given.

TABLE 1 - SUMMARY DESCRIPTION OF ALTERNATIVES

| | |
|---|---|
| I WORLDWIDE TENLEY<br>16 KBPS<br>Modified Autovon CONUS<br>TTC-39 Overseas<br>Predominently Wideband<br>Tri-Tac Type COMSEC | II NARROWBAND<br>8 KBPS<br>Modified Autovon Worldwide<br>Bellfield COMSEC<br>Red/Maroon Interface with Tri-Tac |
| III WIDEBAND<br>16 KBPS<br>Modified Autovon CONUS<br>TTC-39 Overseas<br>Bellfield COMSEC CONUS<br>Tri-Tac COMSEC Overseas | IV HYBRID<br>16 KBPS Overseas<br>8 KBPS Conus<br>Modified Autovon CONUS<br>TTC-39 Overseas<br>Bellfield COMSEC CONUS<br>Tri-Tac COMSEC Overseas |

The AN/TTC-39 is a family of modular and transportable communication switching systemsdesigned to provide secure automatic switching for tactical voice and message traffic. The family consists of hybrid circuit switches varying in size from 450 to 750 terminations by increments of 150 analog or digital terminations and message switches equipped for 25 or 50 terminations.

A more detailed description of the four alternatives are given below:

A. ALTERNATIVE I. The Worldwide Tenley provides for 16 KBPS (Wideband) continuously variable slope delta modulation (CVSD) terminals for all users. Clear, secure voice capability will be provided from the same 16 KBPS terminal. Leased CONUS autovon switches will be modified to emulate certain AN/TTC-39 switch features and the government owned switches overseas will be replaced with AN/TTC-39 type switches. Concentrations of subscribers will be provided access via a new automatic 4-wire Digital Access Exchange (DAX) concentrator. End-to-end encryption will be provided for all calls within the network, except for conferencing and NB/WB conversions requiring red interfaces. Automatic remote electronic cryptographic key distribution will be provided with the Tri-Tac Tenley COMSEC concept in both CONUS and overseas.

B. ALTERNATIVE 2. Alternative 2 provides 8 KBPS Narrowband voice processor terminals and Bellfield COMSEC in both CONUS and overseas portions of the DCS. CONUS Autovon switches will be modified for digital operation and Bernhardt KDC's will be used in CONUS and overseas. A red interface KDC will be required for the DCS Bellfield COMSEC to interoperate with the Tri-Tac Tenley COMSEC. Overseas, the existing government-owned autovon switches will be modified for digital operation. End-to-end voice encryption will be maintained on intra-DCS calls since all users will have compatible terminals. However, calls to Tri-Tac will require 8 to 32 KBPS voice interfaces that will prohibit end-to-end encryption and insert voice degradation.

C. ALTERNATIVE 3. Alternative 3 provides a Worldwide Wideband (16 KBPS) system using Bellfield COMSEC in CONUS and Tenley COMSEC overseas. As opposed to the Tenley alternative, it will not have COMSEC functions at each modified CONUS autovon switch. Instead, up to 3 stand-alone Bernhardt KDC's will be dispersed throughout CONUS to serve the CONUS DCS. CONUS autovon switches will be modified to provide digital service. The CONUS voice terminal will be procured to operate in the Bellfield COMSEC mode. Overseas, this alternative will require a special interface KDC to allow interoperation of the CONUS Bellfield and the overseas Tenley key distribution systems. Voice interoperability with Tri-Tac subscribers and end-to-end encryption will be available.

D. ALTERNATIVE 4. Alternative 4, the Hybrid alternative, provides 8 KBPS Narrowband Voice Terminals with Bellfield COMSEC in CONUS and 16 KBPS voice terminals with Tenley COMSEC overseas. The Bellfield COMSEC in CONUS will be achieved with Bernhardt KDC's. The CONUS secure voice terminals will be the product of a separate Narrowband development. CONUS autovon switches will be modified for digital operation. Overseas, the program will be identical to the Wideband alternative, except that an interface will be required between the two dissimilar voice terminals of each geographic area. This will preclude end-to-end encryption of voice calls between CONUS and overseas DCS or CONUS DCS and Tri-Tac, and will introduce noticeable voice degradation for these calls.

3. METHODOLOGY. The methodologies used in this analysis included cost estimating relationships, regression analysis, learning curve, engineering estimates, analogy, delphi, cost factors, complexity factors, contractor quotes, previous experience, and subjective judgement.

The methodology employed for the investment portion of the ICE consisted of the formulation of the equipment requirements package, research of available cost data, determination of hardware costs by analogy and support costs from historical information and cost estimating guidelines. The cost data elements of the investment analysis include hardware, military construction, engineering, installation and testing, material, initial spares, test equipment, data, training, packing, packaging and transportation.

The operation and support methodology consisted of cost estimating relationships, computer models, expert opinion, analogy, contractor quotes, cost factors, and exponential regression analysis. The cost data elements of the O&S analysis consisted of personnel, consumption, training, integrated logistics support, maintenance, procurement of switch modification, transportation, recurring spares, leasing and utilities.

The methodology for the R&D cost estimates were expert opinion. 64 individual R&D tasks were analyzed using a modified delphi technique and a computer routine. The cost data elements of the R&D analysis included engineering, tooling and prototypes.

A. As an example of the mathematical techniques used in estimating costs, an analysis of CONUS transmission costs is given. These AT&T leased lines will be used for digital rather than the usual analog transmission; thus there was no relevant experience to obtain data.

Two factors were involved in the analysis, the first of which was the increase in the number of digital service areas expected. This is expected to result in a linear decrease in total transmission costs of 2%/year for 10 years. The second factor anticipates a reduction in costs for providing digital transmission due to technological advances and increased equipment production. This decrease is expected to start in 1980 and is expressed by the exponential regression,

$$DC = 1/2 \quad (1 + e^{-t/3})$$

Where $0 \leq t \leq 10$ corresponds to the years 1980 to 1990. This expression results from an exponential regression analysis using all available information on present and past transmission leasing costs.

B. The approach to estimating Operating and Support (O&S) costs was as follows. Operator costs were calculated by multiplying the number of operators required for each equipment by the annual pay and allowance for the operator's grade level.

Maintenance costs were calculated by multiplying the cost per active maintenance man-hour by the total annual maintenance hours per equipment. Total annual maintenance man-hours were calculated by

$$AMMH = HOP\ (MTTR/MTBF),$$

where: AMMH = Annual Maintenance Man-hours.
HOP = Hours of Operation Per Year
MTTR = Mean-time-to Repair
MTBF - Mean-time-between Failure

Depot overhaul costs for labor and material were calculated by multiplying the depot overhaul cost by the overhaul rate to equal the depot cost per unit per year. The overhaul rate indicated how often the unit was expected to be sent back to depot for overhual. The depot overhaul cost was estimated by

$$DOC = 0.809\ (DOR)\ (UC)^{.881}$$

where: DOC = Depot Overhaul Cost/Year
UC = Unit Hardware Cost
DOR = Depot Overhaul Rate
Standard Error = +60%, -37%

C. Cost Estimating Relationships (CER) were used to estimate costs for various equipments. For example, the CER used for the TTC automatic switching equipment was

$$Y_2 = 27284.7 + 0.002\ X_1^2 - 0.125X_2^{1.7} + 24.898X_3^{1.5}$$

where: $Y_2$ = Acquisition Cost

$X_1$ = Weight

$X_2$ = Volume

$X_3$ = Number of Lines

4. UNCERTAINTY ANALYSIS. In all cases of projected cost estimates some degree of uncertainty will exist and it is therefore advisable to state projected cost estimates in terms of most likely value, lowest value, and the most pessimistic (highest) value. The most likely value would be that value normally used in planning, programming and budgeting.

The ratios of high and low values to most likely (taken as 1) are given in Table 2 below for the preferred alternative 1 for R&D and O&S costs.

TABLE 2 - UNCERTAINTY ANALYSIS (ALTERNATIVE 1)

| | LOW | MOST LIKELY | HIGH |
|---|---|---|---|
| R&D | .957 | 1 | 1.024 |
| O&S | .850 | 1 | 2.054 |

The uncertainty in the investment costs was analyzed for the major equipments. The uncertainties are given in terms of percentages of the most likely costs.

TABLE 3 - INVESTMENT UNCERTAINTY ANALYSIS

| EQUIPMENT | UNIT COST | - | + |
|---|---|---|---|
| AN/TTC-39 Switch | $1,860K | 18% | 8% |
| Tenley Family | 298K | 15% | 15% |
| Bellfield Family | 292K | 25% | 25% |
| Loran C | 20K | 5% | .5% |
| Dax (Concentrator) | 58K | 15% | 40% |
| DSVT | 4.3K | 10% | 10% |
| Goldwine Mod | 10.9K | 60% | 10% |
| Conference Directors | 273K | 50% | 50% |
| Transmission Equipment | ----- | 15% | 15% |
| SERVICE AND SUPPORT | | | |
| Engineer, Install and Test | | 10% | 100% |
| Repair Parts | | 15% | 15% |
| Test Equipment | | 20% | 20% |
| Data | | 20% | 50% |
| Packing, Packaging & Transportation | | 15% | 25% |

5. SENSITIVITY ANALYSIS. Cost sentivity analysis is a technique within the context of both individual system and force structure cost analysis. It involves the systematic examination of the effects of changes in total force structure cost resulting from variations in characteristics, size, and composition of force. The variables considered in conducting the sensitivity analysis were the number of subscribers, manning levels, changes in terminals, logistics cost, CONUS, leasing costs, and planning horizons.

6. COST BENEFIT ANALYSIS. By using standard methods of measuring benefits (measures of effectiveness), benefit/cost ratios were calculated for the 4 alternatives. The values are given below:

TABLE 4 - COST BENEFIT ANALYSIS

| Alternative | Benefit/Cost |
|---|---|
| Tenley | 484 |
| Narrowband | 305 |
| Wideband | 378 |
| Hybrid | 296 |

7. SUMMARY COSTS. The table below gives the summary costs in both constant and inflated FY76 dollars. The inflated costs of over a billion dollars is a large but not untypical program for our analysis and evaluation.

TABLE 5 - ICE P2SV GENERAL COST SUMMARY

(Constant 76 $ M)

| ALTERNATIVES | I | II | III | IV |
|---|---|---|---|---|
| R&D | 37.6 | 39.3 | 37.6 | 38.8 |
| Investment | 179.8 | 209.3 | 173.1 | 248.8 |
| O&S | 569.7 | 495.2 | 555.5 | 570.4 |
| TOTAL | 787.1 | 743.8 | 766.2 | 858.0 |

(Inflated 76 $ M)

| | | | | |
|---|---|---|---|---|
| R&D | 44.2 | 45.6 | 44.2 | 45.7 |
| Investment | 235.9 | 279.3 | 230.2 | 329.2 |
| O&S | 1026.2 | 882.3 | 1000.9 | 1017.1 |
| TOTAL | 1306.3 | 1207.2 | 1275.3 | 1392.0 |

8. CONCLUSION. In this presentation I have attempted to give the highlights of the Army's independent analysis of the P2SVP, as well as some of the complementary calculations used in an economic analysis that are needed in the decision acquisition process.

# SOLVING CONTROL PROBLEMS USING DISCRETE CONTROLS

Randy J. Schuetz

Army Materiel Systems Analysis Activity
Attn: DRXSY-RW
Aberdeen Proving Ground, Md. 21005
Formerly, Intern Training Center, DARCOM

Bart Childs

Department of Industrial Engineering
Texas A&M University
Texarkana, Texas 75501

ABSTRACT. The solution of the class of problems governed by a set of first order linear differential equations, subject to a set of linear constraints, and the minimization of a defined quadratic performance index is presented. The number of differential equations must be greater than the number of constraints, otherwise, there is a unique solution and control is not possible. The solution is considered as known once the correct initial conditions are found; a number of initial value methods are available to solve linear differential equations. Only discrete controls are considered here, depicting the real world where continuously variable controls are not always present. Using the above, systems of the open loop type are examined.

The method consists of superposition of linearly independent particular solutions to get the optimal solution. The particular solutions are generated using a power series integration technique on a perturbed set of arbitrarily chosen initial conditions. The superposition constants are determined so that the solution both meets the constraints and minimizes the quadratic performance index. The minimum point is found using a method developed by Childs and Maron for the explicit minimum solution to a set of quadratic equations subject to a set of linear constraints.

1. INTRODUCTION. The solution to a class of control problems with discrete controls is presented. The class of problems examined are those governed by a set of $n$ first order linear ordinary differential equations, subject to a set of $m$ linear constraints $(m<n)$, wherein a given quadratic performance index is to be minimized. The discrete controls appear in the solution as initial values of the differential equations. Those initial values which are unspecified by constraints are determined optimally by minimizing the given quadratic performance index.

The letter $y$ is used for an $n$ element state variable vector which is assumed to be a function of the independent variable $t$, time. The dot $(\dot{})$ is used to denote the total derivative with respect to $t$. The general set of first order linear ordinary differential equations is written as

$$\dot{y} = Ly + f \qquad t \in [o,T] \tag{1.1}$$

where $L$ is a $n$ by $n$ coefficient matrix whose elements may be constants or functions of time, $f$ is an $n$ element vector of forcing functions, and $[o,T]$ is the time interval of interest. The solution of equation (1,1) is subject to the linear equality boundary conditions or constraints

$$q_i\ (y(t_i)) = b_i \qquad i = 1,2,\ldots,\ m<n \tag{1.2}$$

where $q_i$ represents the boundary condition operator that specifies a linear combination of elements of the state vector. The $i$th boundary value, $b_i$, at the specified value of time, $t_i$. A quadratic performance index, $h$, where

$$h = \int_o^T y'\, M\, y\, dt \tag{1.3}$$

is to be minimized. The $n$ by $n$ matrix $M$ is symmetric and known function of time, $t$. The prime ( )' is used to indicate the transpose of a vector or matrix. The above three equations define the basic problem.

The solution to the problem is uniquely defined once the state, $y$, is known at any time $t$. The solution is considered as known once the correct initial conditions, $y(o)$, are known. The $y(o)$ vector gives the desired control parameters, and it can be used with the differential equation (1.1) to generate an accurate solution for $y$ as a function of $t$. This is due to the availability of a variety of initial value differential equation problem solvers for today's digital computers.

The solution method is a superposition of solutions, a "shooting method". [6]

The usual methods of solving similar controls problems involve the use of Lagrange multipliers, Hamiltonians, co-state equations, etc. which are unnecessary in the method presented in this paper. [5] The techniques used in the usual methods require a large amount of mathematical gymnastics in the solution process.

2. A SHOOTING METHOD. A particular solution of equation (1.1) is a solution of a *particular* set of initial conditions. We define such a solution as

$$\dot{p}^{(k)} = L\, p^{(k)} + f \tag{2.1}$$

where the superscript, $k$, is an index which denotes the $k$th particular solution. The state vector, $y$, is determined by the superposition of particular solutions, and is expressed as

$$y = Pa \tag{2.2}$$

where the $k$th column of the matrix $P$ is the state vector $p^{(k)}$ of equation (2.1) and the vector $a$ is the vector of superposition constants. The index, $k$, for the vector $a$ and the columns of $P$ varies from *zero* to $r$, where $r$ is the number of differential equations minus the number of known initial conditions. Equation (2.2) can be rewritten as

$$y = \sum_{k=0}^{r} p^{(k)} a_k \tag{2.3}$$

If equation (2.1) is multiplied by $a_k$ and summed over $k$

$$\sum_{k=0}^{r} \dot{p}^{(k)} a_k = \sum_{k=0}^{r} L\, p^{(k)} a_k + \sum_{k=0}^{r} f\, a_k \tag{2.4}$$

Rewriting after factoring out $L$ and $f$ from the summations (since they are not indexed by k) and substituting equation (2.3) and the derivative of equation (2.3) with respect to $t$ into equation (2.4) gives

$$\dot{y} = L\, y + f \sum_{k=0}^{r} a_k \tag{2.5}$$

Comparing equations (1.1) and (2.5) establishes a constraint which the superposition constants must meet:

$$\sum_{k=0}^{r} a_k = 1 \tag{2.6}$$

The traditional superposition of homogeneous solutions on a single particular solution does not have a similar constraint. However, because we superimpose particular solutions, we need to program only one set of equations for each problem.

*Independence of Solutions and Boundary Value Constraints.* The reason for the

superposition of the particular solutions is to satisfy the boundary conditions or constraints. This requires all $r+1$ subsets of the $r$ particular solutions to be linearly independent. To insure this, $P(o)$ is created using the perturbation strategy: *First*, arbitrary estimates are made of the $r$ unknown values of $y(o)$ and this vector is used for $p^{(o)}(o)$. *Second*, columns 1 through $r$ of are generated by making each column the same as $p^{(o)}(o)$, except that each has one *nonzero* perturbation from one of the estimated elements of $p^{(o)}(o)$. Each estimated element is perturbed in one and only one column. This strategy gives the desired independence.

The boundary conditions are of the form specified in equation (1.2). For control problems, these boundary conditions are usually initial conditions, but this is not required. As stated previously, $r$ denotes the number of elements of $y(o)$ not uniquely specified by equation (1.2), and thus, $(n-r)$ elements of $y(o)$ are uniquely specified. If $m$ is not equal to $(n-r)$, then there are $m-(n-r)$ boundary conditions at times greater than zero. Subsittution of (2.2) into (1.2) gives

$$q_i\,(P(t_i)\,a) = b_i \qquad i = 1,2,\ldots,m \tag{2.7}$$

which can be rewritten for linear operators $q_i$, as

$$\sum_{k=0}^{r} q_i\,(p^{(k)}\,(t_i))\,a_k = b_i \qquad i = 1,2,\ldots,m \tag{2.8}$$

Of these $m$ linear equations, $(n-r)$ specify known values of $y(o)$ and $m-(n-r)$ specify constraints on the unknown values of $y(o)$ in terms of the $(r+1)$ unknown superposition constants, the $a_k$'s. With the addition of constraint equation (2.6), there are $m-(n-r)+1$ constraints with $(r+1)$ unknowns. Since the problem statement declares that $m$ is less than $n$, it is evident that $m-(n-r)+1$

is less than $r+1$, and thus it is an underdetermined system. Therefore, the $a_k$'s are not uniquely specified, and we can choose them to minimize the performance index of equation (1.3).

*Optimizing on the Basis of the Quadratic Performance Index.* The $a$ vector is now included in the performance index by the substitution of equation (2.2) into equation (1.3) which gives

$$h = \int_o^T a' P' M P a \, dt \tag{2.9}$$

The $(r+1)$ by $(r+1)$ matrix $A$ is defined by:

$$A = \int_o^T P' M P \, dt \tag{2.10}$$

It is possible to rewrite equation (2.9) as

$$h(a) = a' A a \tag{2.11}$$

The method that is used to solve for $A$ is to calculate the solution of the initial value problem

$$A = P' M P \qquad A(o) = 0 \tag{2.12}$$

The superposition equation (2.3) requires that $(r+1)*n$ first order linear ordinary differential equations be integrated and the matrix $A$, which is symmetric because the matrix $M$ is symmetric, may be determined by integrating an additional $(r+1) * (r+2)/2$ first order linear ordinary differential equations.

In solving for the optimum $a$ vector, the explicit formula developed by Childs and Maron (1975) is utilized. This formula states that the solution for $a$ such that

$$h(a) = a' A a = minimum \tag{2.13}$$

subject to

$$Ka = c \tag{2.14}$$

is

$$a = a_p - N(N' AN)^{-1} N' A a_p \tag{2.15}$$

where $k$ is $m-(n-r)+1$ by $(r+1)$ and of rank $m-(n-r)+1$, $a_p$ is a particular solution of equation (2.14), and the columns of $N$ form a basis for the null space of $K$. The $(\ )^{-1}$ in equation (2.15) denotes a matrix inverse. By using appropriate matrix operations, it is possible to transform equation (2.14) into the equivalent system

$$\left[\begin{array}{c|c} I & W \end{array}\right] \quad a = d \tag{2.16}$$

This can be used to define $a_p$ and $N$ as

$$a_p = \left[\begin{array}{c} d \\ \hline o \end{array}\right] \qquad \text{and} \qquad N = \left[\begin{array}{c} -W \\ \hline I \end{array}\right] \tag{2.17}$$

The $I$'s in equations (2.16) and (2.17) are identity matrices of appropriate order.

3. AN EXAMPLE. The first problem chosen is

$$\ddot{x} + 0.2\dot{x} + x = u_1 + u_2 t \qquad t \in [0,10] \tag{3.1}$$

subject

$$x(o) = 0 \qquad x(10) = 1 \qquad \dot{x}(o) = 0 \tag{3.2}$$

and

$$h = \int_0^{10} (x^2 + \dot{x}^2)\, dt = minimum \tag{3.3}$$

In state variable form, this can be restated as

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= -y_1 - 0.2y_2 + y_3 + y_4 t \\ \dot{y}_3 &= 0 \\ \dot{y}_4 &= 0 \end{aligned} \tag{3.4}$$

subject to

$$y_1(o) = 0 \qquad y_1(10) = 1 \qquad y_2(o) = 0 \tag{3.5}$$

where

$$h = \int_0^{10} y' M y\, dt \tag{3.6}$$

and

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{3.7}$$

Using an accuracy of $10^{-6}$ and evaluating power series to 10 terms results in the following solution

$$\begin{aligned} y_1(o) &= 0. \\ y_2(o) &= 0. \\ y_3(o) &= -0.105453 \\ y_4(o) &\quad 0.115041 \end{aligned}$$

The $y_3$ and $y_4$ elements are the forcing function constants or control. The solution for $y_1$ and $y_2$ over the interval [0,10] is given in Table 1.

4. CONCLUSIONS. A direct method has been shown for the solution of linear ordinary differential equations subject to minimization of a quadratic performance index and multipoint boundary values. The method avoids the necessity of Lagrange multipliers and other similar tools.

The method can easily be incorporated into boundary value codes. Most problems will have nonlinearities which can be handled in the usual manner [3], [6].

## LIST OF REFERENCES


1. Childs, B. and Porter, H.R., Numerical Solution of Non-Linear Multipoint Boundary Value Problems, in preparation for Applied Mathematics and Computations Series for Addison Wesley.

2. Childs, B. and Maron, M.J., "An Explicit Formula for the Optimum Point of a Quadratic Subject to Linear Equality Constants", Abstract in Notices of the American Mathematical Society, Vol. 21, No. 6, October 1974.

3. Doiron, H., "An Indirect Optimization Method with Improved Convergence Characteristics", a Dissertation presented to the Faculty of the Cullen College of Engineering, University of Houston, in partial fulfillment of the requirements for the degree Doctor of Philosophy, also available as NASA TM X-58088, May 1970.

4. Felberg, E., "Numerical Integration of Differential Equations by Power Series Expansions, Illustrated by Physical Examples", NASA TN No. TN D-2356, October 1964.

5. Lapidus, L. and Luus, R., Optimal Control of Engineering Processes, Blaisdell Publishing Company, Waltham, Massachusetts, 1967.

6. Roberts, S.M. and Shipman, J.S., Modern Analytic and Computational Methods in Science and Mathematics, American Elsevier Publishing Company, Inc., New York, 1972.

TABLE 1

NUMERICAL SOLUTIONS

| Time | $y_1$ | $y_2$ |
|---|---|---|
| 0 | 0 | 0 |
| 0.5 | -0.010 | -0.035 |
| 1.0 | -0.028 | -0.031 |
| 1.5 | -0.035 | 0.006 |
| 2.0 | -0.018 | 0.066 |
| 2.5 | 0.031 | 0.130 |
| 3.0 | 0.111 | 0.186 |
| 3.5 | 0.213 | 0.219 |
| 4.0 | 0.325 | 0.225 |
| 4.5 | 0.434 | 0.205 |
| 5.0 | 0.527 | 0.166 |
| 5.5 | 0.598 | 0.118 |
| 6.0 | 0.646 | 0.075 |
| 6.5 | 0.675 | 0.045 |
| 7.0 | 0.694 | 0.034 |
| 7.5 | 0.713 | 0.043 |
| 8.0 | 0.740 | 0.068 |
| 8.5 | 0.782 | 0.102 |
| 9.0 | 0.841 | 0.135 |
| 9.5 | 0.916 | 0.161 |
| 10.0 | 1.000 | 0.173 |

$y_3 = -0.105$

$y_4 = +0.115$ (constants)

# On Generalized Feller Equation

Siegfried H. Lehnigk
Physical Sciences Directorate, US Army Missile Command
Redstone Arsenal, AL

## ABSTRACT

The generalized Feller equation

$$\ell(z) = Az_{xx} + Bz_x + C_z - z_t = 0, \quad z = z(x,t), \quad x > 0, \quad t > 0,$$

with the coefficients

$$A(x) = \alpha x^{\lambda+1}, \quad \alpha > 0, \quad \lambda \epsilon R, \quad \lambda \neq 1,$$

$$B(x) = \beta_1 x^{\lambda} + \beta_2 x, \quad \beta_{1,2} \epsilon R,$$

$$C(x) = \rho x^{\lambda-1} + \beta_2, \quad \rho = \lambda[\beta_1 - \alpha(1+\lambda)],$$

will be considered. The choice of $\rho$ makes $\ell(z) = 0$ a Fokker-Planck equation.

Solutions of $\ell(z) = 0$ will be derived for given initial and/or boundary conditions. The derivation of initial condition solutions is based on a basic solution of $\ell(z) = 0$ and its adjoint.

---

The complete paper is published elsewhere.

# A PERTURBATION METHOD FOR FREE BOUNDARY PROBLEMS OF ELLIPTIC TYPE*

B. A. Fleishman and Thomas J. Mahar†
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, New York 12181

ABSTRACT. Nonlinear partial differential equations (PDE's) arise in many scientific contexts, and boundary value problems (BVP's) for such equations present formidable computational difficulties. Thus analytical techniques for approximating the solutions of such problems have practical significance.

A formal perturbation method is described here for approximating solutions of certain BVP's for elliptic PDE's containing discontinuous nonlinearities. To illustrate, we treat in detail the BVP

$$P(\varepsilon)\begin{cases} u_{xx} + u_{yy} + f(u) = 0 & \text{in } S: 0 < x < 1,\ -\infty < y < \infty \\ u(0,y) = \varepsilon h(y),\ u_x(1,y) = 0 & \text{for } -\infty < y < \infty \end{cases}$$

where $\varepsilon$ is a small parameter, h is periodic and uniformly bounded, and f is a step-function: $f(u) = 0$ for $u < \mu$, $f(u) = 1$ for $u \geq \mu$ ($\mu$ a positive constant). u and $\partial u/\partial n$ are to be continuous across any "free boundary" $u = \mu$. If $0 < \mu < 1/4$, problem P(0) is shown to have at least one non-trivial solution $u_0 = u_0(x)$ such that $u_0(\bar{x}) = \mu$ $(0 < \bar{x} < 1)$. For $\mu \in (0,1/4)$ an approximate solution $u(x,y)$ of $P(\varepsilon)$ involving a free boundary in S is then sought in the form $u(x,y) = u_0(x) + \varepsilon\tilde{u}(x,y)$, with the free boundary assumed to be $x = \bar{x} + \varepsilon g(y)$.

Two examples are considered, $h(y) = \cos y$ and h a trigonometric polynomial, in which the linear (variational) equation for $\hat{u}$ may be solved by separation of variables.

An unusual feature of our procedure is that this equation for $\tilde{u}$ contains a delta-function coefficient, because in the originial equation f is a step-function in u.

1. INTRODUCTION. Nonlinear partial differential equations (PDE's) arise in many scientific contexts, and boundary value

* Research supported by U. S. Army Research Office.

† Present address: Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, New York 10012 (U.S.A.)

problems (BVP's) for such equations present formidable computational difficulties. Thus analytical techniques for approximating the solutions of such problems have practical significance.

We illustrate here a perturbation method applicable to certain BVP's for elliptic PDE's of the form

$$\Delta u + f(x,u) = 0 \qquad (1)$$

where $x = (x_1,\ldots,x_n)$ is a point in $R^n$, $\Delta$ denotes the Laplacian operator, u is a real scalar variable, and f is a piecewise-continuous function of $x_1,\ldots,x_n$ and u. When f has jump discontinuities with respect to u, among the interfaces across which f changes abruptly there may be so-called "free boundaries" which are not known a priori but must be found along with the solution $u = u(x)$.

Suppose f is a step-function in u and depends also on m of the independent variables, say, $x_1,\ldots,x_m$, where $0 \leq m < n$. Let D be a fixed region in $R^n$ whose bounding surfaces are independent of $x_{m+1},\ldots,x_n$.

Now consider a BVP for (1) on D, denoted by $P(\varepsilon)$, in which a small parameter $\varepsilon$ occurs in the boundary conditions in such a way that the "reduced" problem P(0) does not involve $x_{m+1},\ldots,x_n$. If a solution $u_0 = u_0(x_1,\ldots,x_m)$ of P(0) is obtainable, we seek a solution of $P(\varepsilon)$ in the perturbed form $u = u_0 + \varepsilon\tilde{u}$, with free boundaries (if any) which are perturbations of free boundaries of P(0). As we shall see in the specific problem considered below, for certain boundary data it is easy to find $\tilde{u}$ and the perturbed free boundary.

The unusual mathematical feature of this procedure is that we perturb about a surface of discontinuity, which introduces a delta-function into the (variational) equation satisfied by $\tilde{u}$. Our development is formal; assuming that the solution we seek exists and that it can be closely approximated by an expression of the form $u_0 + \varepsilon\tilde{u}$, etc., we calculate $\tilde{u}$ and the modified free boundary.

Free boundary problems for equations similar to (1) occur in plasma physics; in [1], for example, the authors consider equations of the form $Lu + f(x,u) = 0$, where L is an elliptic operator and f is, however, piecewise-linear in u, not discontinuous. Free boundary problems for equations of the form div (K grad u) = 0, where $K = K(x,u)$ is a piecewise-continuous function (which arise in the equilibrium Stefan problem [2] and govern certain diffusion and metallurgical processes) are also being investigated by the method illustrated here.

Besides occurring naturally, problems with discontinuous nonlinearities are sometimes introduced as approximations (e.g., see [3]) to problems with smooth nonlinearities (which, in general,

can not be solved explicitly). The authors are investigating the feasibility of deriving approximate solutions of BVP's for equations of type (1) in which f is bounded and has smooth dependence on u, by first replacing the smooth function f with one which is a step-function in u, then employing the procedure described here to treat the approximating problem. In this connection it is important to note that if the perturbation procedure is applied directly to an equation of the form (1) containing a smooth nonlinearity f, the variational equation (to be solved for $\tilde{u}$) will always have variable coefficients.

The remainder of this paper (Sections 2, 3, 4 and 5) is devoted to applying the perturbation technique to the particular BVP consisting of equations (2) and (3) below.

2. A PARTICULAR FREE BOUNDARY PROBLEM. Let us denote by $P(\varepsilon)$ the following two-dimensional BVP for a nonlinear PDE in the vertical strip

$$S = \{(x,y): 0 < x < 1,\ -\infty < y < \infty\}:$$

$$P(\varepsilon)\ \begin{cases} \Delta u + f(u) = 0 & \text{in } S \qquad (2) \\ u(0,y) = \varepsilon h(y)\ ,\ u_x(1,y) = 0 & (-\infty < y < \infty) \qquad (3) \end{cases}$$

Here $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, h is a given continuous, bounded, periodic function, $\varepsilon \geq 0$ is a (small) constant, and f is a step-function with given threshold value $\mu \geq 0$:

$$f(u) = \begin{cases} 0 & u < \mu \\ 1 & u \geq \mu \end{cases}$$

(We could also write $f(u) = H(u - \mu)$, where H is the Heaviside unit function.)

Solutions of $P(\varepsilon)$ will be required to be periodic and $C^1$ (therefore bounded) in the closure of S. In particular, then, u and its normal derivative $\partial u/\partial n$ must be continuous across any free boundary (not known a priori), where $u = \mu$.

Suppose that h is bounded by 1, also that $0 \leq \varepsilon < \mu$. Then by continuity, $u < \mu$ at points of S close to the left boundary $x = 0$. If $u < \mu$ throughout S, $f \equiv 0$ in S and $P(\varepsilon)$ is a (linear) BVP for Laplace's equation. For solutions satisfying $u > \mu$ somewhere in S, however, $P(\varepsilon)$ is not linear. Analysis of the "reduced" problem P(0) (see Section 3) suggests that for small positive $\mu$, $P(\varepsilon)$ possesses solutions of both the linear and nonlinear problems.

The nonlinear case of $P(\varepsilon)$ is of interest here. We seek an approximate solution in the form $u = u_0 + \varepsilon\tilde{u}$, where $u_0$ is a (known) solution of the (one-dimensional) nonlinear problem P(0); likewise free boundaries in $P(\varepsilon)$ are assumed to be perturbations of the free boundaries in P(0). In Section 3 we obtain the solution(s) of P(0) for all non-negative values of $\mu$; in par-

ticular, it is shown that when $0 \leq \mu \leq \frac{1}{4}$ there are non-trivial solutions.

In Section 4 we perturb the PDE (2) about $u_0$ and obtain the (linear) variational equation for $\tilde{u}$. In Section 5, taking $h(y) = \cos y$, we solve the BVP for $\tilde{u}$ by separation of variables. The free boundary is determined by substituting for x, in the interface condition

$$u(x,y) = u_0(x) + \varepsilon\tilde{u}(x,y) = \mu \ ,$$

the assumed form $x = \bar{x} + \varepsilon g(y)$, where $u_0(\bar{x}) = \mu$ (that is, $x = \bar{x}$ is the interface for the reduced problem) and g is the periodic function which we must find. Also for more general boundary data (namely, h a trigonometric polynomial) the variables can be separated; in this case we merely sketch the procedure.

3. ANALYSIS OF P(0). When $\varepsilon = 0$, the boundary conditions (3) are both independent of y; thus P(0) reduces to the following one-dimensional problem:

$$P(0) \quad \begin{cases} u'' + f(u) = 0 & \text{in I: } 0 < x < 1 \\ u(0) = 0, \quad u'(1) = 0 \end{cases} \qquad (4)$$

where $' = d/dx$. We shall find all $C^1$ solutions for $\mu \geq 0$.

Note first that all solutions are non-negative, because $u(0) = 0$ and $u'(x) \geq 0$ on I. The latter follows from the facts that $u'' = -f(u) \leq 0$ (wherever u" exists) and $u'(1) = 0$.

When $\mu = 0$, (4) takes the form $u'' = -1$ on I. Then P(0) has the unique solution $u(x) = x - x^2/2$.

For any fixed $\mu > 0$, P(0) has the trivial solution $u(x) \equiv 0$. In order for a non-trivial solution to exist, it is necessary that there be a smallest value $\bar{x}$ in I such that $u(\bar{x}) = \mu$. Then $u(x) < \mu$ for $0 \leq x < \bar{x}$ and (since $u'(x) \geq 0$) $u(x) \geq \mu$ for $\bar{x} \leq x \leq 1$. Therefore a non-trivial solution of P(0) must satisfy

$$\begin{aligned} u'' &= 0 \qquad \text{for} \quad 0 < x < \bar{x} \\ u'' + 1 &= 0 \qquad \text{for} \quad \bar{x} < x < 1 \end{aligned} \qquad (5)$$

plus the boundary and continuity conditions

$$\begin{aligned} &u(0) = 0 \ , \qquad u'(1) = 0 \\ &u(\bar{x}+) = u(\bar{x}-) = \mu, \ u'(\bar{x}+) = u'(\bar{x}-) \end{aligned} \qquad (6)$$

for some $\bar{x}$ in I.

Solving the differential equations (5) on their respective intervals, then subjecting them to conditions (6), we find that for $\mu > 0$,

$$u_0(x) = \begin{cases} (1 - \bar{x})x & (0 \le x \le \bar{x}) \\ x - \frac{1}{2}(x^2 + \bar{x}^2) & (\bar{x} \le x \le 1) \end{cases} \tag{7}$$

is a solution of P(0) provided $\bar{x}$ $(0 < \bar{x} < 1)$ satisfies

$$\bar{x}(1 - \bar{x}) = \mu \ . \tag{8}$$

This quadratic equation has distinct real roots $\bar{x}$ in I when $0 < \mu < 1/4$, the double root $\bar{x} = 1/2$ when $\mu = 1/4$, and complex roots when $\mu > 1/4$.

We can now describe the numbers and types of $C^1$ solutions of P(0) for all non-negative values of $\mu$:

$\mu = 0$: Unique solution: $u(x) = x - x^2/2$.

$0 < \mu < 1/4$: Three solutions: the trivial one plus two solutions given by (7), each corresponding to a different root of (8).

$\mu = 1/4$: Two solutions: the trivial one plus one given by (7) when $\bar{x} = 1/2$.

$\mu > 1/4$: Unique solution: $u(x) \equiv 0$.

4. THE PERTURBATION PROCEDURE. Henceforth our attention is restricted to values of $\mu \in (0,1/4)$.

As seen in Section 3, for each such $\mu$, P(0) possesses two non-trivial solutions in addition to the trivial one. Focussing on the nonlinear case, we have reason to expect (see [4]) that there exists a solution of $P(\varepsilon)$ close to at least one of the non-trivial solutions $u_0(x)$ of P(0).

For fixed $\mu \in (0,1/4)$, let $u_0(x)$ be the solution (7) of $P(\varepsilon)$ corresponding, say, to the smaller root of equation (8); thus, $0 < \bar{x} < 1/2$. (The formal calculation which follows is the same for either root.) We shall assume that the y-periodic solution of $P(\varepsilon)$ close to this $u_0(x)$ can be written, neglecting terms which are $O(\varepsilon^2)$,

$$u(x,y) \approx u_0(x) + \varepsilon\tilde{u}(x,y) \ , \tag{9}$$

where $\tilde{u}$ is a function periodic in y and uniformly bounded in the closed strip.

Similarly, we assume that the solution (9) has a free boundary which may be represented

$$x \approx \bar{x} + \varepsilon g(y), \tag{10}$$

that is, as a perturbation of the "free boundary" $x = \bar{x}$ in $u_0(x)$.

Subtracting $\Delta u_0 + f(u_0) = 0$ from $\Delta u + f(u) = 0$ and noting that (formally)

$$f(u) = f(u_0 + \varepsilon\tilde{u}) \approx f(u_0) + f'(u_0)\cdot\varepsilon\tilde{u}$$

we obtain the variational equation

$$\Delta\tilde{u} + f'(u_0)\tilde{u} = 0 \ ,$$

or

$$\Delta\tilde{u} + \frac{\delta(x - \bar{x})}{u_0'(x)}\,\tilde{u} = 0 \ , \tag{11}$$

where we have used the identities

$$f'(u_0(x)) = H'[u_0(x) - \mu] = \delta[u_0(x) - \mu] = \delta(x - \bar{x})/u_0'(x) \ .$$

From (3), (9) and $u_0(0) = 0$, $u_0'(1) = 0$ follow the boundary conditions on $\tilde{u}$:

$$\tilde{u}(0,y) = h(y) \ , \quad \frac{\partial\tilde{u}}{\partial x}(1,y) = 0 \qquad (-\infty < y < \infty) \tag{12}$$

In the next section two examples are considered in which h actually varies with y in a periodic fashion. First we can gain some confidence in the validity of the perturbation procedure from consideration of the simple example

$$h(y) \equiv \varepsilon \qquad (0 \leq \varepsilon < \mu) \ .$$

In this example $P(\varepsilon)$ is itself a one-dimensional problem; we are still interested in the nonlinear case. Without giving details we point out that if first one solves $P(\varepsilon)$ exactly (by an analysis similar to that of $P(0)$ in the previous section), then seeks an approximate solution in the form $u = u_0 + \varepsilon\tilde{u}$, with interface $x_\varepsilon = \bar{x} + \varepsilon g$ (by solving the BVP (11 - 12)), one finds that the latter expressions agree with the exact representations for u and $x_\varepsilon$ through terms of first order in $\varepsilon$.

5. EXAMPLES. We give two examples in which the linear BVP (11 - 12) can be solved by separation of variables.

EXAMPLE 1: In $P(\varepsilon)$ let

$$h(y) = \cos y \ .$$

Substitution in (11) and (12) of

$$\tilde{u}(x,y) = v(x)\cos y$$

yields the BVP

$$\begin{aligned} &v'' - v + \frac{\delta(x - \bar{x})}{u_0'(x)}\,v = 0 \qquad (0 < x < 1) \\ &v(0) = 1 \ , \ v'(1) = 0 \end{aligned} \tag{13}$$

The differential equation in (13) implies a jump condition at $x = \bar{x}$. Suppose $v(x)$ is a solution continuous on $[0,1]$. Integrating the equation from $\bar{x} - \eta$ to $\bar{x} + \eta$ ($\eta$ small and positive), then letting $\eta \to 0$, we find that the slope of $v(x)$ undergoes a jump at $x = \bar{x}$:

$$v'(\bar{x}+) - v'(\bar{x}-) = - v(\bar{x})/\lambda \ , \tag{14}$$

where

$$\lambda = u_0'(\bar{x}) = 1 - \bar{x} \ .$$

Now solving $v'' - v = 0$ on each of the intervals $0 < x < \bar{x}$ and $\bar{x} < x < 1$ (so that we have four arbitrary constants), then imposing the boundary conditions from (13), the jump condition (14) and the continuity condition $v(\bar{x}+) = v(\bar{x}-)$, we obtain for BVP (13) the continuous solution

$$v(x) = \begin{cases} \cosh x + A \sinh x & (0 \le x \le \bar{x}) \\ B \cosh (1 - x) & (\bar{x} \le x \le 1) \end{cases} \tag{15}$$

where

$$\begin{aligned} A &= B\left[\frac{1}{\lambda} \cosh \bar{x} \cosh (1 - \bar{x}) - \sinh 1\right] \\ B &= \left[\cosh 1 - \frac{1}{\lambda} \sinh \bar{x} \cosh (1 - \bar{x})\right]^{-1} \end{aligned} \tag{16}$$

We seek the free boundary, for the solution $u(x,y)$ given approximately by (9), as a perturbation of $x = \bar{x}$, the free boundary for $u_0(x)$. In other words, it is assumed that $u = \mu$ along a curve

$$x = \bar{x} + \varepsilon g(y) \ , \tag{17}$$

where $g$ is a periodic function and terms of order $\varepsilon^2$ are neglected.

Substitution of $\bar{x} + \varepsilon g(y)$ for $x$ in

$$u_0(x) + \varepsilon v(x) \cos y = \mu$$

gives

$$\begin{aligned} & u_0(\bar{x} + \varepsilon g(y)) + \varepsilon v(\bar{x} + \varepsilon g(y)) \cos y = \mu \ , \\ & u_0(\bar{x}) + u_0'(\bar{x}) \cdot \varepsilon g(y) + \varepsilon v(\bar{x}) \cos y + O(\varepsilon^2) = \mu, \\ & \varepsilon \lambda g(y) + \varepsilon v(\bar{x}) \cos y \approx 0 \ , \end{aligned} \tag{18}$$

where we have used $u_0(\bar{x}) = \mu$, $u_0'(\bar{x}) = \lambda$, and the fact that while $v$ is not differentiable on $[0,1]$ it is Lipschitzian. Finally from (18)

$$g(y) = - (v(\bar{x})/\lambda) \cos y = - \frac{B}{\lambda} \cosh (1 - \bar{x}) \cos y \ . \tag{19}$$

It should be remarked that $u(x,y) = u_0(x) + \varepsilon v(x)\cos y$, where $u_0$ and $v$ are given by (7) and (15) respectively, is not $C^1$ in S, as required. It is only when we adjust the (free) boundary between the left- and right-hand regions, by wiggling the interface, that we obtain a $C^1$ (approximate) solution.

To sum up, for given $\mu \in (0,1/4)$ and $0 < \varepsilon < \mu$ we have derived, by a formal perturbation scheme, an approximate solution of $P(\varepsilon)$, which is $C^1$ and periodic in y, of the form

$$u(x,y) = \begin{cases} (1 - \bar{x})x + \varepsilon(\cosh x + A\sinh x)\cos y, & 0 \le x \le \bar{x} + g(y) \\ x - \frac{1}{2}(x^2 + \bar{x}^2) + \varepsilon B \cosh(1 - x)\cos y, & \bar{x} + g(y) \le x \le 1 \end{cases}$$

where $\bar{x}$ is the smaller root of (8), while A, B and g(y) are given by (16) and (19) respectively.

It may be shown, finally, that the requirement that $\partial u/\partial n$ be continuous across the interface is satisfied to within terms of order $\varepsilon^2$.

EXAMPLE 2: In $P(\varepsilon)$ let

$$h(y) = a_0 + \sum_{n=1}^{N} (a_n \cos ny + b_n \sin ny)$$

where N is a positive integer. Because the treatment is similar to that of the previous example, we shall only touch on the points of difference.

Again we fix $\mu \in (0,1/4)$, choose the root of (8) satisfying $0 < \bar{x} < 1/2$, and require $0 < \varepsilon < \mu$. To insure $|h(y)| \le 1$, let

$$|a_0| + \sum_{n=1}^{N} (|a_n| + |b_n|) \le 1 \quad .$$

Again assuming the approximate solution of $P(\varepsilon)$ to have the form (9) and the free boundary to have the form (10), we are led to the BVP (11 - 12). Instead of $\tilde{u}(x,y) = v(x)\cos y$, however, we now set

$$\tilde{u}(x,y) = a_0 v_0(x) + \sum_{n=1}^{N} (a_n v_n(x)\cos ny + b_n w_n(x)\sin ny) \quad .$$

Substituting this for $\tilde{u}$ in (11) and (12), then separating variables, we find that for $n = 1,\ldots,N$, both $v_n$ and $w_n$ must be solutions of the BVP

$$v'' + \left[\frac{\delta(x - \bar{x})}{u_0'(x)} - n^2\right] v = 0 \qquad (0 < x < 1)$$

$$v(0) = 1 \quad , \quad v'(1) = 0$$

while $v_0$ must be a solution for $n = 0$.

Proceeding as in the previous example, one can obtain the expressions for $u_0(x) + \varepsilon\tilde{u}(x,y)$ to the left and right of the free boundary, also the approximate representation $x = \bar{x} + \varepsilon g(y)$ for the free boundary itself. But we shall omit the details.

REFERENCES.


1. Cenacchi, G., A. Taroni and A. Sestero, Nuovo Cimento 25B (1975) 279-294.

2. Rubenštein, L. I., The Stefan Problem, Translations of Math. Monographs 27, Amer. Math. Soc., Providence, R. I., 1971.

3. Chandra, J., and B. A. Fleishman, Int. J. Non-Linear Mech. 7 (1972) 207-220.

4. Fleishman, B. A., and T. J. Mahar, to appear in Proc. of the Int. Conf. on Nonlinear Systems and Applications, Univ. of Texas at Arlington, 1976.

# DETERMINATION OF PROPAGATION CONSTANTS IN SCATTERING FROM DIELECTRIC-COATED WIRES

Leon Kotin

Communications/Automatic Data Processing Laboratory

US Army Electronics Command

Fort Monmouth, New Jersey

## ABSTRACT

We determine the propagation constants which describe mathematically the behavior of electromagnetic waves reflected from dielectric-coated wires. These are obtained from the roots of two characteristic equations of transcendental type. The roots are the propagation constants of the creeping waves generated by diffraction of plane waves polarized tangentially and normally to the wire axis, respectively. Their real and imaginary parts give the phase and attenuation of the creeping waves around the circumference of the wire.

# DETERMINATION OF PROPAGATION CONSTANTS IN SCATTERING FROM DIELECTRIC-COATED WIRES

Leon Kotin
Communications/Automatic Data Processing Laboratory
U. S. Army Electronics Command, Fort Monmouth, New Jersey 07703

1. Introduction. The effectiveness of many communication systems can be seriously diminished by reflections of electromagnetic signals from obstacles, both natural and man-made. Dielectric-coated wires constitute a man-made obstacle which appears with increasing frequency in military situations. Nor is this obstacle restricted to communications effects. The U. S. A. Board of Aviation Accident Research recently cited the following statistics for a four-year period of daylight operations under peacetime conditions. There were 156 accidents involving low-flying aircraft and electric wires. These resulted in 78 fatalities, 56 injuries, and 6.6 million dollars damage.

In this paper we obtain the propagation constants which describe mathematically the behavior of waves reflected from dielectric-coated wires.

In an attempt to determine reasonably rapid convergent series representations for the scatter field and radar response of dielectric-coated wires, F. Schwering and C. De Santis [6] obtained two complicated characteristic equations of transcendental type. The roots of these equations are the propagation constants of the creeping waves generated by diffraction of plane waves polarized tangentially and normally to the wire axis, respectively. Their real and imaginary parts give the phase and attenuation of the creeping waves around the circumference of the wire.

In the case of tangential polarization of the incident wave, the propagation constants $\nu$ are determined from the characteristic equation [6]

$$U_\nu'' \equiv kH_\nu^{(2)'}(ka)\tilde{W}_\nu(a,b) - k_d H_\nu^{(2)}(ka)\tilde{W}_\nu'(a,b) = 0 \tag{1}$$

where

$$\tilde{W}_\nu(a,b) \equiv J_\nu(k_d a)Y_\nu(k_d b) - J_\nu(k_d b)Y_\nu(k_d a) \tag{2}$$

and

$$\tilde{W}_\nu'(a,b) = \frac{\partial \tilde{W}_\nu(a,b)}{\partial(k_d a)} \tag{3}$$

Here $J_\nu$ and $Y_\nu$ are the Bessel and Neumann functions, $H_\nu^{(2)}$ the Hankel function of the second kind, k the free-space wave number, $k_d$ the wave number of the dielectric material, and a and b the outer and inner radii of the dielectric coat (see Fig. 1).

A more complicated expression appears in $U_\nu^\perp = 0$, the characteristic equation in the case of normal polarization. This will be treated analogously later.

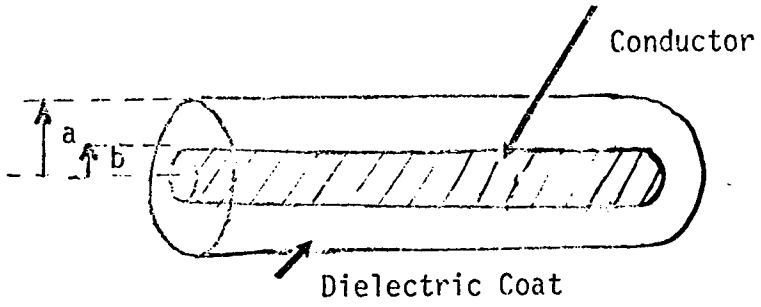



Fig. 1. Wire with Dielectric Coat.

Introducing

$$x = ka\ ,\quad y = k_d a\ ,\quad z = k_d b\ ,\quad H_\nu(x) \equiv H_\nu^{(2)}(x)$$

for simplicity into (1) - (3), we shall obtain $\nu$ as the zeros of the function $U_\nu'' \equiv U_\nu$:

$$U_\nu \equiv xH_\nu'(x)W_\nu(y,z) - yH_\nu(x)W_\nu'(y,z) \tag{4}$$

where

$$W_\nu(y,z) \equiv J_\nu(y)Y_\nu(z) - J_\nu(z)Y_\nu(y) \tag{5}$$

and

$$W_\nu'(y,z) \equiv \frac{\partial W_\nu}{\partial y}\ .$$

Using function-theoretical and analytical techniques, we shall obtain first some general qualitative properties of $\nu$ and then analytical approximations to the large zeros. Finally we shall give numerically the physically significant smallest zeros for several representative values of x, y and z.

2. The symmetry of the zeros. First we show that the function $e^{-i\nu\pi/2}\,U_\nu$ is an even function of $\nu$.

Theorem 1. If $U_\nu$ is defined by (4), then

$$e^{-i\nu\pi/2}U_\nu = e^{i\nu\pi/2}U_{-\nu}$$

Proof. We have [5]

$$Y_\nu(t) = \frac{J_\nu(t)\cos\nu\pi - J_{-\nu}(t)}{\sin\nu\pi} \tag{6}$$

whenever $\nu$ is not an integer. (For integral n, $Y_n(t) = \lim_{\nu\to n} Y_\nu(t)$. In this case, the following argument can be modified by taking limits.)

Then

$$W_\nu(y,z) = \frac{1}{\sin \nu\pi}\Big[J_\nu(y)(\cos \nu\pi\, J_\nu(z) - J_{-\nu}(z)) - J_\nu(z)(\cos \nu\pi\, J_\nu(y) - J_{-\nu}(y))\Big]$$

$$= \frac{1}{\sin \nu\pi}\Big[-J_\nu(y)J_{-\nu}(z) - J_\nu(z)J_{-\nu}(y)\Big] \quad . \tag{7}$$

Thus

$$W_\nu(y,z) = W_{-\nu}(y,z) \tag{8}$$

Since [5, p. 67] $H_{-\nu} = e^{i\nu\pi}H_\nu$, we have from (4)

$$U_{-\nu} = e^{-i\nu\pi}(xH'_\nu(x)W_{-\nu}(y,z) - yH_\nu(x)W'_{-\nu}(y,z), \tag{9}$$

whence from (8)

$$U_{-\nu} = e^{-i\nu\pi}U_\nu \tag{10}$$

This immediately gives us the desired result:

$$e^{-i\nu\pi/2}U_\nu = e^{i\nu\pi/2}U_{-\nu} \tag{11}$$

An obvious consequence is that the zeros are symmetric with respect to the origin in the complex $\nu$-plane.

Corollary. If $\nu$ is a zero of $U_\nu$, so is $-\nu$.

It is interesting to note that this simple theorem yields results which are far less obvious than the above corollary. These results refer to the strict complexity of the zeros and the infinitude of zeros, and appear in the following sections.

3. The strict complexity of the zeros. In the rest of this paper we shall denote the real and imaginary parts of $\nu$ by $\alpha$ and $\beta$, respectively, i.e.,

$$\nu = \alpha + i\beta$$

We now show that neither the real nor imaginary part of any zero of $U_\nu$ is zero.

Theorem 2. If $U_\nu = 0$, then $\alpha\beta \neq 0$.

Proof. Taking complex conjugates of both sides of (10),

$$\overline{U}_{-\nu} = e^{i\overline{\nu}\pi}\,\overline{U}_\nu \tag{12}$$

Since [5] for real argument

$$\overline{H}^{(2)}_\nu = H^{(1)}_{\overline{\nu}}, \quad \overline{J}_\nu = J_{\overline{\nu}}, \quad \overline{Y}_\nu = Y_{\overline{\nu}}, \quad \overline{W}_\nu = W_{\overline{\nu}} \tag{13}$$

where we dropped the dependence on x, y and z, we have from (12) and (4)

$$xH^{(1)\prime}_{-\overline{\nu}} W_{-\overline{\nu}} - yH^{(1)}_{-\overline{\nu}} W'_{-\overline{\nu}} = e^{i\overline{\nu}\pi}(xH^{(1)\prime}_{\overline{\nu}} W_{\overline{\nu}} - yH^{(1)}_{\overline{\nu}}\, W'_{\overline{\nu}}) \tag{14}$$

If $U_\nu = 0$ with $\beta = \mathrm{Im}\nu = 0$, then $\overline{\nu} = \nu$ and we have the simultaneous homogeneous equations

$$\begin{aligned} U_\nu &= xH'_\nu W_\nu - yH_\nu W'_\nu = 0 \\ \overline{U}_\nu &= xH^{(1)\prime}_\nu W_\nu - yH^{(1)}_\nu W'_\nu = 0 \end{aligned} \tag{15}$$

the latter coming from the right-hand side of (14). The determinant of coefficients of $xW_\nu$ and $yW'_\nu$ must then vanish:

$$\Delta \equiv H^{(1)}_\nu H'_\nu - H^{(1)\prime}_\nu H_\nu = 0 \tag{16}$$

This, however, is impossible, since $H^{(1)}_\nu$ and $H_\nu \equiv H^{(2)}_\nu$ are linearly independent solutions of Bessel's equation. Indeed, $\Delta = -4i/\pi x \neq 0$ [5, p. 68]. Thus $\beta \neq 0$ and the zeros are not real.

Applying a similar argument assuming $\alpha = 0$, whence $-\bar{\nu} = \nu$, and taking the left-hand side of (14) give another contradiction. This shows that the zeros cannot be pure imaginary either, completing the proof of the theorem.

4. The infinitude of zeros. We know from physical considerations, of course, that there exist roots of the characteristic equation. We now prove that there are an infinite number of these roots. To this end, we invoke some function-theoretical considerations, such as the concept of order of growth $\omega(f)$ of an entire (or integral) function $f(\nu)$ [1, p. 8; 7, p. 248], defined as the infimum of all exponents $\rho$ such that

$$|f(\nu)| = O(e^{|\nu|^{\rho}}) \text{ as } |\nu| \to \infty$$

Using Poisson's formula [5, p. 79]:

$$J_{\nu}(y) = \frac{2\left(\frac{y}{2}\right)^{\nu}}{\sqrt{\pi}\Gamma\left(\nu + \frac{1}{2}\right)} \int_0^{\pi/2} \cos(y \cos t) \sin^{2\nu} t \, dt , \tag{17}$$

we find easily that when $\alpha \geqq 0$,

$$\begin{aligned} J_{\nu}(y) &\leqq \frac{2\left|\left(\frac{y}{2}\right)^{\nu}\right|}{\sqrt{\pi}\left|\Gamma\left(\nu + \frac{1}{2}\right)\right|} \int_0^{\pi/2} \sin^{2\alpha} t \, dt \\ &\leqq \frac{\sqrt{\pi}\left|e^{\nu \ell n\left(\frac{y}{2}\right)}\right|}{\left|\Gamma\left(\nu + \frac{1}{2}\right)\right|} \\ &\leqq \frac{\sqrt{\pi}}{\left|\Gamma\left(\nu + \frac{1}{2}\right)\right|} e^{|\nu|\left|\ell n \frac{y}{2}\right|} \end{aligned} \tag{18}$$

Thus the order of the integral is at most 1 when $\alpha = \text{Re } \nu \geqq 0$. Moreover the entire function $1/\Gamma(\nu)$ is of order 1 [7, p. 255]. Since the order of the product (or sum) is no greater than that of the greater factor (or term), it follows that the order of $J_\nu(y)$ is no greater than unity when $\text{Re } \nu \geqq 0$.

To eliminate this restriction on the sign of $\alpha = \text{Re } \nu$, we use the facts that

$$\omega\left(H_\nu^{(1)}\right) \leq 1 \qquad [4, \text{ p. } 229] , \tag{19}$$

$$H_\nu^{(1)} = J_\nu + iY_\nu \tag{20}$$

and

$$J_{-\nu} = J_\nu \cos \pi\nu - Y_\nu \sin \pi\nu \tag{21}$$

From (19) and (20), we find that $\omega(Y_\nu) \leqq 1$ for $\alpha \geqq 0$. Then we conclude from (21) and earlier results that $\omega(J_\nu) \leqq 1$, with no restriction on $\alpha$. Moreover, since $J_\nu$, $Y_\nu$, $H_\nu$ and their derivatives can be expressed [5, § 3.1] in terms of $e^{i\nu\pi}$ and the Bessel function J with indices $\pm\nu$, $\pm\nu + 1$ and $\pm\nu - 1$, it follows finally that

Lemma. The order of growth of $U_\nu$ is less than or equal to 1.

Now let $\nu^2 = \lambda$. Then since $e^{-i\nu\pi/2}U_\nu$ is an entire <u>even</u> function of $\nu$ of order $\leqq 1$, the function $f(\lambda) \equiv e^{-i\nu\pi/2}U_\nu$ is an entire function of $\lambda$ whose order is $\leqq \frac{1}{2}$. Consequently [7, pp. 250, 252], $f(\lambda)$ has an infinite number of zeros $\lambda_k$. From the definition of $f(\lambda)$, we conclude

Theorem 3. $U_\nu$ has an infinite number of zeros.

Moreover [7, p. 250] we obtain the following product representation:

$$f(\lambda) = f(o) \prod_{k=1}^{\infty} \left(1 - \frac{\lambda}{\lambda_k}\right) \tag{22}$$

Expressed in terms of $U_\nu$, (22) becomes

$$U_\nu = e^{i\nu\pi} U_o \prod_{k=1}^{\infty} \left(1 - \frac{\nu^2}{\nu_k^2}\right) \tag{23}$$

where the $\nu_k$ are the zeros of $U_\nu$. Note that from Theorem 2, $U_0 \neq 0$, as is required for this product representation to be valid.

5. The large zeros of $U_\nu$. Since there are an infinite number of zeros of the entire function $U_\nu$, the zeros are arbitrarily large. To approximate these when $|\nu| >> \max(x,y,z)$, we first express $U_\nu$ in terms of J alone using standard identities [5, § 3.1], obtaining

$$\begin{aligned} 2i \sin^2 \nu\pi \, U_\nu = x\Big[e^{i\nu\pi}\left(J_{\nu-1}(x) - J_{\nu+1}(x)\right) + J_{-\nu+1}(x) - J_{-\nu-1}(x)\Big] \\ \times \Big[-J_\nu(y)J_{-\nu}(z) + J_\nu(z)J_{-\nu}(y)\Big] + y\Big[e^{i\nu\pi}J_\nu(x) - J_{-\nu}(x)\Big] \\ \times \Big[J_{-\nu}(z)\left(J_{\nu-1}(y) - J_{\nu+1}(y)\right) + J_\nu(z)\left(J_{-\nu+1}(y) - J_{-\nu-1}(y)\right)\Big] \end{aligned} \tag{24}$$

Then using the asymptotic behavior of $J_\mu(t)$:

$$J_\mu(t) = \frac{\left(\frac{t}{2}\right)^\mu}{\Gamma(\mu + 1)} \left(1 + O\left(\frac{t}{\mu}\right)\right) \tag{25}$$

for large $\mu$, and dropping the lower-order terms, we obtain from (24)

$$\nu \ell n(\mp 2y\nu/exz) \sim (n \mp \tfrac{1}{4})\pi i \qquad \text{as } \pm \beta > 0 , \tag{26}$$

implicitly giving approximately the n-th zero for large n.

Since the zeros are symmetric in the $\nu$-plane, we can select $\beta > 0$ and thus drop the lower signs in (26). Rewriting (26) as

$$\nu \sim (n - \tfrac{1}{4})\pi i \,/\, \ell n(-2y\nu/exz) , \tag{27}$$

iterating, and neglecting the lower-order terms, we obtain the following explicit approximation to the large zeros.

Theorem 4. The large zeros of $U_\nu$ in the upper half-plane are given by

$$\nu_n = \frac{-(n - \tfrac{1}{4})\frac{\pi}{2}^2 + (n - \tfrac{1}{4})\pi i \; \ell n \; ((2n - \tfrac{1}{2})\pi y/exz)}{[\ell n((2n - \tfrac{1}{2})\pi y/exz)]^2} \left(1 + O\left(\frac{\ell n \; \ell n \; n}{\ell n \; n}\right)\right) \tag{28}$$

for sufficiently large integers n.

As a consequence,

$$\arg \nu_n \to \frac{\pi}{2} \qquad \text{as } n \to \infty \tag{29}$$

since the real part approaches infinity more slowly than the imaginary part. Furthermore, it can easily be shown from (28) that the distance between consecutive zeros approaches zero.

We remark that this behavior, indeed the asymptotic representation (28), is very similar to that of $H_\nu^{(1)}(x)$, which arises in

the theory of diffraction of electromagnetic waves by a perfectly conducting sphere (cf. [2], [3], [4]).

6. The case of normal polarization. If the incident wave is polarized normally to the axis of the dielectric-coated wire, the characteristic equation is

$$U_\nu^\perp \equiv yH_\nu'(x)\left[J_\nu(y)Y_\nu'(z) - J_\nu'(z)Y_\nu(y)\right] - xH_\nu(x)\left[J_\nu'(y)Y_\nu'(z) - J_\nu'(z)Y_\nu'(y)\right]$$

$$= 0 \qquad (30)$$

Since the treatment of this case is identical to the previously discussed case of parallel polarization, it suffices merely to state the corresponding results.

Theorem 5. $e^{-i\nu\pi/2}U_\nu^\perp = e^{i\nu\pi/2}U_{-\nu}^\perp$.

Corollary. If $\nu$ is a zero of $U_\nu^\perp$, so is $-\nu$.

Theorem 6. If $U_\nu^\perp = 0$, then Re $\nu \cdot$ Im $\nu \neq 0$.

Theorem 7. $U_\nu^\perp$ has an infinite number of zeros.

7. The smallest zeros. Following are a table and curves (Fig. 2) of the smallest zeros of $U_\nu^{\parallel}$ and $U_\nu^\perp$ in the second quadrant of the complex $\nu$-plane for each of several representative values of the parameters x, y, z. These values are x = 0.5(0.5)5, with $y = 1.5x$ and $z = 0.9y$. We recall that $x = ka$, $y = k_d a$, and

$z = k_d b$ where k is the free-space wave number, $k_d$ the wave number of the dielectric coat, and a and b the radii of the coat. The coefficient $1.5 = k_d/k$ is the refractive index of polyethylene and the coefficient $0.9 = b/a$ is the ratio of the two radii.

| | $U_\nu^{\parallel} = 0$ | | $U_\nu^{\perp} = 0$ | |
|---|---|---|---|---|
| x | Re $\nu$ | Im $\nu$ | Re $\nu$ | Im $\nu$ |
| 0.5 | -1.1075 | 1.3605 | -0.7690 | 0.6697 |
| 1.0 | -1.7462 | 1.6556 | -1.3273 | 0.7794 |
| 1.5 | -2.3282 | 1.8668 | -1.8707 | 0.8427 |
| 2.0 | -2.8819 | 2.0365 | -2.4089 | 0.8815 |
| 2.5 | -3.4176 | 2.1804 | -2.9475 | 0.9033 |
| 3.0 | -3.9399 | 2.3063 | -3.4831 | 0.9121 |
| 3.5 | -4.4516 | 2.4184 | -4.0226 | 0.9101 |
| 4.0 | -4.9538 | 2.5198 | -4.5653 | 0.8986 |
| 4.5 | -5.4477 | 2.6118 | -5.1123 | 0.8790 |
| 5.0 | -5.9323 | 2.6959 | -5.6643 | 0.8524 |

These zeros were obtained by J. Herder of ECOM's Math. Support Division using a Burroughs B-5700 and the Bessel routine provided by M. Goldstein of New York University. The following curves were obtained from the above data by C. De Santis of the Communications Research Tech. Area.

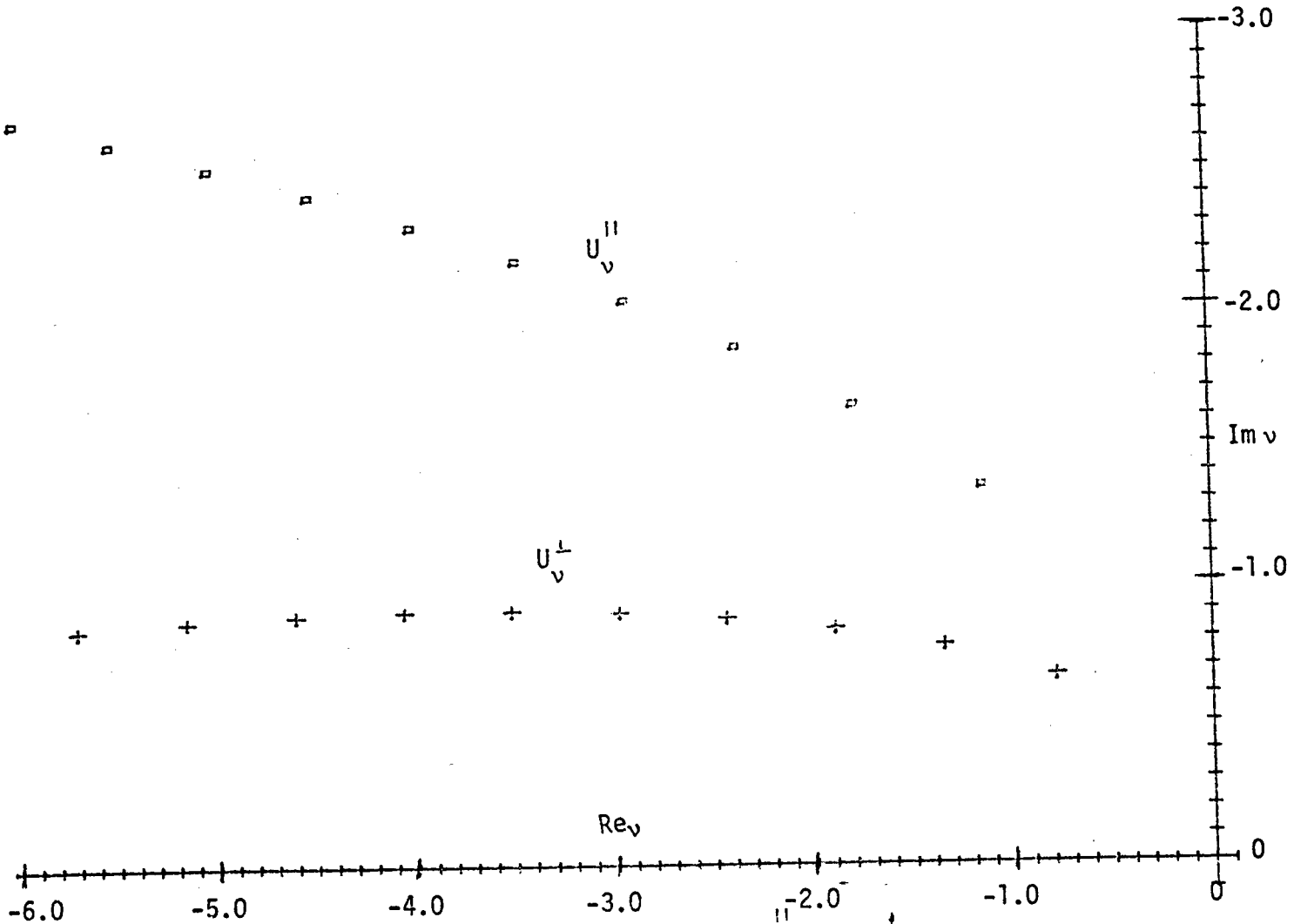



Fig. 2. The Smallest Zeros of $U_\nu^{II}$ and $U_\nu^{\perp}$

Acknowledgment. The author thanks W. Pressman and F. Schwering of the Communications Research Technical Area and J. B. Keller of New York University for informative discussions.

## REFERENCES

[1] R. P. Boas, "Entire Functions," Academic, New York, 1954.

[2] J. B. Keller, S. I. Rubinow, and M. Goldstein, Zeros of Hankel functions and poles of scattering amplitudes, J. Math. Phys. 4 (1963), 829-832.

[3] L. Kotin and W. Magnus, Transcendental equations in electromagnetic theory, New York University Report BR-27 (1958).

[4] W. Magnus and L. Kotin, The zeros of the Hankel function as a function of its order, Numerische Math. 2 (1960), 228-244.

[5] W. Magnus, F. Oberhettinger, and R. P. Soni, "Formulas and Theorems for the Special Functions of Mathematical Physics," Springer, New York, 1966.

[6] F. Schwering and C. De Santis, Radar response of dielectric-coated long wires, ECOM Report, U. S. Army Electronics Command, Fort Monmouth, N. J. (in preparation).

[7] E. C. Titchmarsh, "The Theory of Functions," Oxford University Press, London, 1950.

# ACTIVATION ENERGY ASYMPTOTICS AND UNSTEADY FLAMES*

J. Buckmaster
Mathematics Department and Department of Theoretical and Applied Mechanics
University of Illinois, Urbana, Illinois 61801

G. S. S. Ludford
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca, N.Y. 14853

1. INTRODUCTION. This is a review of some classical problems in laminar flame theory that essentially assumes no knowledge of combustion by the reader.

Laminar flame theory is a branch of fluid mechanics -- essentially the motions of the gases in a flame are governed by the compressible Navier-Stokes equations -- but there are of course some crucial features which are not normally found in classical fluid mechanics. For one thing one is dealing with a mixture of different gases and it is necessary to say something about changes in each of the components of the mixture. Secondly, and most important, there are chemical reactions so that there is a source or sink term in the mass conservation equation for each component. Moreover, heat is released by the chemical reactions so that there is a source term in the overall energy equation. These chemical reactions are extremely sensitive to temperature -- they usually won't take place at all if the temperature is too low (which is fortunate) -- and an essential feature of combustion that helps distinguish it from other branches of aerothermochemistry is that the high temperatures necessary to sustain the reactions are generated by the heat released by the reactions themselves. Provided there is an adequate supply of fuel and oxygen, combustion is a self-sustaining process.

There are two different approaches to the theory of combustion that one can take. One is to insist on being as realistic as possible and retain in the formulation of the problem all the complexities that might play a role in practice. This of course leads to equations of remarkable complexity which can only be solved numerically. Such an approach has its advocates (and is necessary if detailed quantitative results are needed) but a more fruitful approach, given the present state of combustion science, is to strip each problem down to its fundamentals and write down model equations that are clearly inappropriate in reality but nevertheless contain the physical features which are the essence of the problem. The hope is that the equations are simple enough to solve analytically, or, if recourse to a computer is still necessary, simple enough so that useful information can be extracted from the numbers. Quantitative accuracy is sacrificed for qualitative understanding.

Actually there is a third approach to studying combustion problems that has been quite popular but which should be avoided if at all possible. One starts by writing down sensible model equations but then constructs what might be called 'model solutions'. That is, solutions are constructed using ad hoc irrational approximations and as a consequence one can never be sure of the significance of the end results. It isn't clear whether the features of the solution are creatures of the original model or of the irrational approximations. This makes systematic development of the subject difficult and has led to spurious results in the past.

---

*This is a more or less verbatim transcript of a review that was specifically prepared for oral presentation, so that the reader is asked to forgive the colloquial style. The footnotes were not part of the original presentation but have been added for the sake of clarity.

Of course it is clear why such an unsatisfactory approach has been popular -- for many years no rational systematic method of solving the various model equations was known (although the literature is replete with brilliant ad hoc analyses). But in recent years that has changed, and it would probably be fair to say that there has been a revolution in combustion theory. At the heart of this revolution was the realization that combustion theory has its own unique asymptotics which can be exploited using singular perturbation theory. In particular, a combination of Damkohler Number asymptotics and activation energy asymptotics, where appropriate, can often lead to the solution of model equations that were for many years thought too difficult to solve.

What I want to do today is briefly describe the nature of these asymptotic methods, concentrating particularly on activation energy asymptotics; describe the mathematical details of a particularly simple application of activation energy asymptotics; and then describe a perturbation procedure that generates nonlinear solutions for a variety of problems, including a certain class of unsteady problems. In no sense am I going to attempt an exhaustive review.

Let us start by looking at a specific problem.

## 2. QUASI-STEADY FUEL DROP BURNING

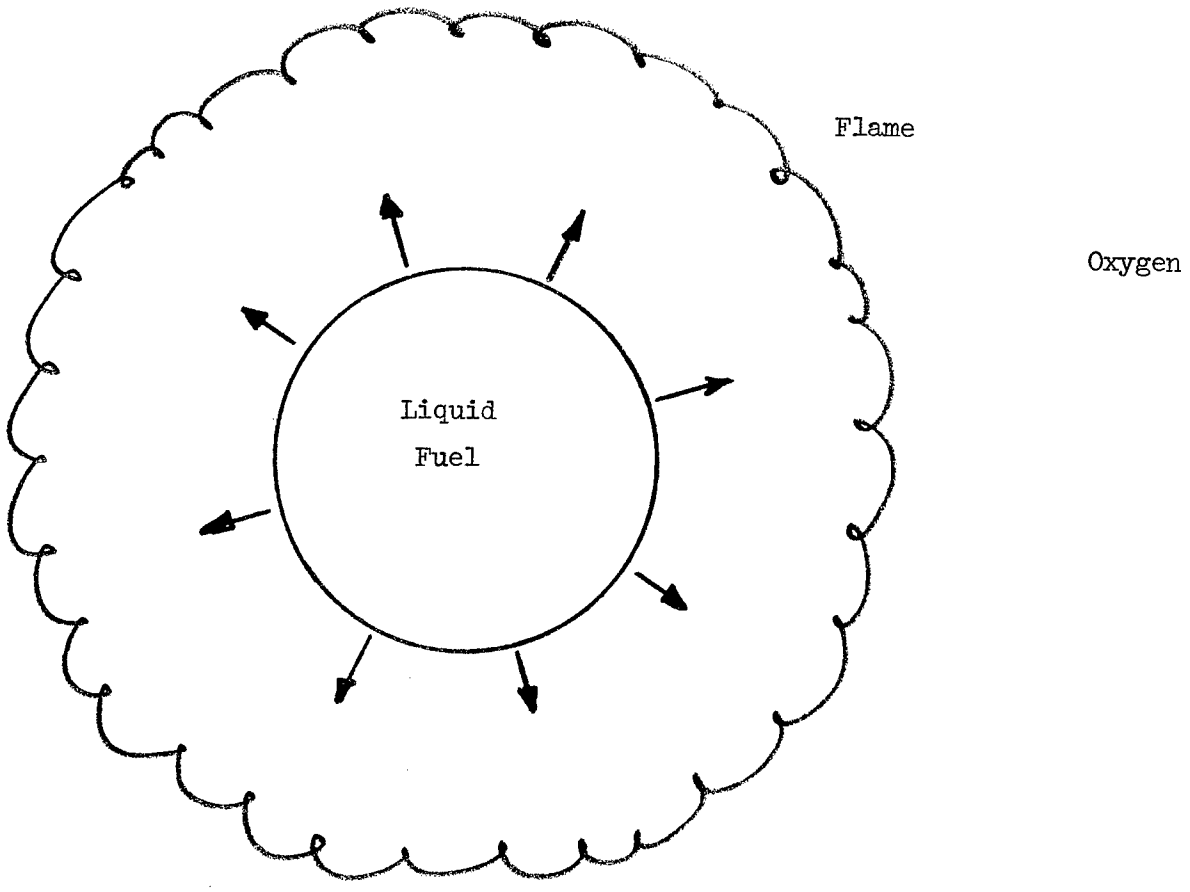



Fig. 1. Burning Fuel Drop

Figure 1 represents an idealized model of a burning fuel drop. The situation is assumed steady and spherically symmetric, conditions never realized in practice which emphasizes that we are examining a highly idealized model.

The ball of fuel, in liquid form, is hot because of the presence of the flame and as a consequence it evaporates, mixes by diffusion with the surrounding atmosphere of oxygen, and then this mixture burns within the flame. Appropriate model equations are,[+]

$$\mathcal{L}Y_0 \equiv \frac{M}{r^2}\frac{dY_0}{dr} - \frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{dY_0}{dr}\right) = -D_1 Y_0 Y_F T^a \exp\left(-\frac{E}{RT}\right) \equiv -\omega$$

$$\mathcal{L}Y_F = -\omega$$

$$\mathcal{L}T = Q\omega .$$

These equations are based on the simple chemical kinetic scheme

$$[\text{Fuel}] + [\text{Oxygen}] \longrightarrow [\text{Product}] .$$

The kinetics of a real flame are much more complicated than this but nevertheless the simple model preserves three essential features -- oxygen is consumed, fuel is consumed, and heat is generated.

Looking at the equation for $Y_0$, the mass fraction of oxygen, we see that there are three terms. The first term is a mass transport term (there is a radial flux of fuel and therefore a mass-averaged radial velocity) and $M$ is a measure of the flux of fuel leaving the surface. It can be regarded as the fundamental unknown of the problem.

The second term is a diffusion term.

The third term, the chemical reaction term, simply indicates that the amount of oxygen consumed depends on how much oxygen is present, how much fuel is present, and the temperature $T$. The most important part of the temperature dependence is the exponential factor -- $R$ is the gas constant and $E$ is a constant known as the activation energy. $E$ tends to be rather large so that the reaction rate is very sensitive to changes in the temperature.

$D_1$ is a parameter that depends on a number of things including the pressure (which is uniform) and is known as the Damkohler Number.

The equation for $Y_F$ is identical to that for $Y_0$, a consequence of assuming equal diffusion coefficients. The energy equation (which is an equation for the temperature since the thermal energy is much larger than the kinetic energy) is very similar (the Lewis number equals one) but the reaction term appears with a positive sign since heat is generated by the reaction, and the amount of heat generated is characterized by the parameter $Q$.

---

[+]Kassoy, D. R. & Williams, F. A. Physics of Fluids, 11, 1343 (1968).

There are appropriate boundary conditions (which have not been written down), 4 at the surface and 3 at infinity, making a total of 7. Since the system is a sixth order one, these conditions are sufficient to determine the three field variables and M, which is a measure of the burning rate.

There are many different ways of characterizing the solution of this problem, and one way is to plot the variation of M with $D_1$.

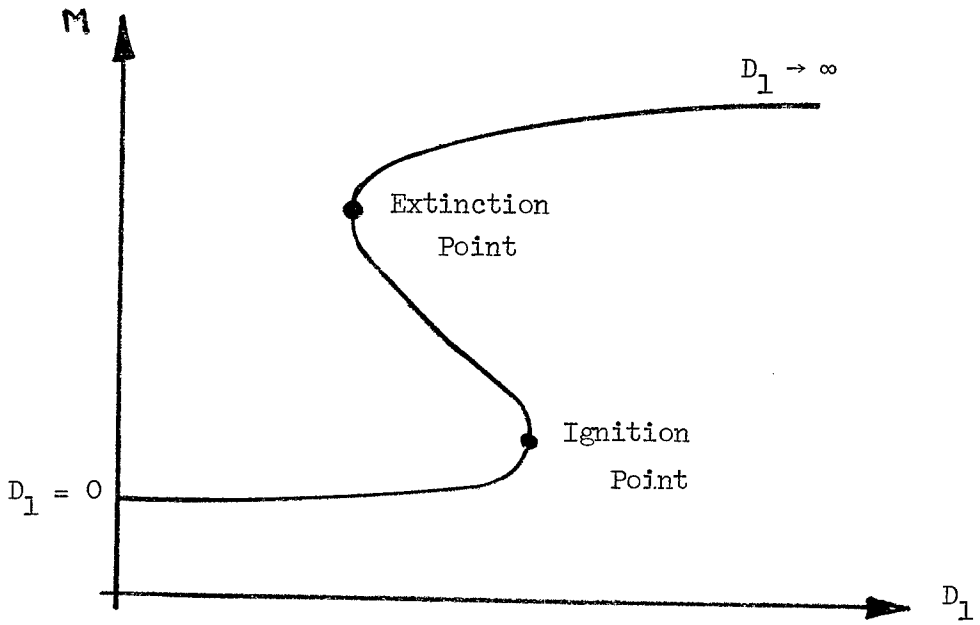



Fig. 2. Burning Response for a Fuel Drop

Figure 2 is typical of the kind of response one gets -- an S shaped curve, and at the risk of oversimplification the turning points are labelled as the ignition point and the extinction point. The reason for this is that if the response is on the lower branch of the curve, where the burning is weak, and $D_1$ is increased (by increasing the pressure, for example) then the response moves to the right until the ignition point is reached whereupon any further increase in $D_1$ causes a jump to the top branch where the burning is strong. A subsequent decrease in $D_1$ moves the response to the left along the strong burning branch until the extinction point is reached where the response drops back to the weak burning branch.[+]

[+]The oversimplification stems from the possibility that the response is forced off of one of the branches by instability before the turning point is reached. This happens in chemical reactor theory where similar S-shaped responses occur (Cohen, D. S. & Poore, A. B. SIAM J. Appl. Math., 27, 416 (1974).

Consider now the ends of the curve. At the left $D_1$ vanishes, whence $\omega$ vanishes and the equations reduce to linear equations which can be easily solved. This so called frozen limit is of little interest since there is no combustion.

The right hand end ($D_1 \rightarrow \infty$) is much more important since typical flames encountered in everyday life often have very large Damkohler numbers. The limit (called the equilibrium limit) is a singular one in which the coefficient of the highest derivative vanishes, and so thin layers (boundary layers or interior layers[+]) can occur. Outside of these layers it is apparent, since $\omega$ must be finite, that as $D_1 \rightarrow \infty$,

$$Y_O Y_F \rightarrow 0,$$

and so $Y_O$ and/or $Y_F$ must vanish. $\omega$ is then the product of something that goes to infinity times something that goes to zero and it is clear from the equations (the equation for $Y_O$ when $Y_O$ vanishes) that this product vanishes in the limit. In this sense there are similarities between the equilibrium limit and the frozen limit, but the possibility of thin layers in the former case is a crucial difference.

Important though Damkohler Number asymptotics may be, it obviously cannot tell us anything about ignition or extinction, so that if we wish to bridge the gap between $D_1 = 0$ and $D_1 \rightarrow \infty$ a different approach is necessary. Activation energy asymptotics is an appropriate tool. More precisely we consider the solution of the equations when

$$\frac{E}{RT_{ref}} \rightarrow \infty$$

where $T_{ref}$ is same reference temperature. This is a realistic limit in many combustion situations, it can be used to solve many important combustion problems, and it is mathematically interesting because the large parameter appears in an unconventional fashion, as the argument of an exponential.

One thing that is immediately clear is that we can not just put $E = \infty$ without doing anything else since that just yields the frozen limit ($\omega = 0$). Bear in mind that we want to determine how the response changes with $D_1$, and the above observation implies that only when $D_1$ is very large can we get away from the frozen limit. What we have to do is write

$$D_1 = \exp(E/RT_*)$$

where $T_*$ is a temperature that characterizes the magnitude of $D_1$ so that

$$\omega \propto \exp\left[\frac{E}{R}\left(\frac{1}{T_*} - \frac{1}{T}\right)\right],$$

and then the behavior of $\omega$ in the limit $E \rightarrow \infty$ depends upon the relative magnitudes of $T$ and $T_*$. There are three possibilities.

(i) In regions where $T > T_*$ the exponential goes to infinity in the limit, so that $Y_O Y_F \rightarrow 0$, $\omega \rightarrow 0$, corresponding to equilibrium.[++]

---

[+]See Buckmaster, J. D. Combustion and Flame, 24, 79 (1975).

[++]Thin layers can occur in such regions of course.

(ii) In regions where $T < T_*$ the exponential vanishes so that $\omega \to 0$, a frozen situation.

(iii) Finally, in transition regions where $T \sim T_*$ (more precisely, $\frac{T-T_*}{T_*} = O\left(\frac{RT_*}{E}\right)$) the exponential can be simplified slightly,

$$\omega \propto \exp\left[\frac{E}{R}\frac{(T-T_*)}{T_*^2}\right],$$

but the important point is that $\omega$ does not vanish so that such a region is a reaction zone. Reaction zones are often thin (but not necessarily so) in which case they are called flame sheets.

Application of activation energy asymptotics to a steady one-dimensional problem such as the fuel drop problem requires, in general, the construction of solutions in the three different kinds of regions and matching them in the usual way (that is, in the sense of matched asymptotic expansions). Usually, the most difficult part of this procedure is deciding what regions are needed and where they are located. As an example, if we ask what is the nature of the solution for a point on the middle branch of the S-shaped response (Fig. 2), it turns out that $T_*$ is the maximum temperature. That is, at some finite value of $r$ the temperature is equal to $T_*$ so that all the reaction occurs in a thin flame sheet located there, and on either side of the sheet the combustion is frozen (Fig. 3).

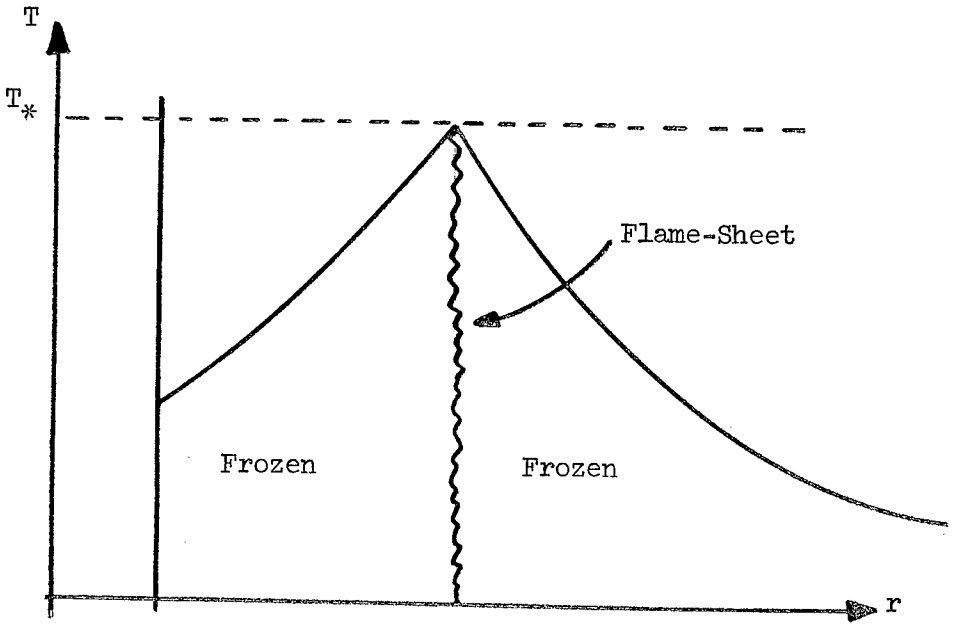



Fig. 3. Typical Temperature Distribution for a Solution on the Middle Branch

In the frozen regions solutions can easily be constructed of the linear governing equations. In the flame sheet the description is nonlinear, but because the sheet is thin the equations are simplified. Matching the flame sheet solution with the solutions in the frozen regions ultimately leads to the complete solution of the problem and in particular, the determination of the burning rate $M^+$. A remark about the nature of the solution on the other two branches will be made later.

3. PREMIXED FLAMES. The fuel drop problem is an example of what is known as a diffusion flame. There are other kinds of flames in which the reactants are supplied as a homogeneous mixture which merely needs to be raised to an adequate temperature to initiate burning. Such flames are called premixed flames, a common example being the inner cone of a bunsen burner flame (observed when the air hole is open which permits oxygen to mix with the gas as it passes up the tube).

If a match is applied to such a mixture, confined within a tube, the mixture will burn and a flame will travel down the tube consuming the mixture as it goes. Under ideal conditions this flame travels as a progressive wave with a more or less well defined wave speed, and one of the classical problems of laminar flame theory is to determine that wave or flame speed. What I want to do now is briefly describe how this can be done using activation energy asymptotics, since this is one of the simplest nontrivial applications of activation energy asymptotics presently known.

For a premixed flame the simplest kind of sensible chemical kinetic scheme is

$$[\text{Mixture}] \Longrightarrow [\text{Product}]$$

at a rate $\omega = BY \exp(-E/RT)$

where $Y$ is the mass fraction of mixture (a preexponential temperature dependence like the $T^a$ that was included in the fuel drop equations could be inserted without essentially changing the subsequent discussion).

The flame is assumed to be one-dimensional and the situation in a flame-fixed frame is shown in Fig. 4.

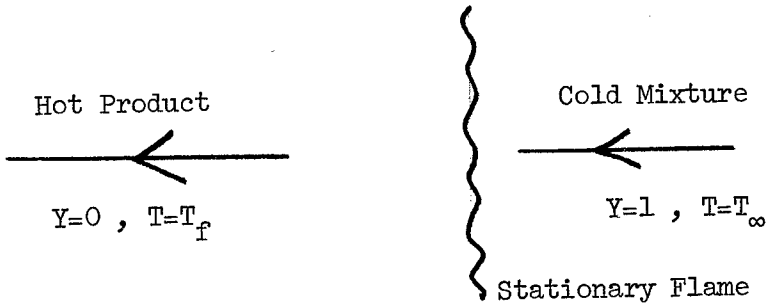



Fig. 4. The One-Dimensional Premixed Flame

[+]The work of A. Linan, Astronautica Acta, 1, 1007 (1974) on the counterflow diffusion flame provides an exhaustive description of calculations of this kind. Kapila A. K., Ludford, G. S. S. & Buckmaster, J. D. Combustion and Flame, 25, 361 (1975) describe similar calculations for a spherical premixed flame.

Cold mixture comes in from the right and passes through the flame where it is burnt and emerges as hot product on the left. The reaction only stops when all the mixture is consumed so that $Y = 0$ on the left and the temperature there is $T_f$ ($> T_\infty$), the so called adiabatic flame temperature.

Appropriate model equations are

$$\rho v \frac{dY}{dx} = \frac{d}{dx}(\rho D \frac{dY}{dx}) - BY \exp(-E/RT)$$

$$\rho v C_p \frac{dT}{dx} = \frac{d}{dx}(\lambda \frac{dT}{dx}) + Q B Y \exp(-E/RT)$$

$$\rho v = -m \text{ (constant)}$$

$$\rho T = \text{constant}$$

which are similar, in many respects, to the fuel drop equations written down earlier. $m$ the constant mass flux, is the fundamental unknown being essentially the flame speed. The equation of state is simply a statement that the pressure is constant, valid for low Mach Number flames.

The flame temperature $T_f$ can be determined without solving this system, for in passing through the flame we know exactly how much of the mixture is consumed (all of it) and we know exactly how much heat is released per unit of mixture consumed (Q). Therefore an overall energy balance requires

$$C_p (T_f - T_\infty) = Q.$$

The system, when appropriately non-dimensionalized, is

$$-\frac{dY}{d\xi} = \frac{1}{L}\frac{d^2Y}{d\xi^2} - \frac{B\lambda}{m^2 C_p} e^{-\theta/\phi}$$

$$-\frac{d\phi}{d\xi} = \frac{d^2\phi}{d\xi^2} + \frac{B\lambda}{m^2 C_p} e^{-\theta/\phi}$$

($\xi \sim x$, $\theta \sim E$, $\phi \sim T$, $L = \frac{\lambda}{\rho D C_p}$ is the Lewis No.)

as $\xi \to +\infty$ $\quad Y \to 1, \quad \phi \to \phi_\infty$

as $\xi \to -\infty$ $\quad Y \to 0, \quad \phi \to 1 + \phi_\infty$ ,

and the essential idea is that this system only has a solution for a unique[+] choice of the parameter $\frac{B\lambda}{\dot{m}^2 C_p}$ and so in this way the flame speed can be determined.

An enormous amount of ingenious effort has been expended over the years on the solution of this problem, and literally dozens of approximate solutions can be found in the literature each purporting to be simpler or more accurate than earlier attempts. Most of this work was rendered obsolete in 1970 by Bush and Fendell[++] who showed how the problem can be solved rationally in the limit of infinite activation energy ($\theta \to \infty$).

Just as for the fuel drop problem we can not just put $\theta = \infty$ in the equations -- it is necessary to let $\frac{B\lambda}{\dot{m}^2 C_p} \to \infty$ at the same time. More precisely we write

$$\frac{B\lambda}{\dot{m}^2 C_p} = \frac{\Omega}{L\,(1+\phi_\infty)^4}\,\theta^2 \exp\left[\frac{\theta}{1+\phi_\infty}\right], \quad \Omega = O(1)$$

a choice partly motivated by the observation that we would expect, on physical grounds, that the flame temperature $(1+\phi_\infty)$ is the maximum temperature and moreover that $\phi$ increases monotonically from $\phi_\infty$ to $(1+\phi_\infty)$ as the flame is traversed. Be that as it may, the problem is to find $\Omega$.

The reaction rate $\omega$ is proportional to

$$\exp\left[\frac{\theta}{1+\phi_\infty} - \frac{\theta}{\phi}\right]$$

so that wherever $\phi$ is less than the flame temperature the reaction is frozen and the governing equations are

$$\frac{d^2\phi}{d\xi^2} + \frac{d\phi}{d\xi} = 0$$

$$\frac{d^2Y}{d\xi^2} + L\frac{dY}{d\xi} = 0$$

---

[+]In actual fact the system as written doesn't have a solution at all, since the upstream state $Y = 1$, $\phi = \phi_\infty$ is not a solution of the equations (the so-called cold boundary difficulty). The problem arises because the temperature dependence of the reaction rate is not accurately modelled by $\exp(-\theta/\phi)$ when $\phi$ is small. A realistic resolution of the difficulty is to introduce a cutoff temperature lying between $\phi_\infty$ and $1+\phi_\infty$ below which the reaction rate is identically zero. No specific choice for this temperature is needed when the activation energy $\theta$ is large, as the subsequent analysis shows.

[++]Bush, W. B. & Fendell, F. E. Combustion Science & Technology, 1, 421 (1970).

with elementary solutions. The location of the origin of coordinates can be chosen so that these equations are valid in $\xi > 0$.

Noting that

$$Y = 0, \qquad \phi = 1 + \phi_\infty$$

is an exact solution of the complete equations, the large scale structure of the flame is obtained by piecing together this exact solution and appropriate solutions of the frozen equations, as shown in Fig. 5.

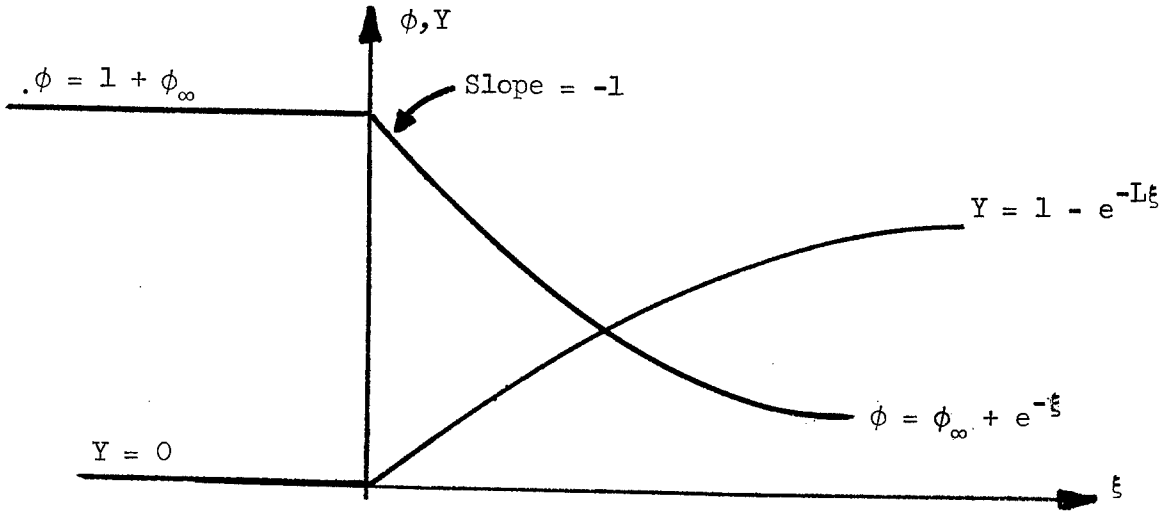



Fig. 5. Large Scale Structure of the Flame

The frozen solutions in $\xi > 0$ are chosen to ensure that the boundary conditions as $\xi \to \infty$ are satisfied and that $\phi$ and $Y$ are continuous at the origin.

To complete the solution it is necessary to analyze the thin region near the origin where the derivatives are smoothed out. The chemical reaction is confined to this region, which is therefore a flame sheet, and the local solution has the form

$$\phi \sim (1 + \phi_\infty) + \frac{1}{\theta}(1 + \phi_\infty)^2 \psi(\ell) + O\left(\frac{1}{\theta^2}\right)$$

$$Y \sim \frac{1}{\theta} y(\ell) + O\left(\frac{1}{\theta^2}\right)$$

$$\xi = \frac{1}{\theta}(1 + \phi_\infty)^2 \ell .$$

In other words, the flame sheet has a thickness of order $O\left(\frac{1}{\theta}\right)$ but gradients there are $O(1)$.

The perturbation quantities satisfy

$$\frac{d^2\psi}{d\ell^2} = \frac{d^2}{d\ell^2}\left[\frac{-y}{L(1+\phi_\infty)^2}\right] = \frac{\Omega\, y\, e^{\psi}}{L(1+\phi_\infty)^2} .$$

It follows that

$$y + L(1+\phi_\infty)^2\,\psi$$

is a linear function $\mathfrak{L}(\ell)$. But then matching (both $y$ and $\psi$ vanish as $\ell \to -\infty$ to match with the solution behind the flame sheet) implies that $\mathfrak{L}$ is identically zero. A problem for $\psi$ alone may then be formulated.

$$0 = \frac{d^2\psi}{d\ell^2} - \Omega\,\psi\, e^{\psi}$$

as $\ell \to -\infty$ $\quad \psi \to 0$

as $\ell \to +\infty$ $\quad \frac{d\psi}{d\ell} \to -1$ .

The latter boundary condition arises from matching ahead of the flame sheet (Fig. 5).

Integrating once,

$$0 = \left(\frac{d\psi}{d\ell}\right)^2 - 2\,\Omega\,(\psi\, e^{\psi} - e^{\psi} + 1)$$

and then applying the condition as $\ell \to \infty$ leads to Bush and Fendell's result

$$\Omega = \frac{1}{2}$$

and completes the determination of the flame speed.

4. THE MODIFIED PREMIXED FLAME. There are two features of Bush and Fendell's solution that I want to emphasize. First of all, one of the reasons that the analysis is so simple is that the chemistry free equations can be so easily solved. This suggests the following question. Suppose that we are concerned with a more complicated problem, one related to the one-dimensional premixed flame but whose description requires additional terms in the governing equations. What additional terms would lead to chemistry-free equations as easy to solve as Bush and Fendell's? Such a question obviously does not have a unique answer, but one possibility is

$$\frac{d^2\phi}{d\xi^2} + \frac{d\phi}{d\xi} = \frac{1}{\theta} f(\xi, \phi, Y, \text{---})$$

$$\frac{d^2Y}{d\xi^2} + L \frac{dY}{d\xi} = \frac{1}{\theta} g(\xi, \phi, Y, \text{---})$$

where f and g are quite arbitrary. Perturbation solutions of these equations can easily be constructed.

The second thing to notice about Bush and Fendell's solution is that the flame speed is extremely sensitive to the value of the maximum temperature. The expression for the flame speed (essentially m) is

$$\frac{B\lambda}{m^2 C_p} = \frac{\theta^2}{2L(1+\phi_\infty)^2} \exp\left[\frac{\theta}{1+\phi_\infty}\right]$$

and it is clear that small changes in the flame temperature $(1+\phi_\infty)$ will generate large changes in the flame speed. Order $0(1/\theta)$ changes in temperature are sufficient to generate $0(1)$ changes in the speed, for example. The significance of the modified equations written down above is that we might expect that the $0(1/\theta)$ perturbation terms can generate $0(1/\theta)$ changes in the maximum temperature and thus lead to solutions quite different from Bush and Fendell's. And yet we would not expect the inclusion of these terms to unduly complicate the analysis.

Let us consider a simple example.

5. EFFECT OF HEAT LOSSES. In any real flame there are heat losses due radiation or conduction to adjacent boundaries. In a one-dimensional formulation these losses can be modelled by adding a term $-K(T-T_\infty)^+$, K = constant, to the energy equation so that

$$\rho v C_p \frac{dT}{dx} = \frac{d}{dx}\left(\lambda \frac{dT}{dx}\right) - K(T-T_\infty) + Q B Y \exp\left(-\frac{E}{RT}\right).$$

The extra term tends to drive the temperature to the reservoir value, and the modified equations are of the type discussed above provided the magnitude of K is such that the non-dimensional term is $O(1/\theta)$.

The analysis is more complicated than Bush and Fendell but no new principles are involved, and defining

---

+Quite general functions of T can in fact be handled by the analysis, see Buckmaster, J. D. Combustion & Flame (in press).

$$H = \text{Flame Speed/Adiabatic Flame Speed}^{+}$$

we find

$$(1+\phi_\infty)^2 \, H^2 \ln H + K' = 0$$

where K' is K non-dimensionalized.

When K' vanishes there are two solutions, H = 1 (Bush and Fendell's result) and H = 0, and for moderate values of K' there are two solutions, but if K' is too large there are no solutions (Fig. 6). This principle has been known for many years

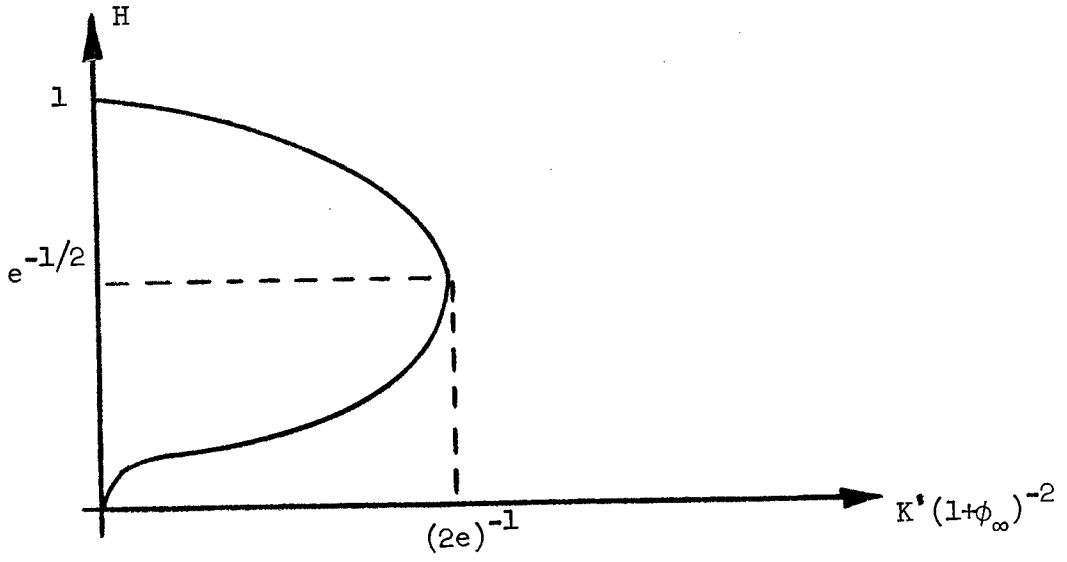



Fig. 6. Flame Speed vs Heat Losses

and is the foundation of the miner's safety lamp invented by Humphrey Davy in the early 19th century. The safety lamp consists of a naked flame surrounded by a wire gauze cage, and if this is carried into a combustible atmosphere, the latter passes through the gauze and burns on contact with the flame. Without the gauze cage the flame would spread through the surrounding atmosphere, usually in a violent (explosive) fashion, but the gauze is such an efficient conductor of heat that the flame can not pass through it. Thus the miner, on seeing the flame flare up, can safely retreat.

Looking again at the response diagram (Fig. 6), recall that as we move around the curve the maximum temperature changes by only an $O(1/\theta)$ amount. Returning to the fuel drop response (Fig. 2), the top branch of the curve, including the extinction point, corresponds to solutions for which the maximum temperature differs by only an

---

+i.e., Bush and Fendell's result.

$O(1/\theta)$ amount from the maximum temperature in the equilibrium ($D_1 \to \infty$) limit. The lower branch, including the ignition point, corresponds to solutions for which the maximum temperature differs by only an $O(1/\theta)$ amount from the maximum temperature in the frozen ($D_1 = 0$) limit. Thus there is an analogy between the C-shaped quenching curve of Fig. 6, and the C-shaped extinction and ignition curves of Fig. 2.

Once the idea of adding $O(1/\theta)$ perturbation terms to systems of flame equations and looking for solutions that differ by an $O(1)$ amount from the unperturbed solution is understood, there are an infinite number of problems that one can examine. One is limited only by one's imagination in conjuring up different kinds of perturbations, and of course any flame can be perturbed, not just the one-dimensional premixed flame. Let us consider some unsteady examples.

6. UNSTEADY ONE-DIMENSIONAL PREMIXED FLAME. Consider the unsteady form of Bush and Fendell's problem, for which the equations are:

$$\rho \frac{\partial Y}{\partial t} + \rho v \frac{\partial Y}{\partial x} = \frac{\partial}{\partial x}\left(\rho D \frac{\partial Y}{\partial x}\right) - BY \exp\left(-\frac{E}{RT}\right)$$

$$\rho C_p \frac{\partial T}{\partial t} + \rho v C_p \frac{\partial T}{\partial x} = \frac{\partial}{\partial x}\left(\lambda \frac{\partial T}{\partial x}\right) + Q B Y \exp\left(-\frac{E}{RT}\right)$$

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0 \qquad \rho T = \text{constant}.$$

These differ from the earlier equations only by the addition of the time derivatives.

Now the steady flame has a characteristic thickness,

$$\sim \frac{\lambda}{m C_p} .$$

There is a characteristic velocity, the flame speed,

$$\frac{m}{\rho_\infty} ,$$

and so we can define a characteristic time

$$\sim \frac{\lambda \rho_\infty}{m^2 C_p} .$$

If we try to solve the unsteady equations -- as an initial value problem, for example -- then most disturbances will change on this time scale and will be governed by the complete system of equations, without simplification. Even without chemistry this system presents a formidable challenge. However, it is conceivable that there are disturbances that change on the much larger time scale

$$t = O\left(\frac{\theta \lambda \rho_\infty}{m^2 C_p}\right),$$

in which case the time derivatives are $O(\frac{1}{\theta})$ terms and so can be handled in the same way as the small heat loss term. Indeed we find

$$2(1+\phi_\infty)^2 H^3 \ln H + b\frac{dH}{d\tau} = 0$$

an equation first derived by Sivashinsky.[+] Here $\tau$ is time, H is the flame speed ratio as before, and b is a parameter that depends upon the Lewis Number L.

$b < 0$ if $L > 1$

$b > 0$ if $L < 1$

$b = 0$ if $L = 1$.

Apparently, when $L = 1$ there are no disturbances that change on the slow time scale, an atypical situation. It is tempting when solving combustion problems to choose $L = 1$, since this often leads to mathematical simplification (the steady one-dimensional premixed flame then has uniform enthalpy, for example) but this temptation is apparently something that should be resisted, at least when dealing with unsteady problems.

There are two possible steady solutions

$H = 0, \quad H = 1,$

and the stability of these solutions depends upon the sign of b:

$b > 0$ $(L < 1)$ $H = 1$ stable, $H = 0$ unstable,

$b < 0$ $(L > 1)$ $H = 1$ unstable, $H = 0$ stable.

Thus if $L > 1$, Bush and Fendell's solution for the one-dimensional flame is unstable. It should be emphasized, of course, that only the predictions of instability are significant. A flame that is stable to the kind of disturbances that we have considered here might well be unstable to other kinds of disturbances.

7. UNSTEADY FLAME WITH HEAT LOSSES. As we saw earlier, when there are heat losses the burning response is multiple valued. Thus it is of interest to add heat losses to the unsteady formulation in the hope of gaining insight into the significance of multivalued responses. The result is

$$2(1+\phi_\infty)^2 H^3 \ln H + 2HK' + b\frac{dH}{d\tau} = 0.$$

Note that in addition to the two steady branches shown earlier in Fig. 6, there is a third steady solution $H = 0$ (Fig. 7).

[+]Sivashinsky, G. I. Int. J. Heat Mass Transfer, 17, 1499 (1974).

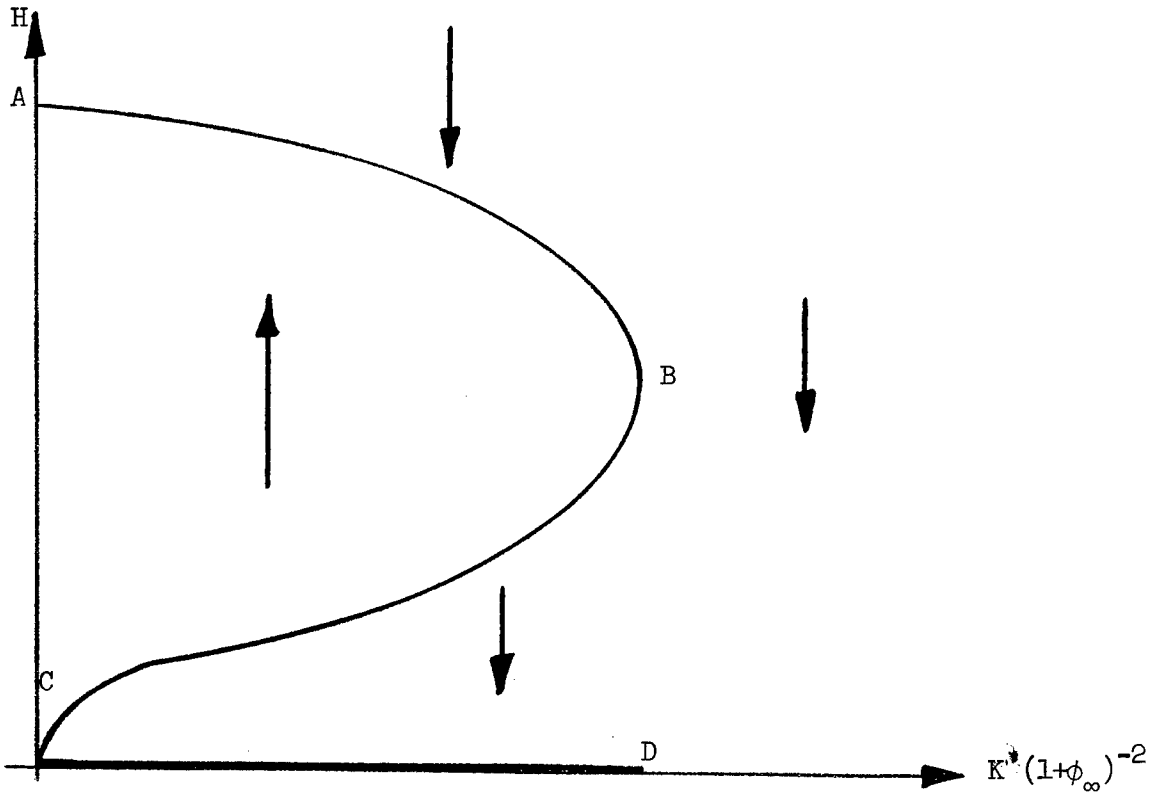



Fig. 7. Stability of a Flame with Heat Losses when $L < 1$

Figure 7 shows stability arrows appropriate when $L < 1$. These indicate the direction the solution will be driven in an unsteady situation. Thus when $L < 1$ the branches AB and CD are stable, whereas CB is unstable. For $L > 1$ the arrows must be reserved.

8. THREE-DIMENSIONAL UNSTEADY FLAMES. The perturbation procedure is not confined to one-dimensional flames. Three-dimensional disturbances can also be treated provided their nature is such that the three-dimensional terms are essentially $O(1/\theta)$. The equations are rather more complicated since the velocity field must be determined and this requires solution of the momentum equation

$$\frac{\partial v}{\partial t} + (v \cdot \nabla) v = -\frac{1}{\rho} \nabla p + \nu \left[\nabla^2 v + \frac{1}{3} \nabla(\nabla \cdot v)\right]$$

in addition to the previous equations.

Permissable disturbances are defined in Fig. 8 (recall that the flame thickness $\sim \frac{\lambda}{m C_p}$).

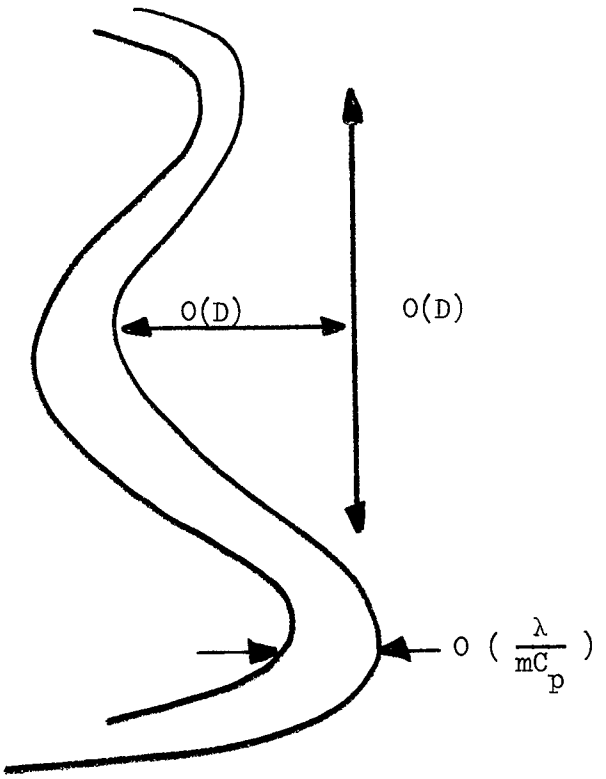



Fig. 8. Allowable Three-Dimensional Disturbances

The time scale is the same long time scale as before.

The result for the flame speed H is[+]

$$(1+\phi_\infty)^2 \left\{ 2 \ln \left(\frac{\partial \chi}{\partial \tau}\right) - \ln \left[1 + \left(\frac{\partial \chi}{\partial \eta}\right)^2 + \left(\frac{\partial \chi}{\partial \xi}\right)^2\right] \right\} \frac{\partial \chi}{\partial \tau}$$

$$+ \frac{b}{\gamma^2} \frac{\partial \gamma}{\partial \tau} + b \left(\frac{\partial^2 \chi}{\partial \eta^2} + \frac{\partial^2 \chi}{\partial \xi^2}\right) = 0 ,$$

$$\gamma = \frac{\partial \chi}{\partial \tau} \left[1 + \left(\frac{\partial \chi}{\partial \eta}\right)^2 + \left(\frac{\partial \chi}{\partial \xi}\right)\right]^{-1} ,$$

$$H = \frac{\partial \chi}{\partial \tau} .$$

[+]This is actually a limiting result only valid when the heat released by the reaction Q is small compared to the enthalpy of the unburnt mixture. In general a single equation governing the flame speed can not be written down when there are three-dimensional disturbances. Nevertheless, many of the qualitative features of the general result are the same as those of the limiting result. The details are in Buckmaster, J. D. Combustion & Flame (to appear).

If we look for perturbations of the one-dimensional steady flame of the form

$$\chi = \tau + \delta\, e^{a\tau} f(\eta,\zeta) , \qquad \delta << 1$$

$$\frac{\partial^2 f}{\partial \eta^2} + \frac{\partial^2 f}{\partial \zeta^2} + k^2 f = 0$$

then

$$a = \frac{1}{2b} \left[ - 2 (1+\phi_\infty)^2 \pm \sqrt{4 (1+\phi_\infty)^4 + 4 b^2 k^2} \right] .$$

If $k$ vanishes, the quantity in square brackets is either zero or negative so that we recover the earlier result that the flame is unstable if $b < 0$ ($L > 1$). But if $k \neq 0$ there is a positive root irrespective of the sign of $b$, so that the one-dimensional flame is also unstable if $L < 1$.

The problem of flame instability is an interesting and a complicated one. Experiment suggests that sometimes instability destroys a flame, sometimes it merely causes it to flicker, and sometimes bifurcation occurs[+]. Most of these observations are presently unexplained but it is possible that the above results will play a role in throwing light on some of these phenomena. In general we can expect activation energy asymptotics to contribute significantly to our understanding of flame instability. For example, Matkowsky and Sivashinsky[++] claim to have explained cellular flames in this way.

I shall conclude by making some additional remarks about the long time scale that plays such an important role in the unsteady problems discussed above. The point is best illustrated by considering a specific problem.

9. SOLID DEFLAGRATION. The burning of a solid is of fundamental interest in the theory of solid propellant rocket motors, and Fig. 9 shows a classical one-dimensional model. The solid is hot, because of the proximity of the flame, and gives off a combustible mixture which burns within the flame. The flame is propagating to the left relative to the gases but the gases are moving to the right and in the steady state the flame is stationary relative to the solid. The burning rate depends upon the pressure and a classical problem is the determination of the steady state burning rate.

The flame is essentially the same as that analyzed by Bush and Fendell. There are differences in the problems, of course, owing to the different boundary conditions, and a solution of the heat conduction equation has to be constructed in the solid (which is being fed to the right in a flame-fixed frame) but the analysis is straightforward and the results have some connection with experimental reality.[+++]

---

[+] At this point Fig. D.1 (p.78), Fig. D.11 (p.86) and Fig. D.10 (p.85) from Markstein, G. H. Non-Steady Flame Propagation Agardograph No. 75, Macmillan, New York, 1964, were shown.

[++] Private communication.

[+++] See Buckmaster, J. D., Kapila, A. K., & Ludford, G. S. S. Astronautica Acta (to appear).

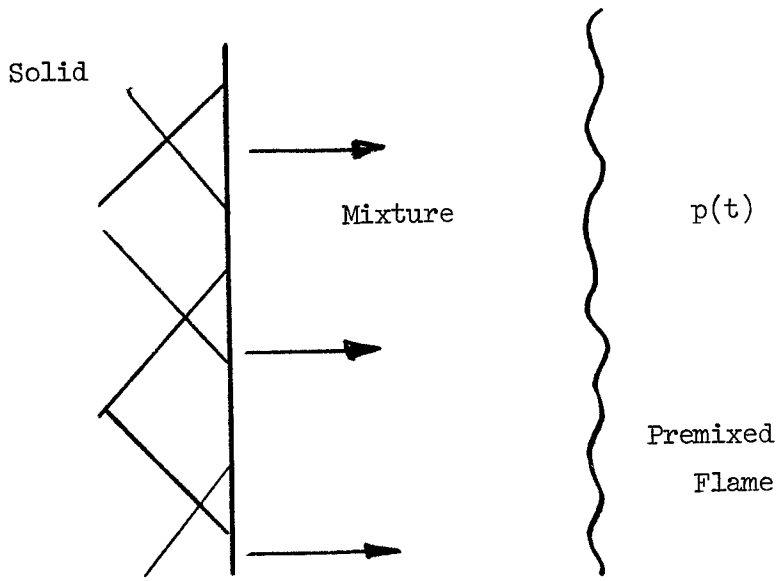



Fig. 9. Burning Solid

A more complicated problem is one for which the pressure varies with time. This also is of interest in the study of solid propellant rocket motors since such motors are often violently unstable. Now if the pressure varies very slowly with time, it is apparent that the response will be quasi-steady. That is, the burning rate will be the steady state value corresponding to the instantaneous value of the pressure. The question then arises: What is the slowest variation in pressure for which there will be a significant lag in the burning response and therefore significant transient effects? The answer is pressures that vary on the long time scale

$$t = O\left(\frac{\theta \lambda \rho_\infty}{m^2 C_p}\right)$$

for these will excite the slowly varying disturbances. Indeed, if the appropriate analysis is carried out we find

$$C_1 \frac{dH}{d\tau} + C_2 H + C_3 \frac{dp}{d\tau} + C_4 p = 0$$

where $H$ is the burning rate, $p$ the pressure, and the $C_j$ are constants. The analysis is inherently a nonlinear one but this is the result for infinitesimal pressure variations.

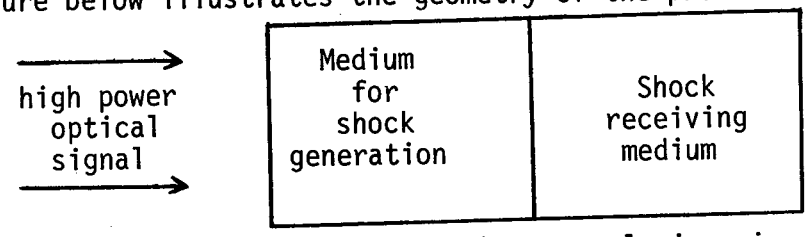
Flames are often subject to external stimuli that change with time and what this example suggests is that provided the steady state solution is known, the unsteady problem can be solved and nontrivial transient effects obtained provided the stimulus changes on the long time scale. This could have application to a variety of important problems.

# A MODEL FOR SHOCK INDUCED STRUCTURAL TRANSFORMATIONS

Paul Harris
Concepts and Effectiveness Division
Nuclear Development and Engineering Directorate
Picatinny Arsenal
Dover, New Jersey 07801

ABSTRACT. The problem of strain propagation in a medium of time and strain (energy) dependent elastic constants is considered. For the elastic constant model considered, analytic and finite difference approximations appear to predict avalanching of the particle velocity in a manner consistent with a dynamic strain induced exothermic structural transformation. The application to enhancement of laser interaction with aerospace materials is discussed.

1. INTRODUCTION. Recent years have seen increasing military interest in the interaction of high power optical signals (lasers) with aerospace materials. A problem of particular interest has been the generation of a shock in an irradiated material in order to produce a dynamic mechanical deformation in an adjacent material. The figure below illustrates the geometry of the problem.




The shock receiving medium could be an explosive, in which case the hardware application might be a detonator or an explosive switch.

For the above type of problem one would obviously like to choose the medium for shock generation so as to maximize the generated shock amplitude. There are essentially two ways in which the shock amplitude can be maximized for a given optical signal: one can maximize the strength of the laser material interaction, or one can attempt to find a generation medium which can act as an amplifier of shock amplitude (the shock being produced in approximately the electromagnetic skin depth of the generation medium). In this paper we will mainly consider some mathematical aspects of the second approach.

2. MATERIAL SELECTION AND PROPERTIES. Some alloys exhibit "anomalously" large Grüneisen parameters as they undergo structural "phase" transformations. Typical alloy examples [1,2] are TiNi and $KTaO_3$. The Grüneisen parameter (proportional to the thermal expansion coefficient) is a measure of the pressure change caused by a change in thermal energy density under constant volume conditions. Since, in the absence of vaporization effects, the laser interaction serves to deposit thermal energy in the skin depth region, an enhanced Grüneisen parameter

is equivalent to an enhanced pressure (shock) amplitude.

TiNi is an appropriate shock generation medium because its metallic properties, even in the absence of Grüneisen effects, serve to produce a small skin depth and thus a strong laser-material interaction. The observed [1] Grüneisen parameter enhancement by a factor of approximately twenty during the near room temperature (martensitic) structural transformation promises enhancement of an already strong laser-material interaction. The practical limitation on the above concept is a $10^{o}$C half maximum width for the spike in the Grüneisen parameter, and that $10^{o}$C temperature rise represents a rather small thermal energy density deposition.

The physics which gives rise to the enhanced Grüneisen parameter also results in exothermic (or endothermic) effects, and different elastic constants on each side of the transition. While the observed[3] exothermicity of approximately 6 Cal/gm is not large, when combined with the observed [1] (approximate) ten percent change in elastic constants, one has a material which promises interesting thermo-mechanical effects. That interest is further raised by the knowledge [3,4] that an applied strain can trigger the transformation.

We thus have a scenario in which a propagating strain wave (shock) can trigger a structural transformation, and thus be amplified in the process. It is that secnario which we will now model and treat below.

3. STRAIN PROPAGATION IN A TRANSFORMING MEDIUM. While there exists a number [2,5] of elegant approaches to the physics of structural phase transitions, those approaches do not yet appear capable of treating the propagating strain condition of interest here. We thus proceed somewhat intuitively.

Consider a one-dimensional strain problem (particle displacement only in the direction of strain propagation) characterized by

$$\rho_o \frac{\partial^2 u}{\partial t^2} - c \frac{\partial^2 u}{\partial x^2} = f \qquad (1)$$

$$c = c_o + \alpha (c_1 - c_o), \qquad (2)$$

where $\rho$ is mass density, u is particle displacement, c is an elastic constant, f denotes a viscosity functional, the subscript zero denotes the undisturbed (prestrain and pretransformation) medium, the subscript one dentoes a final state (transformed) parameter, and $\alpha$ is dependent upon the degree of transformation.

We model $\alpha$ in the form

$$\alpha = \left[1 - \exp\left\{-\left(\frac{\omega}{W}\right)\left(\frac{t}{\tau}\right)\right\}\right], \qquad (3)$$

where $\omega$ is strain energy, W a constant, and $\tau$ is a transformation incubation time. Thus, to first order in $(\omega t)$

$$\alpha = \left(\frac{\omega t}{W\tau}\right), \tag{4}$$

$$\rho_o \frac{\partial^2 u}{\partial t^2} - c_o \left[1 + \left(\frac{c_1}{c_o} - 1\right) \frac{\omega t}{W\tau}\right] \frac{\partial^2 u}{\partial x^2} = f. \tag{5}$$

We will set f = o even though it is known [1] that attenuation is very strong in the presence of phase transformations. We will thus have to keep in mind that any u(x,t) solutions could in practice be of considerably reduced amplitude.

We will now consider two approximations to Eq. (5). The first will be relatively unphysical, but analytically neat. The second will involve the full form of Eq. (5), but will involve a rough finite difference approach.

APPROXIMATION I: We consider

$$\rho_o \frac{\partial^2 u}{\partial t^2} - c_o (1 + \beta t) \frac{\partial^2 u}{\partial x^2} = 0, \quad \beta = \text{const.} \tag{6}$$

Separating variables with u(x,t) = T(t) X (x) gives

$$\frac{\rho_o}{(1 + \beta t)} \frac{1}{T} \frac{\partial^2 T}{\partial t^2} = - m^2 = \frac{c_o}{X} \frac{\partial^2 X}{\partial x^2}. \tag{7}$$

$$X = X_m \exp \left\{ \pm ix \sqrt{\frac{m^2}{c_o}} \right\}, \tag{8}$$

$$\frac{\partial^2 T}{\partial y^2} + \frac{m^2}{\rho_o \beta^2} y T = 0, \qquad y = (1 + \beta t). \tag{9}$$

Eq. (9) is Airy's equation and its solutions can be written as[6]

$$T_m(t) = A_m U_1(Y,1) + B_m U_2 (Y,1), \tag{10}$$

where

$$Y = \left(\frac{m^2}{\rho_o \beta^2}\right)^{1/3} (1 + \beta t), \tag{11}$$

with $U_1$ and $U_2$ being linearly independent and tabulated[6].

We now let $k = 2\pi/\lambda = m/\sqrt{c_o}$ and evaluate $m^2/\rho_o\beta^2$. For $c_1 = c_o/2$ and $\omega = W$ (i.e. the strain energy taken equal to its critical transformation value)

$$\frac{m^2}{\rho_o\beta^2} = \left(\frac{2\pi}{\lambda}\right)^2 v^2\,(4\tau^2), \tag{12}$$

where v is the velocity of sound in the preshocked medium, $\lambda$ is the wavelength of the applied strain disturbance, and $c_1 = c_o/2$ corresponds[3] to an exaggeration of the transition (exothermic) from TiNi (II) to TiNi (III). And using $2\pi v = \omega_o\lambda$, where $\omega_o$ is the angular frequency of the applied disturbance,

$$\frac{m^2}{\rho_o\beta^2} = (2\omega_o\tau)^2 \tag{13}$$

Thus Y becomes

$$Y = (2\omega_o\tau)^{2/3}\left(1 - \frac{t}{2\tau}\right). \tag{14}$$

For a particular $\omega_o$ we can drop the subscript m in Eq. (10) and write

$$u(o,o) = AU_1\left[(2\omega_o\tau)^{2/3},\ 1\right] + BU_2\left[(2\omega_o\tau)^{2/3},\ 1\right], \tag{15}$$

and

$$\left.\frac{\partial u(o,t)}{\partial t}\right|_{t=o} = -\frac{A}{2\tau}(2\omega_o\tau)^{2/3}\,U_1'\left[(2\omega_o\tau)^{2/3},\ 1\right] + \tag{16}$$

$$-\frac{B}{2\tau}(2\omega_o\tau)^{2/3}\,U_2'\left[(2\omega_o\tau)^{2/3},\ 1\right]$$

If we now make the typical "hydrodynamic" approximation of $\omega_o\tau \ll 1$, then from Eqs. (15) and (16)

$$A = u(o,o), \tag{17}$$

$$B = -2(2\omega_o\tau)^{-2/3}\left.\frac{\partial u(o,t)}{\partial t}\right|_{t=o}\tau \simeq -(2\omega_o\tau)^{1/3}\,u(o,o), \tag{18}$$

where [6]

$$U_1(0,1) = 1, \qquad U_2(0,1) = 0 \tag{19a}$$

$$U_1'(0,1) = 0, \qquad U_2'(0,1) = 1 \tag{19b}$$

have been used.

We can thus write

$$u(o,t) \cong u(o,o)\, U_1(Y,1) - (2\omega_o\tau)^{1/3}\, u(o,o)\, U_2(Y,1). \tag{20}$$

We thus predict avalanching of the particle displacement at the boundary, $u(o,t)$, due to the avalanching behavior of $U_1(Y,1)$. The avalanching is strong as it is occuring even in the presence of a harmonic input.

APPROXIMATION II. Here we will consider a crude finite difference version of Eq. (5) written with respect to an almost constant velocity coordinate system.

From Eq. (5)

$$\frac{\partial^2 u}{\partial t^2} = v_o^2\, \frac{\partial^2 u}{\partial x^2} + (v_1^2 - v_o^2) \left\{ \frac{t \left(\frac{\partial u}{\partial x}\right)^2}{\tau\, \eta_o^2} \right\} \frac{\partial^2 u}{\partial x^2}\,, \tag{21}$$

where $v_o^2 \equiv c_o/\rho_o$, $v_1^2 \equiv c_1/\rho_o$, $\omega \equiv M \left(\frac{\partial u}{\partial x}\right)^2$, and $W \equiv M\eta_o^2$ with $\eta_o$ being a critical strain value.

Employing the so-called [7] characteristic stretching transformation

$$\xi \equiv x - Vt, \qquad \zeta \equiv \alpha V t, \tag{22}$$

where $\alpha$ is a dimensionless stretching parameter (we shall neglect terms in $\alpha^2$), and defining

$$\Psi(\xi,\zeta) \equiv \frac{\partial u}{\partial \xi}\,, \tag{23}$$

we arrive at

$$2\alpha V^2 \Psi_\zeta + \left[ (v_o^2 - V^2) + (v_1^2 - v_o^2) \left\{ \frac{\zeta \Psi^2}{\zeta_o \eta_o^2} \right\} \right] \Psi_\xi = o, \tag{24}$$

where $\zeta_o \equiv \alpha V \tau$.

Writing Eq. (24) in crude finite difference form

$$\left[\frac{\Psi(\xi_n, \zeta_{m+1}) - \Psi(\xi_n, \zeta_m)}{\delta}\right] = \tag{25}$$

$$= \left[F_1 - F_2\zeta_m\Psi^2(\xi_n, \zeta_m)\right]\left[\frac{\Psi(\xi_{n+1}, \zeta_m) - \Psi(\xi_n, \zeta_m)}{\Delta}\right]$$

where

$$F_1 \equiv \frac{V^2 - v_o^2}{2\alpha V^2}\ , \tag{26a}$$

$$F_2 \equiv \frac{v_1^2 - v_o^2}{2\alpha V^2 \zeta_o n_o^2}\ . \tag{26b}$$

Rewriting Eq. (25) with the time derivative single-stepped backwards gives

$$\Psi(n,m) - \Psi(n,m-1) = \left[G_1 - G_2\zeta_m\Psi^2(n,m)\right]\Psi(n+1,m) +$$

$$- \left[G_1 - G_2\zeta_m\Psi^2(n,m)\right]\Psi(n,m)\ , \tag{27}$$

where

$$G_1 \equiv \frac{\delta}{\Delta}F_1, \qquad G_2 \equiv \frac{\delta}{\Delta}F_2\ . \tag{28}$$

$$\therefore \Psi(n+1,0) = \left[\frac{1+G_1}{G_1}\right]\Psi(n,o) - \frac{\Psi(n,-1)}{G_1} \tag{29}$$

We now set $\Psi(n,-1)=0$ (equivalent to turning the strain on at t=0, and/or completely neglecting the stretching parameter). With that condition Eq. (29) has a solution

$$\Psi(n,o) = \left[\frac{1+G_1}{G_1}\right]^n \Psi(o,o). \tag{30}$$

Eq. (30) predicts a geometrical avalanching (wave) in position, in support of the temporal avalanching of Eq. (20).

Experimentally it is known [8] that the martensitic transformation in Fe-29.5% Ni propagates at a velocity approximately one third $v_o$. If we thus choose V to be that velocity of propagation of the transformation, then $G_1$ is large ($\alpha$ being small) and negative. Thus the spatial avalanching, while present, does not appear to be as strong as the avalanching in time.

4. DISCUSSION. The two approximations considered above hint strongly that shock amplification can occur in the presence of a structural transformation. Considerably more work is necessary, however, before the prediction of an amplification factor is possible.

In closing we will briefly list what we believe to be the promising approaches for future work.

(a) Modeling. The inclusion of microscopic effects (e.g. soft phonon and interatomic potential effects) in the modeling of $\alpha$.

(b) Attenuation. It is conceivable that known strong attenuation during the transformation process could severly limit the predicted amplification. While experimentally [1] determined attenuation factors in TiNi lead us to believe that this is not the case, $f \neq 0$ must be included at least for completeness.

(c) Soliton propagation. The current fad in spatially bounded non linear propagation effects involves soliton [5,9] physics. It is necessary to seek solutions of Eq. (5) from such a point of view.

(d) Finite differencing. It is necessary to refine the work of approximation II.

## REFERENCES


(1) N.G. Pace and G.A. Saunders, Solid State Commun. 9, 331 (1971).
(2) H.H. Barrett, Phys. Rev. 178, 743 (1969).
(3) F.E. Wang et al, J. Appl. Phys. 36, 3232 (1965).
(4) C.M. Jackson et al, National Aeronautics and Space Administration report NASA-SP 5110 (1972).
(5) J.A. Krumhansl and J.R. Schrieffer, Phys. Rev. B 11, 3535 (1975). T. Schneider and E. Stoll, Phys. Rev. B 13, 1216 (1976).
(6) A.D. Smirnov, Tables of Airy Functions and Special Confluent Hypergeometric Functions (Pergamon, New York, 1960).
(7) A. Jeffrey, Int. J. Non-Linear Mech. 6, 669 (1971).
(8) R.F. Bunshah and R.F. Mehl, Trans. AIME - J. Metals, page 1251 September, 1953).
(9) A.C. Scott et al, Proc. IEEE 61, 1443 (1973).

# SOME NEW METHODS FOR SOLVING LINEAR EQUATIONS[f]

Thomas Kailath
Information Systems Laboratory
Department of Electrical Engineering
Stanford University
Stanford, Ca. 94305

ABSTRACT. It takes of the order of $N^3$ operations to solve a set of $N$ linear equations in $N$ unknowns. When the underlying physical problem has some time- or shift-invariance properties, the coefficient matrix is of Toeplitz (or difference or convolution) type and the equations can be solved with $O(N^2)$ operations. We have shown that with any non-singular $N \times N$ matrix, we can associate an integer $\alpha$ between 1 and $N$ such that it takes $O(N^2\alpha)$ operations to invert the matrix. The number $\alpha$ may be small for many non-Toeplitz matrices of physical interest. Some aspects of this result are discussed here, including extensions to continuous-time kernels and integral equations.

1. INTRODUCTION. Problems in many fields lead ultimately to the solution of linear matrix equations

$$Ra = m ,$$

where $R$ is a given $N \times N$ matrix, say, and $m$ is a given $N \times 1$ vector. The number of operations required to solve such an equation, or to find $R^{-1}$, is of the order of $N^3$ (multiplications and additions). This can be prohibitive if $N$ is large (500 or 1000 or 3000, as can arise in many power system or econonometric calculations). For this, and other reasons, we must often try to bring in any special features or structures that may be present in the original physical problem. In many applications

---

[f]This report is a summary of a talk given at the 22nd Conference of Army Mathematicians, Watervliet Arsenal, New York, May 1976. It was based on work done jointly with B. Friedlander, L. Ljung and M. Morf (see the references).

This work was supported by the Air Force Office of Scientific Research, Air Force Systems Command under Contract AF44-620-74-C-0068, and in part the Joint Services Electronics Program under Contract N00014-75-C-0601, and the National Science Foundation under Contract NSF-Eng 75-18952.

we have the property

$$R = [r_{ij}] = [r_{i-j}]$$

That is, the phenomena are invariant to a change in the time- or space-origin (e.g., as with stationary random processes, or homogeneous media, etc.). In this case, the matrix $R$ is said to be a Toeplitz matrix and has the nice feature that its inverse can be found with only $O(N^2)$ multiplications. Moreover the inverse can be computed recursively, i.e., the $N \times N$ inverse can be easily updated to yield the $(N + 1) \times (N + 1)$ inverse, and Toeplitz matrices also have other useful properties.

Unfortunately, most operations on Toeplitz matrices destroy the Toeplitz property. For example, the inverse of a Toeplitz matrix is not Toeplitz, unless the matrix is also lower- or upper-triangular. So also the product of two Toeplitz matrices is not Toeplitz, unless the matrices are also both lower-triangular or both upper-triangular. However, some reflection will show that in various ways one can regard certain matrices as being "less non-Toeplitz" than others, though present solution methods cannot take advantage of this--they require $O(N^2)$ operations in the Toeplitz case, and $O(N^3)$ otherwise.

By a long process of abstraction and simplification of results originally obtained for certain nonlinear differential equations [1], [2], we have been able to show essentially the following (more precise results are stated later): with any invertible $N \times N$ matrix $R$ we can associate an integer $\alpha$, $1 \leq \alpha \leq N$, such that it takes $O(N^2\alpha)$ operations to compute its inverse. The integer $\alpha$ may be called the displacement rank (or index of nonstationarity) of the matrix and has the property that it is low for matrices that are Toeplitz or near to Toeplitz, while it is high for arbitrary matrices. For example,

i) $\alpha = 1$ for $R = L$ or $U$ or $LU$ or $UL$, where $L$ and $U$ denote lower- and upper-triangular Toeplitz matrices.

ii) $\alpha = 2$ for $R = (L + U)$ and $R = (L + U)^{-1}$

iii) $\alpha \leq 4$ for $R = (L_1 + U_1)(L_2 + U_2)$

iv) $\alpha \leq 3$ for $R = [L_1 + U_1 \vdots L_2 + U_2]$

v) $\alpha \leq n$, if $R$ is the covariance matrix of a linear combination of the components of any n-vector wide-sense Markov random process.

In such cases, $O(N^2\alpha)$ can often be significantly less than $O(N^3)$, thus yielding many advantages, not just for solving a given large set of equations, but also for interactive adjustment of the mathematical model (i.e., of $R$ and $m$) based on actual examination of the now-more-easily determined solution $a$.

We shall outline our major results in Section 2, for matrix equations. A similar, and somewhat simpler, development can be carried out for integral equations, as noted in Section 3. Section 4 contains some concluding remarks on possible extensions and generalizations.

2. THE MATRIX CASE. More details can be found in the paper [ 3], but we note the key definitions and results here.

Definition 1. The (+)-displacement rank of an $N \times N$ matrix $R$ is the smallest integer $\alpha_+(R)$ such that we can write

$$R = \sum_{1}^{\alpha_+(R)} L_i U_i$$

for some lower-triangular Toeplitz matrices $\{L_i\}$ and some upper-triangular Toeplitz matrices $\{U_i\}$.

Definition 2. The (-)-displacement rank of an $N \times N$ matrix $R$ is the smallest integer $\alpha_-(R)$ such that we can write

$$R = \sum_{1}^{\alpha_-(R)} \mathcal{U}_i \mathcal{L}_i$$

for some lower-triangular Toeplitz matrices $\{\mathcal{L}_i\}$ and upper-triangular Toeplitz matrices $\{U_i\}$.

Definition 3. Let

$Z$ = the lower-shift matrix

$$= \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 1 & \ddots & \\ & & \ddots & 0 \\ & & & 1 \end{bmatrix}$$

Lemma 1. Computation of Displacement Ranks

$$\alpha_+(R) = \rho\{\nabla(R)\}$$

where

$$\nabla(R) = R - ZRZ', \qquad (1)$$

and

$$\rho\{A\} = \text{the rank of the matrix } A.$$

Also

$$\alpha_-(R) = \rho\{\Gamma(R)\}$$

where

$$\Gamma(R) = R - Z'RZ \ . \tag{2}$$

The proof follows by using the result of Lemma 2.

Lemma 2. Given two column vectors $x, y$ there is one and only one solution of the functional equation

$$\square(R) = xy' \ , \tag{3a}$$

and this is

$$R = L(x)U(y') \ , \tag{3b}$$

where $'$ denotes transpose, $L(x)$ is a lower-triangular Toeplitz matrix whose first column is $x$, and $U(y')$ is an upper-triangular Toeplitz matrix with first row $y'$.

Proof. For uniqueness, note that

$$\square(R_1) = \square(R_2)$$

implies

$$R_1 - ZR_1Z' = R_2 - ZR_2Z'$$

or

$$R_1 - R_2 = Z(R_1 - R_2)Z' \ ,$$

whose only solution is clearly zero.

The rest amounts to verifying that $\square L(x)U(y') = xy'$, which the reader may find amusing to check by direct computation for $3 \times 3$ matrices. ∎

Lemma 1 now follows easily from the observation that

$$R = \sum_1^{\alpha} L(x_i)U(y_i') \iff \square(R) = \sum_1^{\alpha} x_i y_i' \tag{4}$$

We can now state a first simple, but apparently new, result.

Theorem 1.

$$\alpha_-(R^{-1}) = \alpha_+(R) \ . \tag{5}$$

Therefore,

$$R = \sum_{1}^{\alpha_+(R)} L_i U_i \qquad (6a)$$

implies that $R^{-1}$ has the form

$$R^{-1} = \sum_{1}^{\alpha_+(R)} \mathcal{U}_i \mathcal{L}_i \qquad (6b)$$

Proof. We give the simple proof (suggested by S-Y. Kung) because it shows that the result is quite general and depends very little on the nature of the entries of R--for example, they could themselves be matrices. We note that

$$\begin{aligned} \alpha_-(R^{-1}) &= \rho\{R^{-1} - Z'R^{-1}Z\} \\ &= \rho\{(R^{-1} - Z'R^{-1}Z)R\} \\ &= \rho\{I - Z'R^{-1}ZR\} \end{aligned}$$

since rank is unaffected by multiplication by a nonsingular matrix. Now by a well-known matrix result that

$$\rho\{I - AB\} = \rho\{I - BA\}$$

we can continue the above chain as

$$\begin{aligned} \alpha_-(R^{-1}) &= \rho\{I - ZRZ'R^{-1}\} \\ &= \rho\{(I - ZRZ'R^{-1})R\} \\ &= \rho\{R - ZRZ'\} \\ &= \alpha_+(R) \ . \ \blacksquare \end{aligned}$$

Example. If T is a symmetric Toeplitz matrix, then $\alpha_+(T) = 2 = \alpha_-(T)$ since we have the representations

$$\begin{aligned} T &= T_+ \cdot I + I \cdot T_+' \\ &= I \cdot T_+ + T_+' \cdot I \ , \end{aligned}$$

where

$T_+$ = the lower-triangular part of the matrix T.

The fact that

$$\alpha_+(T) = 2 = \alpha_-(T)$$

can also be seen by checking that

$$\Delta(T) = T - ZTZ'$$

$$= \begin{bmatrix} t_1 & t_2 & \cdots & t_N \\ t_2 & & & \\ \vdots & & \bigcirc & \\ t_N & & & \end{bmatrix}, \quad \text{for all } N \geq 2$$

and

$$\Gamma(T) = \begin{bmatrix} & & & t_N \\ & & & t_{N-1} \\ & & \bigcirc & \vdots \\ & & & t_2 \\ t_N & \cdots & t_2 & t_1 \end{bmatrix}, \quad N \geq 2 .$$

Now it turns out to have been well-known in many contexts (see the discussion in [ 4 ]) that there exist two lower-triangular Toeplitz matrices A and B such that

$$T^{-1} = A'A - B'B \tag{7}$$

so that

$$\alpha_-(T) = 2 = \alpha_+(T) . \blacksquare$$

Remark. Notice that the displacement ranks seem to identify a better property of matrices than their being Toeplitz. The class of Toeplitz matrices is not closed under inversion, unlike the $(\pm)$-displacement ranks and the corresponding representations ( 6 ).

Theorem 2. The inverse of an $N \times N$ matrix $R$ can be found with $O(N^2\alpha)$ multiplications, where $\alpha$ is an integer such that $\alpha_+ \leq \alpha \leq \alpha_+ + 2$. This

can be done via certain recursive formulas called the generalized Szegö-Levinson recursions.

The recursions are a bit too complicated to describe here, but we may note that for Toeplitz matrices they are equivalent to the well-known recursions for the Szegö polynomials orthogonal on the unit circle (see, e.g., [5, Ch. 11 ] or [ 6 ]). These were rediscovered in the statistics literature by Levinson [7 ] and by Durbin [ 8 ] for recursively solving the so-called Yule-Walker normal equations [ 9 ].

For other results in the matrix case, we refer to [ 3 ], [ 10]-[11], and instead turn briefly here to an examination of the integral operator case.

3. INTEGRAL EQUATIONS. The Fredholm integral equation of the second kind

$$a(t) + \int_0^T K(t,s)a(s)\,ds = m(t), \quad 0 \le t \le T \tag{8}$$

has been extensively studied, see, e.g., the recent monograph [12]. Except for the handful of cases where explicit analytic solution is possible, the generic technique is to replace the integral equation by some approximating set of $N$ linear equations

$$Ra = m .$$

This can be done in various ways--use of degenerate kernels, projection (Galerkin and collocation) methods, etc. For example in the degenerate kernel method we replace $K(t,s)$ by the function

$$K_N(t,s) = \sum_1^N \phi_i(t)\psi_i(s) \tag{9}$$

for some suitably chosen functions $\{\phi_i(\cdot),\psi_i(\cdot)\}$. In any case, the resulting set of linear equations will in general require $O(N^3)$ operations for their solution and this may be prohibitively large. More significant however is the observation that such approximation methods will generally destroy any nice structure that might have been present in the original problem.

For example, if the kernel was of Toeplitz (also called displacement or convolution) type,

$$K(t,s) = K(t - s) , \text{ say}$$

then in general

$$K_N(t,s) \neq \text{Toeplitz for } N < \infty .$$

This is bad, because the Toeplitz property can be exploited to find a nice solution of the integral equation. Briefly, first define

$H(t,s)$ = the Fredholm resolvent of $K(t,s)$

as the solution of the integral equation

$$H_T(t,s) + \int_0^T H_T(t,r)K(r,s)\,dr = K(t,s) , \quad 0 \le t,s \le T .$$

In operator notation, we can write this as

$$H + HK = K$$

or as

$$(I - H)(I + K) = I$$

which shows that the original equation

$$(I + K)a = m$$

can be resolved as

$$a = (I + K)^{-1}m = (I - H)m$$

or

$$a(t) = m(t) - \int_0^T H_T(t,s)m(s)\,ds .$$

Therefore the basic problem is to find $H(t,s)$. Now even though $K(t - s)$ may be Toeplitz, this will not in general be true of its resolvent $H_T(t,s)$ (for $T < \infty$). Nevertheless $H_T(t,s)$ is not a completely arbitrary kernel, but should in some sense be close to a Toeplitz kernel (after all, its resolvent is Toeplitz).

We can quantify this intuitive feeling in the following way (the analog of the method used in Section 2). Define the operator

$$\nabla K(t,s) = \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right)K(t,s) , \tag{10}$$

and note that

$$\nabla K(t - s) \equiv 0 .$$

If $K(t,s)$ is not Toeplitz $\nabla K(t,s) \not\equiv 0$, but it will be some function of two variables, which we can write as

$$\Delta K(t,s) = \sum_{1}^{\alpha} \phi_i(t)\psi_i(s) \tag{11}$$

for some functions $\{\phi_i(\cdot),\psi_i(\cdot)\}$ and some integer $\alpha$, possibly even infinite. However let us define the displacement rank of $K(t,s)$ as the smallest integer $\alpha(K)$ such that the representation (11) is possible.

Examples. i) $K$ is Toeplitz, $\alpha(K) = 0$.

ii) $K(t,s) = \min(t,s)$, the covariance of the simplest nonstationary random process, the Wiener process. Clearly $\Delta K(t,s) \equiv 1$ and $\alpha = 1$.

iii) $K(t,s) = ts - \min(t,s)$, the covariance of the so-called Brownian bridge process. Now $\Delta K(t,s) = s + t - 1$ and $\alpha = 2$. ■

We can show the following result, analogous to Theorem 1 in the matrix case.

Theorem 3. $\alpha(H_T(t,s)) \leq \alpha(K(t,s) + 2$.

Example. When $K$ is Toeplitz, $\alpha(K) = 0$. However even though its resolvent $H_T(t,s)$ is not Toeplitz, there exist two functions $A_T(\cdot)$, $B_T(\cdot)$ such that

$$\Delta H_T(t,s) = A_T(t)A_T(s) - B_T(t)B_T(s), \tag{12}^{\dagger}$$

so that

$$\alpha(H_T(t,s)) = 2 .$$

Moreover the functions $A_T(\cdot)$ and $B_T(\cdot)$, of one variable, can be determined more easily than functions of two variables. In fact they can be obtained via the differential equations

$$\left(\frac{\partial}{\partial T} + \frac{\partial}{\partial t}\right)A_T(t) = - B_T(t)B_T(T) , \quad 0 \leq t \leq T \tag{13a}$$

$$\frac{\partial}{\partial T} B_T(t) = - A_T(t)B_T(t) , \quad 0 \leq t \leq T \tag{13b}$$

with certain easily determined boundary conditions $A_T(0)$ and $B_T(T)$.

---

[†] This is the analog of (7) in the matrix case.

The point is that these differential equations can be solved by a simple recursive procedure, which needs only proportional to $N^2$ operations, where $N$ is the number of points in $[0,T]$ used in any discretization procedure.

We call (13) Krein-Szegö-Levinson equations because they are exactly the recursions found by Krein [13] for the continuous analogs of the Szegö polynomials on the unit circle.

Theorem 4. If $K(t,s)$ has displacement rank $\alpha$, $H_T(t,s)$ can be found with $\alpha$ times as much computation as in the Toeplitz case. The solution is found recursively via a set of generalized Krein-Szegö-Levinson equations.

Proofs and further results can be found in the papers [14]-[15]. However, we might draw explicit attention to the fact that though we are using a degenerate-kernel representation in (11), this is for $\dot{\wedge}K(t,s)$ and not for $K(t,s)$. Even though $\dot{\wedge}K(t,s)$ is degenerate it can be seen by integration that, in operator notation,

$$K = \sum_{1}^{\alpha+1} L_i U_i$$

where the $\{L_i\}$ and $\{U_i\}$ are lower- and upper-Volterra operators. Therefore $K$ can be very far from a degenerate kernel. The feature of our method is that it preserves any "Toeplitz-like" structure that may be present in $K(t,s)$. This thought is pursued a bit further in Section 4.

4. CONCLUDING REMARKS. We have taken Toeplitz kernels as basic because they, or things close to them, arise in many applications of interest to us. However in other problems, other "nice" kernels may be more basic. For example, we might have Hankel kernels

$$K(t,s) = K(t + s) \text{ , say .}$$

Integral equations with such kernels can be solved efficiently, and therefore it may be of interest to classify kernels in terms of their degree of "non-Hankelness". This can clearly be done as above by using the operator

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial s}\right)K(t,s) \text{ ,}$$

which gives zero when applied to Hankel kernels. Similar results can also be obtained for basic kernels of the form $K_1(t - s) + K_2(t + s)$.

Furthermore we could also define "second" and higher-order operators of the type

$$\mathcal{J}^2\{K(t,s)\} = \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right)^2 K(t,s)$$

and so on. It is easy to find examples where these are particularly appropriate.

As a final comment, we should express our feeling that the basic ideas described above should be adaptable to a variety of different situations. Also there is clearly some quite general algebraic structure lurking behind our results, which some of the people in this audience may be better equipped to identify than we can.

## REFERENCES


1. T. Kailath, "Some New Algorithms for Recursive Estimation in Constant Linear Systems," IEEE Trans. on Inform. Theory, vol. IT-19, pp. 750-760, November 1973.

2. T. Kailath and L. Ljung "A Scattering Theory Framework for Fast Algorithms," Proceedings 4th International Symposium on Multivariate Analysis, Dayton, Ohio, June 1975. To appear as Multivariate Analaysis - IV, ed. by P. R. Krishnaiah, Amsterdam: North Holland, 1976.

3. B. Friedlander, M. Morf, T. Kailath and L. Ljung, "New Inversion Formulas for Matrices Classified in Terms of Their Distance from Toeplitz Matrices," submitted to the IEEE Decision and Control Conference, 1976; also submitted to the SIAM Journal on Applied Mathematics.

4. T. Kailath, A. Vieira and M. Morf, "Inverses of Toeplitz Operators, Innovations, and Orthogonal Polynomials," to appear SIAM Review; see also Proc. 1975 IEEE Decision & Control Conference, pp. 749-754, Houston, Texas, December 1975.

5. G. Szegö, Orthogonal Polynomials, Providence, R.I.: Amer. Math. Soc., 1939.

6. U. Grenander and G. Szegö, Toeplitz Forms and Their Applications, Berkeley, Ca.: University of California Press, 1958.

7. N. Levinson, "The Wiener RMS (Root-Mean-Square) Error Criterion in Filter Design and Prediction," J. Math. Phys., vol. 25, pp. 261-278, January 1947.

8. J. Durbin, "The Fitting of Time-Series Models," Rev. Intern. Statist. Inst., vol. 28, pp. 229-249, 1959.

9. T. W. Anderson, The Statistical Analysis of Time-Series, New York: J. Wiley, 1971.

10. B. Friedlander, T. Kailath, M. Morf and L. Ljung, "Levinson- and Chandrasekhar-Type Equations for a General Discrete-Time Linear Estimation Problem," submitted for publication.

11. B. Friedlander, "Scattering Theory and Linear Least Squares Estimation," Ph.D. Dissertation, Dept. of Electrical Engineering, Stanford University, Stanford, Ca., August 1976.

12. K. E. Atkinson, A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind, Philadelphia, Pa.: SIAM, 1976.

13. M. G. Krein, "The Continuous Analogues of Theorems on Polynomials Orthogonal on the Unit Circle," Dokl. Akad. Nauk SSSR, vol. 104, pp. 637-640, 1955.

14. T. Kailath, L. Ljung and M. Morf, "A New Approach to the Determination of Fredholm Resolvents of Non-Displacement Kernels," submitted for publication.

15. T. Kailath, L. Ljung and M. Morf, "Recursive Input-Output and State-Space Solutions for Continuous-Time Linear Estimation Problems," submitted to the IEEE Decision and Control Conference, 1976; also submitted to the IEEE Trans. on Automatic Control.

# AN EXACT SOLUTION TO AN ELASTIC-PLASTIC DEFORMATION PROBLEM IN A RADIALLY STRESSED ANNULAR PLATE

Peter C. T. Chen
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, NY 12189

ABSTRACT. An exact solution to the small strain contained plastic deformation problem in an annular plate under internal pressure is obtained on the basis of the deformation theory of Hencky, the Mises yield criterion and a modified Ramberg-Osgood law. Expressions for the stresses, strains and displacement are given. Some numerical results have been worked out and assessed by using the Budianky's criterion for the acceptability of the deformation theory.

1. INTRODUCTION. The problem is a partly plastic, annular plate radially stressed by uniform pressure. The material is assumed to be elastic-plastic and obeying the Mises yield condition. For ideally plastic materials, the stress solution for this problem was first obtained by Mises [1] and the corresponding two strain solutions were recently obtained by the present author on the basis of both $J_2$ deformation and flow theories [2]. The numerical results obtained by using these two theories indicate that the strain differences are very small and compressibility of the material should be considered. However, there is no published solution for strain-hardening materials, which is the purpose of the present investigation.

In the present paper, an exact elastic-plastic solution for strain-hardening materials is given on the basis of $J_2$ deformation theory together with a modified Ramberg-Osgood law [3]. Exact solutions based on this particular model were given recently to an infinite sheet having a circular hole under uniform external tension [3] and internal pressure [4]. This paper considers annular plates of arbitrary inner and outer radii. Some numerical results are presented and the limitations of the solution are discussed.

2. BASIC EQUATIONS. Assuming small strains and neglecting inertia forces in the axisymmetric state of plane stress, the radial and tangential stresses, $\sigma_r$ and $\sigma_\theta$, must satisfy the equilibrium equation,

$$\sigma_\theta = (\partial/\partial r)\ (r\sigma_r) \quad ; \tag{1}$$

and the corresponding strains, $\varepsilon_r$ and $\varepsilon_\theta$, are given in terms of the radial displacement, $u$, by

$$\varepsilon_r = \partial u/\partial r \quad , \quad \varepsilon_\theta = u/r \quad . \tag{2}$$

We shall assume that the material is elastic-plastic, isotropic, obeying the simple deformation theory and the strains are related to the stresses by

$$\varepsilon_r = E^{-1}(\sigma_r - \nu\sigma_\theta) + (E_s^{-1} - E^{-1})\left(\sigma_r - \frac{1}{2}\sigma_\theta\right) \tag{3}$$

$$\varepsilon_\theta = E^{-1}(\sigma_\theta - \nu\sigma_r) + (E_s^{-1} - E^{-1})\left(\sigma_\theta - \frac{1}{2}\sigma_r\right) , \tag{4}$$

where $E$, $\nu$ are elastic moduli and $E_s$ is the secant modulus on the effective stress-strain curve with $E_s = \sigma/\varepsilon$ and

$$\sigma = (\sigma_r^2 + \sigma_\theta^2 - \sigma_r\,\sigma_\theta)^{1/2} \quad . \tag{5}$$

If a modified uniaxial relation of the Ramberg-Osgood type is assumed [3, 4], we have

$$E_s^{-1} = E^{-1} \text{ for } \sigma \lesssim \sigma_y;\ E_s^{-1} = E^{-1}(\sigma/\sigma_y)^{n-1} \text{ for } \sigma \gtrsim \sigma_y \tag{6}$$

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 $\varepsilon_z$ can be determined by

$$\varepsilon_r + \varepsilon_\theta + \varepsilon_z = E^{-1}(1-2\nu)(\sigma_r + \sigma_\theta) \quad , \tag{7}$$

which holds in the elastic as well as plastic region.

The boundary conditions on the problem are

$$\sigma_r(a, t) = -P \ , \quad \sigma_r(b, t) = 0 \quad . \tag{8}$$

Where a, b and P are the inner, outer radii and internal pressure, respectively. 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

$$\alpha = a/b, \quad \xi = r/b, \quad \beta = \rho/b, \quad p = P/\sigma_y \quad ,$$

$$S_r = \sigma_r/\sigma_y, \quad S_\theta = \sigma_\theta/\sigma_y, \quad S = \sigma/\sigma_y \quad ,$$

$$e_r = E\varepsilon_r/\sigma_y \ , \quad e_\theta = E\varepsilon_\theta/\sigma_y, \quad e_z = E\varepsilon_z/\sigma_y \quad . \tag{9}$$

where $r = \rho$ locates the elastic-plastic boundary.

3. ELASTIC REGION. For small pressure ($p \leq p^*$), the plate will be elastic throughout ($\alpha \leq \xi \leq 1$) and the solution is

$$\left.\begin{matrix} S_r \\ S_\theta \end{matrix}\right\} = p(\alpha^{-2}-1)^{-1}\,(1 \mp \xi^{-2}) \quad ,$$

$$\left.\begin{matrix} e_r \\ e_\theta \end{matrix}\right\} = p(\alpha^{-2}-1)^{-1}[(1-\nu) \mp (1+\nu)\xi^{-2}] \quad .$$

$$e_z = -2p\,(\alpha^{-2}-1)^{-1}\,\nu \quad . \tag{10}$$

The critical value $p^*$ to cause incipient deformation is

$$p^* = (1-\alpha^2)\,[3+\alpha^4]^{-1/2} \tag{11}$$

For values of p larger than $p^*$, the plate becomes plastic in the inner region ($\alpha \leq \xi \leq \beta$) and is 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

$$\left.\begin{matrix} S_r \\ S_\theta \end{matrix}\right\} = (1 \mp \xi^{-2})/(1+3\beta^{-4})^{1/2} \quad ,$$

$$\left.\begin{matrix} e_r \\ e_\theta \end{matrix}\right\} = [(1-\nu) \mp (1+\nu)\xi^{-2}]/(1+3\beta^{-4})^{1/2} \quad ,$$

$$e_z = -\,2\nu/(1+3\beta^{-4})^{1/2} \quad . \tag{12}$$

4. PLASTIC REGION ($\alpha \leq \xi \leq \beta$, $p^{**} \geq p \geq p^*$). Following Nadai for isotropic problems [5], we introduce the parametric representation ($0 \leq \phi \leq \pi/2$)

$$S_r = -S\cos\phi/\sin(\pi/3)$$

$$S_\theta = -S\cos(\phi + 2\delta)/\sin(\pi/3) \tag{13}$$

which satisfies equation (5) identically and leads to the following equation upon substituting into the equation of equilibrium,

$$\xi^{-1}d\xi = [\sin(\pi/3)(\tan(\pi/6) + \tan\phi)]^{-1}(\tan\phi d\phi - S^{-1}dS) \quad . \quad (14)$$

By the extended Mitchell theorem [6], the stress solution for the present problem is independent of $\nu$. So choose $\nu = 1/2$ and then equations (3), (4), (6) and (9) lead to

$$e_r = (S_r - S_\theta/2\ )S^{n-1} \quad ,$$

$$e_\theta = (S_\theta - S_r/2\ )S^{n-1} \quad . \quad (15)$$

The compatibility equation follows from (2) and (9) as

$$e_r = (\partial/\partial\xi)(\xi e_\theta) \quad . \quad (16)$$

Substituting (15) into (16) with the aid of (13), we can obtain

$$\xi^{-1}d\xi = [-\sin(\pi/3)(\cot(\pi/6) + \cot\phi)]^{-1}\ (\cot\phi d\phi + nS^{-1}dS) \quad . \quad (17)$$

Combining (14) and (17) yields

$$S^{-1}dS = (\tan\phi + \tan(\pi/6))/(1 - n\ \tan\phi\ \tan(\pi/6))\ .\ d\phi \quad (18)$$

which can be integrated with the known condition at the elastic-plastic boundary. Since S and $\phi$ are functions of $\xi$ and $\beta$, the notation $S_{\xi\beta} = S(\xi,\beta)$, $\phi_{\xi\beta} = \phi(\xi,\beta)$ are introduced [2]. After some manipulation, the relation between $S_{\xi\beta}$ and $\phi_{\xi\beta}$ is given by

$$S_{\xi\beta} = \left[\frac{n\ \sin\phi_{\beta\beta} - \sqrt{3}\ \cos\phi_{\beta\beta}}{n\ \sin\phi_{\xi\beta} - \sqrt{3}\ \cos\phi_{\xi\beta}}\right]^{\mu} \exp\left[\frac{(n-1)\sqrt{3}}{n^2+3}\ (\phi_{\beta\beta}\ \phi_{\epsilon\beta})\right] \quad (19)$$

where $\mu = (n + 3)/(n^2 + 3)$,

and

$$\tan\phi_{\beta\beta} = (\beta^2/\sqrt{3} + \sqrt{3})/(1-\beta^2) \quad (20)$$

follows from (12) and (13) at $\xi = \beta$.

Substituting (18) into (14) and carrying out the integration with the known condition at the elastic-plastic boundary, we have

$$(\beta/\xi)^2 = F(\phi_{\xi\beta})$$

and

$$F(\phi_{\xi\beta}) = \frac{\sin(\phi_{\xi\beta}+\pi/6)}{\sin(\phi_{\beta\beta}+\pi/6)} \times \left[\frac{n\sin\phi_{\beta\beta} - \sqrt{3}\cos\phi_{\beta\beta}}{n\sin\phi_{\xi\beta} - \sqrt{3}\cos\phi_{\xi\beta}}\right]^{\frac{4n}{n^2+3}}$$

$$\times \exp\left[\frac{(n^2-1)\,3}{n^2+3}(\phi_{\beta\beta} - \phi_{\xi\beta})\right] \qquad (21)$$

from which $\phi_{\xi\beta}$ can be solved as a function of $\xi$ and $\beta$. At the inside surface, $\xi = \alpha$, $\phi = \phi_{\alpha\beta}$, thus the expression relating $\alpha,\beta$ and p can be written parametrically as

$$p = S_{\alpha\beta}\cos\phi_{\alpha\beta}/\sin(\pi/3)$$

$$(\beta/\alpha)^2 = F(\phi_{\alpha\beta}) \quad , \qquad (22)$$

where $S_{\alpha\beta}$ and $F(\phi_{\alpha\beta})$ are given by (19) and (21), respectively. By examining (19) and (21), it can be found that $S_{\alpha\beta}$, p, $\beta/\alpha \to \infty$ as $\phi_{\alpha\beta} \to \phi_o = \tan^{-1}(\sqrt{3}/n)$ for finite n. It should be noted that for the present problem we always have $\phi_{\alpha\beta} \leq \phi_{\xi\beta} \leq \phi_{\beta\beta}$ and $\phi_o \leq \phi_{\alpha\beta} \leq \phi_{\alpha\alpha} \leq \phi_{\beta\beta} \leq \phi_{11} = \pi/2$.

Now we have completed the stress solution which is given by (13), (19), (20), (21) and (22).

The solution for the strains in the plastic region ($\alpha \leq \xi \leq \beta$, $p > p^*$) of an elastic-plastic (finite n) plate can be obtained from (3), (4) and (7), using (6), (9) and the above stress solution. After some manipulation, the equations for the dimensionless strains can be written as

$$e_r = -S_{\xi\beta}^n \sin(\phi_{\xi\beta}+\pi/3) - S_{\xi\beta}\cos(\phi_{\xi\beta}+\pi/3)(\tfrac{1}{2}-\nu)/\sin(\pi/3)$$

$$e_\theta = S_{\xi\beta}^n \sin\phi_{\xi\beta} - S_{\xi\beta}\cos\phi_{\xi\beta}(\tfrac{1}{2}-\nu)/\sin(\pi/3)$$

$$e_z = [S_{\xi\beta}^n - (1-2\nu)S_{\xi\beta}]\cos(\phi_{\xi\beta}+\pi/6) \qquad (23)$$

where $S_{\xi\beta}$ and $\phi_{\xi\beta}$ can be evaluated as functions of $\xi$ and $\beta$ by equations (19), (20) and (21).

5. DISCUSSION OF RESULTS. Since the deformation theory is used, the validity of the above solution should be assessed by applying Budiansky's criterion [7] which requires the following inequality to be satisfied.

$$[(ns^{n-1}-1)/(s^{n-1}-1)]^{1/2} \geq (n \tan\phi - \sqrt{3})/(\sqrt{3} \tan\phi + 1) \quad (24)$$

For any values of n, the ranges of S and $\phi$ over which the inequality may not be valid can be determined. In the present case, the above inequality is satisfied except over a certain range of S and $\phi$ for n>17 [4].

Another limitation of the above solution is due to the small strain assumption. In the case of annular plates with arbitrary ratio of inner to outer radius $\alpha$ , there may exist two types of plastic flow. Full plastic flow with complete yielding may happen for larger values of $\alpha$ . In the case of a flat ring with smaller values of $\alpha$ , it is impossible to obtain complete yielding in it through applying a pressure on its inner boundary. The outer portion of the ring must remain strained elastically and a case of partial plastic flow with thickening will occur. Neither full plastic flow for larger $\alpha$ nor partial plastic flow with thickening for smaller $\alpha$ will be permitted under the assumption of small strain.

Some numerical results have been worked out for the 2219-T87 aluminum plate with geometric ratio b/a = 3. The material constants [4] are n = 9, $\nu$ = 0.3, E = $10.5 \times 10^6$psi, $\sigma y = 5.5 \times 10^4$psi. The effect of $\rho/a$ on the radial and tangential stress distributions are shown in Figures 1 and 2, respectively. The corresponding strain distributions for the radial, tangential and axial components are shown in Figs. 3, 4 and 5, respectively. Finally it should be noted that the validity of the above results based on the deformation theory have been assessed by applying Budiansky's criterion. The range of S and $\phi$ for the above stresses and strains satisfy the inequality (24).

## 6. 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. Budiansky, B., "An Exact Solution to an Elastic-Plastic Stress Concentration Problem," Prikladnaya Mathematika k Physik, Vol. 35, No. 1, 1971, pp. 40-48.

4. Hsu, Y. C., and Forman, R. G., "Elastic-Plastic Analysis of an Infinite Sheet Having a Circular Hole Under Pressure," ASME paper No. 75-APM-15, to be published in Journal of Applied Mechanics, Trans. ASME, Series E.

5. Nadai, A., Theory of Flow and Fracture of Solids, McGraw-Hill, New York, Vol. 1, 1950, Chapter 33.

6. Budiansky, B., "Extension of Mitchell's Theorm to Problems of Plasticity and Creep," Quarterly of Applied Mathematics, Vol. 16, 1958, pp. 307-309.

7. Budiansky, B., "A Reassessment of Deformation Theories of Plasticity," Journal of Applied Mechanics, Vol. 26, Trans. ASME, Vol. 81, 1959, pp. 259-264.

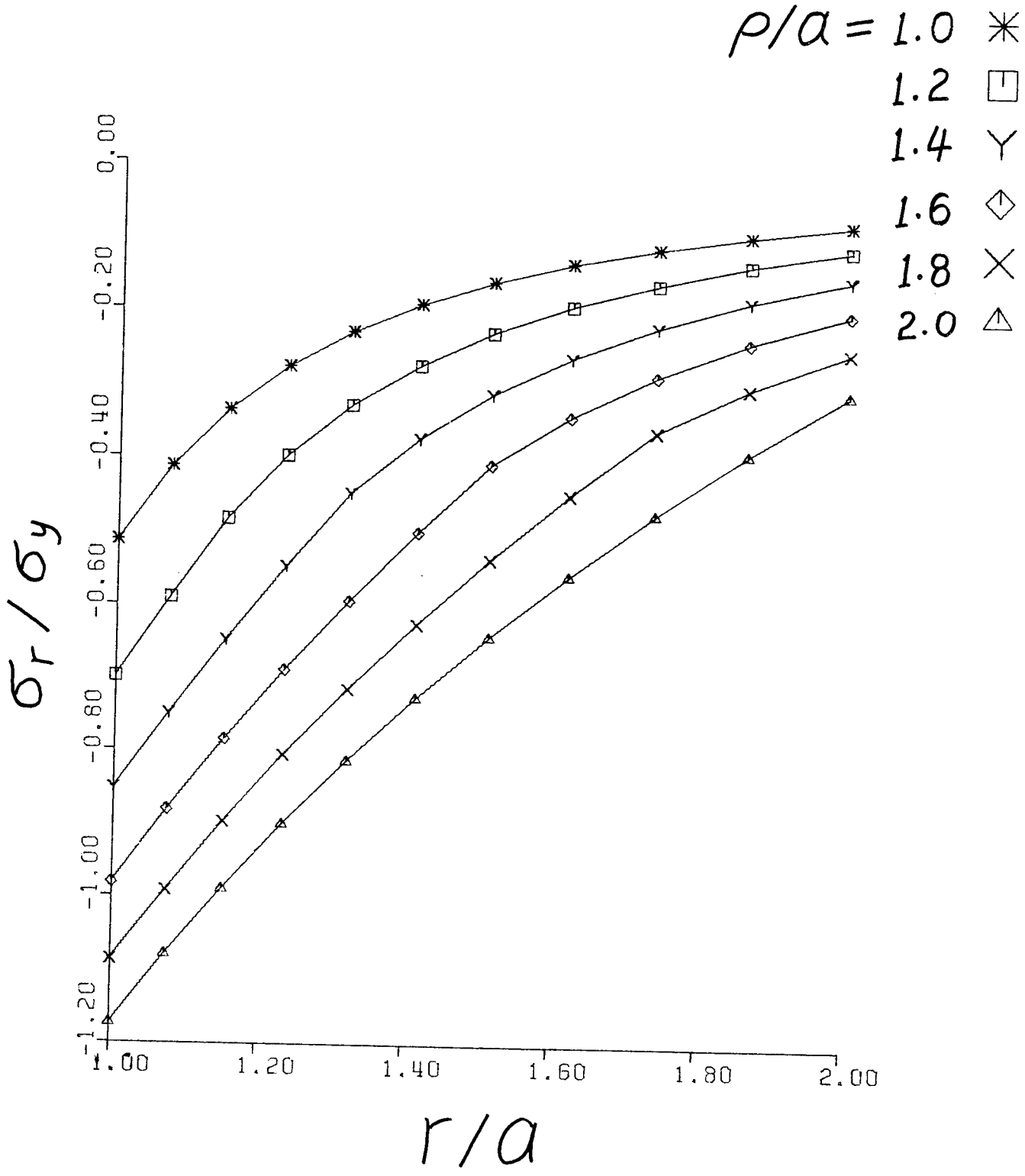



Fig. 1. The Radial Stress Distribution

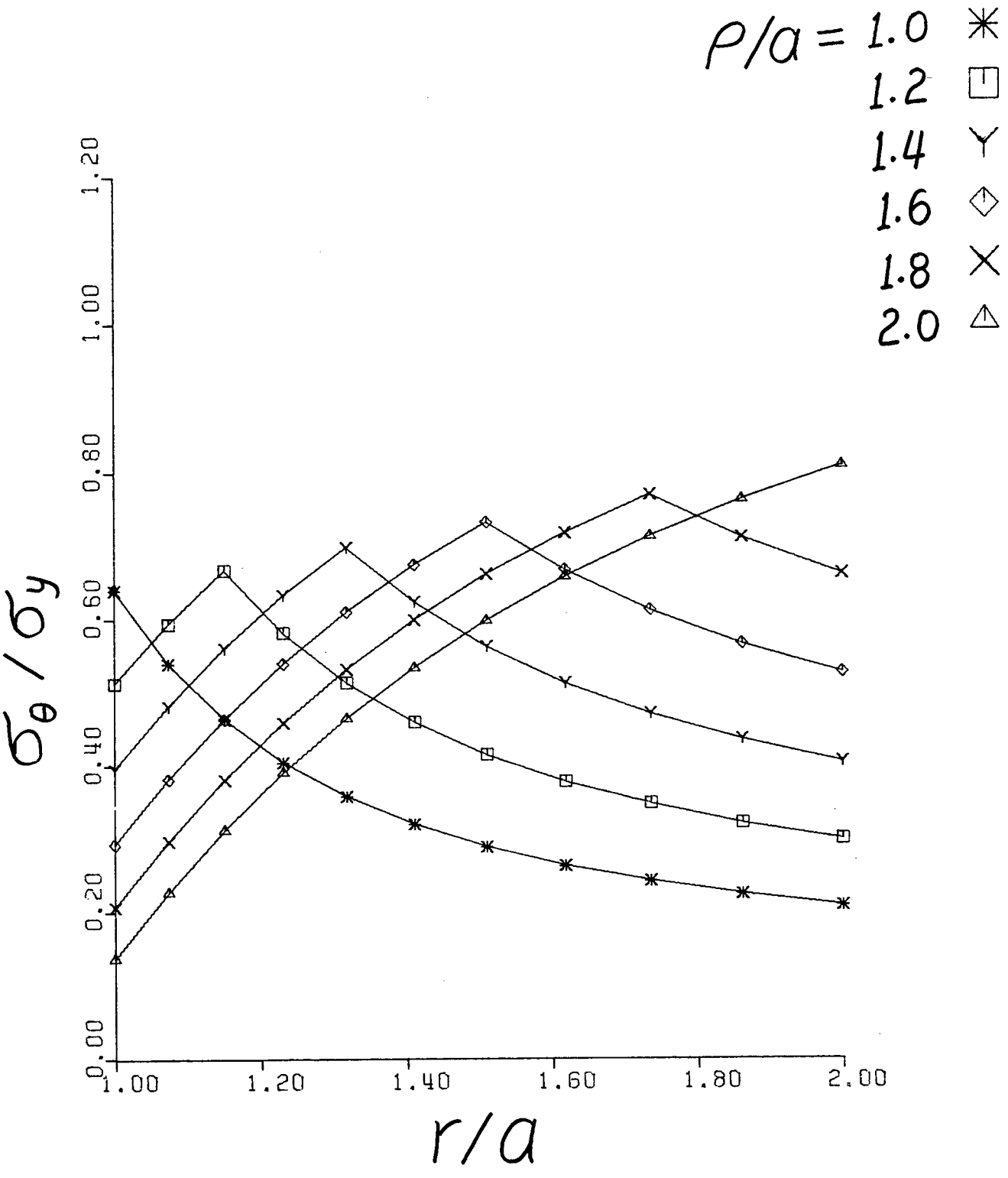



Fig. 2. The Tangential Stress Distribution

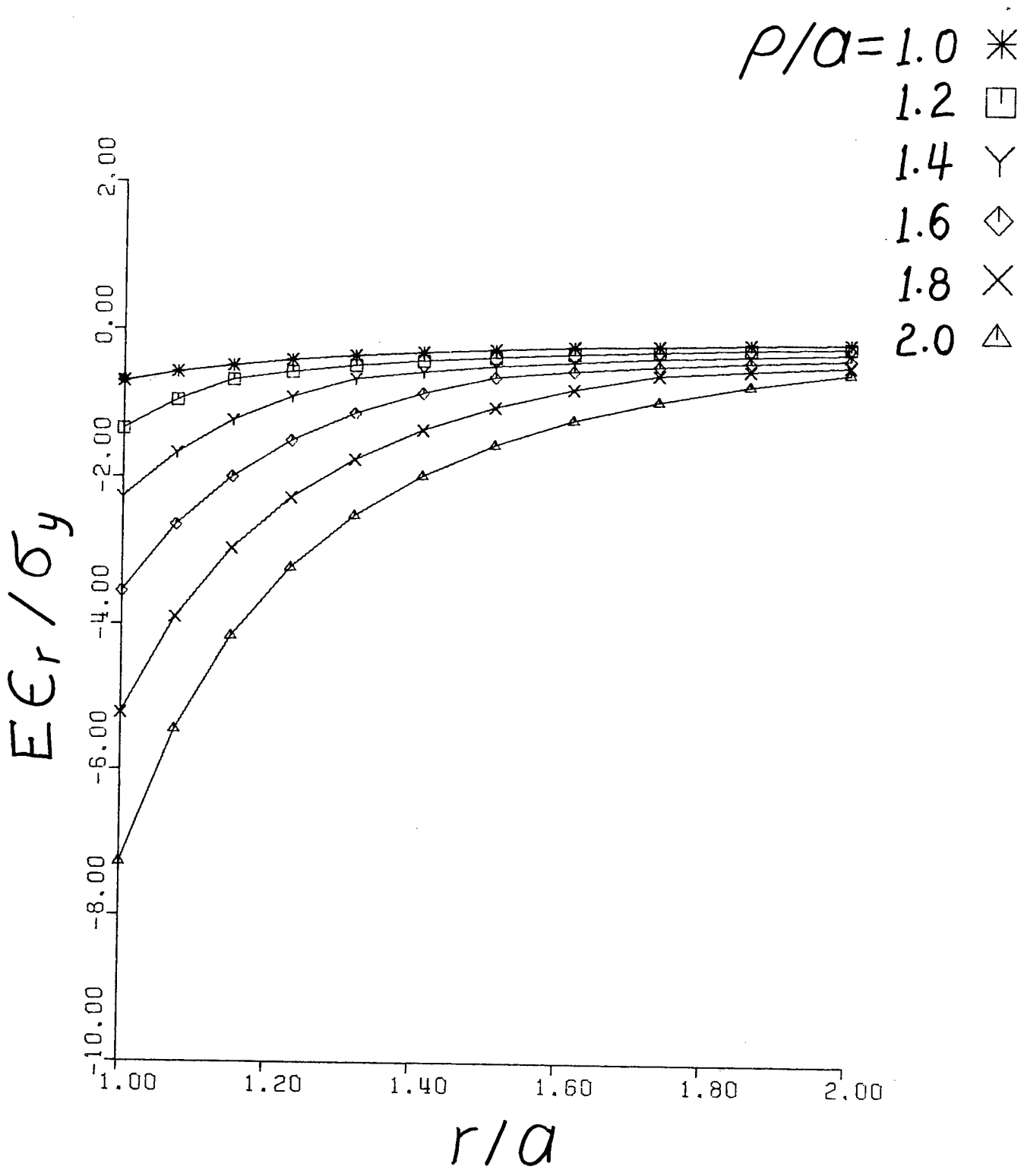



Fig. 3. The Radial Strain Distribution

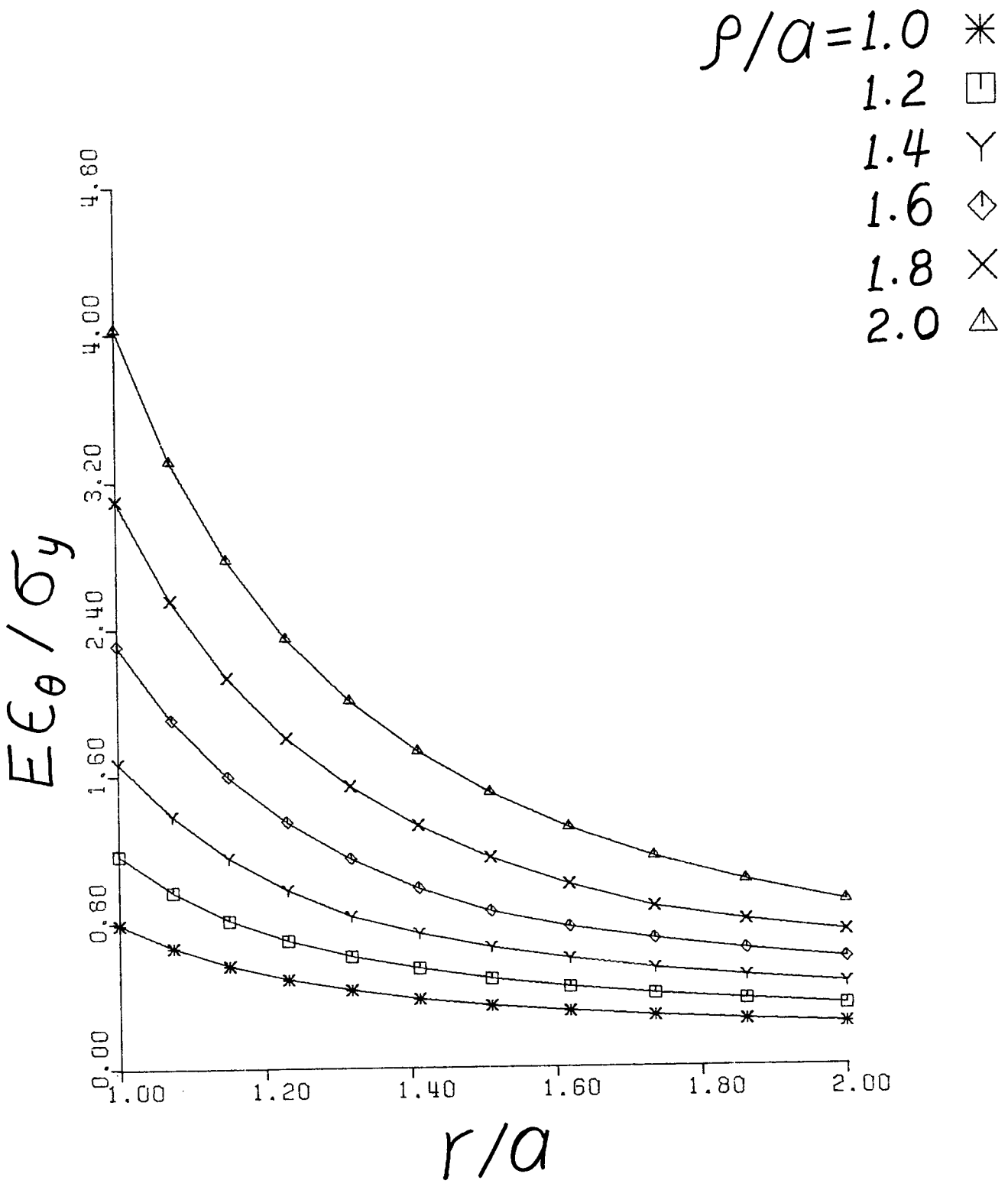



Fig. 4. The Tangential Strain Distribution

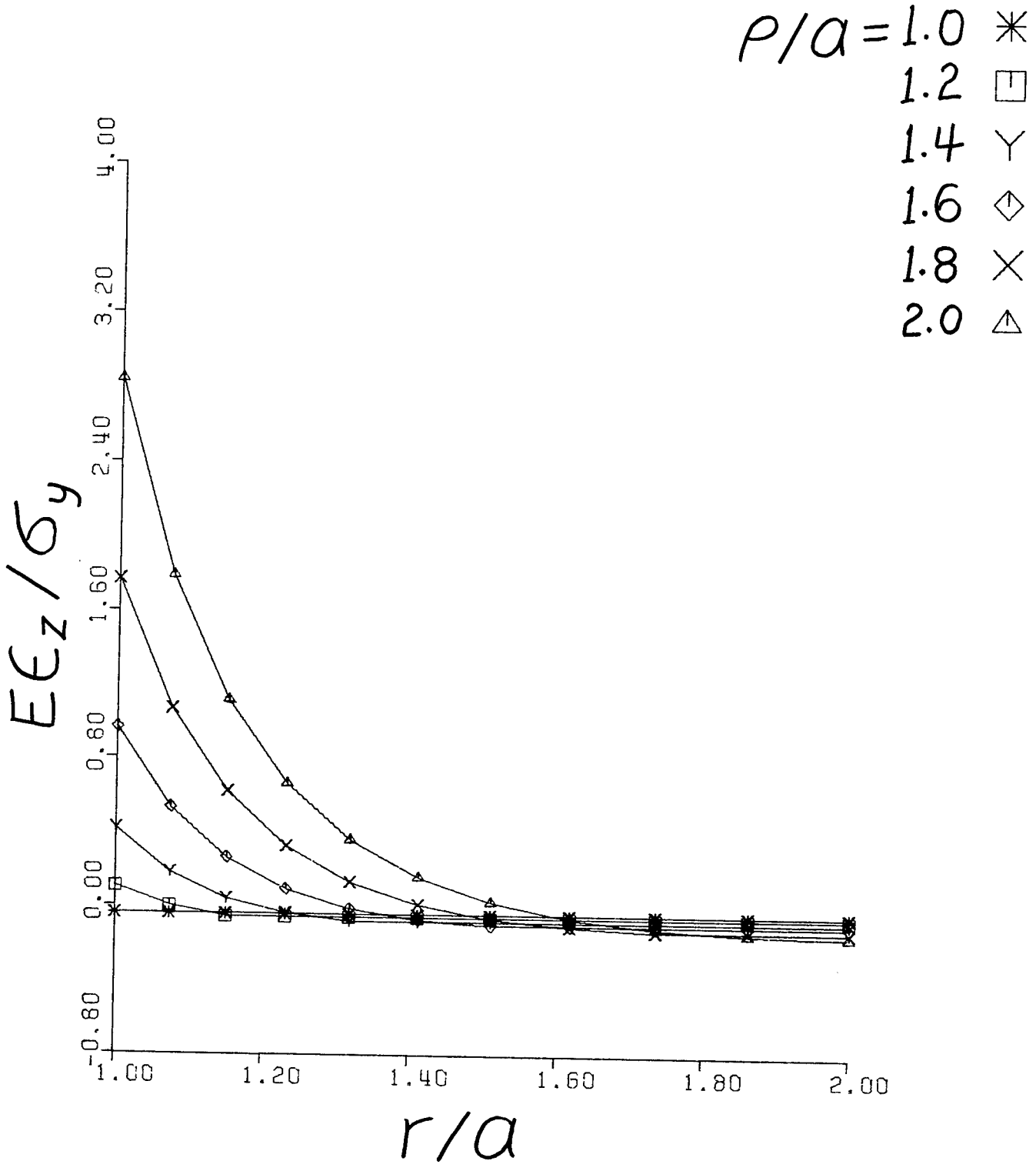



Fig. 5. The Axial Strain Distribution

# AN EFFECTIVE STIFFNESS VISCOELASTIC COMPOSITE BEAM THEORY

Charles R. Thomas
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. Viscoelasticity in the individual beam layers is modeled according to the standard linear model and the Timoshenko beam theory with the resulting equations utilized in deriving a microstructure or effective stiffness viscoelastic laminated beam theory. A time harmonic wave propagation along the length coordinate of the viscoelastic composite beam has been utilized to illustrate an application of the derived theory and to point out the influence of the various viscoelastic and geometric parameters involved.

The first task in deriving the viscoelastic laminated beam theory was to formulate energies for individual viscoelastic layers in terms of the Timoshenko beam theory in a form suitable for developing the composite theory. A goal of the direct derivation of the beam theory, instead of the intermediate step of developing a viscoelastic laminated continuum theory which must then be reduced to a beam theory, was accomplished by the introduction of a gross rotation term for the laminated beam into the derivation of individual layer energy relations. The final result was an energy conservation law for the individual beam layers in terms of kinetic energy, potential energy, and dissipation energy.

The viscoelastic laminated beam is composed of a number of alternating plane, parallel layers of two homogeneous, isotropic viscoelastic materials which are respectively termed the reinforcing layer and the matrix layer. To obtain the total energy for the viscoelastic composite beam, the individual layer kinetic, potential, and dissipation energies were summed over the n layer pairs of which the composite beam was composed. The discrete system thus obtained was then converted to a continuous system by means of a smoothing operation, that is a replacement of the resulting energy summations by weighted integrations over beam thickness. A reduction of one variable from the formulation was made possible through a continuity condition resulting from continuity of displacement across layer interfaces. The final result of the derivational work was a set of three flexure equations of motion and corresponding boundary conditions for viscoelastic laminated composite standard linear model Timoshenko beams.

Time harmonic waves of the form

$$(w,\psi,\phi) = (hW,\Psi,\Phi)e^{-\alpha y}e^{ip(y/c-t)}$$

were passed through the three equations of motion and resulted in a characteristic equation in terms of p, the circular frequency; c, the phase velocity; $\alpha$, the attenuation coefficient; and the numerous viscoelastic and geometric parameters involved.

1. INTRODUCTION. A great deal of work has been accomplished in the area of elastic laminated effective stiffness or microstructure continuum theories and approximate plate and beam theories. By the same token, little has been accomplished with viscoelastic counterparts to these theories.

An elastic continuum theory which included effective stiffness for both the reinforcing and matrix layers of a laminated continuum was developed by Sun, Achenbach, and Herrmann [1, 2]. The continuum theory was utilized by Thomas [3] to study the simple thickness modes for laminated media with layering both parallel and perpendicular to the plate free surfaces. Sun [4] deduced a two dimensional theory for laminated plates from the three dimensional continuum theory. Velocity correction coefficients were introduced into the two dimensional theory by Thomas [5] and flexural and extensional vibrations for plate strips and rectangular plates were studied by Thomas [6, 7] according to this theory and compared to similar results from effective modulus plate theories. A microstructure theory for an elastic, laminated composite beam was developed by Sun [8] and the approach utilized in this paper will be followed in deriving a viscoelastic, laminated composite beam theory. Thomas [9] showed that the flexure beam theory in reference [8] is directly obtainable through a simple reduction of the existing flexure equations for composite plates [4, 5].

A continuum theory for a viscoelastic laminated composite was developed by Grot and Achenbach [10], however the equations developed were not applied to any problems of wave propagation or vibration. It is certainly theoretically possible to start with the equations in reference [10], to make appropriate series expansions and derive a plate theory, and to then follow reference [9] to make a direct reduction to a viscoelastic beam theory. However, for convenience and simplicity of analysis, the approach in the current report will be to begin with the viscoelastic Timoshenko beam equations and work towards a viscoelastic laminated beam equation in the manner of reference [8]. With somewhat guarded conclusions, Stern, Bedford, and Yew [11] have demonstrated a definite need for an effective stiffness type formulation for viscoelastic laminates.

The current approach to obtaining a viscoelastic laminated beam theory will be a viscoelastic development which mirrors the elastic development given by Sun [8]. Surprisingly, the real difficulty is in obtaining the energies for a single layer modeled as a viscoelastic Timoshenko beam. The most pleasing and straightforward development of suitable viscoelastic Timoshenko beams results from a utilization of viscoelastic constitutive relations of the differential form; it is these equations which yield a viscoelastic development which closely mirrors Sun's [8] elastic derivation.

2. THE ENERGY PRINCIPLE. As Sun [8] does in the development of an elastic laminated beam theory, the first task in deriving a viscoelastic laminated beam theory is to formulate energies for individual viscoelastic layers in terms of the Timoshenko [12] beam theory. In the past, Lee [13] developed viscoelastic Timoshenko beam equations for viscoelastic extensional strain but the shear strain was left elastic. Pan [14] extended the analysis to include viscoelastic shear strains. The current objective is to develop the viscoelastic Timoshenko beam equations in a form more suitable to the development of a viscoelastic composite beam theory. A first goal will be the development of a single layer energy principle suitable for a direct application in the derivation of a multilayer energy principle.

The development of an approximate theory such as for laminated elastic plates has originally been a two step procedure. In the first instance, the Mindlin plate theory [15] in its first order approximation was utilized to develop a continuum theory for laminated composites. Then to obtain a laminated plate theory a first order approximation is made on those variables which came from the first order part of the Mindlin theory as in Sun [4] and Thomas [5] - this explanation will become clear shortly. Now in developing an elastic laminated beam theory, Sun [8] has made both of these approximations simultaneously to obtain a flexure theory for laminated beams. Actually, Thomas [9] has shown that the flexure beam theory is directly obtainable from the existing flexure plate theory.

The current objective is to immediately derive a viscoelastic laminated beam theory and to not have to develop a viscoelastic laminated continuum theory first. In making the various zero and first order expansions of displacement, terms which lead to an extension theory are also maintained since the second expansion of extensional displacements leads to a flexure term. The first order displacements which will result in the Timoshenko beam equations [12] for flexure as well as an extensional equation for beams are

$$v(y,z,t) = \bar{v}(y,t) - z\phi(y,t)$$

$$w(y,z,t) = \bar{w}(y,t) - z\Phi(y,t). \qquad (1)$$

the zero order terms in (1) are $\bar{v}$ and $\bar{w}$ and a first order expanison of these two displacements results in the expressions

$$\bar{v}(y,t) = v_\alpha^k(y,t) - z_\alpha^k \psi_\alpha(y,t)$$

$$\bar{w}(y,t) = w_\alpha^k(y,t) - z_\alpha^k \Psi_\alpha(y,t) \qquad (2)$$

where the subscript $\alpha = 1, 2$ will later denote whether a stiff or soft laminated beam layer is indicated and the superscript k which layer pair is indicated. While absolutely necessary at this point, the notation in (2) jumps into the laminate notation while seeming to be at the single layer stage of development. See Sun, Achenbach, and Herrmann [1] or Sun [8] if clarification is required.

Combining equations (1) and (2) and extracting only those terms which result in flexural motion results in the displacement relations

$$v(y,z,t) = -z_\alpha^k \psi_\alpha(y,t) - z\phi(y,t)$$

$$w(y,z,t) = w_\alpha^k(y,t) \qquad (3)$$

where $\psi_\alpha(y,t)$ represents the gross rotation in the laminated beam, $w_\alpha^k(y,t)$ represents the transverse deflection, and $\phi(y,t)$ represents the individual layer rotation. The various displacements and rotations on the right side of (2) represent the reduction from a laminated continuum theory to a laminated beam theory; thus, from continuity of displacement and rotation at laminate interfaces, it is clear that the notation may be simplified to $w(y,t) = w_\alpha^k(y,t)$ and $\psi(y,t) = \psi_\alpha(y,t)$ for $\alpha = 1, 2$ and for all values of k. Hence with these notational simplifications in mind, the final form of the first order flexure displacement expansion is

$$v(y,z,t) = -z_\alpha^k \psi(y,t) - z\phi(y,t)$$

$$w(y,z,t) = w(y,t) \qquad (4)$$

where these equations are valid only when eventually utilized in developing a laminated beam theory. Equations (4) may be reduced to those for a homogeneous or single layered beam by setting $\psi(y,t) = 0$; this being done, equations (4) reduce to those given by Brunelle [16] for flexure of a beam.

The non-zero strain-displacement relations are

$$\varepsilon_y = \frac{\partial v}{\partial y}$$

$$\varepsilon_{yz} = \frac{1}{2}\left[\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right]$$

The non-zero stress equations of motion which pertain to the problem are

$$\sigma_{yz,y} = \rho\ddot{w}$$

$$\sigma_{y,y} + \sigma_{yz,z} = \rho\ddot{v} \qquad (6)$$

From the appendix and equations (A-17) the constitutive equations for a special case of the standard linear model are

$$(1 + C\frac{\partial}{\partial t})\sigma_{yz} = (2kG + 2k^*G^*\frac{\partial}{\partial t})\varepsilon_{yz}$$

$$(1 + C\frac{\partial}{\partial t})\sigma_y = (E + E^*\frac{\partial}{\partial t})\varepsilon_y \qquad (7)$$

where shear correction coefficients k and $k^*$ have now been introduced in a manner similar to that of Timoshenko [12] and Mindlin and Deresiewicz [17].

The procedure involved in deriving the theory will be to manipulate the left sides of equations (6) until they are of the form of the left sides of equations (7). Thus, taking the first time derivatives of (6) and multiplying by the viscoelastic constant C results in the equations

$$C\,\dot{\sigma}_{yz,y} = \rho C\,\dddot{w}$$

$$C\,\dot{\sigma}_{y,y} + C\,\dot{\sigma}_{yz,z} = \rho C\,\ddot{\ddot{v}} \qquad (8)$$

which when added to their counterparts in equation (6) become

$$\sigma_{yz,y} + C\,\dot{\sigma}_{yz,y} = \rho\ddot{w} + \rho C\,\dddot{w}$$

$$\sigma_{y,y} + C\,\dot{\sigma}_{y,y} + \sigma_{yz,z} + C\,\dot{\sigma}_{yz,z} = \rho\ddot{v} + \rho C\,\ddot{\ddot{v}} \qquad (9)$$

Multiplying the first equation of (8) by w and the second equation by v, integrating over the beam volume and time, and finally adding the final answers results in the equation

$$\int_A \int_o^{\ell} \int_o^t \left[ \begin{array}{l} (\sigma_{yz,y} + C\,\dot{\sigma}_{yz,y})\dot{w} + (\sigma_{y,y} + C\,\dot{\sigma}_{y,y})\dot{v} \\ + (\sigma_{yz,z} + C\,\dot{\sigma}_{yz,z})\dot{v} \end{array} \right] dA\, dy\, dt$$

$$= \int_A \int_o^{\ell} \int_o^t \rho \left[ \ddot{v}\,\dot{v} + C\,\ddot{\dot{v}}\,\dot{v} + \ddot{w}\,\dot{w} + C\,\ddot{\dot{w}}\,\dot{w}) \right] dA\, dy\, dt \qquad (10)$$

After several integrations by parts, equation (10) may be expressed as

$$\int_A \int_o^t \left[ (\sigma_{yz} + C\,\dot{\sigma}_{yz})\dot{w} + (\sigma_y + C\,\dot{\sigma}_y)\dot{v} \right]_o^{\ell} dA\, dt$$

$$+ \int_A \int_o^{\ell} \int_o^t \frac{d}{dz} \left[ (\sigma_{yz} + C\,\dot{\sigma}_{yz})\dot{v} \right] dA\, dy\, dt$$

$$- \int_A \int_o^{\ell} \int_o^t \left[ \begin{array}{l} (\sigma_{yz} + C\,\dot{\sigma}_{yz}) \left(\frac{\partial \dot{w}}{\partial y} + \frac{\partial \dot{v}}{\partial z}\right) \\ + (\sigma_y + C\,\dot{\sigma}_y) \frac{\partial \dot{v}}{\partial y} \end{array} \right] dA\, dy\, dt$$

$$= \int_A \int_o^{\ell} \int_o^t \rho \left[ (\ddot{w} + C\,\ddot{\dot{w}})\,\dot{w} + (\ddot{v} + C\,\ddot{\dot{v}})\dot{v} \right] dA\, dy\, dt \qquad (11)$$

it is immediately clear that

$$\int_A \int_o^{\ell} \int_o^t \frac{d}{dz} \left[ (\sigma_{yz} + C\,\dot{\sigma}_{yz})\dot{v} \right] dA\, dy\, dt = 0 \qquad (12)$$

since both beam surfaces are stress free and that

$$\int_A \int_o^t \left[ (\sigma_{yz} + C\,\dot{\sigma}_{yz})\,\dot{w} + (\sigma_y + C\,\dot{\sigma}_y)\dot{v} \right]_o^{\ell} dA\, dt = 0 \qquad (13)$$

since the boundary terms will be satisfied at the beam ends. Applying equations (5) and (7) to equation (11) and taking into account equations (12) and (13) results in

$$\int_A \int_o^{\ell} \int_o^t \begin{bmatrix} (2kG\varepsilon_{yz} + 2k^*G^*\dot{\varepsilon}_{yz}) & (2\dot{\varepsilon}_{yz}) \\ +(E\varepsilon_y + E^*\dot{\varepsilon}_y)\dot{\varepsilon}_y & \end{bmatrix} \quad dA\ dy\ dt$$

$$= \int_A \int_o^{\ell} \int_o^t \rho[(\ddot{w} + C\dddot{w})\dot{w} + (\ddot{v} + C\dddot{v})\dot{v}] \quad dA\ dy\ dt \qquad (14)$$

But, from the chain rule of partial differentiation it is clear that

$$\frac{\partial}{\partial t}[\varepsilon^2] = \varepsilon\dot{\varepsilon} + \dot{\varepsilon}\varepsilon \qquad (15)$$

or that

$$\varepsilon\dot{\varepsilon} = \frac{1}{2}\frac{d}{dt}(\dot{\varepsilon}^2) \qquad (16)$$

Similarly, the fact that an indefinite integral can be defined as a definite integral with a variable upper limit

$$\int g(t)dt = \int_a^t g(t)dt + \text{const.} \qquad (17)$$

immediately results, after taking a time derivative of both sides, in the equation

$$\frac{d}{dt}\int_a^t g(t)dt = g(t) \qquad (18)$$

which for $g(t) = \dot{\varepsilon}^2$ results in the relationship

$$\dot{\varepsilon}^2 = \frac{d}{dt}\int_o^t (\dot{\varepsilon})^2 dt \qquad (19)$$

A direct application of relations (16) and (19) to equation (14) with an introduction of equations (4) and (5) results in the equation

$$\int_0^{\ell}\int_0^t \frac{d}{dt}\left[\begin{array}{l}\frac{1}{2}AkG(\frac{\partial w}{\partial y}-\phi)^2 + Ak^*G^*\int_0^t(\frac{\partial \dot{w}}{\partial y}-\dot{\phi})^2 d\tau \\ + \frac{AE}{2}(z_\alpha^k)^2(\frac{\partial}{\partial y})^2 + AE^*\int_0^t(z_\alpha^k)^2(\frac{\partial \dot{\psi}}{\partial y})^2\, d\tau \\ + \frac{EI}{2}(\frac{\partial \phi}{\partial y})^2 + E^*I\int_0^t(\frac{\partial \dot{\phi}}{\partial y})^2\, d\tau\end{array}\right] dy\, dt$$

$$+\int_0^{\ell}\int_0^t \frac{\rho}{2}\frac{d}{dt}\left[\begin{array}{l}A\dot{w}^2 - 2AC\int_0^t \ddot{w}^2 d\tau + A(z_\alpha^k)^2\ddot{\psi}^2 \\ + I\dot{\phi}^2 - 2AC\int_0^t (z_\alpha^k)^2\ddot{\psi}^2 d\tau \\ - 2IC\int_0^t \ddot{\phi}^2 d\tau\end{array}\right] dy\, dt = 0 \qquad (20)$$

after an integration over the beam area where

$$A = bd \quad , \quad I = \frac{bd^3}{12} \qquad (21)$$

with b being the beam width and d being the beam thickness.

Following Anderson [18], a conservation law is sought in the existance of a quantity H such that

$$H = \text{constant}, \qquad (22)$$

such that obviously

$$\frac{dH}{dt} = 0 \qquad (23)$$

where

$$H = T + U + V \qquad (24)$$

with the quantities T, U, and V being called the kinetic energy, the potential energy, and the dissipation energy. From a comparison of equations (20), (23), and (24) it is clear that the various energies may be defined as

$$T = \int_0^{\ell}\int_0^t T^*\, dy\, dt$$

$$U = \int_0^{\ell} \int_0^{t} U^* \, dy \, dt$$

$$V = \int_0^{\ell} \int_0^{t} V^* \, dy \, dt \tag{25}$$

and from equation (20) it is clear that the energies are

$$T^* = \frac{\rho}{2} [A\dot{w}^2 + A(z_\alpha^k)^2 \dot{\psi}^2 + I\dot{\phi}^2]$$

$$U^* = \frac{1}{2} AkG(\frac{\partial w}{\partial y} - \phi)^2 + \frac{AE}{2}(z_\alpha^k)^2(\frac{\partial \psi}{\partial y})^2 + \frac{EI}{2}(\frac{\partial \phi}{\partial y})^2$$

$$V^* = \int_0^{\tau} \left[ \begin{array}{l} Ak^*G^*(\frac{\partial \dot{w}}{\partial y} - \dot{\phi})^2 + E^*I(\frac{\partial \dot{\phi}}{\partial y})^2 \\ \quad + AE^*(z_\alpha^k)^2(\frac{\partial \dot{\psi}}{\partial y})^2 - \rho AC\, \ddot{w}^2 \\ \quad - \rho AC\, (z_\alpha^k)^2 \ddot{\psi}^2 - \rho IC\, \ddot{\phi}^2 \end{array} \right] d\tau \tag{26}$$

3. THE LAMINATED BEAM THEORY. The laminated beam, Figure 1, is composed of a number of alternating plane, parallel layers of two homogeneous, isotropic viscoelastic materials which are respectively termed the reinforcing layer and the matrix layer. The reinforcing layer is the stiffer of the two layer combination and is indicated by the subscript "1" while the softer matrix layer is indicated by the subscript "2". The elastic constants, the viscoelastic constants, the layer density, and the thickness for the reinforcing and matrix layers respectively are $E_1$, $G_1$, $E_1^*$, $G_1^*$, $C_1$, $\rho_1$, $d_1$, and $E_2$, $G_2$, $E_2^*$, $G_2^*$, $C_2$, $\rho_2$, $d_2$.

The basic variables involved are w, the transverse deflection; $\psi$, the gross rotation of the stiff layer; and $\phi_2$, the rotation of the soft layer. The midplane positions for the kth pair of neighboring reinforcing matrix layers are $y_1^k$ and $y_2^k$ respectively as indicated in Figure 1, with the layer midplanes taken perpendicular to the z-axis. The width of the beam is b and the total or gross thickness is h.

From equation (26), the kinetic, potential, and dissipative energies in the individual layers are

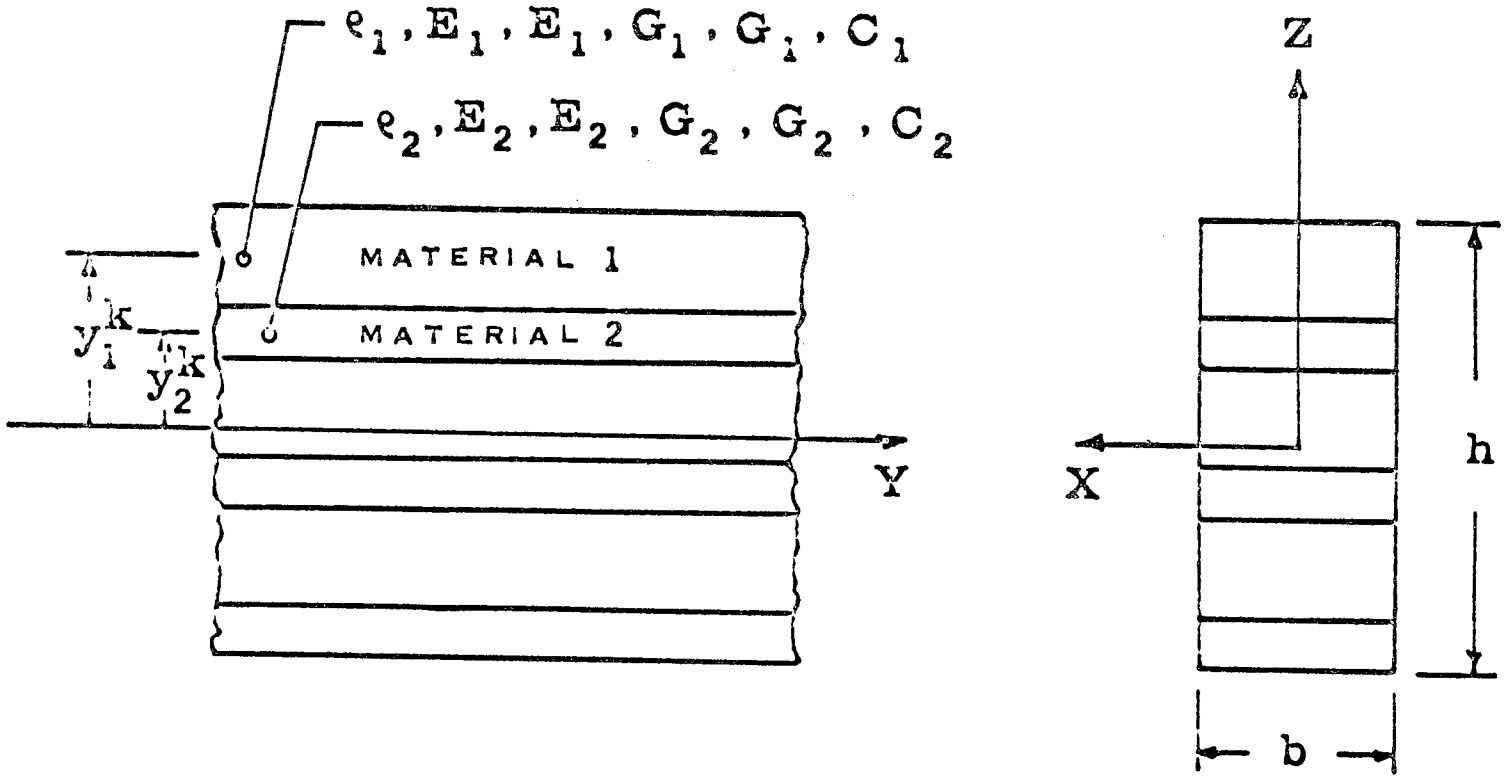



Figure 1 - The Beam Coordinates

$$T_\alpha^{*k} = \frac{\rho_\alpha}{2}[A_\alpha \dot{w}^2 + A_\alpha (z_\alpha^k)^2 \dot{\psi}^2 + I_\alpha \dot{\phi}_\alpha^2]$$

$$U_\alpha^{*k} = \frac{1}{2} A_\alpha k_\alpha G_\alpha (\frac{\partial w}{\partial y} - \phi_\alpha)^2$$

$$+ \frac{A_\alpha E_\alpha}{2}(z_\alpha^k)^2 (\frac{\partial \psi}{\partial y})^2 + \frac{E_\alpha I_\alpha}{2}(\frac{\partial \phi_\alpha}{\partial y})^2$$

$$V_\alpha^{*k} = \int_0^\tau \left[ \begin{array}{l} A_\alpha k_\alpha^* G_\alpha^* (\frac{\partial \dot{w}}{\partial y} - \dot{\phi}_\alpha)^2 + E_\alpha^* I_\alpha (\frac{\partial \dot{\phi}_\alpha}{\partial y})^2 \\ + A_\alpha E_\alpha^* (z_\alpha^k)^2 (\frac{\partial \dot{\psi}}{\partial y})^2 - \rho_\alpha A_\alpha C_\alpha \ddot{w}^2 \\ - \rho_\alpha A_\alpha C_\alpha (z_\alpha^k)^2 \ddot{\psi}^2 - \rho_\alpha I_\alpha C_\alpha \ddot{\phi}_\alpha^2 \end{array} \right] d\tau, \tag{27}$$

where $\alpha = 1, 2$ respectively gives the reinforcing and matrix layer energies.

Now, the three energies are summed over the n layer pairs to determine the total energies for the composite beam

$$T^* = \sum_{k=1}^{k=n} (T_1^{*k} + T_2^{*k})$$

$$U^* = \sum_{k=1}^{k=n} (U_1^{*k} + U_2^{*k})$$

$$V^* = \sum_{k=1}^{k=n} (V_1^{*k} + V_2^{*k}) \; . \tag{28}$$

It is now convenient to convert the discrete system (28) to a continuous system by utilization of a smoothing operation, that is to replace the summations in (28) by weighted integrations over the thickness variable z.

The result of the smoothing operation is the energies

$$T^* \simeq \int_{-h/2}^{h/2} \frac{1}{(d_1+d_2)} (T_1^* + T_2^*)\, dz$$

$$U^* \simeq \int_{-h/2}^{h/2} \frac{1}{(d_1+d_2)} (U_1^* + U_2^*)\, dz$$

$$V^* \simeq \int_{-h/2}^{h/2} \frac{1}{(d_1+d_2)} (V_1^* + V_2^*)\, dz \qquad (29)$$

where after smoothing

$$z = z_1^k = z_2^k. \qquad (30)$$

Carrying out the integrations in (29) in terms of (27) and taking into account (30) results in the energies

$$T^* = \frac{1}{2}(\rho_1 A_1 + \rho_2 A_2) \frac{h}{(d_1+d_2)} \dot{w}^2 + \frac{1}{24}(\rho_1 A_1 + \rho_2 A_2) \frac{h^3}{(d_1+d_2)} \dot{\psi}^2$$

$$+ \frac{1}{2} \rho_1 I_1 \frac{h}{(d_1+d_2)} \dot{\phi}_1^{\,2} + \frac{1}{2} \rho_2 I_2 \frac{h}{(d_1+d_2)} \dot{\phi}_2^{\,2}$$

$$U^* = \frac{1}{2} A_1 k_1 G_1 \frac{h}{(d_1+d_2)} \left(\frac{\partial w}{\partial y} - \phi_1\right)^2 + \frac{1}{2} A_2 k_2 G_2 \frac{h}{(d_1+d_2)} \left(\frac{\partial w}{\partial y} - \phi_2\right)^2$$

$$+ \frac{1}{24}(A_1 E_1 + A_2 E_2) \frac{h^3}{(d_1+d_2)} \left(\frac{\partial \psi}{\partial y}\right)^2 + \frac{1}{2} E_1 I_1 \frac{h}{(d_1+d_2)} \left(\frac{\partial \phi_1}{\partial y}\right)^2$$

$$+ \frac{1}{2} E_2 I_2 \left(\frac{\partial \phi_2}{\partial y}\right)^2$$

$$V^* = \int_o^t \left[ \begin{array}{l} A_1 k_1^* G_1^* \frac{h}{(d_1+d_2)} (\frac{\partial \dot{w}}{\partial y} - \dot{\phi}_1)^2 + A_2 k_2^* G_2^* \frac{h}{(d_1+d_2)} (\frac{\partial \dot{w}}{\partial y} - \dot{\phi}_2)^2 \\ + \frac{1}{12}(A_1 E_1^* + A_2 E_2^*) \frac{h^3}{(d_1+d_2)} (\frac{\partial \dot{\psi}}{\partial y})^2 + E_1^* I_1 \frac{h}{(d_1+d_2)} (\frac{\partial \dot{\phi}_1}{\partial y})^2 \\ + E_2^* I_2 \frac{h}{(d_1+d_2)} (\frac{\partial \dot{\phi}_2}{\partial y})^2 - (\rho_1 A_1 C_1 + \rho_2 A_2 C_2) \frac{h}{(d_1+d_2)} \ddot{w}^2 \\ - \frac{1}{12}(\rho_1 A_1 C_1 + \rho_2 A_2 C_2) \frac{h^3}{(d_1+d_2)} \ddot{\psi}^2 - \rho_1 I_1 C_1 \frac{h}{(d_1+d_2)} \ddot{\phi}_1^2 \\ - \rho_2 A_2 C_2 \frac{h}{(d_1+d_2)} \ddot{\phi}_2^2 \end{array} \right] d\tau. \quad (31)$$

At this point, continuity of displacement at the interface of the kth pair of layers must be considered. Applying equation (4) to a multilayer beam results in the equation

$$v_\alpha(y,z,t) = -z_\alpha^k (y,t) - z\phi_\alpha (y,t) \quad (32)$$

and with the aid of Figure 2 it is clear that

$$v_1 = -z_1^k \psi + \frac{d_1}{2}\phi_1 \quad , \quad v_2 = -z_2^k \psi - \frac{d_2}{2}\phi_2 \quad (33)$$

at the interface between layers 1 and 2. It is also clear from Figure 2 that

$$z_2^k = z_1^k - \frac{1}{2}(d_1+d_2) \quad (34)$$

and that equations (33) describe the same interface such that

$$v_1 = v_2 \, . \quad (35)$$

From equations (35) applied to equations (33) it is clear that the continuity condition is

$$\psi = \eta\phi_1 + (1-\eta)\phi_2 \quad (36)$$

where

$$\eta = \frac{d_1}{(d_1+d_2)} \quad , \quad (1-\eta) = \frac{d_2}{(d_1+d_2)} \, . \quad (37)$$

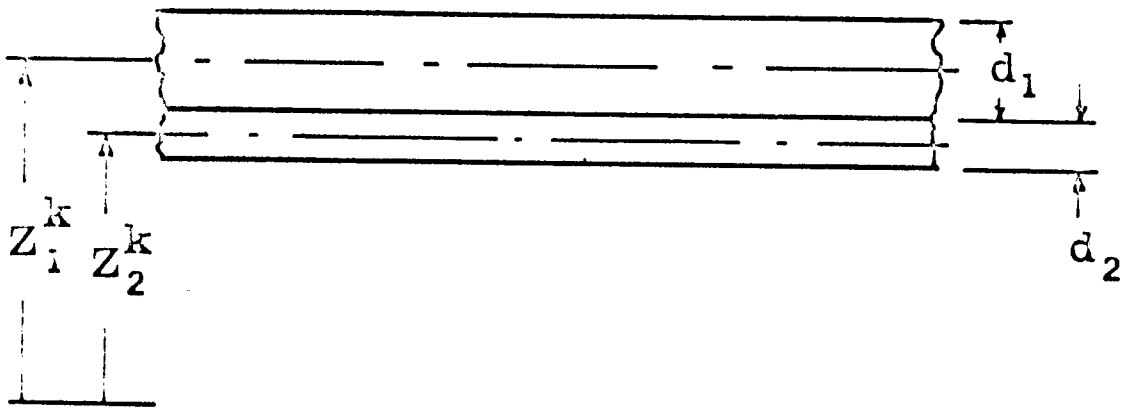



Figure 2 - The Layer Midplanes

Following Sun [8], the variable $\phi_2$ is eliminated such that

$$\phi_2 = \frac{\psi - \eta\phi}{(1-\eta)} \tag{38}$$

where for convenience the notation $\phi = \phi_1$ has been introduced.

Expression (38) is directly substituted into equations (31) and the dimensionless variable

$$\xi = h/(d_1 + d_2) \tag{39}$$

is introduced to yield the energy expressions

$$T^* = \frac{1}{2}\xi(\rho_1 A_1 + \rho_2 A_2)\,\dot{w}^2 + \frac{1}{2}I_b[\eta\rho_1 + (1-\eta)\rho_2]\dot{\psi}^2$$
$$+ \frac{1}{2}\xi\rho_1 I_1\dot{\phi}^2 + \frac{1}{2}\xi\rho_2 I_2\left(\frac{\dot{\psi}}{(1-\eta)} - \frac{\eta}{(1-\eta)}\dot{\phi}\right)^2$$

$$U^* = \frac{1}{2}\xi A_1 k_1 G_1\left(\frac{\partial w}{\partial y} - \phi\right)^2 + \frac{1}{2}\xi A_2 k_2 G_2\left(\frac{\partial w}{\partial y} - \frac{\psi}{(1-\eta)} + \frac{\eta}{(1-\eta)}\phi\right)^2$$
$$+ \frac{I_b}{2}(\eta E_1 + (1-\eta)E_2)\left(\frac{\partial\psi}{\partial y}\right)^2 + \frac{1}{2}E_1 I_1\left(\frac{\partial\phi}{\partial y}\right)^2$$
$$+ \frac{1}{2}\xi E_2 I_2\left(\frac{1}{(1-\eta)}\frac{\partial\psi}{\partial y} - \frac{\eta}{(1-\eta)}\frac{\partial\phi}{\partial y}\right)^2$$

$$V^* = \int_o^t \left[\begin{array}{l} \xi A_1 k_1^* G_1^*\left(\frac{\partial\dot{w}}{\partial y} - \dot{\phi}\right)^2 + \xi A_2 k_2^* G_2^*\left(\frac{\partial\dot{w}}{\partial y} - \frac{\dot{\psi}}{(1-\eta)} + \frac{\eta\dot{\phi}}{(1-\eta)}\right)^2 \\ + I_b[\eta E_1^* + (1-\eta)E_2^*]\left(\frac{\partial\dot{\psi}}{\partial y}\right)^2 + \xi E_1^* I_1\left(\frac{\partial\dot{\phi}}{\partial y}\right)^2 \\ + \xi E_2^* I_2\left(\frac{1}{(1-\eta)}\frac{\partial\dot{\psi}}{\partial y} - \frac{\eta}{(1-\eta)}\frac{\partial\dot{\phi}}{\partial y}\right)^2 - \xi(\rho_1 A_1 C_1 + \rho_2 A_2 C_2)\ddot{w} \\ - I_b[\eta\rho_1 C_1 + (1-\eta)\rho_2 C_2]\ddot{\psi}^2 - \xi\rho_1 I_1 C_1\,\ddot{\phi}^2 \\ - \xi\rho_2 I_2 C_2\left(\frac{1}{(1-\eta)}\ddot{\psi} - \frac{\eta}{(1-\eta)}\ddot{\phi}\right)^2 \end{array}\right] d\tau \tag{40}$$

where

$$I_b = \frac{bh^3}{12} \tag{41}$$

Now, all the squares of the various sums in equations (40) are expanded out to yield the final forms of the energy expressions as

$$T^* = \frac{1}{2}\xi a_4 \dot{w}^2 + \frac{1}{2}\xi a_9 \dot{\psi}^2 + \frac{1}{2}\xi a_{13} \dot{\phi}^2 - \xi a_{10} \dot{\psi}\dot{\phi}$$

$$U^* = \frac{1}{2}\xi a_1 \left(\frac{\partial w}{\partial y}\right)^2 - \xi a_3 \phi \frac{\partial w}{\partial y} + \frac{1}{2}\xi a_{12} \phi^2 - \xi a_2 \psi \frac{\partial w}{\partial y}$$

$$+ \frac{1}{2}\xi a_6 \psi^2 - \xi a_8 \phi\psi + \frac{1}{2}\xi a_5 \left(\frac{\partial \psi}{\partial y}\right)^2 + \frac{1}{2}\xi a_{11} \left(\frac{\partial \phi}{\partial y}\right)^2$$

$$- \xi a_7 \frac{\partial \psi}{\partial y} \frac{\partial \phi}{\partial y}$$

$$V^* = \int_o^t \left[ \begin{array}{l} \xi b_1 \left(\frac{\partial \dot{w}}{\partial y}\right)^2 - 2\xi b_3 \dot{\phi} \frac{\partial \dot{w}}{\partial y} + \xi b_{12} \dot{\phi}^2 - 2\xi b_2 \dot{\psi} \frac{\partial \dot{w}}{\partial y} + \xi b_6 \dot{\psi}^2 \\ -2\xi b_8 \dot{\phi}\dot{\psi} + \xi b_5 \left(\frac{\partial \dot{\psi}}{\partial y}\right)^2 + \xi b_{11} \left(\frac{\partial \dot{\phi}}{\partial y}\right)^2 - 2\xi b_7 \frac{\partial \dot{\psi}}{\partial y} \frac{\partial \dot{\phi}}{\partial y} \\ -\xi b_4 \ddot{w}^2 - \xi b_9 \ddot{\psi}^2 - \xi b_{13} \ddot{\phi}^2 + 2\xi b_{10} \dddot{\psi}\dddot{\phi} \end{array} \right] d\tau \quad (42)$$

where the constants $a_i$ are

$$a_1 = A_1 k_1 G_1 + A_2 k_2 G_2$$

$$a_2 = A_2 k_2 G_2/(1-\eta)$$

$$a_3 = A_1 k_1 G_1 - A_2 k_2 G_2/(1-\eta)$$

$$a_4 = \rho_1 A_1 + \rho_2 A_2$$

$$a_5 = \frac{I_b}{\xi}[\eta E_1 + (1-\eta)E_2] + \frac{E_2 I_2}{(1-\eta)^2}$$

$$a_6 = A_2 k_2 G_2/(1-\eta)^2 = a_2/(1-\eta)$$

$$a_7 = \frac{\eta}{(1-\eta)^2} E_2 I_2$$

$$a_8 = \frac{\eta}{(1-\eta)^2} A_2 k_2 G_2 = \eta a_6$$

$$a_9 = \frac{I_b}{\xi}[\eta\rho_1 + (1-\eta)\rho_2] + \frac{\rho_2 I_2}{(1-\eta)^2}$$

$$a_{10} = \frac{\eta}{(1-\eta)^2}\rho_2 I_2$$

$$a_{11} = E_1 I_1 + \frac{\eta^2}{(1-\eta)^2} E_2 I_2$$

$$a_{12} = A_1 k_1 G_1 + \frac{\eta^2}{(1-\eta)^2} A_2 k_2 G_2$$

$$a_{13} = \rho_1 I_1 + \frac{\eta^2}{(1-\eta)^2}\rho_2 I_2 \qquad (43)$$

which corresponds to the elastic constants given by Sun [8] for elastic laminated beams and where the constants $b_i$ are

$$b_1 = A_1 k_1^* G_1^* + A_2 k_2^* G_2^*$$

$$b_2 = A_2 k_2^* G_2^*/(1-\eta)$$

$$b_3 = A_1 k_1^* G_1^* - \frac{\eta}{(1-\eta)} A_2 k_2^* G_2^*$$

$$b_4 = \rho_1 A_1 C_1 + \rho_2 A_2 C_2$$

$$b_5 = \frac{I_b}{\xi}[\eta E_1^* + (1-\eta)E_2^*] + \frac{E_2^* I_2}{(1-\eta)^2}$$

$$b_6 = A_2 k_2^* G_2^*/(1-\eta)^2 = b_2/(1-\eta)$$

$$b_7 = \frac{\eta}{(1-\eta)^2} E_2^* I_2$$

$$b_8 = \frac{\eta}{(1-\eta)^2} A_2 k_2^* G_2^* = \eta b_6$$

$$b_9 = \frac{I_b}{\xi}[\eta\rho_1 C_1 + (1-\eta)\rho_2 C_2] + \frac{1}{(1-\eta)^2}\rho_2 I_2 C_2$$

$$b_{10} = \frac{\eta}{(1-\eta)^2}\rho_2 I_2 C_2$$

$$b_{11} = E_1^* I_1 + \frac{\eta^2}{(1-\eta)^2} E_2^* I_2$$

$$b_{12} = A_1 k_1^* G_1^* + \frac{\eta^2}{(1-\eta)^2} A_2 k_2^* G_2^*$$

$$b_{13} = \rho_1 I_1 C_1 + \frac{\eta^2}{(1-\eta)^2}\rho_2 I_2 C_2 \qquad (44)$$

which corresponds to the viscoelastic contribution of the current analysis of viscoelastic laminated beams. It should be noted that the author [19] has evaluated viscoelastic shear correction constants in another paper and based on this evaluation it is clear that $k_1 = k_2 = k_1^* = k_2^* = \pi^2/12$.

Now, from equations (22-25) in conjunction with equations (42) it is easy to form energy principle (23), that is dH/dt=0, which upon various integrations by parts and a gathering of common factors of $\dot{w}, \dot{\psi}$, and $\dot{\phi}$ results in equation (23) becoming

$$\frac{dH}{dt} = \int_0^t \int_0^{\ell} \dot{w} \left[ \begin{array}{l} - a_1 \frac{\partial^2 w}{\partial y^2} + a_2 \frac{\partial \psi}{\partial y} + a_3 \frac{\partial \phi}{\partial y} + a_4 \ddot{w} - b_1 \frac{\partial^2 \dot{w}}{\partial y^2} + b_2 \frac{\partial \dot{\psi}}{\partial y} \\ + b_3 \frac{\partial \dot{\phi}}{\partial y} + b_4 \dddot{w} \end{array} \right] dtdy$$

$$+ \int_0^t \int_0^{\ell} \dot{\psi} \left[ \begin{array}{l} - a_2 \frac{\partial w}{\partial y} - a_5 \frac{\partial^2 \psi}{\partial y^2} + a_6 \psi + a_7 \frac{\partial^2 \phi}{\partial y^2} - a_8 \phi + a_9 \ddot{\psi} \\ - a_{10} \ddot{\phi} - b_2 \frac{\partial \dot{w}}{\partial y} - b_5 \frac{\partial^2 \dot{\psi}}{\partial y^2} + b_6 \dot{\psi} + b_7 \frac{\partial^2 \dot{\phi}}{\partial y^2} - b_8 \dot{\phi} \\ + b_9 \dddot{\psi} - b_{10} \dddot{\phi} \end{array} \right] dtdy$$

$$+ \int_0^t \int_0^{\ell} \dot{\phi} \left[ \begin{array}{l} - a_3 \frac{\partial w}{\partial y} + a_7 \frac{\partial^2 \psi}{\partial y^2} - a_8 \psi - a_{10} \ddot{\psi} - a_{11} \frac{\partial^2 \phi}{\partial y^2} \\ + a_{12} \phi + a_{13} \ddot{\phi} - b_3 \frac{\partial \dot{w}}{\partial y} + b_7 \frac{\partial^2 \dot{\psi}}{\partial y^2} - b_8 \dot{\psi} - b_{10} \dddot{\psi} \\ - b_{11} \frac{\partial^2 \dot{\phi}}{\partial y^2} + b_{12} \dot{\phi} + b_{13} \dddot{\phi} \end{array} \right] dtdy$$

$$+ \int_0^t \delta\dot{w} \left[ a_1 \frac{\partial w}{\partial y} - a_2\psi - a_3\phi + b_1 \frac{\partial \dot{w}}{\partial y} - b_2\dot{\psi} - b_3\dot{\phi} \right]_0^{\ell} dt$$

$$+ \int_0^t \delta\dot{\psi} \left[ a_5 \frac{\partial \psi}{\partial y} - a_7 \frac{\partial \phi}{\partial y} + b_5 \frac{\partial \dot{\psi}}{\partial y} - b_7 \frac{\partial \dot{\phi}}{\partial y} \right]_0^{\ell} dt$$

$$+ \int_0^t \delta\dot{\phi} \left[ - a_7 \frac{\partial \psi}{\partial y} + a_{11} \frac{\partial \phi}{\partial y} - b_7 \frac{\partial \dot{\psi}}{\partial y} + b_{11} \frac{\partial \dot{\phi}}{\partial y} \right]_0^{\ell} dt = 0. \quad (45)$$

The viscoelastic equations of motion and boundary conditions for laminated beams are now obtained by applying the first lemma of the calculus of variations to equation (45). Thus, the three equations of motion are

$$a_1 \frac{\partial^2 w}{\partial y^2} - a_2 \frac{\partial \psi}{\partial y} - a_3 \frac{\partial \phi}{\partial y} + b_1 \frac{\partial^2 \dot{w}}{\partial y^2} - b_2 \frac{\partial \dot{\psi}}{\partial y} - b_3 \frac{\partial \dot{\phi}}{\partial y} = a_4 \ddot{w} + b_4 \dot{\ddot{w}}$$

$$a_2 \frac{\partial w}{\partial y} + a_5 \frac{\partial^2 \psi}{\partial y^2} - a_6 \psi - a_7 \frac{\partial^2 \phi}{\partial y^2} + a_8 \phi + b_2 \frac{\partial \dot{w}}{\partial y} + b_5 \frac{\partial^2 \dot{\psi}}{\partial y^2} - b_6 \dot{\psi}$$

$$- b_7 \frac{\partial^2 \dot{\phi}}{\partial y^2} + b_8 \dot{\phi} = a_9 \ddot{\psi} - a_{10} \ddot{\phi} + b_9 \dot{\ddot{\psi}} - b_{10} \dot{\ddot{\phi}}$$

$$a_3 \frac{\partial w}{\partial y} - a_7 \frac{\partial^2 \psi}{\partial y^2} + a_8 \psi + a_{11} \frac{\partial^2 \phi}{\partial y^2} - a_{12} \phi + b_3 \frac{\partial \dot{w}}{\partial y} - b_7 \frac{\partial^2 \dot{\psi}}{\partial y^2} + b_8 \dot{\psi}$$

$$+ b_{11} \frac{\partial^2 \dot{\phi}}{\partial y^2} - b_{12} \dot{\phi} = - a_{10} \ddot{\psi} + a_{13} \ddot{\phi} - b_{10} \dot{\ddot{\psi}} + b_{13} \dot{\ddot{\phi}} \quad (46)$$

and the corresponding boundary conditions are

$$a_1 \frac{\partial w}{\partial y} - a_2 \psi - a_3 \phi + b_1 \frac{\partial \dot{w}}{\partial y} - b_2 \dot{\psi} - b_3 \dot{\phi} = 0,$$

or

$$w = 0 \qquad \text{on} \qquad y = 0, \ell \quad (47\text{-a})$$

$$a_5\frac{\partial\psi}{\partial y} - a_7\frac{\partial\phi}{\partial y} + b_5\frac{\partial\dot{\psi}}{\partial y} - b_7\frac{\partial\dot{\phi}}{\partial y} = 0,$$

or

$$\psi = 0 \qquad \text{on} \qquad y = 0,\ell \tag{47-b}$$

$$a_7\frac{\partial\psi}{\partial y} - a_{11}\frac{\partial\phi}{\partial y} + b_7\frac{\partial\dot{\psi}}{\partial y} - b_{11}\frac{\partial\dot{\phi}}{\partial y} = 0,$$

or

$$\phi = 0 \qquad \text{on} \qquad y = o,\ell \ . \tag{47-c}$$

4. WAVE PROPAGATION. Following Sun [8], but with a visco-elastic counterpart, assume flexural wave propagation in the y-direction of the form

$$\begin{aligned} \omega &= hWe^{-\alpha y}\, e^{ip(y/c-t)} \\ \psi &= \Psi e^{-\alpha y}\, e^{ip(y/c-t)} \\ \phi &= \Phi e^{-\alpha y}\, e^{ip(y/c-t)} \end{aligned} \tag{48}$$

Where $\alpha$ is the attenuation coefficient, p is the circular frequency, and c is the phase velocity. It is also convenient at this time to introduce some additional relationships as

$$\begin{aligned} p &= 2\pi\omega \\ \omega &= \frac{c}{\lambda} \\ \lambda &= \frac{2\pi c}{p} = c\tau \\ \tau &= \frac{2\pi}{p} \\ K &= \frac{2\pi}{\lambda} \\ \beta &= \lambda\alpha \end{aligned} \tag{49}$$

Where $\omega$ is the frequency, $\lambda$ is the wave length, $\tau$ is the period, K is the wave number and $\beta$ is the attenuation constant.

Now, equations (48) are passed through differential equations (46) to obtain the characteristic equations for wave propagation. At the same time, the following dimensionless parametric, elastic, and viscoelastic dimensionless variables are introduced

$$\alpha_1 = \frac{A_1}{h^2}$$

$$\alpha_2 = \frac{A_2}{h^2}$$

$$\beta = \lambda\alpha$$

$$\gamma = \frac{G_1}{G_2}$$

$$\gamma_1^* = p\frac{G_1^*}{G_2}$$

$$\gamma_2^* = p\frac{G_2^*}{G_2}$$

$$\theta = \frac{\rho_1}{\rho_2}$$

$$V = \frac{C}{\sqrt{G_2/\rho}}$$

$$\overline{C} = \frac{C_1 C}{\lambda}$$

$$\overline{C} = \frac{C_2 C}{\lambda}$$

$$\overline{\lambda} = \frac{\lambda}{h}$$

$$\xi = \frac{h}{(d_1+d_2)}$$

$$\varepsilon_b = \frac{I_b}{h^4}$$

$$\varepsilon_1 = \frac{I_1}{h^4}$$

$$\varepsilon_2 = \frac{I_2}{h^4}$$

$$\delta_1 = \frac{E_1}{G_2}$$

$$\delta_2 = \frac{E_2}{G_2}$$

$$\alpha = \frac{\beta}{h\bar{\lambda}}$$

$$\delta_1^* = \frac{pE_1^*}{G_2}$$

$$\delta_2 = \frac{pE_2^*}{G_2} \quad . \qquad (50)$$

The final form of the characteristic equation for viscoelastic wave propagation is

$$\text{DET} \begin{vmatrix} (R_{11} + i\ I_{11}) & (R_{12} + i\ I_{12}) & (R_{13} + i\ I_{13}) \\ (R_{12} + i\ I_{12}) & (R_{22} + i\ I_{22}) & (R_{23} + i\ I_{23}) \\ (R_{13} + i\ I_{13}) & (R_{23} + i\ I_{23}) & (R_{33} + i\ I_{33}) \end{vmatrix} = 0 \qquad (51)$$

Where

$$R_{11} = -a_{11}\,\beta^2 + b_{11}\,\beta - c_{11}V^2 + d_{11}$$

$$R_{12} = R_{21} = -\,b_{12}\,\beta + d_{12}$$

$$R_{13} = R_{31} = -\,b_{13}\,\beta + d_{13}$$

$$R_{22} = a_{22}\,\beta^2 - b_{22}\,\beta + c_{22}V^2 - d_{22}$$

$$R_{23} = R_{32} = -\,a_{23}\,\beta^2 + b_{23}\,\beta - c_{23}V^2 + d_{23}$$

$$R_{33} = a_{33}\,\beta^2 - b_{33}\,\beta + c_{33}V^2 - d_{33}$$

$$I_{11} = A_{11}\,\beta^2 + B_{11}\,\beta + \bar{c}_{11}V^2 - D_{11}$$

$$I_{12} = I_{21} = B_{12}\,\beta + D_{12}$$

$$I_{13} = I_{31} = B_{13}\,\beta + D_{13}$$

$$I_{22} = -\,A_{22}\,\beta^2 - B_{22}\,\beta - \bar{c}_{22}V^2 + D_{22}$$

$$I_{23} = I_{32} = A_{23}\,\beta^2 + B_{23}\,\beta + \bar{c}_{23}V^2 - D_{23}$$

$$I_{33} = -\,A_{33}\,\beta^2 - B_{33}\,\beta - \bar{c}_{33}V^2 + D_{33} \qquad (52)$$

The constants introduced into equation (52) are defined as

$$a_{11} = \frac{\alpha_1 k_1 \gamma}{\bar{\lambda}^2} + \frac{\alpha_2 k_2}{\bar{\lambda}^2}$$

$$b_{11} = \frac{4\pi\alpha_1 k_1^* \gamma_1^*}{\bar{\lambda}^2} + \frac{4\pi\,\alpha_2 k_2^* \gamma_2^*}{\bar{\lambda}^2}$$

$$c_{11} = \frac{4\pi^2 \theta\alpha_1}{\bar{\lambda}^2} + \frac{4\pi^2\alpha_2}{\bar{\lambda}^2}$$

$$d_{11} = \frac{4\pi^2\alpha_1 k_1\gamma}{\bar{\lambda}^2} + \frac{4\pi^2\alpha_2 k_2}{\bar{\lambda}^2}$$

$$b_{12} = \frac{\alpha_2 k_2}{(1-\eta)\bar{\lambda}}$$

$$d_{12} = \frac{2\pi\alpha_2 k_2^* \gamma_2^*}{(1-\eta)\bar{\lambda}}$$

$$b_{13} = \frac{\alpha_1 k_1\gamma}{\bar{\lambda}} - \frac{\eta\alpha_2 k_2}{(1-\eta)\bar{\lambda}}$$

$$d_{13} = \frac{2\pi\alpha_1 k_1^* \gamma_1^*}{\bar{\lambda}} - \frac{2\pi\eta}{(1-\eta)\bar{\lambda}} \alpha_2 k_2^* \gamma_2^*$$

$$a_{22} = \frac{\varepsilon_b \eta\delta_1}{\xi\,\bar{\lambda}^2} + \frac{\varepsilon_b(1-\eta)\delta_2}{\xi\,\bar{\lambda}^2} + \frac{\varepsilon_2\delta_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$b_{22} = \frac{4\pi\varepsilon_b \eta\delta_1^*}{\xi\,\bar{\lambda}^2} + \frac{4\pi\varepsilon_b(1-\eta)\delta_2^*}{\xi\,\bar{\lambda}^2} + \frac{4\pi\varepsilon_2\delta_2^*}{(1-\eta)^2\bar{\lambda}^2}$$

$$c_{22} = \frac{4\pi^2\varepsilon_b \eta\theta}{\xi\,\bar{\lambda}^2} + \frac{4\pi^2\varepsilon_b(1-\eta)}{\xi\,\bar{\lambda}^2} + \frac{4\pi^2\varepsilon_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$d_{22} = \frac{4\pi^2\varepsilon_b \eta\delta_1}{\xi\,\bar{\lambda}^2} + \frac{4\pi^2\varepsilon_b(1-\eta)\delta_2}{\xi\,\bar{\lambda}^2}$$

$$+ \frac{4\pi^2\varepsilon_2\delta_2}{(1-\eta)^2\bar{\lambda}^2} + \frac{\alpha_2 k_2}{(1-\eta)^2}$$

$$a_{23} = \frac{\eta\varepsilon_2\delta_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$b_{23} = \frac{4\pi\eta\varepsilon_2\delta_2^*}{(1-\eta)^2\bar{\lambda}^2}$$

$$c_{23} = \frac{4\pi^2\eta\varepsilon_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$d_{23} = \frac{4\pi^2\eta\delta_2\varepsilon_2}{(1-\eta)^2\bar{\lambda}^2} + \frac{\eta\alpha_2 k_2}{(1-\eta)^2}$$

$$a_{33} = \frac{\delta_1\varepsilon_1}{\bar{\lambda}^2} + \frac{\eta^2\delta_2\varepsilon_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$b_{33} = \frac{4\pi\delta_1^*\varepsilon_1}{\bar{\lambda}^2} + \frac{4\pi\eta^2\delta_2^*\varepsilon_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$c_{33} = \frac{4\pi^2\theta\varepsilon_1}{\bar{\lambda}^2} + \frac{4\pi^2\eta^2\varepsilon_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$d_{33} = \frac{4\pi^2\delta_1\varepsilon_1}{\bar{\lambda}^2} + \frac{4\pi^2\eta^2\delta_2\varepsilon_2}{(1-\eta)^2\bar{\lambda}^2} + \alpha_1 k_1\gamma + \frac{\eta^2\alpha_2 k_2}{(1-\eta)^2}$$

$$A_{11} = \frac{b_{11}}{4\pi}$$

$$B_{11} = 4\pi a_{11}$$

$$\bar{c}_{11} = \frac{8\pi^3\theta\alpha_1\bar{c}_1}{\bar{\lambda}^2} + \frac{8\pi^3\alpha_2\bar{c}_2}{\bar{\lambda}^2}$$

$$D_{11} = \pi b_{11}$$

$$B_{12} = \frac{d_{12}}{2\pi}$$

$$D_{12} = 2\pi b_{12}$$

$$B_{13} = \frac{d_{13}}{2\pi}$$

$$D_{13} = 2\pi b_{13}$$

$$A_{22} = \frac{b_{22}}{4\pi}$$

$$B_{22} = 4\pi a_{22}$$

$$\bar{c}_{22} = \frac{8\pi^3 \varepsilon_b \eta\theta \bar{c}_1}{\xi\, \bar{\lambda}^2} + \frac{8\pi^3 \varepsilon_b (1-\eta)\bar{c}_2}{\xi\, \bar{\lambda}^2} + \frac{8\pi^3 \varepsilon_2 \bar{c}_2}{(1-\eta)^2 \bar{\lambda}^2}$$

$$D_{22} = \frac{4\pi^2 \varepsilon_b \eta \delta_1^*}{\xi\, \bar{\lambda}^2} + \frac{4\pi^2 \varepsilon_b (1-\eta)\delta_2^*}{\xi\, \bar{\lambda}^2} + \frac{4\pi^2 \varepsilon_2 \delta_2^*}{(1-\eta)^2 \bar{\lambda}^2} + \frac{\alpha_2 k_2^* \gamma_2^*}{(1-\eta)^2}$$

$$A_{23} = \frac{b_{23}}{4\pi}$$

$$B_{23} = 4\pi a_{23}$$

$$\bar{c}_{23} = 2\pi \bar{c}_2 c_{23}$$

$$D_{23} = \frac{4\pi^2 \eta \varepsilon_2 \delta_2^*}{(1-\eta)^2 \bar{\lambda}^2} + \frac{\eta \alpha_2 k_2^* \gamma_2^*}{(1-\eta)^2}$$

$$A_{23} = \frac{\delta_1^* \epsilon_1}{\bar{\lambda}^2} + \frac{\eta^2 \delta_2^* \epsilon_2}{(1-\eta)^2 \bar{\lambda}^2}$$

$$B_{33} = \frac{4\pi\delta_1\epsilon_1}{\bar{\lambda}^2} + \frac{4\pi\eta^2\delta_2\epsilon_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$\bar{c}_{33} = \frac{8\pi^3\theta\epsilon_1\bar{c}_1}{\bar{\lambda}^2} + \frac{8\pi^3\eta^2\epsilon_2\bar{c}_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$D_{33} = \frac{4\pi^2\delta_1^*\epsilon_1}{\bar{\lambda}^2} + \frac{4\pi^2\eta^2\delta_2^*\epsilon_2}{(1-\eta)^2\bar{\lambda}^2}$$

$$+ \alpha_1 k_1^* \gamma_1^* + \frac{\eta^2 \alpha_2 k_2^* \gamma_2^*}{(1-\eta)^2} \quad . \tag{53}$$

A numerical solution to characteristic equation (51) is possible if it is recast as the following function

$$f(\beta,V) = \text{ABS}\ (\text{DET}|R_{ij} + iI_{ij}|). \tag{54}$$

Using a numerical technique such as the Rosenbrock [21] optimization procedure a solution is obtained when

$$f(\beta,V) = 0 \quad . \tag{55}$$

5. SUMMARY. An energy principle has been formulated for viscoelastic Timoshenko beams according to the standard linear model with the stipulation, and hence additional terms, that the energy principle be utilized in building a viscoelastic laminated beam theory. The Timoshenko model considered has accounted for both viscoelastic extensional and viscoelastic shear strains. To later incorporate the single layer energy principle into the development of a laminated beam theory, a term which accounts for the beam's gross rotation was included in the single layer development.

Using the single layer energies developed, a viscoelastic laminated beam theory composed of a number of alternating, plane, parallel layers of two homogeneous, isotropic viscoelastic materials, termed the reinforcing layer and the matrix layer, was derived. In deriving the theory, the individual layer kinetic, potential, and dissipative energies were summed over n layer pairs to obtain the total energy of the composite beam; these results are converted to a continuous system by utilization of a smoothing operation or weighted integration. The number of independent variables in the total composite beam energies is reduced from four to three thru the introduction of a condition for continuity at layer interfaces. A direct application of the energy principle developed to the composite beam energies results in a set of three equations of motion and their corresponding boundary conditions for viscoelastic, laminated composite beams.

Flexural wave propagation has been considered by passing viscoelastic harmonic waves through the derived equations of motion. Numerical solutions are possible by applying the Rosenbrock optimization procedure to the resulting characteristic equation. A lack of computation funds precludes the presentation of numerical results at the present writing.

## REFERENCES


1. C. T. SUN, J. D. ACHENBACH and G. HERRMANN (1968) Journal of Applied Mechanics, 35, 467. Continuum Theory for a Laminated Medium.

2. J. D. ACHENBACH, C. T. SUN and G. HERRMANN (1968) Journal of Applied Mechanics, 35, 689. On the Vibrations of a Laminated Body.

3. C. R. THOMAS (1972) Journal of Sound and Vibration, 23(3), 341-361. Simple Thickness Modes for Laminated Composite Materials.

4. C. T. SUN (1971) Journal of Applied Mechanics, 38, 231-238. Theory of Laminated Plates.

5. C. R. THOMAS (1972) Journal of Sound and Vibration, 25(3), 407-431. Velocity Corrected Theory of Laminated Plates Applied to Free Plate Strip Vibrations.

6. C. R. THOMAS (1973) Journal of Sound and Vibration, 31(2), 195-211. Extensional Vibrations of Simply Supported Composite Plate Strips.

7. C. R. THOMAS (1975) Journal of the Acoustical Society of America, 57(3), 655-659. Flexural and Estensional Vibrations of Simply Supported Laminated Rectangular Plates.

8. C. T. SUN (1971) Journal of Applied Mechanics, 38, 947-954. Microstructure Theory for a Composite Beam.

9. C. R. THOMAS (1973) Watervliet Arsenal Technical Report, R-WV-T-6-45-73. Flexure Equations of Motion for Laminated Composite Beams.

10. R. A. GROT and J. D. ACHENBACH (1970) Acta Mechanica, 9, 245-263. Linear Isothermal Theory for a Viscoelastic Laminated Composite.

11. M. STERN, A. BEDFORD, and C. H. YEW (1971) Journal of Applied Mechanics, 38(2), 448-454. Wave Propagation in Viscoelastic Laminates.

12. S. P. TIMOSHENKO (1922) Philosophical Magazine, Ser 6, Vol 43, 125-131. On the Transverse Vibrations of Bars of Uniform Cross-Section.

13. H. C. LEE (1960) Journal of Applied Mechanics, 27, 551-556. Forced Lateral Vibration of a Uniform Cantilever Beam with Internal and External Damping.

14. H. PAN (1966) Jour. Eng. Mech. Div., Proc. Amer. Soc. Civil Eng., 213-234. Vibration of a Viscoelastic Timoshenko Beam.

15. R. D. MINDLIN (1955) Signal Corps Engineering Laboratories, Fort Monmouth, New Jersey (AD-88471). An Introduction to the Mathematical Theory of Vibrations of Elastic Plates.

16. E. J. BRUNELLE (1970) J. Composite Materials, 4, 404-416. The Statics and Dynamics of a Transversely Isotropic Timoshenko Beam.

17. R. D. MINDLIN and H. DERESIEWICZ (1953) Columbia University Technical Report No. 10. Timoshenko's Shear Coefficient for Flexural Vibrations of Beams.

18. G. L. ANDERSON (1975) Journal of Sound and Vibration, 39(1), 55-76. Stability of a Rotating Cantilever Subjected to Dissipative, Aerodynamic, and Transverse Follower Forces.

19. C. R. THOMAS (1976) Watervliet Arsenal Technical Report, WVT-TR-76009. Simple Thickness Modes and Shear Correction Coefficients for Viscoelastic Timoshenko Beams.

20. Y. C. FUNG (1965) Prentice-Hall, Inc. "Foundations of Solid Mechanics".

21. H. H. ROSENBROCK (1960) Computer Journal 3, 175-184. An Automatic Method for Finding the Greatest or Least Value of a Function.

APPENDIX. The present objective is to derive a set of constitutive relations which can be utilized in conjunction with the basic equations for a Timoshenko beam. While constitutive equations may be formulated in either integral or differential form, preliminary work in the direction of formulation of a viscoelastic beam theory for laminated composite materials indicates that the differential form of constitutive relations will be most useful. The differential constitutive relations will be utilized in the present development.

The general form of the differential constitutive equations is adapted from Fung [20] where the stress-strain relations are of the form

$$P_1(D)\sigma'_{ij} = Q_1(D)e'_{ij}$$
$$P_2(D)\sigma_{kk} = Q_2(D)e_{kk} \qquad \text{(A-1)}$$

where $P_i(D)$ and $Q_i(D)$ are given by

$$P_1(D) = \sum_{k=o}^{k=n_1} a_k D^k$$

$$P_2(D) = \sum_{k=o}^{k=n_2} c_k D^k$$

$$Q_1(D) = \sum_{k=o}^{k=m_1} b_k D^k$$

$$Q_2(D) = \sum_{k=o}^{k=m_1} d_k D^k \qquad \text{(A-2)}$$

with D being the time-derivative operator of the form

$$D^i f = \frac{\partial^i f(t)}{\partial t^i} \qquad \text{(A-3)}$$

and where $\sigma'_{ij}$ and $e'_{ij}$ are the components of the stress and strain deviators

$$\sigma'_{ij} = \sigma_{ij} - \frac{1}{3}\delta_{ij}\sigma_{kk}$$

$$e'_{ij} = e_{ij} - \frac{1}{3}\delta_{ij}e_{kk} \qquad \text{(A-4)}$$

in which $\sigma_{ij}$ and $e_{ij}$ are the components of stress and strain.

Now, assume equations (A-1) to have the form of the standard linear model

$$(1 + \bar{A}\frac{\partial}{\partial t})\sigma = (\bar{B} + \bar{C}\frac{\partial}{\partial t})\varepsilon \qquad \text{(A-5)}$$

where $\sigma$ is stress and $\varepsilon$ is strain. Comparing the form of (A-5) with equations (A-1) it is clear that to have the form of the standard linear model it must be true that

$$n_1 = m_1 = n_2 = m_2 = 1 \qquad \text{(A-6)}$$

and operators (A-2) in light of (A-6) reduce to

$$P_1(D) = a_o + a_1 D$$

$$Q_1(D) = b_o + b_1 D$$

$$P_2(D) = C_o + C_1 D$$

$$Q_2(D) = d_o + d_1 D \quad . \qquad \text{(A-7)}$$

As will be subsequently seen, the only non-zero stresses and strains for a Timoshenko beam with its y-axis along the length and its z-axis through the thickness are $\sigma_y$ and $\sigma_{yz}$ & $\varepsilon_y$ and $\varepsilon_{yz}$. Thus, from equation (A-4) the non-zero stress and strain deviators are

$$\sigma'_y = \frac{2}{3}\sigma_y \quad , \quad \sigma'_{yz} = \sigma_{yz}$$

$$\varepsilon'_y = \frac{2}{3}\varepsilon_y \quad , \quad \varepsilon'_{yz} = \varepsilon_{yz} \qquad \text{(A-8)}$$

Now, a direct substitute of equations (A-2), (A-7), and (A-8) into equation (A-1) results in

$$[1 + (a_1/a_0)D]\sigma_{yz} = [(b_0/a_0) + (b_1/a_0)D]\varepsilon_{yz}$$

$$[1 + (a_1/a_0)D]\sigma_y = [(b_0/a_0) + (b_1/a_0)D]\varepsilon_y$$

$$[1 + (C_1/C_0)D]\sigma_y = [(d_0/C_0) + (d_1/C_0)D]\varepsilon_y \quad . \qquad \text{(A-9)}$$

There are thus two equations for stress-strain in the y- coordinate

$$D_1\sigma_y = D_2\varepsilon_y$$

$$D_3\sigma_y = D_4\varepsilon_y \quad , \qquad \text{(A-10)}$$

where

$$D_1 = 1-(a_1/a_0)D$$

$$D_2 = (b_0/a_0) + (b_1/a_0)D$$

$$D_3 = 1 + (C_1/C_0)D$$

$$D_4 = (d_0/C_0) + (d_1/C_0)D \quad , \qquad \text{(A-11)}$$

and they must be combined to form a single constitutive equation

$$2D_1D_3\sigma = (D_2D_3 + D_1D_4)\varepsilon_y \quad . \qquad \text{(A-12)}$$

Now, from both the right and left sides of equation (A-12) it is clear that the constitutive equation is of the form

$$(1 + \bar{a}D + \bar{b}D^2)\sigma_y = (1 + \bar{C}\,D + \bar{d}D^2)\varepsilon_y \quad , \qquad \text{(A-13)}$$

but it would now be desireable to have the form of the standard linear model as in equation (A-5), if possible. This can be achieved if the restriction is now made that

$$D_1 = D_3 = 1 + (a_1/a_0)D \qquad \text{(A-14)}$$

such that equation (A-12) now becomes

$$[1 + (a_1/a_0)D]\sigma_y = \frac{1}{2}[(b_0/a_0 + d_0/C_0) + (b_1/a_0 + d_1/C_0)D]\varepsilon_y \quad . \qquad \text{(A-15)}$$

As a final step, define the constants

$$C = a_1/a_o$$

$$E = \frac{1}{2}(b_o/a_o + d_o/C_o)$$

$$E^* = \frac{1}{2}(b_1/a_o + d_1/C_o)$$

$$2G = b_o/a_o$$

$$2G^* = b_1/a_o \qquad \text{(A-16)}$$

with the final form of the constitutive equation thus being

$$(1 + C\frac{\partial}{\partial t})\sigma_{yz} = (2G + 2G^*\frac{\partial}{\partial t})\varepsilon_{yz}$$

$$(1 + C\frac{\partial}{\partial t})\sigma_y = (E + E^*\frac{\partial}{\partial t})\varepsilon_y \quad . \qquad \text{(A-17)}$$

# USING FAST TRANSFORMS TO COMPUTE THE WEIGHT DISTRIBUTION OF A LINEAR CODE

Bart F. Rice
Department of Defense
Fort George G. Meade, Maryland

ABSTRACT. N. J. Patterson, in an unpublished note, observed that the weight distribution of a linear code could be computed using a Fast Hadamard Transform. In this paper, we expand on Patterson's rather brief exposition, providing a proof that the method actually produces the weight distribution and making a comparison of the storage and time involved using Patterson's method and the "brute force" approach.

The weight distribution of a linear code contains a lot of information about the code, including its minimum distance and the probabilities of decoding error and failure if the decoding algorithm decodes all patterns of $\leq t$ errors and nothing else (cf. [3]). It is not surprising, therefore, that there has been much effort expended in investigation of weight enumeration of linear codes. In the case of linear binary codes, a method for computing weight distributions involving Fast Hadamard transforms [1] in an unpublished note by N. J. Patterson has certain computational advantages over the "brute force" technique of weight enumeration (in which a basis for the code is chosen and every possible linear combination of the basis codewords is taken in an unimaginative way, with the weight of each codeword recorded as the codeword is derived). In this paper we expand on Patterson's rather brief discussion, providing a proof that the method actually computes the weight distribution of a linear binary code and making a comparison of this technique with the brute force approach.

Let A be a (n,k) linear code over GF(q), with "weight enumerator polynomial"

$$W_A(x,y) = \sum_{i=0}^{n} A_i \, x^{n-i} \, y^i ,$$

where $A_i$ is the number of codewords $v \in A$ with weight $w(v)=i$. Let $A^\perp$ denote the dual of A. MacWilliams' Identity states that

$$|A^\perp| W_A(x,y) = W_{A^\perp}(x+(q-1)y, x-y). \tag{1}$$

If A is a (n,k) linear binary code then (1) becomes

$$2^{n-k} W_A(x,y) = W_{A^\perp}(x+y, x-y). \tag{2}$$

For convenience, we will assume that A is binary. The method is quite general, though, and has obvious extensions to cases when $q>2$.

Assume that $k > \frac{n}{2}$, or that A has rate $k/n > \frac{1}{2}$. If $k/n \leq \frac{1}{2}$, the following procedure should be modified by interchanging A and $A^{\perp}$ and replacing k by n-k. Let H denote an (n-k)xn parity check matrix for A, say

$$H = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{n-k-1} \end{bmatrix},$$

where the rows $v_i$, $0 \leq i \leq n-k-1$, are vectors in $GF(2)^n$ which constitute a basis for $A^{\perp}$. Write

$$H^t = \begin{bmatrix} v_0^t & v_1^t & \cdots & v_{n-k-1}^t \end{bmatrix} = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{bmatrix},$$

where $u_i = (u_{i0}, u_{i1}, \ldots, u_{i,n-k-1})$ is the binary (n-k)-tuple comprising the i-th row of $H^t$. Suppose $0 \leq s \leq 2^{n-k}-1$, say $s = \sum_{j=0}^{n-k-1} s_j 2^j$. Let

$$b_s = \sum_{i=0}^{n-1} (-1)^{s \cdot u_i} = \sum_{i=0}^{n-1} (-1)^{\sum_{j=0}^{n-k-1} s_j u_{ij}}.$$

Notice that if we define $f: V = GF(2)^{n-k} \to C$ = complex numbers by

$$f(v) = \begin{cases} 1 & v=u_i \text{ for some } i,\ 0 \leq i \leq n-1; \\ 0 & \text{otherwise,} \end{cases}$$

then $b_s = \sum_{v \in V} f(v)\ (-1)^{v \cdot s}$. Therefore, $b_s$ is an n-dimensional Hadamard transform [1] of f. Now, the vector

$$(s\cdot u_0, s\cdot u_1, \ldots, s\cdot u_{n-1}) = \begin{pmatrix} s\cdot u_0 \\ s\cdot u_1 \\ \vdots \\ s\cdot u_{n-1} \end{pmatrix}^t = \left( \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{bmatrix} \quad \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{n-k-1} \end{bmatrix} \right)^t$$

$$= (s_0, s_1, \ldots, s_{n-k-1})\, H = \sum_{i=0}^{n-k-1} s_i v_i = V(s),$$

which may be termed the "s-th codeword" in $A^{\perp}$. Clearly, as s runs through all the integers from 0 to $2^{n-k}-1$, V(s) runs through all the codewords of $A^{\perp}$.

We have just shown that $s\cdot u_i$ is the i-th coordinate of V(s), and thus $b_s$ = the # of 0-coordinates in V(s) minus the # of 1-coordinates

$$= n - 2w(V(s)).$$

Hence $w(V(s)) = (n-b_s)/2$. Thus, we can compute the weight distribution of $A^{\perp}$ (and subsequently, using MacWilliams' Identity, of A) via the Fast Algorithm:

Step 0. Select a basis $\{v_0, v_1, \ldots, v_{n-k-1}\}$ for $A^{\perp}$. Let $B_i$ denote the coefficients of $W_{A^{\perp}}$, initialized to 0, $0 \le i \le n$.

Step 1. Compute the "bulges" $b_s$, $0 \le s \le 2^{n-k}-1$, using a Fast Hadamard Transform.

Step 2. For each s, $0 \le s \le 2^{n-k}-1$, let $i = (n-b_s)/2$ and replace $B_i$ by $1+B_i$.

Step 3. Use the equation (MacWilliams 1963)

$$\sum_{i=0}^{n-r} \binom{n-i}{r} A_i = 2^{k-r} \sum_{i=0}^{r} \binom{n-i}{n-r} B_i, \quad 0 \le r \le n,$$

to compute the coefficients $A_i$, $0 \le i \le n$.

A glaring disadvantage of this method is that all of the $2^{n-k}$ bulges $b_s$ must be saved. If not enough storage is available, then the algorithm must be modified. The advantage is that the work factor of the method is $(n-k)2^{n-k}$. By contrast, the brute force method requires the computation of $(s_0, s_1, \ldots, s_{n-k-1})H$ for each of the $2^{n-k}$ vectors $s = (s_0, s_1, \ldots, s_{n-k-1}) \in GF(2)^{n-k}$. This could be accomplished by the following:

Brute Force Algorithm:

Step 0. Select a basis $\{v_0, v_1, \ldots, v_{n-k-1}\}$ for $A^{\perp}$. Let $B_i$ denote the coefficients of $W_{A^{\perp}}$, initialized to 0, $0 \le i \le n-1$, and let $s = 0 = (0, 0, \ldots, 0)$.

Step 1. Let $i = 0$, $v = (0, 0, \ldots, 0)$

Step 1.1. If $s_i=1$, replace v by $v+v_i$. When $q>2$, this requires n additions. When $q=2$ these n additions can be accomplished by several mod 2 additions, the exact number depending on word size of the machine used to implement the algorithm.

Step 1.2. Replace i by $1+i$.

Step 1.3. If $i\leq n-1$, go to step 1.1. Otherwise, go to step 2.

Step 2. Compute $\alpha$ = weight of v and replace $B_\alpha$ by $1+B_\alpha$.

Step 3. Replace s by $1+s$. If $s\leq 2^{n-k}-1$, go to step 1. Otherwise stop.

On the average, the vectors s in the Brute Force Algorithm will have density $(n-k)/2$. Thus, the work factor for this algorithm is $n(n-k)2^{n-k-1}$. That is, the extra cost in time is proportional to n. The advantage of this method is, of course, that the only storage required is for the arrays $\{A_i\}$, $\{B_i\}$ and H. If A is cyclic, with parity check polynomial $h(x)$ (of degree k), then (regarding a vector in $GF(2)^n$ as a polynomial of degree $\leq n-1$), we may take $v_0=h(x)$, $v_1=xh(x)$, ..., $v_{n-k-1} = x^{n-k-1}h(x)$, so that only $v_0$ need be saved. (In this case, $(s_0, s_1, \ldots, s_{n-k-1})H = h(x) \sum_{i=0}^{n-k-1} s_i x^i$.)

In conclusion, using a Hadamard transform to compute the weight distribution of a (n,k) linear code results in a time saving proportional to n at a cost in storage of approximately $2^{n-k}$ words. The technique is particularly advantageous for high rate codes.

REFERENCES


1 Ahmed, N; Rao, K.R.; Abdussattar, A.L.; "BIFORE or Hadamard Transform", IEEE Transactions on Audio and Electroacoustics; Vol. 19, No. 3 (September, 1971); 225-234.

2 Assmus, E.F., Jr.; Mattson, H.F., Jr; "Coding and Combinatorics", SIAM Review, Vol. 16, No. 3, July 1974; 349-388.

3 Berlekamp, Elwyn R; Algebraic Coding Theory, McGraw-Hill, New York (1968).

4 Gleason, Andrew M.; "Weight Polynomials and the MacWilliams Identities", Actes. Congrès Internat. Math., 3 (1970), 211-215.

5 Patterson, N.J.; unpublished note.

# FACTORIAL AND HADAMARD SERIES FOR BESSEL FUNCTIONS OF ORDERS ZERO AND ONE

Alexander S. Elder
Emma M. Wineholt
Propulsion Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. Bessel functions of orders zero and one for moderate and large positive arguments have been programmed in FORTRAN using factorial series for $J_n(x)$, $Y_n(x)$ and $K_n(x)$ and Hadamard series for $I_n(x)$. A subroutine to calculate Stirling numbers of the first kind was developed for use in the factorial series. The recurrence relation was modified and the resulting Stirling numbers scaled so that the entire range of the computer was utilized; e.g., $10^{-150} < S < 10^{150}$ instead of $10^{0} < S < 10^{150}$. In this way, more terms of the series can be calculated and higher accuracy obtained. For use in the Hadamard series, a subroutine to calculate incomplete gamma functions was developed. Various algorithms were necessary to encompass the required range of arguments.

These programs were devised to verify the accuracy (for moderate and large arguments) of our previously developed Bessel function subroutine. These programs replace the asymptotic series with convergent series, which, of course, is desirable. Extension of the program to complex arguments is now in progress.

1. INTRODUCTION. Factorial series derived from the Laplace integral converge rapidly for large values of the argument, and, thus, are preferable to the corresponding asymptotic series. However, the traditional algorithm leads to very large numbers and must be modified if it is to be useful for numerical work. One procedure for scaling the large Stirling numbers which occur in the analysis is derived below.

Factorial series based on a Laplace integral evaluated between finite limits will generally diverge, so that an alternate procedure is required. One method, due to Hadamard, is to expand the Laplace integral in a series of incomplete gamma functions. The resulting series converge rapidly for large values of the argument. In practice, expansions in terms of the Kummer function are more convenient for computation. These functions are closely related to the incomplete gamma function.

Computer programs based on these algorithms will be used to check the accuracy of the BRL subroutines for Bessel functions of complex argument and integral order. This is necessary as tables are not available for a sufficient range of order and argument to make a detailed check by comparison.

2. FACTORIAL SERIES. The factorial series are used to calculate $K_n(x)$, $J_n(x)$, $Y_n(x)$.

$K_n(x)$ can be expressed in terms of the Whittaker function as[1]

$$K_n(x) = \left(\frac{\pi}{2x}\right)^{1/2} W_{o,n}(2x),$$

where the asymptotic expansion for the Whittaker function is[2]

$$W_{o,n}(2x) = e^{-x} \left\{ 1 + \sum_{m=1}^{\infty} \frac{[n^2-(-1/2)^2]\,[n^2-(-3/2)^2]\,\ldots\,[n^2-(1/2-m)^2]}{m!\,(2x)^m} \right\}$$

This asymptotic expansion was derived from a Laplace integral evaluated between zero and infinity and involves only negative integral powers of the argument.

For n = 0,

$$K_o(x) = \left(\frac{\pi}{2x}\right)^{1/2} e^{-x} \left\{ 1 - \frac{1^2}{1!\,(8x)} + \frac{1^2\cdot 3^2}{2!\,(8x)^2} - \frac{1^2\cdot 3^2\cdot 5^2}{3!\,(8x)^3} + \ldots \right\}$$

$$= \left(\frac{\pi}{2x}\right)^{1/2} e^{-x} \left\{ \sum_{j=o}^{k} \frac{A_j}{x^j} \right\}$$

For n = 1,

$$K_1(x) = \left(\frac{\pi}{2x}\right)^{1/2} e^{-x} \left\{ 1 + \frac{1\cdot 3}{1!\,(8x)} - \frac{1^2\cdot 3\cdot 5}{2!\,(8x)^2} + \frac{1^2\cdot 3^2\cdot 5\cdot 7}{3!\,(8x)^3} - \ldots \right\}$$

$$= \left(\frac{\pi}{2x}\right)^{1/2} e^{-x} \left\{ \sum_{j=o}^{k} \frac{B_j}{x^j} \right\}$$

A computer tabulation of the first fifty of these coefficients is shown in Table I.

---

1 *Handbook of Mathematical Functions*, NBS55, U.S. Government Printing Office, 1964, p. 377.

2 *Modern Analysis*, E. J. Whittaker and G. N. Watson, University Press, Cambridge, England, 1927, p. 343.

| I | AH(I)/(I-1)! | BH(I)/(I-1)! |
|---|---|---|
| 1 | 0.100000000000000E 01 | 0.100000000000000E 01 |
| 2 | -0.125000000000000E 00 | 0.375000000000000E 00 |
| 3 | 0.351562500000000E-01 | -0.585937500000000E-01 |
| 4 | -0.122070312500000E-01 | 0.170898437500000E-01 |
| 5 | 0.467300415039063E-02 | -0.600814819335938E-02 |
| 6 | -0.189256668090820E-02 | 0.231313705444336E-02 |
| 7 | 0.795140862464905E-03 | -0.939711928367615E-03 |
| 8 | -0.342803075909615E-03 | 0.395542010664940E-03 |
| 9 | 0.150645882968092E-03 | -0.170732000697171E-03 |
| 10 | -0.671862039780535E-04 | 0.750904632695892E-04 |
| 11 | 0.303177745450967E-04 | -0.335091192340542E-04 |
| 12 | -0.138121266264335E-04 | 0.151275672575224E-04 |
| 13 | 0.634254773036746E-05 | -0.689407361996464E-05 |
| 14 | -0.293202095523644E-05 | 0.316658263165535E-05 |
| 15 | 0.136316535482613E-05 | -0.146414056629473E-05 |
| 16 | -0.636901146338206E-06 | 0.680825363327048E-06 |
| 17 | 0.298858399233895E-06 | -0.318139586281243E-06 |
| 18 | -0.140768510711813E-06 | 0.149299935603438E-06 |
| 19 | 0.665283277862541E-07 | -0.703299465168972E-07 |
| 20 | -0.315364545496475E-07 | 0.332411277685473E-07 |
| 21 | 0.149896710531293E-07 | -0.157583721327770E-07 |
| 22 | -0.714218736970249E-08 | 0.749058675359042E-08 |
| 23 | 0.341061581781506E-08 | -0.356924911166692E-08 |
| 24 | -0.163196999789118E-08 | 0.170450199779746E-08 |
| 25 | 0.782339784145318E-09 | -0.815630838789800E-09 |
| 26 | -0.375679564346582E-09 | 0.391013424115830E-09 |
| 27 | 0.180684642541690E-09 | -0.187770314798227E-09 |
| 28 | -0.870272909635814E-10 | 0.903113396791883E-10 |
| 29 | 0.419734622392911E-10 | -0.434997699570835E-10 |
| 30 | -0.202692893602046E-10 | 0.209804924956504E-10 |
| 31 | 0.979963836984338E-11 | -0.101318295010245E-10 |
| 32 | -0.474303516833861E-11 | 0.489854451812020E-11 |
| 33 | 0.229798664344921E-11 | -0.237093860038411E-11 |
| 34 | -0.111443911484997E-11 | 0.114872954915304E-11 |
| 35 | 0.540951252872135E-12 | -0.557099051465333E-12 |
| 36 | -0.262802950502473E-12 | 0.270420427328632E-12 |
| 37 | 0.127776781778835E-12 | -0.131376127744436E-12 |
| 38 | -0.621733446036719E-13 | 0.638767239078820E-13 |
| 39 | 0.302739840197069E-13 | -0.310812902602324E-13 |
| 40 | -0.147513520096024E-13 | 0.151345040098518E-13 |
| 41 | 0.719243655405693E-14 | -0.737452355542546E-14 |
| 42 | -0.350904046930157E-14 | 0.359568344385223E-14 |
| 43 | 0.171299459984542E-14 | -0.175427157815494E-14 |
| 44 | -0.836694563539963E-15 | 0.856381494446786E-15 |
| 45 | 0.408893411120479E-15 | -0.418293259651984E-15 |
| 46 | -0.199928685770698E-15 | 0.204421465226220E-15 |
| 47 | 0.978030155285417E-16 | -0.999525323533448E-16 |
| 48 | -0.478665845012651E-16 | 0.488959734152708E-16 |
| 49 | 0.234372789238237E-16 | -0.239306953222199E-16 |
| 50 | -0.114807037377268E-16 | 0.117174192787109E-16 |

Table I. Coefficients for Asymptotic Series

These series can be summed by convergent factorial series using an algorithm described by Wasow:[3]

$$x^{-p} = \sum_{r=p-1}^{\infty} \frac{\Gamma^r_{r-p+1}}{x(x+1)(x+2)\ \ldots\ (x+r)}$$

where $\Gamma$ denotes the Stirling numbers of the first kind.

Now, $K_o(x) = \left(\frac{\pi}{2x}\right)^{1/2} e^{-x}\, S_o$

where $S_o = 1 - \frac{1^2}{1!(8x)} + T_o$

$$T_o = \sum_{j=2}^{k} \frac{A_j}{x^j} = \frac{A_2}{x^2} + \frac{A_3}{x^3} + \ldots$$

Applying Wasow's algorithm to these terms,

$$\frac{A_2}{x^2} = A_2 \left( \frac{\Gamma_o^{\,1}}{x(x+1)} + \frac{\Gamma_1^{\,2}}{x(x+1)(x+2)} + \frac{\Gamma_2^{\,3}}{x(x+1)(x+2)(x+3)} + \ldots \right),$$

$$\frac{A_3}{x^3} = A_3 \left( \frac{\Gamma_o^{\,2}}{x(x+1)} + \frac{\Gamma_1^{\,3}}{x(x+1)(x+2)} + \frac{\Gamma_2^{\,4}}{x(x+1)(x+2)(x+3)} + \ldots \right),$$

Therefore, $T_o$ can be expressed as

$$T_o = \sum_{r=1}^{\infty} \frac{V_{o,r}}{x(x+1)\ \ldots\ (x+r)}$$

where $V_{o,r} = A_2\, \Gamma^r_{r-1} + A_3\, \Gamma^r_{r-2} + A_4\, \Gamma^r_{r-3} + \ldots$

These coefficients can be calculated and stored in the memory of the computer for recall on demand.

The calculations for these coefficients, involving Stirling numbers, lead to very large numbers in the computation of high-order terms.

Since the Stirling numbers are always greater than or equal to one, we modified them for optimal use of the full range of the computer.

---

[3] *Asymptotic Expansions for Ordinary Differential Equations*, W. Wasow, Interscience Publishers, John Wiley, NY, 1965, p. 330.

The Stirling numbers were modified in the following way:

$$S_1^1 = F,\ F = \text{scale factor, such as } 10^{125}$$

$$S_1^n = S_1^{n-1}/(n-1)$$

$$S_k^n = S_{k-1}^{n-1} + S_k^{n-1}/(n-1)$$

The scale factor and the number of modified Stirling numbers which can be calculated are machine-dependent. The computers at BRL have a range from $10^{-155}$ to $10^{155}$, single precision, which is larger than the range of most computers. As can be seen from Table II, for $F = 10^{125}$ and $n = 150$, the modified Stirling numbers range from $10^{-135}$ to $10^{125}$. The process of scaling the Stirling numbers in this way must then be reversed in calculating each term of the factorial series.

By this transformation, we obtained accurate results (15 significant digits) for $x \geq 6$ by summing 150 terms. Similar accuracy could be obtained on most computers using double precision.

Similarly, $K_1(x) = \left(\frac{\pi}{2x}\right)^{1/2} e^{-x} S_1$

where $S_1 = 1 + \frac{1 \cdot 3}{1!(8x)} + T_1$

$$T_1 = \sum_{r=1}^{\infty} \frac{V_{1,r}}{x(x+1) \ . \ . \ . \ (x+r)}$$

where $V_{1,r} = B_2\Gamma_{r-1}^r + B_3\Gamma_{r-2}^r + B_4\Gamma_{r-3}^r + \ . \ . \ .$

The results for $K_1(x)$ were equally accurate.

The asymptotic series for the ordinary Bessel functions, $x \leq 25$, are:[4]

$$J_o(x) = \left(\frac{2}{\pi x}\right)^{1/2} [P_o(x) \cos(x - \tfrac{\pi}{4}) - Q_o(x) \sin(x - \tfrac{\pi}{4})]$$

$$J_1(x) = \left(\frac{2}{\pi x}\right)^{1/2} [P_1(x) \cos(x - \tfrac{3\pi}{4}) - Q_1(x) \sin(x - \tfrac{3\pi}{4})]$$

---

[4] *Bessel Functions, Part I, published by British Association for the Advancement of Science, University Press, Cambridge, England, 1937, p. 202.*

| | | | |
|---|---|---|---|
| 1 | 0.262541431038901-135 | 0.293390049185972-131 | 0.162469629570885-127 |
| 4 | 0.594416307877133-124 | 0.161631807256184-120 | 0.348403288776085-117 |
| 7 | 0.620095188981095-114 | 0.937263899085794-111 | 0.122805989253782-107 |
| 10 | 0.141690995161720-104 | 0.145745824899596-101 | 0.1349946811499E-98 |
| 13 | 0.113521536148685E-95 | 0.872724632875019E-93 | 0.616964862448953E-90 |
| 16 | 0.403105100240183E-87 | 0.244485921578491E-84 | 0.138176751796042E-81 |
| 19 | 0.730187886089758E-79 | 0.361878183982837E-76 | 0.168651283284991E-73 |
| 22 | 0.740917060108629E-71 | 0.307506122625604E-68 | 0.120810725011220E-65 |
| 25 | 0.450102130711339E-63 | 0.159290231850353E-60 | 0.536288624778063E-58 |
| 28 | 0.172006626264944E-55 | 0.526244225186191E-53 | 0.153759178190215E-50 |
| 31 | 0.429520654391673E-48 | 0.114831115329136E-45 | 0.294089524864916E-43 |
| 34 | 0.722149952063180E-41 | 0.170161172544539E-38 | 0.385045773882694E-36 |
| 37 | 0.837326275585929E-34 | 0.175105193623330E-31 | 0.352369350644976E-29 |
| 40 | 0.682727806061237E-27 | 0.127434483546970E-24 | 0.229266634594032E-22 |
| 43 | 0.397758943581573E-20 | 0.665766457973381E-18 | 0.107555387883677E-15 |
| 46 | 0.167774015729906E-13 | 0.252791612712235E-11 | 0.368043389831136E-09 |
| 49 | 0.517937210655300E-07 | 0.704743584443506E-05 | 0.927441656806729E-03 |
| 52 | 0.118075771165955E 00 | 0.145466542291410E 02 | 0.173458453446450E 04 |
| 55 | 0.200240793190783E 06 | 0.223832004301461E 08 | 0.242317917476252E 10 |
| 58 | 0.254107442106383E 12 | 0.258158475955319E 14 | 0.254129580523208E 16 |
| 61 | 0.242426888397106E 18 | 0.224137435817219E 20 | 0.200863910466421E 22 |
| 64 | 0.174495356761975E 24 | 0.146958814419687E 26 | 0.119996196936958E 28 |
| 67 | 0.950002380342511E 29 | 0.729268853726516E 31 | 0.542840153293270E 33 |
| 70 | 0.391821615456175E 35 | 0.274247724384982E 37 | 0.186138969273117E 39 |
| 73 | 0.122508995716187E 41 | 0.781856783017240E 42 | 0.483840847139701E 44 |
| 76 | 0.290319945584933E 46 | 0.168899572458664E 48 | 0.952645306311109E 49 |
| 79 | 0.520899483162194E 51 | 0.276096523091475E 53 | 0.141844276858788E 55 |
| 82 | 0.706254632117985E 56 | 0.340768370204050E 58 | 0.159312784381071E 60 |
| 85 | 0.721564602846948E 61 | 0.316568545939851E 63 | 0.134510966747108E 65 |
| 88 | 0.553438118324812E 66 | 0.220455227036434E 68 | 0.850010949437138E 69 |
| 91 | 0.317167294679423E 71 | 0.114501970441310E 73 | 0.399846121274463E 74 |
| 94 | 0.135025685209439E 76 | 0.440823733512481E 77 | 0.139095522997561E 79 |
| 97 | 0.424060880287624E 80 | 0.124873374341855E 82 | 0.355050280369826E 83 |
| 100 | 0.974387656698037E 84 | 0.258006379428166E 86 | 0.658888038138901E 87 |
| 103 | 0.162215177246175E 89 | 0.384836375802625E 90 | 0.879348340649532E 91 |
| 106 | 0.193432814981343E 93 | 0.409407659018037E 94 | 0.833293416906864E 95 |
| 109 | 0.163005483966049E 97 | 0.306267211155246E 98 | 0.552343491325650E 99 |
| 112 | 0.955495084811848+100 | 0.158431163016062+102 | 0.251599210263367+103 |
| 115 | 0.382364608112042+104 | 0.555606274242301+105 | 0.771214486559342+106 |
| 118 | 0.102158281160150+108 | 0.129004865734295+109 | 0.155126469001113+110 |
| 121 | 0.177415994653956+111 | 0.192739054938949+112 | 0.198618366362429+113 |
| 124 | 0.193865171243942+114 | 0.178944908244716+115 | 0.155930607458550+116 |
| 127 | 0.128034645087299+117 | 0.988621557189148+117 | 0.716280594664574+118 |
| 130 | 0.485782753348748+119 | 0.307581787025027+120 | 0.181290891228498+121 |
| 133 | 0.991494859813093+121 | 0.501359095737511+122 | 0.233459123119021+123 |
| 136 | 0.996588263274595+123 | 0.388006214031004+124 | 0.136972391170442+125 |
| 139 | 0.435467363384560+125 | 0.123698885459199+126 | 0.311019186727229+126 |
| 142 | 0.684403248787668+126 | 0.129991588080456+127 | 0.209419799774947+127 |
| 145 | 0.279759712120229+127 | 0.300552651141317+127 | 0.248534593164330+127 |
| 148 | 0.147742753080902+127 | 0.558451392197723+126 | 0.100000000000000+126 |

Table II. Modified Stirling Numbers for n = 150

$$Y_o(x) = \left(\frac{2}{\pi x}\right)^{1/2} [P_o(x) \sin(x - \tfrac{\pi}{4}) + Q_o(x) \cos(x - \tfrac{\pi}{4})]$$

$$Y_1(x) = \left(\frac{2}{\pi x}\right)^{1/2} [P_1(x) \sin(x - \tfrac{3\pi}{4}) + Q_1(x) \cos(x - \tfrac{3\pi}{4})] ,$$

where $P_o(x) \sim 1 - \frac{1^2 \cdot 3^2}{2!(8x)^2} + \frac{1^2 \cdot 3^2 \cdot 5^2 \cdot 7^2}{4!(8x)^4} - \ldots$

$$= \sum_{j=o}^{k} \frac{C_j}{x^{2j}}$$

and $Q_o(x) \sim - \frac{1^2}{1!(8x)} + \frac{1^2 \cdot 3^2 \cdot 5^2}{3!(8x)^3} - \frac{1^2 \cdot 3^2 \cdot 5^2 \cdot 7^2 \cdot 9^2}{5!(8x)^5} + \ldots$

$$= \sum_{j=o}^{k} \frac{D_j}{x^{2j+1}}$$

Note that $C_o = |A_o|$, $C_1 = -|A_2|$, . . . , $C_j = (-1)^j |A_{2j}|$

and $D_o = -|A_1|$, $D_1 = |A_3|$, . . . , $D_j = (-1)^{j+1} |A_{2j+1}|$

Similarly,

$$P_1 \sim \sum_{j=o}^{k} \frac{E_j}{x^{2j}} \quad \text{and} \quad Q_1 \sim \sum_{j=o}^{k} \frac{F_j}{x^{2j+1}}$$

And, again, $E_o = |B_o|$, $E_1 = -|B_2|$, . . . , $E_j = (-1)^j |B_{2j}|$

$F_o = |B_1|$, $F_1 = -|B_3|$, . . . , $F_j = (-1)^j |B_{2j+1}|$

For the ordinary Bessel functions, $x > 25$,

$J_o(x) = G(x) \sin(x) + H(x) \cos(x)$

$J_1(x) = M(x) \sin(x) - N(x) \cos(x)$

$Y_o(x) = H(x) \sin(x) - G(x) \cos(x)$

$Y_1(x) = -N(x) \sin(x) - M(x) \cos(x)$

where $G(x) = (\pi x)^{-1/2} [P_o(x) - Q_o(x)]$

$H(x) = (\pi x)^{-1/2} [P_o(x) + Q_o(x)]$

$M(x) = (\pi x)^{-1/2} [P_1(x) + Q_1(x)]$

$N(x) = (\pi x)^{-1/2} [P_1(x) - Q_1(x)]$

So, for $x > 25$, the same coefficients are merely arranged in a different manner.

As before, the results obtained were accurate to 15 significant digits for $x \geq 6$ by summing 150 terms. A sample tabulation of the ordinary Bessel functions from the computer is shown in Table III.

We attempted to calculate $I_n(x)$ in the same manner but the factorial series diverged.

3. HADAMARD SERIES. The factorial series for calculating $I_o(x)$ and $I_1(x)$ diverge since the Laplace integrals representing these functions are taken between finite limits and, therefore, cannot be expanded according to the previous algorithm. The Hadamard series, useful for large x, was used instead and has been programmed.

$I_n(x)$ can be expressed by:[5]

$$I_n(x) = \frac{(x/2)^n}{\Gamma(n+1/2)\Gamma(1/2)} \int_o^\pi e^{x \cos\theta} \sin^{2n} \theta \, d\theta$$

After expansion and term-by-term integration, the Hadamard series can then be written in the form

$$I_n(x) = \frac{e^x (2x)^{-1/2}}{\Gamma(n+1/2)\Gamma(1/2)} \sum_{m=o}^{\infty} \frac{(1/2-n)_m \, \gamma(n+m+1/2, 2x)}{m!(2x)^m}$$

where $\gamma$ denotes the incomplete gamma function and $(1/2-n)_m$ denotes Pochhammer's symbol.

---

[5] *Theory of Bessel Functions*, 2nd Ed., G. N. Watson, Macmillan Co., N.Y., 1948, p. 204.

[6] *Handbook of Mathematical Functions*, NBS 55, U.S. Government Printing Office, 1964, pp. 262, 504.

| X | K | J0 | J1 | Y0 | Y1 | |
|---|---|---|---|---|---|---|
| 6 | 150 | 0.150645257250995E 00 | -0.276683858127566E 00 | -0.288194683981577E 00 | -0.175010344300406E 00 | FACT.SERIES |
| 6 | | 0.150645257250997E 00 | -0.276683858127566E 00 | -0.288194683981579E 00 | -0.175010344300398E 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 7 | 150 | 0.300079270519555E 00 | -0.468282348234564E-02 | -0.259497439672093E-01 | -0.3026672370241855 00 | FACT.SERIES |
| 7 | | 0.300079270519556E 00 | -0.468282348234592E-02 | -0.259497439672093E-01 | -0.3026672370241855 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 8 | 150 | 0.171650807137554E 00 | 0.234636346853915E 00 | 0.223521489387566E 00 | -0.158060461731248E 00 | FACT.SERIES |
| 8 | | 0.171650807137554E 00 | 0.234636346853915E 00 | 0.223521489387566E 00 | -0.158060461731247E 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 9 | 141 | -0.903336111828761E-01 | 0.245311786573325E 00 | 0.249936698285025E 00 | 0.104314575196716E 00 | FACT.SERIES |
| 9 | | -0.903336111828762E-01 | 0.245311786573325E 00 | 0.249936698285025E 00 | 0.104314575196716E 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 10 | 112 | -0.245935764451348E 00 | 0.434727461688615E-01 | 0.556711672835995E-01 | 0.249015424206954E 00 | FACT.SERIES |
| 10 | | -0.245935764451349E 00 | 0.434727461688616E-01 | 0.556711672835995E-01 | 0.249015424206954E 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 11 | 93 | -0.171190300407196E 00 | -0.176785298956722E 00 | -0.168847323892079E 00 | 0.163705537414943E 00 | FACT.SERIES |
| 11 | | -0.171190300407196E 00 | -0.176785298956722E 00 | -0.168847323892079E 00 | 0.163705537414943E 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 12 | 79 | 0.476893107968335E-01 | -0.223447104490628E 00 | -0.225237312634361E 00 | -0.570992182608965E-01 | FACT.SERIES |
| 12 | | 0.476893107968336E-01 | -0.223447104490627E 00 | -0.225237312634362E 00 | -0.570992182608967E-01 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 13 | 69 | 0.206926102377068E 00 | -0.703180521217785E-01 | -0.782078645278760E-01 | -0.210081408420693E 00 | FACT.SERIES |
| 13 | | 0.206926102377068E 00 | -0.703180521217787E-01 | -0.782078645278759E-01 | -0.210081408420693E 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 14 | 61 | 0.171073476110459E 00 | 0.133375154698793E 00 | 0.127192568582184E 00 | -0.166644841856172E 00 | FACT.SERIES |
| 14 | | 0.171073476110458E 00 | 0.133375154698793E 00 | 0.127192568582184E 00 | -0.166644841856172E 00 | SUBROUTINE |
| X | K | J0 | J1 | Y0 | Y1 | |
| 15 | 56 | -0.142244728267808E-01 | 0.205104038613523E 00 | 0.205464296038918E 00 | 0.210736280368735E-01 | FACT.SERIES |
| 15 | | -0.142244728267808E-01 | 0.205104038613522E 00 | 0.205464296038919E 00 | 0.210736280368736E-01 | SUBROUTINE |

Table III. Computer Tabulation of Ordinary Bessel Functions

$$(a)_n = a(a+1)(a+2) \ldots (a+n+1), \quad (a)_o = 1$$

Each term in the expansion of these series contains the incomplete gamma function, which is expressed below in terms of the Kummer function.

$$\gamma(a,x) = a^{-1} x^a e^{-x} M(1, 1+a, x) ,$$

where M denotes the Kummer function.

Hence, after substituting and simplifying, we have

$$I_n(x) = \frac{e^{-x}(2x)^n}{\Gamma(n+1/2)\Gamma(1/2)} \sum_{m=o}^{\infty} \frac{(1/2-n)_m \, M\,(1,\ n+m+3/2,\ 2x)}{(n+m+1/2)\ m!} \quad .$$

The solution of these series is straightforward and presented no problems in overflowing the memory of the computer. The calculation of the Kummer function required many terms (250 terms for x=75) to get the required accuracy. The solutions of the Hadamard series seem to have the correct convergent behavior. A sample computer tabulation is shown in Table IV.

The results were good but not as accurate for moderate argument as we had hoped. We obtained 15 significant digits for $x \geq 17$ by summing 25 terms or less in the Hadamard series. This is not much better than the asymptotic series given by *

$$I_o(x) = \frac{e^x}{(2\pi x)^{1/2}} \left\{ 1 + \frac{1^2}{1!(8x)} + \frac{1^2 \cdot 3^2}{2!(8x)^2} + \frac{1^2 \cdot 3^2 \cdot 5^2}{3!(8x)^3} + \ldots \right\}$$

$$I_1(x) = \frac{e^x}{(2\pi x)^{1/2}} \left\{ 1 - \frac{1 \cdot 3}{1!(8x)} - \frac{1^2 \cdot 3 \cdot 5}{2!(8x)^2} - \frac{1^2 \cdot 3^2 \cdot 5 \cdot 7}{3!(8x)^3} - \ldots \right\}$$

When the asymptotic series were programmed, we obtained 15 significant digits for $x \geq 19$. However, the Hadamard series does provide an independent check on the accuracy of the asymptotic series used in our Bessel function subroutine.

4. FUTURE PLANS. Extension of the factorial series program to complex variables is now in progress. After thorough checking of the accuracy and speed of computation, a decision will be made on which series will be used in our Bessel function subroutine.

Besides the immediate value in verifying the accuracy of the subroutine, we are developing insight and practical numerical procedures for calculating Laplace integrals for both finite and infinite limits of integration. These integrals arise in the solution of ordinary differential equations.

---

* Reference 4, p. 271.

| X | I0(X) | I1(X) | M | |
|---|---|---|---|---|
| 17.0 | 0.235497022316828E 07 | 0.228462158380809E 07 | 25 | Hadamard Series |
| 17.0 | 0.235497022316829E 07 | 0.228462158380808E 07 | | Subroutine |
| 18.0 | 0.621841242078099E 07 | 0.604313324211564E 07 | 22 | |
| 18.0 | 0.621841242078101E 07 | 0.604313324211563E 07 | | |
| 19.0 | 0.164461904406117E 08 | 0.160073737858370E 08 | 21 | |
| 19.0 | 0.164461904406117E 08 | 0.160073737858370E 08 | | |
| 20.0 | 0.435582825595536E 08 | 0.424549733851279E 08 | 19 | |
| 20.0 | 0.435582825595536E 08 | 0.424549733851278E 08 | | |
| 21.0 | 0.115513961922158E 09 | 0.112729199137776E 09 | 18 | |
| 21.0 | 0.115513961922158E 09 | 0.112729199137775E 09 | | |
| 22.0 | 0.306692993640365E 09 | 0.299639606877380E 09 | 17 | |
| 22.0 | 0.306692993640365E 09 | 0.299639606877379E 09 | | |
| 23.0 | 0.815142122512893E 09 | 0.797220026089651E 09 | 17 | |
| 23.0 | 0.815142122512892E 09 | 0.797220026089650E 09 | | |
| 24.0 | 0.216861908824138E 10 | 0.212294789328732E 10 | 16 | |
| 24.0 | 0.216861908824138E 10 | 0.212294789328731E 10 | | |
| 25.0 | 0.577456060646631E 10 | 0.565786512987871E 10 | 15 | |
| 25.0 | 0.577456060646631E 10 | 0.565786512987870E 10 | | |
| 26.0 | 0.153889767056608E 11 | 0.150900726423417E 11 | 15 | |
| 26.0 | 0.153889767056608E 11 | 0.150900726423416E 11 | | |

Table IV. Computer Tabulation of $I_o(x)$ and $I_1(x)$

# FINITE AND INFINITE INHOMOGENEOUS LADDER NETWORKS

C. C. Yang
Mathematics Science Staff, Plasma Physics Division
Naval Research Laboratory, Washington, DC

T. N. Lee
The George Washington University
Washington, DC

## INTRODUCTION

It is well known that by given sufficient spectral data, the entries of a continued fraction expansion relate intimately to the density function of the inverse Sturm-Liouville problem.[1,2] The investigation of the pole-zero distribution of a continued fraction with each of its entries a different complex function is significant because of the simple implementation analytically and numerically. However, although traced back to 19th century, the literature shows very little of this kind of study. A recent paper by Lee and Brown,[3] sheds some light on the pole-zero distribution pattern of the immitance function of finite inhomogeneous ladder networks by using the chain matrix parameter method.

In this paper, continued fractions with complex function entries are first studied in a general setting. The pole-zero distribution region is described by a conventional root locus equation and is found to be bounded in the corresponding complex plane. The applications of the theorems are illustrated by examples.

## PRELIMINARY DEFINITIONS

Let $R^+$ denote the positive real line, C the whole complex plane and P.R. any positive real rational function. Two polynomials are said to

be relative prime polynomials or simply r.p., if they do not have any common factor.

A. The set of arcs in C satisfying the root locus[4] equation

$$1 + \frac{k}{G(s)} = 0, \; k \in \left\{R^{+} \cup 0\right\},$$

is denoted by $\left[G(s), k\right]$. Therefore, $\left[G(s), k\right]$ starts from the set of zeros of $G(s)$ at $k = 0$ and ends at the set of poles of $G(s)$ at $k = \infty$.

B. Suppose $P(s) = k_p \prod_{i=1}^{N} \left[G(s) + p_i\right]$ and

$$Q(s) = k_q \prod_{i=1}^{N(\text{or } N-1)} \left[G(s) + q_i\right].$$

If $0 < p_i < q_i < p_{i+1} \; \forall i = 1, 2, \cdots, N-1$, then the zeros of $P(s)$ and $Q(s)$ are said to alternate with respect to $\left[G(s), k\right]$. The zeros of $P(\omega)$ and $Q(\omega)$ alternate on the negative $\omega = G(s)$ axis of the $\omega$-plane and thus the zeros of $P(s)$ and $Q(s)$ alternate along each locii of

$$1 + \frac{k}{G(s)} = 0, \; k \in \left\{R^{+} \cup 0\right\} \quad \text{in s-plane.}$$

C. We shall denote the following continued fraction expansion, or C.F., by $F_N\left[f_i z(s), g_i y(s)\right]$ if

$$F_N(s) = Z_N + \cfrac{1}{Y_N + \cfrac{1}{Z_{N-1} + \cfrac{1}{Y_{N-1} + \cdots + \cfrac{1}{Z_i + \cfrac{1}{Y_i + \cdots + \cfrac{1}{Z_1 + \cfrac{1}{Y_1}}}}}}}, \tag{1}$$

where $Z_i = f_i z(s)$, $Y_i = g_i y(s)$, $\forall i = 1, 2, \cdots, N$, are the entries of the C.F. and $z(s)$, $y(s)$ are two different complex functions of s.

POLE-ZERO DISTRIBUTION OF FINITE CONTINUED FRACTION OF ARBITRARY COMPLEX FUNCTION ENTRIES

Consider C.F. $F_N\left[f_i z(s),\ g_i y(s)\right] = A_N/C_N$, then multiplying both sides by $y(s)$ yields

$$A_N/y(s)^{-1} C_N = f_N \omega + \cfrac{1}{g_N + \cfrac{1}{f_{N-1}\omega + \cdots + \cfrac{1}{f_1 \omega + \cfrac{1}{g_1}}}}, \qquad (2)$$

where $\omega = z(s)\, y(s)$. Therefore, $A_N$ and $y^{-1} C_N$ are functions of $\omega$.

Lemma 1: In the C.F. $F_N\left[f_i z(s),\ g_i y(s)\right] = A_N/C_N$, if $f_i,\ g_i \in R^+$, $\forall i = 1, 2, \cdots, N$, then

$a_1$: the zeros of $A_N(\omega)$ and $y^{-1} C_N(\omega)$ interlace on the negative real axis of $\omega$-plane with $0 < \alpha_i < \gamma_i < \alpha_{i+1}$ for $i = 1, 2, \cdots, N-1$, where $-\alpha_i$ and $-\gamma_i$ are the zeros of $A_N(\omega)$ and $y^{-1} C_N(\omega)$ respectively,

$b_1$: $A_N(\omega)\big|_{\omega=0} = 1$ and $y^{-1} C_N(\omega)\big|_{\omega=0} = \sum_{i=1}^{N} g_i$.

Proof: By elementary property of two-element-kind R-L ladder networks, $a_1$ follows immediately from the expression of (2). To show $b_1$, mathematical induction is used. Suppose the expression holds for N = n case, then

$$A_{n+1}(\omega)/y^{-1} C_{n+1}(\omega) = f_{n+1}\omega + \cfrac{1}{g_{n+1} + A_n(\omega)/y^{-1} C_n(\omega)},$$

which yields,

$$A_{n+1}(\omega) = f_{n+1}\,\omega\left[g_{n+1}A_n(\omega) + y^{-1}C_n(\omega)\right] + A_n(\omega)$$

and

$$y^{-1}C_{n+1}(\omega) = g_{n+1}A_n(\omega) + y^{-1}C_n(\omega)\,.$$

Hence,

$$A_{n+1}(\omega)\big|_{\omega=0} = A_n(\omega)\big|_{\omega=0} = 1\,,$$

and

$$y^{-1}C_{n+1}(\omega)\big|_{\omega=0} = g_{n+1} + \sum_{i=1}^{n} g_i = \sum_{i=1}^{n+1} g_i\,.$$

This completes the proof of the lemma.

The following corollary is the direct consequence of the above lemma.

Corollary 1: If the same hypothesis of the foregoing lemma holds for the C.F. $F_N\left[f_i z(s),\ g_i y(s)\right]$, then

$a_2$: the zeros of $A_N(s)$ and $C_N(s)$ alternate with respect to $[z(s)y(s),\ k]$,

$b_2$: $A_N(s)\big|_{\{s\,|\,z(s)y(s)\,=\,0\}} = 1$ and $y(s)^{-1}C_N(s)\big|_{\{s\,|\,z(s)y(s)\,=\,0\}} = \sum_{i=1}^{N} g_i$.

The following facts are observed

$a_3$: Consider $F_N\left[f_i z(s),\ g_i y(s)\right] = A_N(s)/C_N(s)$, then

$$A_N(s) = \left(\prod_{i=1}^{N} \alpha_i\right)^{-1} \prod_{i=1}^{N} \left[z(s)y(s) + \alpha_i\right], \qquad (3)$$

$$y^{-1} C_N(s) = \left( \prod_{i=1}^{N-1} \gamma_i \right)^{-1} \left( \sum_{i=1}^{N} g_i \right) \prod_{i=1}^{N-1} \left[ z(s)y(s) + \gamma_i \right] . \qquad (4)$$

$b_3$: if $z(s) = n_a(s)/d_a(s)$ and $y(s) = n_b(s)/d_a(s)$, where $n_a(s)$ and $d_a(s)$, $n_b(s)$ and $d_b(s)$ are r.p., then we have

Case 1: $n > m$, where n = degree of $(n_a(s)n_b(s))$ , m = degree of

$$(d_a(s)d_b(s))\, A_N(s) = \left( \prod_{i=1}^{N} \alpha_i \right)^{-1} \prod_{i=1}^{nN} \left( s - z_{a_i} \right)$$

and

$$C_N(s) = \left( \prod_{i=1}^{N-1} \gamma_i \right)^{-1} \left( \sum_{i=1}^{N} g_i \right) d_a(s)n_b(s) \prod_{i=1}^{n(N-1)} \left( s - z_{c_i} \right),$$

where $z_{a_i}$ and $z_{c_i}$ alternate with respect to $\left[ n_a(s)n_b(s)/d_a(s)d_b(s),\ k \right]$ .

Case 2: for $m > n$, then the above explicit forms remain the same except for the upper running indices of the product of the factors, using m instead of n.

In what follows the decomposition theorem pertinent to the synthesis of a finite ladder network is established.

Theorem 1: Let $Z(s) = A(s)/C(s)$ be a rational function.

$Z(s) = F_N \left[ f_i z(s),\ g_i y(s) \right]$ , $f_i,\ g_i \epsilon R^+,\ \forall i < N$; $z(s) = n_a(s)/d_a(s)$ , $y(s) = n_b(s)/d_b(s)$, $n_a(s)$ and $d_a(s)$, $n_b(s)$ and $d_b(s)$ are r.p., iff. A(s) and C(s) satisfy the following conditions,

$a_4$: the zeros of A(s) and $y(s)^{-1}$ C(s) alternate with respect to $\left[ z(s)y(s),\ k \right]$ ,

$b_4$: $$\left[d_a(s)d_b(s)\right]^{-1} A(s) \Big|_{\{s \mid n_a(s)n_b(s) = 0\}} = 1$$

and

$$\left[n_a(s)n_b(s)\right]^{-1}\left[d_a(s)d_b(s)\right]^{-(N-1)} C(s) \Big|_{\{s \mid n_a(s)n_b(s) = 0\}} = \sum_{i=1}^{N} g_i > 0 ,$$

$c_4$: for $n > m$, $n$ = degree of $(n_a(s)n_b(s))$, $m$ = degree of $(d_a(s)d_b(s))$, $A(s)$ and $y(s)^{-1} C(s)$ are polynomials of degree $nN$ and $n(N-1)$; for $m > n$, $A(s)$ and $y(s)^{-1} C(s)$ are polynomials of degree $mN$.

Proof. The "only if" part: It follows trivially from Lemma 1 and its corollary.

The "if" part: Let $A(s)$ and $C(s)$ satisfy condition $a_4$ through $c_4$. It follows from definition B,

$$A(s) = \left(\prod_{i=1}^{N} \alpha_i\right)^{-1} \prod_{i=1}^{N} \left[n_a(s)n_b(s) + \alpha_i d_a(s)d_b(s)\right]$$

and

$$C(s) = \left(\prod_{i=1}^{N-1} \gamma_i\right)^{-1} \left(\sum_{i=1}^{N} g_i\right) n_b(s)d_a(s) \prod_{i=1}^{N-1} \left[n_a(s)n_b(s) + \gamma_i d_a(s)d_b(s)\right] ,$$

where

$$0 < \alpha_i < \gamma_i < \alpha_{i+1}, \quad \forall i = 1, 2, \cdots, N - 1.$$

Therefore,

$$A(s)/C(s) = \frac{\left(\prod_{i=1}^{N} \alpha_i\right)^{-1} \prod_{i=1}^{N} \left[z(s)y(s) + \alpha_i\right]}{y(s)\left(\prod_{i=1}^{N-1} \gamma_i\right)^{-1} \left(\sum_{i=1}^{N} g_i\right) \prod_{i=1}^{N-1} \left[z(s)y(s) + \gamma_i\right]} ,$$

which yields,

$$A(\omega)/y^{-1}C(\omega) = A(s)/y(s)^{-1}C(s)\,\Big|_{z(s)y(s)\,=\,\omega}$$

$$= \frac{a_N \omega^N + a_{N-1}\omega^{N-1} + \cdots\cdots\cdots + a_1\omega + 1}{c_{N-1}\omega^{N-1} + c_{N-2}\omega^{N-2} + \cdots\cdots + c_1\omega + c_o} \quad ,$$

where

$$a_N = \left(\prod_{i=1}^{N} \alpha_i\right)^{-1}, \quad b_{N-1} = \left(\prod_{i=1}^{N-1} \gamma_i\right)^{-1}\left(\sum_{i=1}^{N} g_i\right), \quad b_o = \left(\sum_{i=1}^{N} g_i\right),$$

and

$$a_i,\ b_i > 0,\ \forall i \ .$$

Write,

$$A(\omega)/y^{-1}C(\omega) = f_N\omega + \frac{1}{g_N + \left[A^*/y^{-1}C^*\right]^{-1}} \quad ,$$

where $f_N = a_N/b_{N-1} > 0$, $\quad g_N = b_{N-1}/a_N\left[a_{N-1}/a_N - b_{N-2}/b_N\right] > 0$ ,

and

$$A^* = A(\omega) - f_N\omega y^{-1}C(\omega) = a^*_{N-1}\omega^{N-1} + \cdots + a^*_1\omega + 1 \ , \qquad (5)$$

$$y^{-1}C^* = y^{-1}C(\omega) - g_N A^* = b^*_{N-2}\omega^{N-2} + \cdots + b^*_1\omega \ , \qquad (6)$$

then

$$a^*_i,\ b^*_i > 0 \ .$$

Moreover, the interlacing zeros of $A(\omega)$ and $y^{-1}C(\omega)$ in the negative real

axis of $\omega$-plane implies that the zeros of $A^*$ and $y^{-1}C^*$ alternate on the same negative real axis, by hypothesis $a_4$ through $c_4$ and Fig. 1 shows the locations of the zeros of $A^*$. Therefore, the zeros of $y^{-1}C^*$ and $A^*$ interlace on the negative $\omega$-axis starting with the first zero belonging to $A^*$ as shown in Fig. 2 following the same argument. Hence, $A^*$ and $y^{-1}C^*$ can then be written as

$$A^* = k_{a^*} \prod_{i=1}^{N-1} \left(\omega + \alpha_i^*\right) ,$$

and

$$y^{-1}C^* = k_{c^*} \prod_{i=1}^{N-2} \left(\omega + \gamma_i^*\right) , \quad i = 1, 2, \cdots, N-1, \quad 0 < \alpha_i < \gamma_i < \alpha_{i+1} ,$$

where

$$k_a^* = \left( \prod_{i=1}^{N-1} \alpha_i^* \right)^{-1} ,$$

followed from

$$A^*\Big|_{\omega=0} = A(\omega) - f_N \omega y^{-1} C(\omega)\Big|_{\omega=0} = 1 ,$$

and

$$k_c^* = \left( \prod_{i=1}^{N-2} \gamma_i^* \right)^{-1} \left( \sum_{i=1}^{N-1} g_i \right) ,$$

followed from

$$y^{-1}C^*\Big|_{\omega=0} = y^{-1}C(\omega) - g_N A^*\Big|_{\omega=0} = \sum_{i=1}^{N} g_i - g_N = \sum_{i=1}^{N-1} g_i .$$

It is easily seen that $A^*$ and $C^*$ satisfy conditions $a_4$ through $c_4$ by simply substituting back $\omega=z(s)y(s)$ to $A^*$ and $C^*$, except for the degree of $A^*(s)$ and $y^{-1}C^*(s)$ are n or m degree less than that of corresponding degree of $A(s)$ and $y(s)^{-1}C(s)$ respectively.

Therefore, this process is continued until N = 1, in this case

$$A_1(\omega)/y^{-1}C_1(\omega) = a_1\omega + 1/c_1 ,$$

where $a_1 = f_1 > 0$ and $c_1 = g_1 > 0$ .

Q.E.D.

BOUNDEDNESS OF THE MODULUS OF ZEROS AND POLES OF $F_N\left[f_i z(s),\ g_i y(s)\right]$

In what follows the uniform bound is found.

Lemma 2: In the C.F. $F_N\left[f_i z(s),\ g_i y(s)\right]$ , if $f_i,\ g_i \epsilon R^+, \forall i,$ then $\partial/\partial\omega\left\{y(s)F_N\left[f_i z(s),\ g_i y(s)\right]\right\}_{z(s)y(s)=\omega} > 0, \forall\omega$, real and $\omega \neq \gamma_i$ where $\gamma_i$ are the poles of $y(s)F_N\left[f_i z(s),\ g_i y(s)\right]_{z(s)y(s)=\omega}$ in the $\omega$-plane.

Proof: Straightforward computation shows

$$\partial/\partial\omega\left[yF_N(\omega)\right] = \frac{\partial/\partial\omega\left[yF_{N-1}(\omega)\right] + f_N\left\{g_N\left[yF_{N-1}(\omega)\right] + 1\right\}^2}{\left\{g_N\left[yF_{N-1}(\omega)\right] + 1\right\}^2} , \forall\omega \neq y_i ,$$

where $-\gamma_i$ are the poles of $yF_N(\omega)$. The foregoing relation implies that if $\partial/\partial\omega\left[yF_{N-1}(\omega)\right] > 0$, then $\partial/\partial\omega\left[yF_N(\omega)\right] > 0$. But N = 1, 2..are trivially true and hence the lemma.

Lemma 3: In the C.F. $F_N\left[f_i z(s),\ g_i y(s)\right]$ , if $f_i,\ g_i \epsilon R^+, \forall i,$ then

$$\partial/\partial f_i \left\{ y(s) F_N \left[ f_i z(s),\ g_i y(s) \right] \right\}_{z(s)y(s) = \omega} < 0$$

and

$$\partial/\partial g_i \left\{ y(s) F_N \left[ f_i z(s),\ g_i y(s) \right] \right\}_{z(s)y(s) = \omega} < 0, \forall \omega,$$

real and nonpositive.

Proof: It follows from (2),

$$\partial/\partial f_N \left[ yF_N(\omega) \right] < 0, \forall \omega \in R^+ ,$$

$$\partial/\partial g_N \left[ yF_N(\omega) \right] = - \frac{1}{\left[ yF_N^*(\omega) \right]^2} < 0, \forall \omega .$$

Let

$$yF_j^*(\omega) = g_j + \frac{1}{f_{j-1}\omega + \dfrac{1}{g_{j-1} + \cdots + \dfrac{1}{f_1 \omega + \dfrac{1}{g_1}}}} \tag{7}$$

Then, simple computation yields, for any $k < N$

$$\partial/\partial f_k \left[ yF_N(\omega) \right] = (-1)^{2(N-k)} \frac{\omega}{\left[ yF_N^*(\omega) \right]^2 \left[ yF_{N-1}(\omega) \right]^2 \cdots \left[ yF_k(\omega) \right]^2}$$

$$\partial/\partial g_k \left[ yF_N(\omega) \right] = (-1)^{2(N-k)+1} \frac{1}{\left[ yF_N^*(\omega) \right]^2 \cdots \left[ yF_k(\omega) \right]^2 \left[ yF_k^*(\omega) \right]^2}$$

It is obvious that the right hand sides of the above equations are nonpositive for $\omega \in R^+$ This completes the proof of the lemma.

Theorem 2: Let there be two C.F., $F_N\left[f_i z(s),\ g_i y(s)\right]$ and $F_N^*\left[f_i^* z(s),\ g_i^* y(s)\right]$. If $f_i > f_i^* \epsilon R^+$, $g_i > g_i^* \epsilon R^+$, $i < N$, then $\alpha_i < \alpha_i^*$, $\forall i < N$ and $\gamma_j < \gamma_j^*$, $\forall j < N - 1$, where $-\alpha_i$, $-\alpha_i^*$ and $-\gamma_j$, $-\gamma_j^*$ are the zeros and the poles of $yF_N(\omega)$ and $yF_N^*(\omega)$ respectively.

Proof: Since we have by Lemma 2, $yF_N(\omega)$ is a monotonically increasing function of $\omega \neq -\gamma_i$, by Lemma 3, same function is an nonincreasing function of $f_i$ and $g_i$, $\forall \omega$, real and nonpositive. Therefore all the zeros of $yF_N(\omega)$ shift to the right on the real $\omega$-axis as all the entries $f_i$ and $g_i$ increase in value, as shown in Fig. 3. This gives $\alpha_i < \alpha_i^*$, $\forall i < N$. The result of the poles of $yF_N(\omega)$ and $yF_N^*(\omega)$ follows by using the same argument to the function of $\left[yF_N(\omega)\right]^{-1}$ and it is omitted here.

Q.E.D.

It is noted that if the entries of the C.F. $F_N\left[f_i z(s),\ g_i y(s)\right] = A_N/C_N$ are uniform, $f_i = f \epsilon R^+$, $g_i = g \epsilon R^+$, $i < N$, then we have

$$A_N(\omega) = \sinh (N + 1)\, a(\omega) - \sinh Na(\omega)/\sinh a(\omega) \tag{8}$$

$$y^{-1} C_N(\omega) = g \sinh Na(\omega)/\sinh a(\omega) , \tag{9}$$

where

$$\cosh a(\omega) = 2 + fg\omega/2 .$$

Lemma 4: Let the entries of the C.F. $F_N\left[f_i z(s),\ g_i y(s)\right] = A_N/C_N$ be uniform as defined above, then

$a_5$:
$$\gamma_k = 2\left(1 - \cos\frac{k\pi}{N}\right)\Big/fg, \quad \forall k < N - 1 \tag{10}$$

and

$b_5$:
$$2\left[1 - \cos\frac{(2k-1)\pi}{2N}\right]\Big/fg < \alpha_k < 2\left(1 - \cos\frac{k\pi}{N}\right)\Big/fg, \quad \forall k < N. \tag{11}$$

Proof: Substituting the following identities into (8) and (9),

$$\cosh Na = 2^{N-1}\prod_{k=1}^{N}\left[\cosh a - \cos\frac{(2k-1)\pi}{2N}\right]$$

$$\sinh Na = 2^{N-1}\sinh a\prod_{k=1}^{N-1}\left[\cosh a - \cos\frac{k\pi}{N}\right],$$

results in $a_5$ and $b_5$ of the lemma by using the same argument as in Theorem 1 concerning to the sum of two polynomials with interlacing zeros on the real axis.

As a consequence, the following theorem is established.

Theorem 3: Let $-\alpha_i$, $-\gamma_i$ be the zeros and the poles of the C.F. $F_N\left[f_i z(s), g_i(s)\right]\Big|_{z(s)y(s)=\omega}$. If $f_i, g_i \in R^+$, then

$$0 < 2\left[1 - \cos\frac{(2k-1)\pi}{2N}\right]\Big/\underline{f}\,\underline{g} < \alpha_k < \gamma_k =$$

$$2\left(1 - \cos\frac{k\pi}{N}\right)\Big/\underline{f}\,\underline{g} < \alpha_N < 4/\underline{f}\,\underline{g} \quad \forall k < N - 1, \tag{12}$$

where

$$\underline{f} = \inf_{i<N} f_i, \quad \underline{g} = \inf_{i<N} g_i, \quad \bar{f} = \sup_{i<N} f_i \text{ and } \bar{g} = \sup_{i<N} g_i.$$

Proof: The above result follows immediately from Theorem 2 and Lemma 4, since the zeros and the poles of the C.F. $y(s)F_N\left[f_i z(s), g_i y(s)\right]\Big|_{z(s)y(s)=\omega}$ shift to the right on the negative real $\omega$-axis by the increasing in value of all its entries $f_i$, $g_i$ $\forall i$. Therefore, these zeros and poles are bounded in modulus by that of the zeros and the poles of the two corresponding C.F. of uniform entries each with $\underline{f} = \inf_{i<N} f_i$, $\underline{g} = \inf_{i<N} g_i$ and $\bar{f} = \sup_{i<N} f_i$, $\bar{g} = \sup_{i<N} g_i$, respectively.

Q.E.D.

ASYMPTOTIC DISTRIBUTION OF THE POLES AND THE ZEROS OF THE SEQUENCE OF THE CONTINUED FRACTIONS

Let $\left\{F_N\left[{}_N f_i z(s), {}_N g_i y(s)\right]\right\}$ be defined as a sequence of C.F. for $N = 1, 2, \cdots$ Now for each fixed N, the entries of the corresponding C.F. are ${}_N f_i = {}_N a_i/N^{\alpha}$ and ${}_N g_i = {}_N c_i/N^{\beta}$, $i < N$ In what follows the result pertaining to the integrated networks are derived.

Theorem 4: If ${}_N a_i$ and ${}_N c_i$, for all N and i, of the above defined sequence are bounded away from zero, then ${}_N\alpha_k$ and ${}_N\gamma_k = O(N^{\alpha+\beta})$ for sufficient large N and k, where $-{}_N\alpha_k$ and $-{}_N\gamma_k$ are the kth zero and pole of the corresponding C.F. $F_N\left[{}_N f_i z(s), {}_N g_i y(s)\right]$ in the $\omega = z(s)y(s)$ plane.

Proof: Since ${}_N a_i$ and ${}_N c_i$ are bounded away from zero for all N and i, we choose

$$\bar{a} = \sup_{i<N} {}_N a_i \text{ for all N and } \bar{c} = \sup_{i<N} {}_N c_i \text{ for all N .}$$

Hence a sequence of uniform C.F. $\left\{F_N\left[\bar{f}_N z(s),\ \bar{g}_N y(s)\right]\right\}$ with $\bar{f}_N = a/N^{\alpha}$ and $\bar{g}_N = \bar{c}/N^{\beta}$ has the following relationship by Lemma 4,

$$2\left[1 - \cos\frac{(2k-1)\pi}{2N}\right]\frac{N^{\alpha+\beta}}{\bar{a}\bar{c}} < {}_N\bar{\alpha}_k < {}_N\bar{\gamma}_k\ ,\quad \forall N,\ k < N-1\ ,$$

where ${}_N\bar{\alpha}_k$ and ${}_N\bar{\gamma}_k$ are the kth zero and pole of the uniform C.F. $F_N\left[\bar{f}_N z(s),\ \bar{g}_N y(s)\right]$ in the $\omega$-plane.

It follows from Theorem 3, we have,

$$2\left[1 - \cos\frac{(2k-1)\pi}{2N}\right]\frac{N^{\alpha+\beta}}{\bar{a}\bar{c}} < {}_N\alpha_k < {}_N\gamma_k\ ,\quad \forall N,\ k < N-1\ .$$

The conclusion of the theorem follows.

Q.E.D.

REMARK

Theorem 4 is used to investigate the asymptotic behavior of the zeros and the poles of the nonuniform C.F. in $\omega$-plane as well as the convergence of $A_N(\omega)$ and $y^{-1}C_N(\omega)$ as $N \to \infty$. It follows in particular that if ${}_Nf_i = {}_Na_i/N$ and ${}_Ng_i = {}_Nc_i/N$, ${}_Na_i$, ${}_Nc_i$ bounded away from zero, then ${}_N\alpha_k$ and ${}_N\gamma_k = O(N^2)$, as $N \to \infty$, where ${}_N\alpha_k$, ${}_N\gamma_k$ are the zeros and the poles of the corresponding C.F. This result consistent with the result obtained from solving the transmission line equations for the distributed networks.

EXAMPLES AND APPLICATIONS

Example 1: Let the entries of the C.F. be $(f_3, f_2, f_1) = (8/15, 8/5, 48/5)$ and $(g_3, g_2, g_i) = (5/8, 5/16, 1/16)$; $z(s) = s/(s - 1)$ and $y(s) = 1/(s - 1)$.

Simple computation yields

$$F_3\left[f_i z(s), g_i y(s)\right] = A_3/C_3 =$$

$$\frac{8(15\,s^6 - 67\,s^5 + 142\,s^4 - 179\,s^3 + 142\,s^2 - 67\,s + 15)}{15(8\,s^5 - 34\,s^4 + 63\,s^3 - 63\,s^2 + 34\,s - 8)},$$

and $\left[z(s)y(s), k\right] = \left[s/(s - 1)^2, k\right]$ which satisfies the root locus equation of $1 + \frac{k(s - 1)^2}{s} = 0$, $k \in \left\{\text{OUR}^+\right\}$ and is shown in Fig. 4.

It follows that

$$A_3 = (1/15)\left[s - (1 + j\sqrt{3}/2)\right]\left[s - (1 - j\sqrt{3}/2)\right]\left[s - (5 + j\sqrt{11}/6)\right]$$
$$\left[s - (5 - j\sqrt{11}/6)\right]\left[s - (9 + j\sqrt{19}/10)\right]\left[s - j\sqrt{19}/10)\right]$$

$$y^{-1}C_3 = (1/8)\left[s - (3 + j\sqrt{7}/4)\right]\left[s - (3 - j\sqrt{7}/4)\right]\left[s - (7 + j\sqrt{15}/8)\right]$$
$$\left[s - (7 - j\sqrt{15}/8)\right].$$

The zeros of $A_3$ and $y^{-1}C_3$ are shown in Fig. 5. As can be seen that they alternate with respect to $\left[s/(s - 1)^2, k\right]$.

In example given below, Theorem 1 is used to realize a ladder network with a given immitance function.

Example 2: Let the poles and the zeros of a driving point impedance $Z(s)$ be specified at $-1, -2, -3 \pm j\sqrt{7}/2, -3 \pm j\sqrt{15}/2$ and at $-3 \pm j\sqrt{3}/2, -3 \pm j\sqrt{11}/2, -3 \pm j\sqrt{19}/2$, respectively.

Synthesis procedures:

1) Construct the pole-zero plot for $Z(s)$, as shown in Fig. 6.

2) Find an arc as shown in Fig. 7 passing through all these singularities. This arc is described by $\left[(s+1)(s+2), k\right]$ by inspection hence, let $z(s) = s + 1$ and $y(s) = s + 2$.

3) Multiplying out, results in

$$Z(s) = \frac{k_a(s^6 + 9\,s^5 + 42\,s^4 + 117\,s^3 + 206\,s^2 + 213\,s + 105)}{k_c(s^5 + 8\,s^4 + 31\,s^3 + 68\,s^2 + 84\,s + 48)}.$$

4) Since $A(s)\Big|_{\{s\,|\,(s+1)(s+2)=0\}} = 1$, yields $k_a = 1/15$, and $(s+2)^{-1}C(s)\Big|_{\{s\,|\,(s+1)(s+2)=0\}} = 4$, yields $k_c = 1/2$ (note that the number 4 is arbitrarily assumed which happens to be the total capacitance of the ladder network.), therefore, we have

$$(s+2)Z(s)\Big|_{(s+1)(s+2)=\omega} = \frac{2(\omega+1)(\omega+2)(\omega+3)}{15(\omega+2)(\omega+4)}.$$

5) Hence C.F. gives $(f_3, f_2, f_1) = (2/15, 6/15, 24/15)$ and $(g_3, g_2, g_1) = (15/6, 15/12, 1/4)$. Fig. 8 shows the corresponding network.

## CONCLUSION AND SUMMARY

The complete pole-zero pattern of a continued fraction of nonuniform entries is established using arcs in the s-plane defined by a simple root locus method.

A process of decomposition of rational functions satisfying the foregoing pole-zero patterns into continued fractions is used to synthesize general inhomogeneous ladder networks.

The analysis and synthesis results established are being extended to the case of infinite ladder networks ($N \to \infty$) and the problem of the transition between lumped and distributed networks.

## ACKNOWLEDGEMENT

The authors wish to thank Dr. David Brown of the University of Wisconsin, for his criticism and Dr. C. E. Carroll of the University of Pennsylvania for his idea in the proof of Lemma 2.

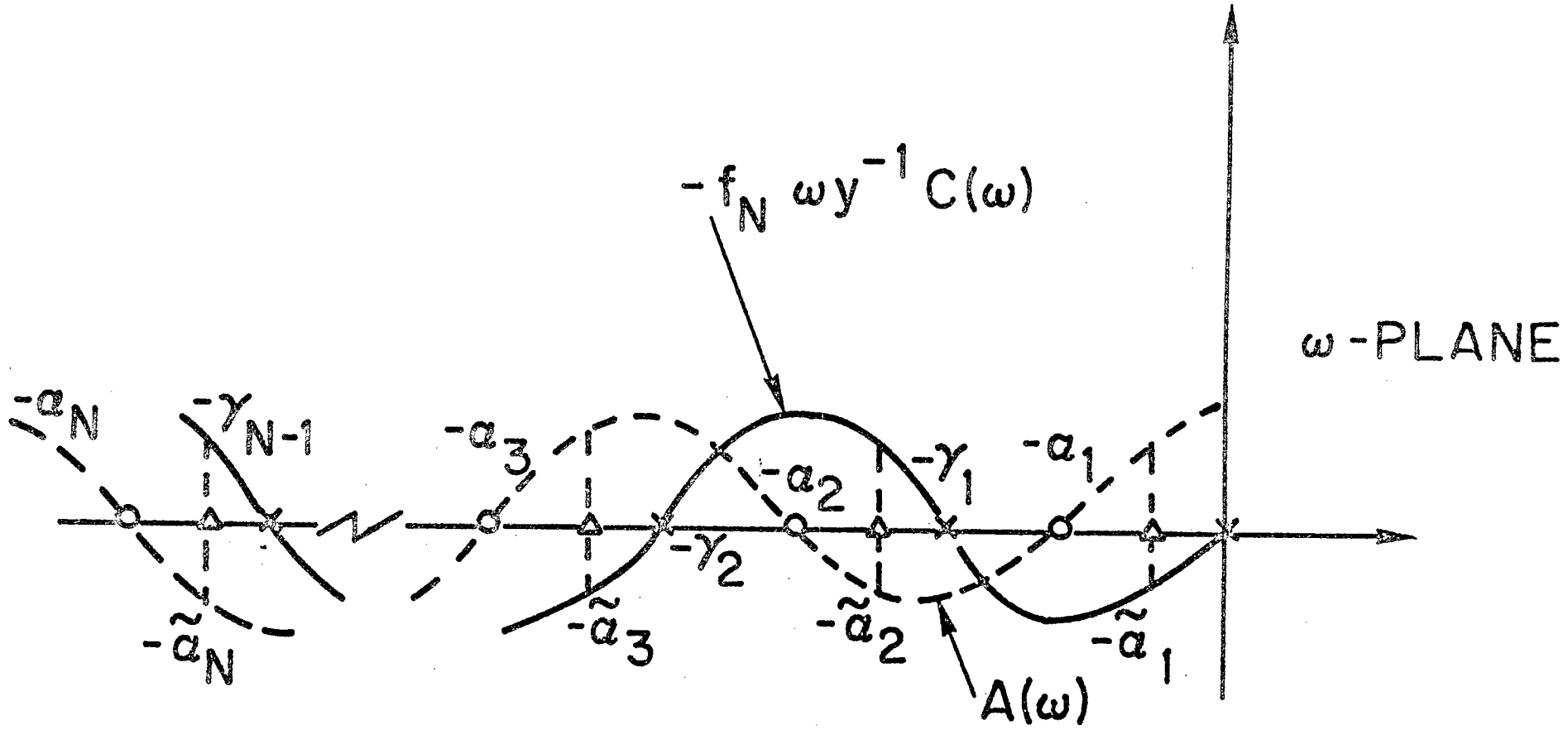
## REFERENCES

1. Bellman, R., "A Note on an Inverse Problem in Mathematical Physics," Quarterly J. Mech. Appl. Math. 19 (1961).

2. Anderson, L. E., "On the Defective Determination of the Wave Operator from Given Spectral Data in the Case of a Difference Equation Corresponding to a Sturm-Liouville Differential Equation," J. Math. Anal. and Appl. 29 (1970).

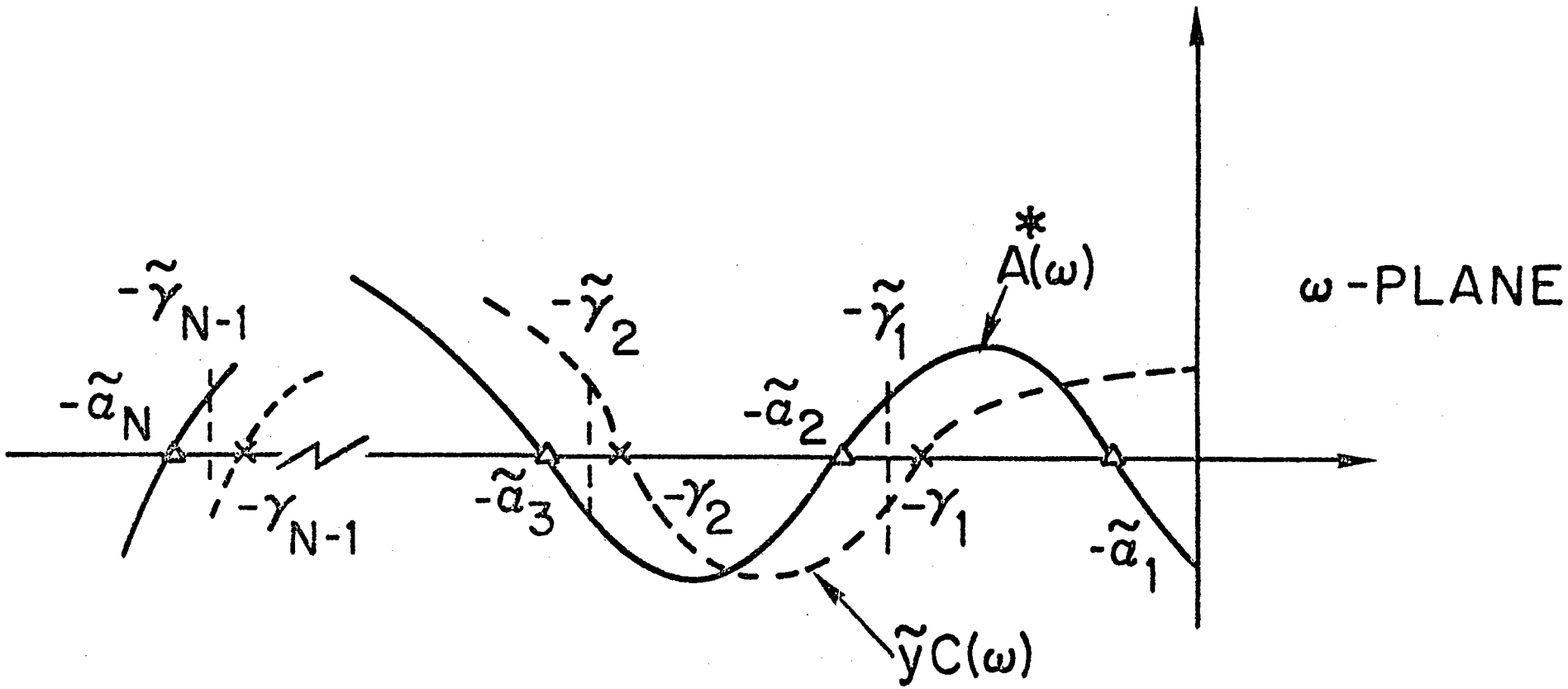
3. Lee and Brown, "Decomposition Theorem," Allerton Proc. on Ckt. and Syst. Th. (1974).

4. Evans, W., "Control System Dynamics," McGraw Hill, pp. 96-120 (1954).



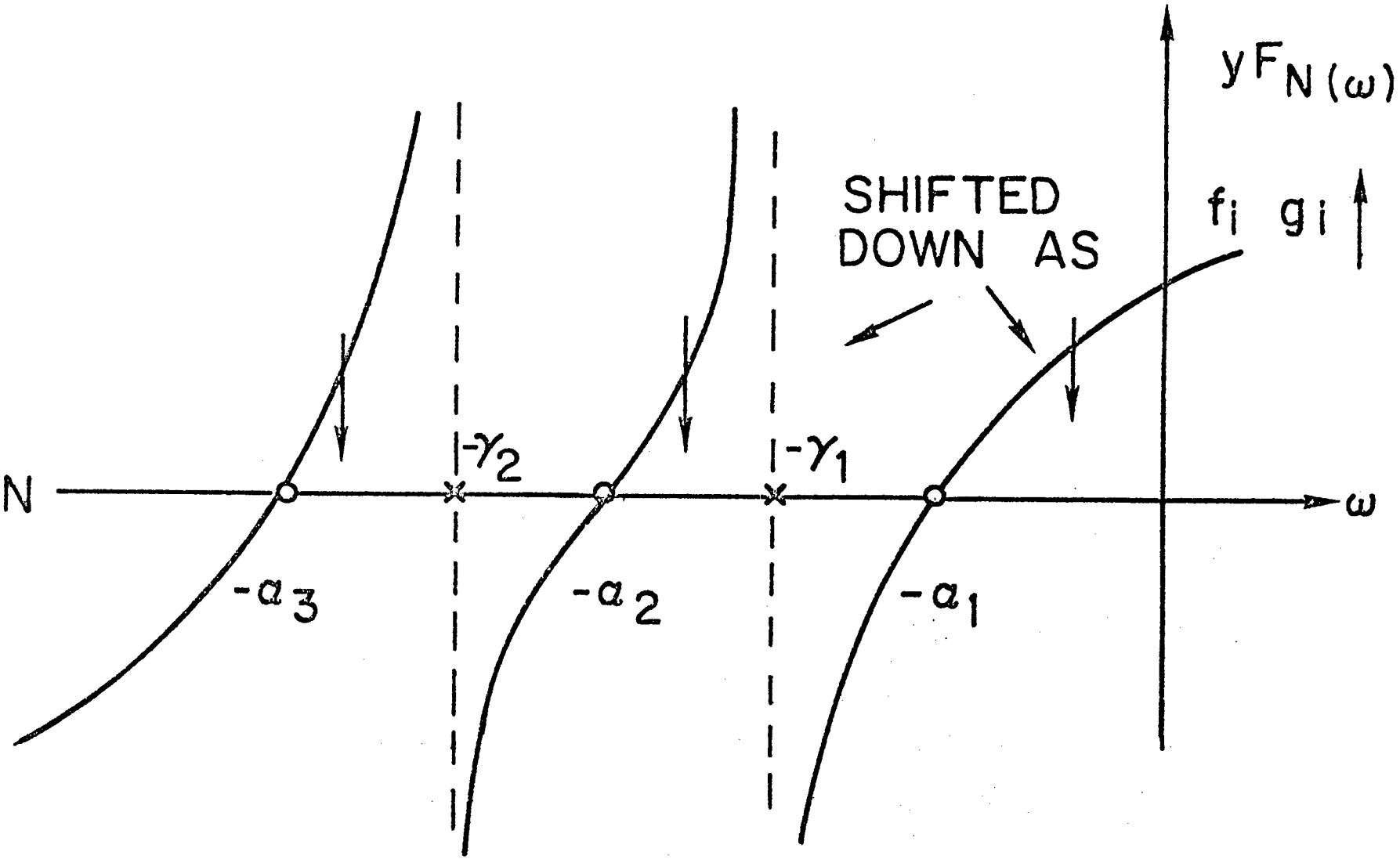

Locations of the zeros of A*

Figure 1



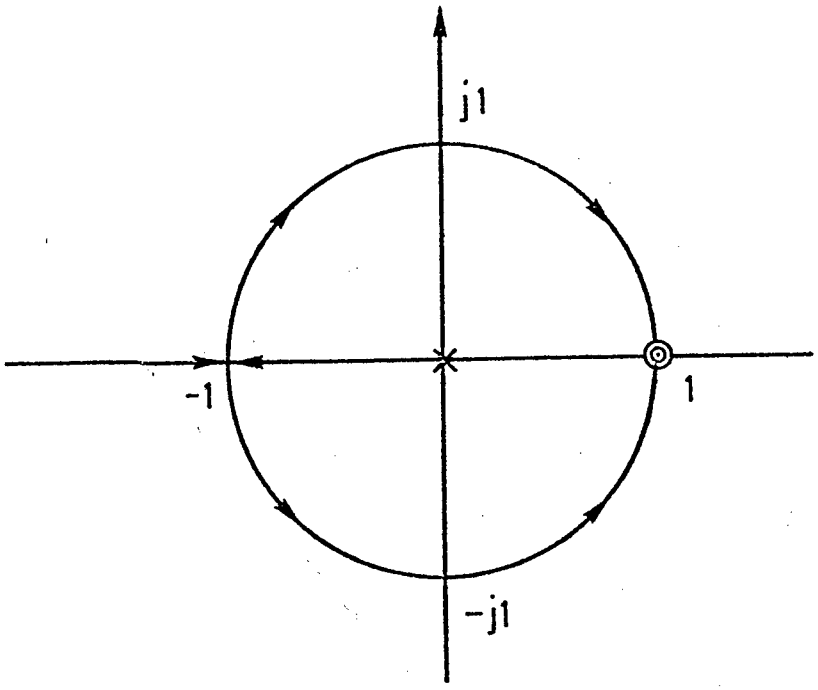

The interlace of the zeros of $y^{-1}C^*$ and $A^*$

Figure 2



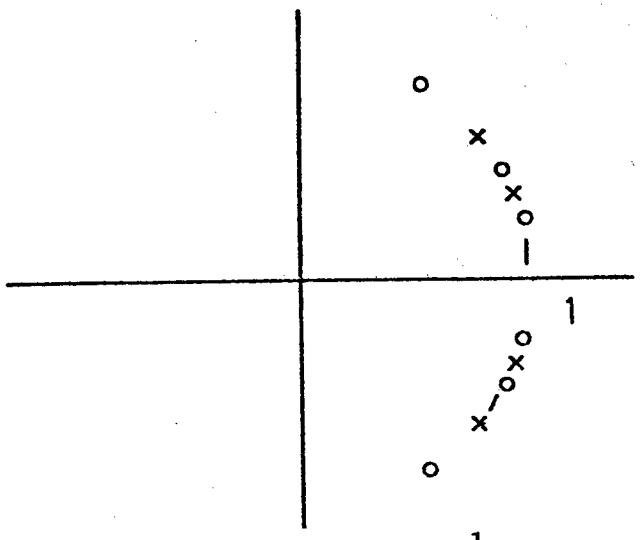

The shifted zeros of $yF_N(\omega)$ as $f_i$ and $g_i$ increase in value

Figure 3



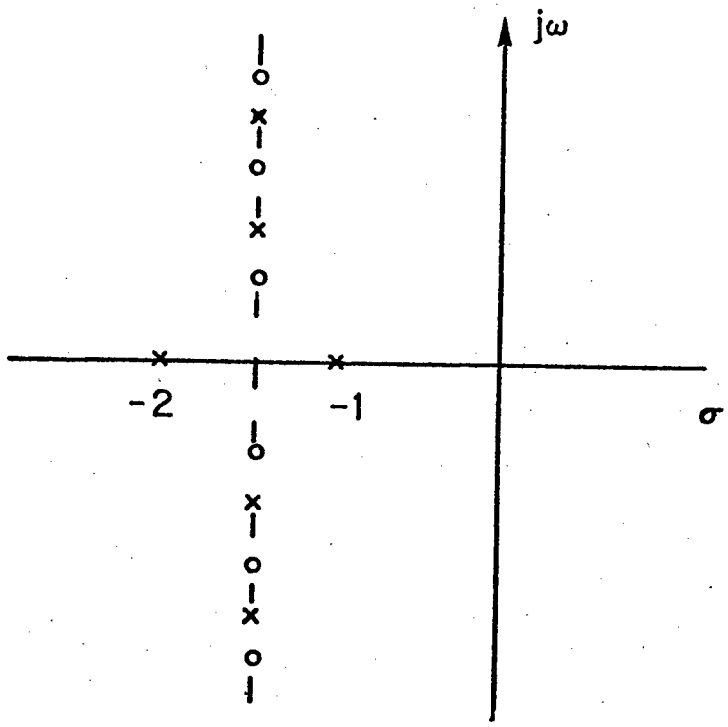

The root locus equation of $1 + k(s - 1)^2/s = 0$

Figure 4



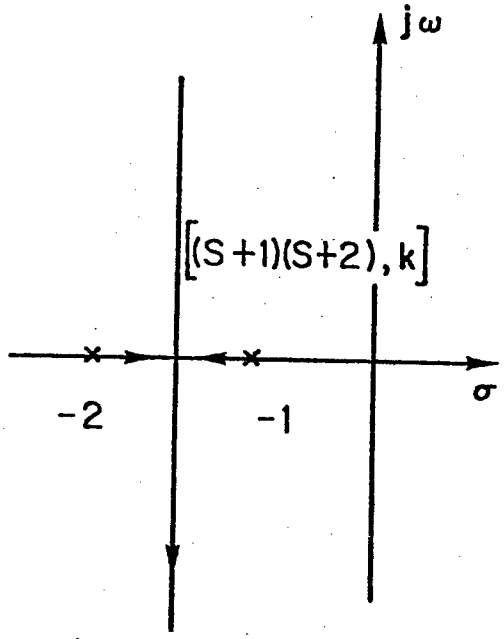

The zeros of $A_3$ and $y^{-1}C_3$

Figure 5



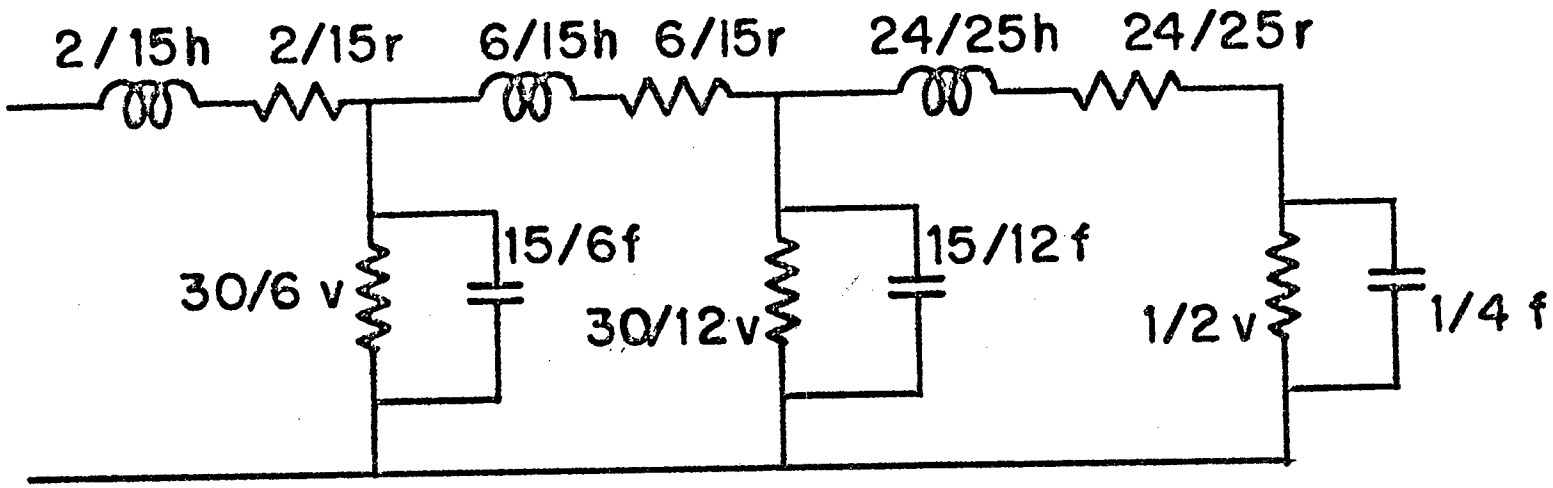

Location of the poles and zeros of Z(s)

Figure 6




Locations of all the singularities

Figure 7




The corresponding network

Figure 8

# AUTOMATIC NUMERICAL INTEGRATION USING VP-SPLINES

Royce W. Soanes, Jr.
Research Directorate
Benet Weapons Laboratory
Watervliet Arsenal
Watervliet, New York 12189

ABSTRACT. A method of exploiting VP (variable power) splines for the purpose of automatic numerical quadrature is presented. The essence of the adaptive method given here is to select mesh points near the node where an upper bound on the local area discrepancy between the trapezoidal estimate and the local VP spline estimate of the integral is a maximum. A comparison is made with Gaussian quadrature for an integral containing a parameter.

1. INTRODUCTION. The term "Automatic Integrator" refers to numerical integration algorithms which adapt themselves to the particular situation at hand. Automatic integrators are particularly handy for obtaining dependable integral estimates during computation on a problem which may involve many integrals and whose nature may change from time to time as the parameters involved fluctuate. They are also useful in situations where the integrand may be expensive (time consuming) to evaluate as is the case with multidimensional integrals.

The basic philosophy behind the automatic integration in this article will be to spend some computational overhead time in monitoring the region of the integrand where the VP spline interpolater is making the most significant contribution to the integral estimate (relative to the linear interpolater) and evaluate the integrand in these significant regions.

As increasingly more information is accumulated about the integrand, it will be possible for the algorithm to gradually abandon evaluation of the integrand over large regions of uniform behavior and transfer its attention to regions where the integrand behaves more abruptly. This process will generally produce a nonuniform mesh and it will be necessary to have on hand an interpolater which is smooth but stable. Variable power splines satisfy this requirement since they are twice differentiable and they may be given some local derivative control which renders them less likely to inject interpolatory oscillations.

2. SUMMARY OF BASIC VP SPLINE FORMULAS. The interpolatory functions used here are the VP (variable power) splines given on the $i$th subinterval by Eq. (1).

(1) $k_i y_i(x) = a_i+b_i r_i+c_i r_i^{m_i}+d_i(1-r_i)^{n_i}$

where $k_i = m_i+n_i-m_i n_i$ ,

$r_i = (x-x_i)/\ell_i$, and

$\ell_i = x_{i+1}-x_i$

The four parameters $a_i$, $b_i$, $c_i$ and $d_i$ may be eliminated in favor of $y_i$, $y_{i+1}$, $y_i'$ and $y_{i+1}'$ .

(2) $a_i = k_i y_i+\ell_i(m_i q_i-(m_i-1)y_i'-y_{i+1}')$

(3) $b_i = \ell_i(-m_i n_i q_i+m_i y_i'+n_i y_{i+1}')$

(4) $c_i = \ell_i(n_i q_i-y_i'-(n_i-1)y_{i+1}')$

(5) $d_i = \ell_i(-m_i q_i+(m_i-1)y_i'+y_{i+1}')$

where $q_i = (y_{i+1}-y_i)/\ell_i$

If second derivative continuity is enforced at the interior nodes and the curvature is set equal to zero at the end points, the following tridiagonal system of equations may be obtained.

(6) $(m_1-1)y_1'+y_2' = m_1 q_1$

(7) $A_i y_{i-1}'+B_i y_i'+C_i y_{i+1}' = D_i \quad (1<i<N)$

(8) $y_{N-1}'+(n_{N-1}-1)y_N' = n_{N-1} q_{N-1}$

The coefficients in Eq. (7) are given by equations (9-12).

(9) $A_i = m_{i-1}(m_{i-1}-1)/(k_{i-1}\ell_{i-1})$

$$(10)\quad C_i = n_i(n_i-1)/(k_i\ell_i)$$

$$(11)\quad B_i = (n_{i-1}-1)A_i+(m_i-1)C_i$$

$$(12)\quad D_i = n_{i-1}A_iq_{i-1}+m_iC_iq_i$$

The solution to the system described by equations (6-12) yields the nodal derivatives which insure continuity of the second derivative of the interpolater.

All that is needed now to completely define the interpolater is the setting of the nonlinear parameter vectors m and n. The values of $m_{i-1}$ and $n_i$ are set by obtaining a VP spline over the restricted node set $[x_{i-1}, x_i, x_{i+1}]$. Setting the end curvatures equal to zero and setting $y_i$ equal to the slope of the line through $(x_i,y_i)$ which makes equal angles with the linear interpolater on the left and right of $x_i$ yields Eq. (13).

$$(13)\quad n_i/m_{i-1} = (\ell_i/\ell_{i-1})\left[(q_{i-1}^2+1)/(q_i^2+1)\right]^{1/2}$$

Equation (13) sets the m's and n's while assuming a lower bound of L on them i.e., either $n_i = L$ or $m_{i-1} = L$. This lower bound L must be greater than 2 and it need not be greater than 3. Values of L greater than 3 tend to produce too much flattening of the interpolater between nodes.

3. INTEGRATION FORMULA. If the VP spline is integrated over the $i$th subinterval, we may obtain Eq. (14) after some rearrangement and simplification.

$$(14)\quad \int_{x_i}^{x_{i+1}} y_i(x)dx = (\ell_i/2)(y_i+y_{i+1}) + \Delta_i$$

where

$$\Delta_i = \ell_i^2\left[(1+k_i)\left\{m_i(y_i'-q_i) + n_i(q_i-y_{i+1}')\right\} + 2(y_{i+1}'-y_i')\right] / \left\{2k_i(m_i+1)(n_i+1)\right\}$$

The quantity $\Delta_i$ is the discrepancy between the trapezoidal estimate and the VP spline estimate of the integral over the $i$th subinterval. This expression for $\Delta_i$ is not dependent on the existence of second derivatives.

If $q_i$ is between $q_{i-1}$ and $q_{i+1}$ and $y'_k$ is between $q_{k-1}$ and $q_k$ ($k = i, i+1$) and $m_i = n_i = m$, the maximum value that $|\Delta_i|$ may take on is $\ell_i^2|q_{i-1}-q_{i+1}|/(6+4\sqrt{2})$ for an $m$ of $1 + \sqrt{2}$.

4. SIGNIFICANT NODES. An initial mesh over the desired interval of integration must be assumed. This mesh may be uniform, or prior analytic knowledge of the integrand may prompt the insertion of a node or two near an abruptness in the integrand. In any case, the initial mesh may be uniform or non-uniform and may contain as few as three points.

The relative significance of the various points in the sample must be determined first. This will be done by considering the behavior of a VP spline with zero end curvatures over the restricted node set $[x_{i-1}, x_i, x_{i+1}]$. Enforcement of second derivative continuity at node $i$ yields Eq. (15).

$$(15) \quad R_i = n_i/m_{i-1} = (\ell_i/\ell_{i-1})(q_{i-1}-y'_i)/(y'_i-q_i)$$

This equation implies Eq. (13) with $y'_i$ selected as previously mentioned. If Eq. (14) is used with the conditions for zero end curvatures, two simple integral formulas may be obtained for the three point VP spline.

$$(16) \quad \int_{x_{i-1}}^{x_i} y_{i-1}(x)dx = (\ell_{i-1}/2)(y_{i-1}+y_i) + u_i$$

$$(17) \quad \int_{x_i}^{x_{i+1}} y_i(x)dx = (\ell_i/2)(y_i+y_{i+1}) + v_i$$

where

$$u_i = (\ell_{i-1}^2/2)(q_{i-1}-y'_i)/(m_{i-1}+1)$$

and

$$v_i = (\ell_i^2/2)(y'_i-q_i)/(n_i+1)$$

The two area discrepancy terms $u_i$ and $v_i$ will be used to determine the significant points of the sample.

At this point, we want to notice the effect of $y_i'$ on $u_i$ and $v_i$ as it varies between the left and right difference quotients $q_{i-1}$ and $q_i$ which are taken to be the reasonable limits for the assignment of $y_i'$ locally.

From Eq. (15) we see that as $y_i'$ approaches $q_i$, $R_i$ approaches infinity. The value L is therefore assigned to $m_{i-1}$ as $n_i$ becomes infinite. The quantity $u_i$ therefore approaches its extreme value $u_i^*$ as $v_i$ approaches zero.

$$(18) \quad u_i^* = (\ell_{i-1}^2/2)(q_{i-1}-q_i)/(L+1)$$

Similarly, as $y_i'$ approaches $q_{i-1}$, $R_i$ approaches zero. Hence, $n_i$ is assigned the value L as $m_{i-1}$ becomes infinite. We therefore have $v_i$ approaching its extreme value $v_i^*$ while $u_i$ approaches zero.

$$(19) \quad v_i^* = (\ell_i^2/2)(q_{i-1}-q_i)/(L+1)$$

These extreme values of $u_i$ and $v_i$ gives us the significance weights that we will assign to the nodes in the sample.

$$(20) \quad w_i = (\ell_{i-1}^2+\ell_i^2)\,|\,q_i-q_{i-1}|$$

5. INTEGRATION ALGORITHM. The weight given by Eq. (20) is proportional to the sum of $|u_i^*|$ and $|v_i^*|$; it is an easily calculated measure of the possible disagreement which may exist between the VP spline estimate of the integral locally and the linear estimate. It behooves us, therefore, to examine the integrand more closely near the node where $w_i$ is presently the largest. An algorithm for automatic integration may therefore be summarized by the following procedural outline.

I. Generate an initial (not necessarily uniform) mesh over the interval of integration, evaluate the integrand and compute the trapezoidal estimate of the integral.

II. Compute the $i$th nodal significance weight according to Eq. (20) for $1<i<N$.

III. Find the node where $w_i$ is the largest.

IV. If the maximum weight is less than a given fraction of the trapezoidal estimate or if the number of functional evaluations exceeds a given amount, skip to VII., otherwise continue to V.

V. Evaluate the integrand at the midpoint of the $i^{th}$ ($(i-1)^{th}$) subinterval if $\ell_i$ is larger (smaller) than $\ell_{i-1}$.

VI. Update the x and y arrays and the trapezoidal estimate and recalculate the three appropriate nodal weights. Return to step III.

VII. Set the m's and n's according to Eq. (13).

VIII. Compute the nodal derivatives using equations (6-12).

IX. Compute the VP spline integral estimate using Eq. (14).

6. A TEST CASE. The following integral containing a parameter is considered here as a test case; it is obtained from a Weibull probability density.

$$(21) \quad \int_0^2 k(b)x^{b-1}e^{-x^b}dx = 1$$

where $k(b) = b/(1-e^{-2^b})$

As b becomes large, the integrand will become a tall spike centered near 1. The performance of VP spline adaptive integration will be compared with that of 32 point Gauss-Legendre quadrature. It is obvious that any quadrature formula using a constant mesh may be defeated by this integral if b is chosen large enough. The purpose of the comparison is therefore not to belabor this fact but to indicate that the adaptive method is capable of handling even this pathological case accurately and stably.

The following error table was computed for an L of 2.5. Only 32 functional evaluations were made for the VP spline integral estimates.

| b | VP | GAUSS |
|---|---|---|
| 2 | .00012 | -.00000000000000036 |
| 4 | - .000019 | -.00000000000000014 |
| 6 | .00016 | .0000000000060 |
| 8 | .00028 | .00000030 |
| 10 | .00054 | -.000043 |
| 12 | .00048 | .00011 |
| 14 | .00014 | .0032 |
| 16 | .00010 | .0069 |
| 18 | .00041 | .0034 |
| 20 | .00066 | -.012 |
| 22 | .00013 | -.04 |
| 24 | .000018 | -.079 |
| 26 | .00048 | -.13 |
| 28 | .00034 | -.18 |
| 30 | .00037 | -.23 |
| 40 | .00071 | -.47 |
| 50 | .00019 | -.60 |
| 70 | .00031 | -.78 |
| 90 | .00043 | -.90 |
| 110 | .0010 | -.95 |
| 130 | .00074 | -.98 |
| 150 | .00062 | -.99 |

For well behaved integrands, Gaussian quadrature seems to be unbeatable - as evidenced by the early entries in the table. The Gaussian accuracy deteriorates, however, as its mesh becomes less capable of detecting the spike. By the time b has reached a value of 150, Gaussian quadrature has "lost" 99% of the integral value. Adaptive VP spline integration, although not as accurate as Gaussian for small values of b, displays a uniform error pattern which is independent of b over a considerable range.

Needless to say, a much better parametric study than has been done here could be done for a variety of integrands. Fortran listings of relevant subroutines are given here as an appendix for those interested in using adaptive integration in a practical setting or for those who might be able to do a more complete parametric study.

```
      SUBROUTINE VPSD (NPNTS,X,Y,D,M,N)                                 VPSD0001
C     VPSD - VARIABLE POWER SPLINE DERIVATIVES                          VPSD0002
C     VPSD SOLVES A TRIDIAGONAL SYSTEM FOR THE NODAL                    VPSD0003
C     DERIVATIVES WHICH INSURE SECOND DERIVATIVE CONTINUITY             VPSD0004
C     OF THE VARIABLE POWER (VP) SPLINE                                 VPSD0005
C     NPNTS,X,Y=DATA                                                    VPSD0006
C     D=NODAL DERIVATIVES                                               VPSD0007
C     M,N=VARIABLE POWERS                                               VPSD0008
      IMPLICIT REAL*8 (A-H,O-Z)                                         VPSD0009
      DIMENSION X(1), Y(1), D(1), M(1), N(1)                            VPSD0010
      DOUBLE PRECISION M, N, KI, KIM                                    VPSD0011
      DIMENSION A(100), B(100), C(100)                                  VPSD0012
      NP=NPNTS                                                          VPSD0013
      NM=NP-1                                                           VPSD0014
      DXIM=X(2)-X(1)                                                    VPSD0015
      QIM=(Y(2)-Y(1))/DXIM                                              VPSD0016
      KIM=1.-(M(1)-1.)*(N(1)-1.)                                        VPSD0017
      B(1)=M(1)-1.                                                      VPSD0018
      C(1)=1.                                                           VPSD0019
      D(1)=M(1)*QIM                                                     VPSD0020
C     DEFINE TRIDIAGONAL SYSTEM                                         VPSD0021
      DO 1 I=2,NM                                                       VPSD0022
      DXI=X(I+1)-X(I)                                                   VPSD0023
      QI=(Y(I+1)-Y(I))/DXI                                              VPSD0024
      KI=1.-(M(I)-1.)*(N(I)-1.)                                         VPSD0025
      A(I)=M(I-1)*(M(I-1)-1.)/(KIM*DXIM)                                VPSD0026
      C(I)=M(I)*(N(I)-1.)/(KI*DXI)                                      VPSD0027
      B(I)=(N(I-1)-1.)*A(I)+(M(I)-1.)*C(I)                              VPSD0028
      D(I)=N(I-1)*A(I)*QIM+M(I)*C(I)*QI                                 VPSD0029
      DXIM=DXI                                                          VPSD0030
      QIM=QI                                                            VPSD0031
      KIM=KI                                                            VPSD0032
    1 CONTINUE                                                          VPSD0033
      A(NP)=1.                                                          VPSD0034
      B(NP)=N(NM)-1.                                                    VPSD0035
      D(NP)=N(NM)*QI                                                    VPSD0036
C     REDUCE MATRIX BELOW THE DIAGONAL                                  VPSD0037
      DO 2 I=1,NM                                                       VPSD0038
      Q=A(I+1)/B(I)                                                     VPSD0039
      B(I+1)=B(I+1)-Q*C(I)                                              VPSD0040
      D(I+1)=D(I+1)-Q*D(I)                                              VPSD0041
    2 CONTINUE                                                          VPSD0042
C     BACK SUBSTITUTE                                                   VPSD0043
      DO 3 J=1,NM                                                       VPSD0044
      I=NP-J                                                            VPSD0045
      D(I+1)=D(I+1)/B(I+1)                                              VPSD0046
      D(I)=D(I)-C(I)*D(I+1)                                             VPSD0047
    3 CONTINUE                                                          VPSD0048
      D(1)=D(1)/B(1)                                                    VPSD0049
      RETURN                                                            VPSD0050
      END                                                               VPSD0051
```

```
      SUBROUTINE ADVPSI (F,NPNTS,X,Y,MINEX,MAXEX,NTOT,YINT,TOL,L)          ADVP0001
C     ADVPSI - ADAPTIVE VARIABLE POWER SPLINE INTEGRATION                  ADVP0002
C     F=INTEGRAND                                                          ADVP0003
C     NPNTS,X=INITIAL MESH                                                 ADVP0004
C     Y=INTEGRAND VALUES                                                   ADVP0005
C     MINEX,MAXEX=MINIMUM AND MAXIMUM NO. OF EXTRA POINTS TO BE GENERATEADVP0006
C     NTOT=TOTAL NO. OF POINTS IN THE SAMPLE                               ADVP0007
C     YINT=DEFINITE INTEGRAL                                               ADVP0008
C     ADVPSI MAY EASILY BE CHANGED TO YIELD THE INDEFINITE                 ADVP0009
C     INTEGRAL AS WELL                                                     ADVP0010
C     TOL=STOPPING TOLERANCE                                               ADVP0011
C     IF TOLERANCE IS MET, TOL WILL BE GREATER THAN THE MAXIMUM LOCAL      ADVP0012
C     INTEGRAL DISCREPANCY DIVIDED BY THE TRAPEZOIDAL ESTIMATE             ADVP0013
C     THIS ROUTINE CALLS VPSD AND MNSET                                    ADVP0014
      IMPLICIT REAL*8 (A-H,O-Z)                                            ADVP0015
      DIMENSION X(1), Y(1)                                                 ADVP0016
      DIMENSION B(100), M(100), N(100), W(100)                             ADVP0017
      DOUBLE PRECISION M, N, L, KI                                         ADVP0018
      COMMON /PLT/ YIN(100)                                                ADVP0019
      TOLLP=TOL*(L+1.)                                                     ADVP0020
C     EVALUATE INTEGRAND OVER INITIAL MESH                                 ADVP0021
      DO 1 I=1,NPNTS                                                       ADVP0022
    1 Y(I)=F(X(I))                                                         ADVP0023
      TIN=0.                                                               ADVP0024
      NEX=0                                                                ADVP0025
      NTOT=NPNTS                                                           ADVP0026
      NM=NTOT-1                                                            ADVP0027
C     COMPUTE TRAPEZOIDAL ESTIMATE OF INTEGRAL                             ADVP0028
      DO 2 I=1,NM                                                          ADVP0029
    2 TIN=TIN+(X(I+1)-X(I))*(Y(I)+Y(I+1))                                  ADVP0030
      DXIM=X(2)-X(1)                                                       ADVP0031
      QIM=(Y(2)-Y(1))/DXIM                                                 ADVP0032
C     COMPUTE SIGNIFICANCE WEIGHTS FOR EACH NODE                           ADVP0033
      DO 3 I=2,NM                                                          ADVP0034
      DXI=X(I+1)-X(I)                                                      ADVP0035
      QI=(Y(I+1)-Y(I))/DXI                                                 ADVP0036
      W(I)=(DXIM**2+DXI**2)*DABS(QI-QIM)                                   ADVP0037
      DXIM=DXI                                                             ADVP0038
      QIM=QI                                                               ADVP0039
    3 CONTINUE                                                             ADVP0040
C     FIND MOST SIGNIFICANT NODE                                           ADVP0041
    4 WMX=W(2)                                                             ADVP0042
      IMX=2                                                                ADVP0043
      DO 7 I=2,NM                                                          ADVP0044
      IF (WMX-W(I)) 5,6,6                                                  ADVP0045
    5 WMX=W(I)                                                             ADVP0046
      IMX=I                                                                ADVP0047
    6 CONTINUE                                                             ADVP0048
    7 CONTINUE                                                             ADVP0049
C     CHECK FOR EXIT                                                       ADVP0050
      IF (NEX-MINEX) 10,8,8                                                ADVP0051
    8 IF (WMX-TOLLP*DABS(TIN)) 21,21,9                                     ADVP0052
    9 IF (NEX-MAXEX) 10,21,21                                              ADVP0053
   10 CONTINUE                                                             ADVP0054
C     ISI=INDEX OF MOST SIGNIFICANT SUBINTERVAL                            ADVP0055
      ISI=IMX                                                              ADVP0056
```

```
      IF (X(IMX-1)+X(IMX+1)-2.*X(IMX)) 11,11,12                          ADVP0057
   11 ISI=IMX-1                                                          ADVP0058
   12 CONTINUE                                                           ADVP0059
C     COMPUTE POINT TO BE INSERTED                                       ADVP0060
      XI=(X(ISI)+X(ISI+1))/2.                                            ADVP0061
      YI=F(XI)                                                           ADVP0062
C     SHIFT ARRAYS TO RIGHT OF ISI                                       ADVP0063
      K=NTOT                                                             ADVP0064
      NSH=NTOT-ISI                                                       ADVP0065
      DO 13 I=1,NSH                                                      ADVP0066
      X(K+1)=X(K)                                                        ADVP0067
      Y(K+1)=Y(K)                                                        ADVP0068
      W(K+1)=W(K)                                                        ADVP0069
      K=K-1                                                              ADVP0070
   13 CONTINUE                                                           ADVP0071
      I=ISI+1                                                            ADVP0072
      X(I)=XI                                                            ADVP0073
      Y(I)=YI                                                            ADVP0074
      NM=NTOT                                                            ADVP0075
      NEX=NEX+1                                                          ADVP0076
      NTOT=NTOT+1                                                        ADVP0077
C     RECALCULATE NEIGHBORING WEIGHTS AND COMPUTE ADDITIONAL             ADVP0078
C     CONTRIBUTION TO TRAPEZOIDAL ESTIMATE OF INTEGRAL                   ADVP0079
      IMID=I                                                             ADVP0080
      I1=I-1                                                             ADVP0081
      I2=I+1                                                             ADVP0082
      IF (I1-1) 14,14,15                                                 ADVP0083
   14 I1=2                                                               ADVP0084
   15 IF (I2-NTOT) 17,16,16                                              ADVP0085
   16 I2=NM                                                              ADVP0086
   17 DXIM=X(I1)-X(I1-1)                                                 ADVP0087
      QIM=(Y(I1)-Y(I1-1))/DXIM                                           ADVP0088
      DO 20 I=I1,I2                                                      ADVP0089
      DXI=X(I+1)-X(I)                                                    ADVP0090
      QI=(Y(I+1)-Y(I))/DXI                                               ADVP0091
      W(I)=(DXIM**2+DXI**2)*DABS(QI-QIM)                                 ADVP0092
      IF (I-IMID) 19,18,19                                               ADVP0093
   18 TL=DXIM*(Y(I-1)+Y(I))                                              ADVP0094
      TR=DXI*(Y(I)+Y(I+1))                                               ADVP0095
      TM=(DXIM+DXI)*(Y(I-1)+Y(I+1))                                      ADVP0096
      TIN=TIN+(TL+TR-TM)                                                 ADVP0097
   19 DXIM=DXI                                                           ADVP0098
      QIM=QI                                                             ADVP0099
   20 CONTINUE                                                           ADVP0100
      GO TO 4                                                            ADVP0101
C     SET M'S AND N'S                                                    ADVP0102
   21 CALL MNSET (NTOT,X,Y,M,N,L)                                        ADVP0103
C     COMPUTE NODAL DERIVATIVES                                          ADVP0104
      CALL VPSD (NTOT,X,Y,D,M,N)                                         ADVP0105
C     COMPUTE INTEGRAL OF VP SPLINE                                      ADVP0106
      YINT=0.                                                            ADVP0107
      DO 22 I=1,NM                                                       ADVP0108
      DXI=X(I+1)-X(I)                                                    ADVP0109
      QI=(Y(I+1)-Y(I))/DXI                                               ADVP0110
      T=Y(I)+Y(I+1)                                                      ADVP0111
```

```
      KI=M(I)+N(I)-M(I)*N(I)                                          ADVP0112
      E1=(KI+1.)*(M(I)*(D(I)-QI)+N(I)*(QI-D(I+1)))                    ADVP0113
      E2=2.*(D(I+1)-D(I))                                             ADVP0114
      E3=DXI/(KI*(M(I)+1.)*(N(I)+1.))                                 ADVP0115
      CI=DXI*(T+E3*(E1+E2))                                           ADVP0116
      YINT=YINT+CI                                                    ADVP0117
   22 CONTINUE                                                        ADVP0118
      YINT=YINT/2.                                                    ADVP0119
      RETURN                                                          ADVP0120
      END                                                             ADVP0121-
```

```
      SUBROUTINE MNSET (NPNTS,X,Y,M,N,L)                              MNSE0001
C     MNSET - MNSET SETS THE M'S AND N'S FOR A VP SPLINE              MNSE0002
C     NPNTS,X,Y=DATA                                                  MNSE0003
C     M,N=VARIABLE POWERS FOR A VP SPLINE                             MNSE0004
C     L=LOWER BOUND ON M'S AND N'S                                    MNSE0005
C     L MUST BE GREATER THAN 2 AND NEED NOT BE GREATER THAN 3         MNSE0006
C     THE LOCAL DERIVATIVE USED IS THE SLOPE OF A LINE THROUGH        MNSE0007
C     THE POINT WHICH MAKES EQUAL ANGLES WITH THE LINEAR INTERPOLATER MNSE0008
C     ON THE LEFT AND RIGHT OF THE POINT                              MNSE0009
      IMPLICIT REAL*8 (A-H,O-Z)                                       MNSE0010
      DIMENSION X(1), Y(1), M(1), N(1)                                MNSE0011
      DOUBLE PRECISION M, N, L                                        MNSE0012
      NM=NPNTS-1                                                      MNSE0013
      N(1)=L                                                          MNSE0014
      M(NM)=L                                                         MNSE0015
      DXIM=X(2)-X(1)                                                  MNSE0016
      QIM=(Y(2)-Y(1))/DXIM                                            MNSE0017
      DO 4 I=2,NM                                                     MNSE0018
      DXI=X(I+1)-X(I)                                                 MNSE0019
      QI=(Y(I+1)-Y(I))/DXI                                            MNSE0020
      R=(DXI/DXIM)*DSQRT((1.+QIM**2)/(1.+QI**2))                      MNSE0021
      IF (R-1.) 1,1,2                                                 MNSE0022
    1 N(I)=L                                                          MNSE0023
      M(I-1)=L/R                                                      MNSE0024
      GO TO 3                                                         MNSE0025
    2 M(I-1)=L                                                        MNSE0026
      N(I)=L*R                                                        MNSE0027
    3 DXIM=DXI                                                        MNSE0028
      QIM=QI                                                          MNSE0029
    4 CONTINUE                                                        MNSE0030
      RETURN                                                          MNSE0031
      END                                                             MNSE0032-
```

# TIME EVOLUTION OF AN ORTHOGONAL MATRIX

James M. Wilkes
Army Materiel Test and Evaluation Directorate
White Sands Missile Range, NM

ABSTRACT. The usual method of computing a rotation matrix as a function of the Euler angles is discussed. On a digital computer these angles must be obtained by a numerical integration of the angle derivatives, which are functions of the angular velocity components of the rotating coordinate system. The numerical integration in effect imposes a rotational motion with constant angular velocity over a time interval of length equal to the integration step-size. This constancy of the angular velocity is exploited to formulate a simple second-order differential equation for the orthogonal matrix describing the rotation. The equation is easily solved exactly, and gives an expression for the matrix at the end of an integration interval as a function of the matrix, and of the angular velocity components, at the beginning of the interval. The second method avoids some of the difficulties of the Euler angle method, and can be usefully applied in digital simulations of rigid-body motion.

1. INTRODUCTION. A mathematical model of the motion of a rigid body requires information regarding the relation between two cartesian coordinate systems, one of which is rotating with respect to the other. This information is contained in the nine elements of the matrix R describing the change of basis from one coordinate system to the other. The physical requirement that the magnitude of a vector be invariant under a change of basis due to a rotation, imposes the following mathematical condition [1] on R:

$$RR^T = I = R^T R \ , \tag{1.1}$$

where I is the identity matrix, and the T-superscript denotes the matrix transpose. This condition is referred to as the orthogonality condition, and R is said to be an orthogonal matrix.

Equation (1.1) represents nine linear equations in the nine elements of R, which would uniquely determine those elements but for the fact that $RR^T = I = R^TR$ is a symmetric matrix. Due to this symmetry, only six of the equations are linearly independent. The three undetermined elements serve to parameterize the (infinite number of) different rotation matrices, and the set of all such matrices constitutes the three parameter group of orthogonal matrices.

A popular choice for the parameters is a set of three angular coordinates $\theta_1$ , $\theta_2$ , and $\theta_3$, known as the Euler angles [2]. With this choice the matrix R can be written as a product of three separate rotations, through each of the three Euler angles. At least two potential difficulties accompany this parameterization. The first is a matter of economy of computation. Once the values of the Euler angles have been determined, one still must compute the matrix elements of R as sums and products of trigonometric functions of the angles. Such computations can become very time-consuming, and therefore expensive, on a digital computer. The second problem is of a mathematical nature. It can be shown that for a given sequence of Euler rotations, the angular velocity components $\omega_i$, i = 1,2,3, in the rotating basis, can be expressed as linear functions of the Euler angle derivatives $\dot{\theta}_i$, i = 1,2,3. That is, at any time t, one has relations of the following form:

$$\omega_i(t) = \Sigma G_{ij}(\theta_2(t),\theta_3(t))\ \dot{\theta}_j(t) \ , \qquad i = 1,2,3, \tag{1.2}$$

where all summations are understood to be from 1 to 3, on repeated indices of the summand. (The coefficient matrix G depends, in general, only upon the last two rotation angles of the rotation sequence.) To determine the angles, one must first solve (1.2) for the derivatives of the angles, and then integrate these derivatives. The solution of (1.2) for the derivatives involves inverting the matrix G. However, for certain values of the Euler angles, the determinant of G vanishes, hence $G^{-1}$ does not exist, and the Euler angle method fails for those values of the angles.

The following observations are important for developing an alternate method of computing a rotation matrix. In a digital model all integrations are performed numerically. Typically, a numerical method requires for the computation of the value of a variable the previously calculated value of the variable and its derivative. For illustrative purposes, consider a numerical integration based on a first-order Taylor's series. Assuming the values $\theta_i(0)$ and $\dot{\theta}_i(0)$, $i = 1,2,3$, to have been computed at the beginning of an integration interval (which we take for convenience to be $t = 0$), this method computes the following values for the Euler angles at the end of an integration interval of step-size $\tau$ :

$$\theta_i(\tau) = \theta_i(0) + \tau\dot{\theta}_i(0) \quad , \quad i = 1,2,3. \tag{1.3}$$

For values of the Euler angles for which the coefficient matrix G in (1.2) is non-singular, we find from (1.2):

$$\dot{\theta}_i(0) = \Sigma G^{-1}_{ij}\,(\theta_2(0),\theta_3(0))\;\omega_j(0) \quad , \quad i = 1,2,3. \tag{1.4}$$

Substituting (1.4) into (1.3) then yields for the new values of the angles

$$\theta_i(\tau) = \theta_i(0) + \tau\Sigma\; G^{-1}_{ij}\,(\theta_2(0),\theta_3(0))\;\omega_j(0) \quad , \quad i = 1,2,3\,. \tag{1.5}$$

In (1.5) the angular velocity dependence of the new values involves only the <u>previous</u> values $\omega_j(0)$. Since the elements of $R(\tau)$ can be constructed as functions of the $\theta_i(\tau)$, the values $\omega_j(0)$ are the best values of the angular velocity components available for computing $R(\tau)$. Hence, for digital computation purposes the angular velocity components can be considered to have the constant values $\omega_j(0)$ on time intervals equal in length to the integration step-size, that is, for all $t\varepsilon[0,\tau]$.

This constancy of the angular velocity on integration intervals allows us to formulate and solve a simple second-order differential equation for R. The solution

allows a direct computation of $R(\tau)$ as a function of the initial matrix $R(0)$, and the angular velocity components $\omega_j(0)$, $j = 1,2,3$. For the case $R(0) = I$ (that is, when the two coordinate systems initially coincide), the result is the well-known expression [3,4] for the matrix describing rotations about an arbitrary fixed axis. Although the method we describe is thus fairly well-known (it was in fact developed for, and is being successfully applied in, a large digital missile simulation [5]), the derivation given in Section 3 is believed to be new and, in our opinion, much more straight-forward than the geometrical arguments given in the usual derivations [3,4].

2. SOME PROPERTIES OF ANTISYMMETRIC MATRICES. By definition, an antisymmetric matrix A is a square matrix satisfying the identity $A^T = -A$. From this identity one can easily deduce the following general form for a 3×3 antisymmetric matrix:

$$A = \begin{bmatrix} 0 & a_3 & -a_2 \\ -a_3 & 0 & a_1 \\ a_2 & -a_1 & 0 \end{bmatrix} . \tag{2.1}$$

Introducing the Levi-Civita permutation symbol $\varepsilon_{ijk}$ ($\varepsilon_{123} = 1$, $\varepsilon_{ijk} = 1$ (-1) for even (odd) permutations of 1,2,3, and $\varepsilon_{ijk} = 0$ if any two indices are the same), the matrix elements of A can be written concisely as

$$A_{ij} = \Sigma\varepsilon_{ijk}\, a_k , \quad i,j = 1,2,3. \tag{2.2}$$

By taking the product of A with itself, we obtain the matrix elements of $A^2$ in the form

$$A^2_{ij} = -a^2\, \delta_{ij} + a_i a_j , \tag{2.3}$$

where $\delta_{ij}$ is the Kronecker delta symbol ($\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$),

and where $a^2 \equiv a_1^2 + a_2^2 + a_3^2$ .

Defining a symmetric matrix $S(\underline{a})$ by

$$S_{ij}(\underline{a}) = a_i a_j \ , \tag{2.4}$$

one can write $A^2$ as :

$$A^2 = -a^2 I + S(\underline{a}) \ . \tag{2.5}$$

It is easy to show, using (2.2) and (2.4), that $AS(\underline{a}) = 0$, hence, multiplying both sides of (2.5) by A gives the very useful identity:

$$A^3 = -a^2 A \ . \tag{2.6}$$

3. THE DIFFERENTIAL EQUATION FOR R. Assuming the elements of R to be differentiable functions of time on the interval $[0,\tau]$, we differentiate both sides of (1.1) to obtain

$$\dot{R}(t)\ R^T(t) + R(t)\ \dot{R}^T(t) = 0, \tag{3.1}$$

where $\dot{R}$ is the matrix containing the derivatives of the elements of R, and we note that $\dot{I} = 0$. Defining a new matrix $\Omega$ by

$$\Omega(t) \equiv \dot{R}(t)\ R^T(t) \ , \tag{3.2}$$

we obtain from (3.2) and (3.1), and the identity $(AB)^T = B^T A^T$ :

$$\Omega(t) = \dot{R}(t)\ R^T(t) = -R(t)\dot{R}^T(t) = -[\dot{R}(t)R^T(t)]^T = -\Omega^T(t) \ ,$$

and it follows that the matrix $\Omega$ is antisymmetric. By (2.1) $\Omega$ can be written in the general form:

$$\Omega(t) = \begin{bmatrix} 0 & \omega_3(t) & -\omega_2(t) \\ -\omega_3(t) & 0 & \omega_1(t) \\ \omega_2(t) & -\omega_1(t) & 0 \end{bmatrix} \tag{3.3}$$

It is demonstrated in several textbooks [6,7] that the elements of $\Omega$ , defined by (3.2), can be identified with the components in the rotating basis of the angular velocity vector. As discussed in the Introduction, the best available values of these components on the interval $[0,\tau]$ are the previously computed values $\omega_j(0)$.
Setting

$$\Omega = \Omega(0) , \quad \omega_j = \omega_j(0), \quad j = 1,2,3, \tag{3.4}$$

and multiplying both sides of (3.2) by R(t), using the orthogonality condition (1.1), we obtain the following first-order differential equation for R:

$$\dot{R}(t) = \Omega R(t) . \tag{3.5}$$

Since $\Omega$ is a constant matrix on $[0,\tau]$, (3.5) can be differentiated to yield:

$$\ddot{R}(t) = \Omega\dot{R}(t) = \Omega^2 R(t) , \tag{3.6}$$

where $\dot{R}(t)$ has been replaced by (3.5) in the last equation of (3.6). Multiplying (3.6) by $\Omega$ now gives

$$\ddot{\Omega R}(t) - \Omega^3 R(t) = \Omega\ddot{R}(t) + \omega^2 \Omega R(t) = 0 \tag{3.7}$$

where we have used (2.6) for $\Omega^3$, and where

$$\omega^2 \equiv \omega_1^2 + \omega_2^2 + \omega_3^2 . \tag{3.8}$$

Since $\omega^2$ is a scalar, it commutes with $\Omega$ , and (3.7) can be written as

$$\Omega[\ \ddot{R}(t) + \omega^2 R(t)] = \Omega C_o = 0 \quad , \tag{3.9}$$

where we have defined

$$\ddot{R}(t) + \omega^2\ R(t) \equiv C_o \quad . \tag{3.10}$$

Equation (3.10) is the familiar equation for a forced harmonic oscillator, except that the "dependent variable" is here a matrix function R, and the "forcing function" is an as yet undetermined matrix $C_o$ . It is easy to show, using (3.5), (3.6), and (2.6), that $\dot{C}_o = 0$, so that $C_o$ is in fact a constant matrix. Furthermore, using (2.5), one can show that $C_o = 0$ implies that $\Omega = 0$, which, from (3.5) corresponds to the trivial solution $R(t) = R(0)$, $t\varepsilon[0,\tau]$. By direct substitution one can then verify that the non-trivial solutions of (3.10) have the general form:

$$R(t) = C_o/\omega^2 + C_1 \sin\omega t + C_2 \cos\omega t \quad , \tag{3.11}$$

where $C_1$ and $C_2$ are arbitrary constant matrices. To determine the constant matrices in (3.11), we evaluate R and its first two derivatives (found by differentiating (3.11)) at $t = 0$, and compare the results with (3.5) and (3.6) evaluated at $t = 0$. The results are

$$C_o/\omega^2 = R(0) + \Omega^2 R(0)/\omega^2 \quad ,$$

$$C_1 = \Omega R(0)/\omega \quad ,$$

$$C_2 = -\Omega^2 R(0)/\omega^2 \quad .$$

Substituting these expressions into (3.11), we obtain the following solution for the rotation matrix at time $t = \tau$ :

$$R(\tau) = [\, I + (\Omega/\omega)\sin\omega\tau + (\Omega^2/\omega^2)\,(1-\cos\omega\tau)\,]\, R(0) \ . \qquad (3.12)$$

It is convenient to define a "transition" matrix $\lambda$ by

$$\lambda(\tau) = I + (\Omega/\omega)\sin\omega\tau + (\Omega^2/\omega^2)\,(1-\cos\omega\tau) \ . \qquad (3.13)$$

If the matrix $R(0)$ is known, then $\lambda(\tau)$ defines the transition over the interval of length $\tau$, to the new matrix

$$R(\tau) = \lambda(\tau)\, R(0) \ . \qquad (3.14)$$

If the two coordinate systems initially coincide, so that $R(0) = I$, then $R(\tau) = \lambda(\tau)$. Using (2.2) and (2.3) in (3.13), we obtain the matrix elements of $\lambda$ in the form

$$\lambda_{ij}(\tau) = \delta_{ij}\cos\omega\tau + \Sigma\varepsilon_{ijk}(\omega_k/\omega)\sin\omega\tau + (\omega_i\omega_j/\omega^2)\,(1-\cos\omega\tau) \ ,$$

which is a slightly simplified form of equation (19) of Ref. 4 for the elements of the matrix describing a rotation through the angle $\omega\tau$, about an axis defined by the direction cosines $\omega_i/\omega$ , $i = 1,2,3$.

4. CONCLUSION. The transition matrix method described in this paper eliminates the inversion singularity problem of the Euler angle method, as well as the numerical integration of the Euler angle derivatives required by that method. Also, the only trigonometric functions to be computed in (3.12) are $\sin\omega\tau$ and $\cos\omega\tau$ , hence computation time should be reduced by the transition matrix method. If so desired, the Euler angles can be recovered at any time from the rotation matrix, for they are simply inverse trigonometric functions of the matrix elements. We remark that (3.12) is approximately valid on any interval for which the angular

velocity is approximately constant, that is, on any interval where the angular acceleration is "small". It would appear that this formalism has significant advantages over the usual Euler angle method.

REFERENCES.


1. J. T. Cushing, Applied Analytical Mathematics for Physical Scientists, (Wiley, New York,1975), pp. 54-55.

2. H. Goldstein, Classical Mechanics, (Addison-Wesley, 1950), pp. 107-109.

3. H. Jeffreys and B. S. Jeffreys, Methods of Mathematical Physics, (Cambridge University Press, 1956), 3rd ed., p. 96.

4. A. Palazzolo, Am. J. Phys., **44**, 63 (1976).

5. SAD-SB-9-1, "H-1 Six Degree of Freedom Interceptor Missile Simulation User's Manual (U)", Raytheon Corp., October 1972, Unclassified.

6. H. C. Corben and Philip Stehle, Classical Mechanics, (Wiley, New York, 1960), 2nd ed., pp. 141-142.

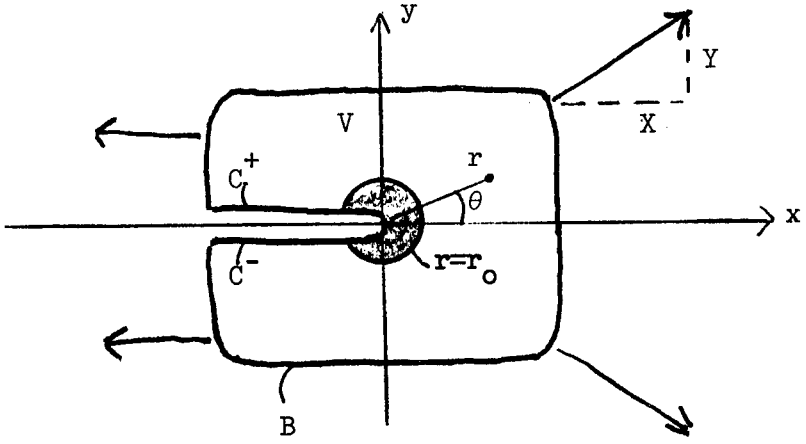
7. C. Broxmeyer, Inertial Navigation Systems, (McGraw-Hill, New York, 1964), pp. 25-26.

# THE WEIGHT FUNCTIONS OF MODE I OF THE PENNY-SHAPED AND OF THE ELLIPTIC CRACK

Hans F. Bueckner
Mathematician, Large Steam Turbine-Generator Department
General Electric Company
Schenectady, N. Y.

ABSTRACT. Fundamental fields and weight functions are presented in closed form by algorithm and formula.

1. INTRODUCTION. STATES OF PLANE STRAIN. During the last three decades the analysis of stress fields near the edges of cracks has grown into a discipline of its own. Various methods for the computation of stress intensity factors have been developed. The use of weight functions is one of them. Originally proposed for states of plane strain [1], the method can be extended to three-dimensional fields [2, 3, 4]. In the sequel we shall do this for the configurations of the penny-shaped and of the elliptic crack. The analysis is within the frame of the classical theory of elasticity. Using a rectangular cartesian coordinate system x,y,z we denote the respective displacements by u,v,w and the stresses by $\sigma_x$, $\tau_{xy}$ etc. in the familiar manner. It is useful to begin with a review of states of plane strain within a cylindrical elastic body V with generators parallel to the z-axis. Figure 1 shows its cross-section in the (x,y)-plane. V has mirror symmetry with respect to the (x,z)-plane. In the same plane a crack with faces $C^+$, $C^-$ extends from the z-axis in the direction of the negative x-axis. The boundary of V consists of the crack faces and of a cylindrical surface B. Let B be attacked by a load of tractions, the latter acting with components X,Y in x- and y-direction respectively and with X,Y the same along a generator. Assuming mirror symmetry of the distribution of tractions with respect to the (x,z)-plane and imposing the constraint w = 0 we obtain a state of deformation in V where u,v do not depend on z (plane strain) and where a suitable disposition of rigid body motion makes u an even and v an odd function of y (mode I). Let $x = r\cos\theta$, $y = r\sin\theta$ define polar coordinates $r,\theta$. With their aid the asymptotic behavior near r = 0 of the relevant field quantities can be described as follows:




plane strain
mode I

Figure 1

$$\left.\begin{aligned} \sigma &= \frac{k}{\sqrt{2r}}\, f(\theta)\cos\tfrac{1}{2}\theta \quad \text{with a suitable constant } k \text{ and where} \\ f(\theta) &= 1 - \sin\tfrac{1}{2}\theta \sin\tfrac{3}{2}\theta \qquad \text{for } \sigma = \sigma_x \\ f(\theta) &= 1 + \sin\tfrac{1}{2}\theta \sin\tfrac{3}{2}\theta \qquad \text{for } \sigma = \sigma_y \\ f(\theta) &= \sin\tfrac{1}{2}\theta \cos\tfrac{3}{2}\theta \qquad \text{for } \sigma = \tau_{xy}\ ; \end{aligned}\right\} \tag{1.1}$$

furthermore

$$\left.\begin{aligned} u &= \frac{k}{2\mu}\sqrt{\frac{1}{2}r}\ (\kappa - \cos\theta)\cos\tfrac{1}{2}\theta \\ v &= \frac{k}{2\mu}\sqrt{\frac{1}{2}r}\ (\kappa - \cos\theta)\sin\tfrac{1}{2}\theta \end{aligned}\right\} \tag{1.2}$$

$\kappa = 3 - 4\nu$, $\nu$ = Poisson's ratio

$\mu$ = shear modulus .

The constant k is known as stress intensity factor. The asymptotic relations (1.1), (1.2) stay valid if a bounded and smooth distribution of tractions on $C^+$, $C^-$ is admitted in accord with the symmetry of mode I. It is customary to consider the term $r^{-1/2}$ in (1.1) as a point singularity in the (x,y)-plane at the "crack tip" r = 0. Nevertheless the singularity is along the whole z-axis as a singular line (the edge of the crack). This should be kept in mind.

Although the stresses are unbounded near r = 0 the energy of deformation per unit length in z-direction is bounded in general. More precisely it is bounded within any cylinder $r = r_o$ of sufficiently small radius $r_o$. If unbounded the cause is not asymptotic behavior in accord with (1.1) but singular behavior of the stress field at points $r \neq 0$ of load application. The latter happens for concentrated loads. If B is smooth and if the tractions are bounded and smoothly distributed then the energy per unit length is bounded. In practical mechanics no other situations are encountered. The singular behavior (1.1) of the stresses notwithstanding, we are justified to denote the field responding to the applied tractions as a <u>regular</u> field.

Let now a field of plane strain and of mode I have the property that

$$u,v = O(r^{-1/2}),\ \sigma = O(r^{-3/2}) \quad \text{near } r = 0 \tag{1.3}$$

We shall call such a field <u>fundamental</u> if it goes without body forces and if it displays no surface tractions. It is not difficult to construct such a field.

Let $t \neq 0$ be an arbitrary constant. We set up

$$\left.\begin{aligned} u &= u_s + u_r, \quad v = v_s + v_r \qquad \text{where} \\ u_s &= tr^{-1/2}\left(\tfrac{1}{2}\cos\tfrac{5}{2}\theta + (\kappa - \tfrac{3}{2})\cos\tfrac{1}{2}\theta\right) \\ v_s &= tr^{-1/2}\left(\tfrac{1}{2}\sin\tfrac{5}{2}\theta - (\kappa + \tfrac{3}{2})\sin\tfrac{1}{2}\theta\right) \end{aligned}\right\} \qquad (1.4)$$

and where $u_r, v_r$ are the displacements $u, v$ of a suitably chosen regular field. It so happens that the displacements $u_s, v_s$ create a stress field without body forces; no tractions are induced on $C^+$, $C^-$ while a system of self-equilibrated tractions shows up on B. We choose $u_r, v_r$ so as to compensate the tractions on B. This establishes $u, v$ by (1.4) as the displacements of a fundamental field. The asymptotic laws (1.3) can be rewritten in the vein of (1.2), (1.1). The details follow from the explicit form of $u_s, v_s$ in (1.4). It has been shown in [1] that the construction (1.4) yields <u>all</u> fundamental fields in V of mode I. The energy of deformation per unit length is infinite. More precisely the energy is already infinite within any cylinder $r = r_o$, no matter how small $r_o > 0$. We can dispose of t by normalizing the fundamental field. If $t(\kappa + 1) = 1$ then

$$v = |x|^{-\frac{1}{2}} \text{ on } C^+, \quad v = -|x|^{-\frac{1}{2}} \text{ on } C^- \qquad (1.5)$$

$$\text{near } x = 0.$$

We shall write $u = u_f$, $v = v_f$ if the fundamental field is normalized by (1.5); setting $u = u_r$, $v = v_r$ we shall characterize a generic regular field, i.e. the meaning of $u_r, v_r$ will not be restricted to (1.4). Let us now consider the mixed energy of deformation (per unit length) W associated with $u_f, v_f$ and $u_r, v_r$. To be on the safe side we exclude the cylindrical domain $r < r_o$ from V. In the remaining portion the mixed energy can be assumed to exist. By Betti's theorem two representations $W = W_{rf}$, $W = W_{fr}$ of the mixed energy are available. Here $W_{rf}$ is the work of the tractions of the regular field through the displacements of the fundamental one; $W_{fr}$ is the work of the tractions of the fundamental field through the displacements of the regular one. In either case the tractions on the cylinder $r = r_o$ must be taken into account. We can write

$$-W'_{rf} + W'_{fr} = W''_{rf} - W''_{fr} \qquad (1.6)$$

where primes refer to the cylinder $r = r_o$ and double primes to the boundary of V outside that cylinder; the latter includes B and part of $C^+$, $C^-$. Since the fundamental field exhibits no tractions on B, $C^+$, $C^-$ we find $W''_{fr} = 0$. For sufficiently small $r_o$ the left-hand side of (1.6) can be evaluated with the aid of the asymptotic relations (1.1), (1.2) for the regular field and (1.3), (1.4) for the

fundamental field. In this context (1.4) must be supplemented by formulas for the stresses to which $u_s, v_s$ give rise. Without going into any further detail we observe that specified stresses $\sigma_r, \sigma_f$ and displacements $w_r, w_f$ of regular and fundamental field respectively obey the order relations

$$r_o \sigma_r \cdot w_f = O(1), \quad r_o \sigma_f \cdot w_r = O(1) \text{ as } r_o \to o \tag{1.7}$$

on the cylinder $r = r_o$. Since $W'_{rf}$ and $W'_{fr}$ are representable as line integrals over the circle $r = r_o$ the asymptotic relations determine the left-hand side of (1.6) in the limit $r_o \to 0$. The final result of this procedure is

$$k = \frac{1}{\sqrt{2}\pi} \int_{\mathcal{L}} (X_r u_f + Y_r v_f)\, ds \tag{1.8}$$

for the stress intensity factor k of the regular field. $X_r, Y_r$ are the components of traction of that very field, and the integration in (1.8) is over the line $\mathcal{L}$ which bounds the cross section of V in the (x,y)-plane, ds being the length element of $\mathcal{L}$. $\mathcal{L}$ is the projection of B as well as of $C^+$, $C^-$ onto that plane. Details of the derivation of (1.8) can be found in [1,2]; a different derivation is in [3]. It is possible to extend (1.8) to regular fields with body forces. In the special case that the tractions appear exclusively on $C^+$, $C^-$ in the form of a pressure distribution formula (1.8) specializes into

$$k = \frac{\sqrt{2}}{\pi} \int_{\mathcal{L}^+} m(s)p(s)ds; \quad \mathcal{L}^+ = \text{projection of } C^+ \tag{1.8'}$$

$$p = \text{applied pressure}, \quad m = v_f \text{ on } C^+ .$$

We call the displacements $u_f, v_f$ weight functions. They permit to represent k as a weighted sum of the tractions $X_r, Y_r$. The use of a formula of type (1.8) for the computation of the stress intensity factor k is advantageous in two respects:

(1) $u_f, v_f$ depend exclusively on the shape of V; thus geometry and loading appear independently in (1.8).

(2) the effort to calculate $u_f, v_f$ is not higher than the effort to calculate the displacements of some regular field.

At a specified point s' the value of m(s') in formula (1.8') can be interpreted as the stress intensity factor of a regular field responding to concentrated pressure

$$p(s) = \frac{\pi}{\sqrt{2}} \delta(s-s')$$

where $\delta(\ldots)$ is Dirac's Delta function. For this reason one could be inclined to classify $m(s)$ as Green's function. Unfortunately the interpretation makes the function $m(s)$ an abstract from infinitely many fields, each characterized by a different point $s$ of load concentration. To compute $m$ that way would sacrifice the advantage (2) which rests on the circumstance that $m(s)$ is a boundary displacement of only one field. The term "weight function" was chosen in order to avoid the misleading suggestions associated with the concept of Green's function.

For some plane strain configurations of mode I in which the crack faces alone are loaded by some pressure distribution $p(s)$ integral equations have been found [2] which link $p(s)$ to the crack opening displacement $v(s) = v_r$ on $C^+$ in the form

$$p(s) = \frac{d}{ds}\int_a^b L(s,t)q(t)dt; \quad q(t) = \frac{2\mu}{\pi(\kappa+1)} \cdot v(t) \tag{1.9}$$

The interval $(a,b)$ is identical with $\mathcal{L}^+$; $L(s,t)$ is a Cauchy type singular integral operator. The integral is taken as Cauchy principal value. An example is

$$p(s) = -\frac{d}{ds}\int_{-1}^{0} \frac{2q(t)dt}{t-s} \tag{1.10}$$

for the Griffith crack ( $-1 \leq x \leq 0$; $y = 0$) in an infinite solid. The homogeneous case $p(s) \equiv 0$ admits the solution $q(t) \equiv 0$ only if one insists that $q(t)$ be bounded. If one drops this condition then $q(t) = cm(t)$ with $c$ as constant coefficient becomes a solution. For the Griffith crack the homogeneous equation admits two solutions associated with the crack tips $x = 0$, $x = -1$, namely

$$m(s) = \left|\frac{1+s}{s}\right|^{1/2}, \quad m(s) = \left|\frac{s}{1+s}\right|^{1/2}. \tag{1.11}$$

2. FIELDS IN THREE DIMENSIONS. Let us now generalize the states of plane strain of mode I into states of three dimensions. We shall assume a plane crack in the $(x,y)$-plane. Figure 2a shows an elliptic crack as example. The faces are denoted by $C^+$, $C^-$ and the contour by $C'$. In Figure 2b an infinite crack occupying the half-plane $x \leq 0$ is represented. We shall assume that the displacement field has mirror symmetry with respect to the $(x,y)$-plane; more precisely $u,v$ are to be the same at points $(x,y,z)$ and $(x,y,-z)$ while $w$ changes sign without change of absolute value. This is the generalization of mode I of plane strain. Finally we confine the attention to those fields which can be derived from a Boussinesq-Papkovich potential $G(x,y,z)$. This potential is harmonic, i.e.

$$\nabla^2 G = G_{xx} + G_{yy} + G_{zz} = 0 . \tag{2.1}$$

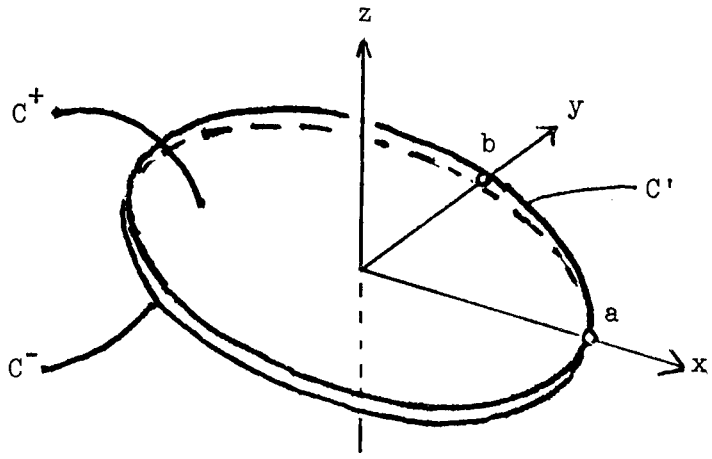



Figure 2a

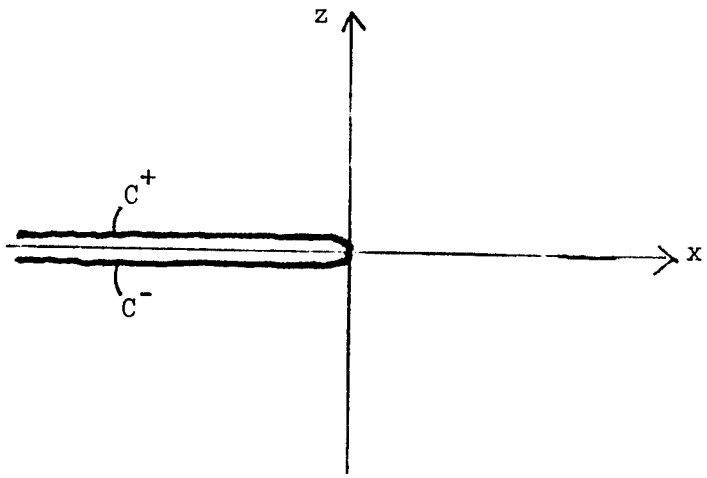



Figure 2b

Here and in what follows coordinate-denoting subscripts indicate partial derivatives. The displacements and stresses are derived as follows:

$$\left.\begin{aligned} u = -zG_{xz} - (1-2\nu)G_x, \quad v = -zG_{yz} - (1-2\nu)G_y, \\ w = -zG_{zz} + 2(1-\nu)G_z \end{aligned}\right\} \tag{2.2}$$

$$\left.\begin{aligned} \sigma_x &= -2\mu\left[(zG_{xx})_z + 2\nu G_{yy}\right], \quad \sigma_y = -2\mu\left[(zG_{yy})_z + 2\nu G_{xx}\right], \\ \sigma_z &= -2\mu\left[zG_{zzz} - G_{zz}\right], \\ \tau_{yz} &= -2\mu z G_{yzz}, \quad \tau_{zx} = -2\mu z G_{xzz}, \\ \tau_{xy} &= -2\mu\left[zG_{xyz} + (1-2\nu)G_{xy}\right]. \end{aligned}\right\} \tag{2.3}$$

For the sake of a first orientation let us consider the configuration of Figure 2b. It admits in particular states of plane strain, and the asymptotic relations of the preceding section apply if the roles of y and z are exchanged. In order to exhibit a more general class of states we set up

$$G(x,y,z) = F(x,z)\cos\lambda y \tag{2.4}$$

with some real constant $\lambda \geq 0$. The case $\lambda = 0$ is that of plane strain. The function $F(x,z)$ must satisfy

$$F_{xx} + F_{zz} - \lambda^2 F = 0 \, . \tag{2.5}$$

Defining polar coordinates $\rho$, $\phi$ by means of

$$\rho e^{i\phi} = x + iz \tag{2.6}$$

we can rewrite (2.5) in the form

$$\rho^2 F_{\rho\rho} + \rho F_\rho - F_{\phi\phi} - \lambda^2\rho^2 F = 0 \, .$$

It admits the product solution

$$\left.\begin{aligned} &F = F^*(\lambda\rho)\cos\tfrac{3}{2}\phi \quad \text{with} \\ &F^*(t) = 3\sqrt{\tfrac{1}{2}\pi}\; I_{3/2}(t) = 3\sqrt{t}\cdot\frac{d}{dt}\frac{\sinh t}{t} \end{aligned}\right\} \, . \tag{2.7}$$

$I_{3/2}$ is the modified Bessel function of type I and of fractional order 3/2 . Altogether we can write

$$\left.\begin{aligned} &G(x,y,z) = g(x,z)h(\lambda\rho)\cos\lambda y \quad \text{with} \\ &\quad g(x,z) = (\lambda\rho)^{3/2}\cos\tfrac{3}{2}\phi = \mathrm{Re}\left[\lambda(x+iz)\right]^{3/2}, \\ &\quad h(t) \;\; = \frac{3}{t}\frac{d}{dt}\frac{\sinh t}{t} \, . \end{aligned}\right\} \tag{2.8}$$

h(t) admits an expansion in even powers of t; moreover h(0) = 1. The function g(x,z) itself is a Boussinesq-Papkovich potential. It describes a state of plane strain and of mode I. The field of displacements and stresses is regular. The y-axis represents the edge of the crack. As we let $\rho \to 0$ we approach the edge. Asymptotic relations of type (1.1) and (1.2) after exchange of the roles of y,z become valid with an intensity factor k depending on y. This is due to the preponderance of g(x,z) in the representation (2.8) of G(x,y,z). Locally the behavior of the field near an edge point y is given by the field of plane strain of g(x,z) but modified by the factor $\cos\lambda y$. We list in particular:

$$k = k(y) = k(0)\cos\lambda y \tag{2.9}$$

$$\left.\begin{aligned} u &= \frac{k}{2\mu}\sqrt{\frac{1}{2}\rho}\,(\kappa - \cos\phi)\cos\tfrac{1}{2}\phi\,, \quad w = \frac{k}{2\mu}\sqrt{\tfrac{1}{2}\rho}\cdot(\kappa - \cos\phi)\sin\tfrac{1}{2}\phi \\ v &= O(\rho^{3/2}) \end{aligned}\right\} \tag{2.10}$$

$$\left.\begin{aligned} \sigma &= \frac{k}{\sqrt{2\rho}}\,f(\phi)\cos\tfrac{1}{2}\phi \qquad \text{where} \\ f(\phi) &= 1 - \tfrac{1}{2}\sin\tfrac{1}{2}\phi\sin\tfrac{3}{2}\phi \qquad \text{for } \sigma = \sigma_x \\ f(\phi) &= 1 + \tfrac{1}{2}\sin\tfrac{1}{2}\phi\sin\tfrac{3}{2}\phi \qquad \text{for } \sigma = \sigma_z \\ f(\phi) &= 2\nu \qquad\qquad\qquad\quad \text{for } \sigma = \sigma_y \\ f(\phi) &= \sin\tfrac{1}{2}\phi\cos\tfrac{3}{2}\phi \qquad\quad \text{for } \sigma = \tau_{zx} \end{aligned}\right\} \tag{2.11}$$

Furthermore the stresses $\tau_{xy}$, $\tau_{yz}$ stay bounded. The special potential (2.8) induces no tractions on the faces of the crack. This is obvious inasmuch as $\tau_{zx}$, $\tau_{yz}$ are concerned. As for $\sigma_z$ we derive from (2.3) that $\sigma_z = 2\mu G_{zz} = -2\mu(G_{xx} + G_{yy})$ on $C^+$, $C^-$. But G = 0 on the faces and $\sigma_z = 0$ follows. The displacements are unbounded as $\rho \to \infty$. For this reason the use of (2.8) must be confined to domains of bounded $\rho$.

Still with regard to Figure 2b let us consider

$$G(x,y,z) = \mathrm{Erfc}(q)e^x\cos y\,; \quad q = \sqrt{2\rho}\cdot\cos\tfrac{1}{2}\phi\,. \tag{2.12}$$

The function G is harmonic. Writing for simplicity Erfc(q) = Q(q) and observing that q as well as the product $e^x\cos y$ are harmonic functions we find

$$\nabla^2 G = e^x \cos y\,(\nabla^2 Q + 2Q_x)\ , \tag{2.13}$$

$$\nabla^2 Q = Q''(q)(q_x^2 + q_z^2) = -\,2qQ'(q)(q_x^2 + q_y^2) = -\,qQ'(q)/\rho \tag{2.14}$$

$$Q_x = Q'(q)q_x = qQ'/2\rho \tag{2.15}$$

and altogether $\nabla^2 G = 0$ as asserted. The potential (2.12) is periodic in y with the period $2\pi$. Inspite of the factor $e^x$ the potential as a whole and all of its partial derivatives go to zero as $\rho \to \infty$. For small $\rho$ we may use

$$G = e^x \cos y\,(1 - \frac{2}{\sqrt{\pi}}\,q + O(q^3)) \tag{2.16}$$

in order to determine the asymptotic behavior of displacements and stresses as we approach the edge of the crack. The function q can be taken as Boussinesq-Papkovich potential; as such it leads to a state of plane strain. The state has displacements

$$u = u_s = \frac{-\sqrt{2}}{4\sqrt{\rho}}\left[\frac{1}{2}\cos\frac{5}{2}\phi + (\kappa - \frac{3}{2})\cos\frac{1}{2}\phi\right]$$

$$w = w_s = \frac{-\sqrt{2}}{4\sqrt{\rho}}\left[\frac{1}{2}\sin\frac{5}{2}\phi - (\kappa + \frac{3}{2})\sin\frac{1}{2}\phi\right]. \tag{2.17}$$

A comparison with (1.4) shows that $u_s, w_s$ have the asymptotic properties of the displacements of a fundamental field of plane strain and of mode I. Going back to (2.16) we can expect the potential q to dominate the behavior of G in the approach $\rho \to 0$. More precisely we find

$$u = a(y)u_s, \quad w = a(y)w_s \quad \text{with } a(y) = -\frac{2}{\sqrt{\pi}}\cos y \tag{2.18}$$

as asymptotic representations of u,w in the case of G.

The field of G has vanishing shearing stresses $\tau_{zx}$, $\tau_{yz}$ on the (x,y)-plane. We assert that $\sigma_z$ vanishes on the faces of the crack. As before we find

$$\sigma_z = -\,2\mu\Delta G \quad \text{with } \Delta G = G_{xx} + G_{yy} \text{ on } C^+,\ C^- \tag{2.19}$$

But

$$G = e^x \cos y \quad \text{on } C^+,\ C^-$$

and $\Delta G = 0$ follows. The displacements and stresses go to zero as $\rho \to \infty$.

The potential $G(x,y,z)$ of (2.12) gives rise to other potentials $G(\lambda x, \lambda(y-y'), \lambda z)$ where $\lambda$, $y'$ are real constants and also $\lambda \geq 0$. These potentials can be linearly combined in a finite number of terms, the combination coefficients to be real. All combinations form a real linear space of infinite dimension. Each potential of this space yields a field of displacements and stresses which we now designate as fundamental field. This is a generalization of fundamental fields of plane strain and justified by the asymptotic relations of type (2.18) as well as by the absence of tractions on the crack faces.

We turn next to Figure 2a and disregard temporarily that the crack is to be elliptic. More generally we admit as crack contour C' any rectifiable Jordan curve of continuous tangent. The Boussinesq-Papkovich potentials associated with this crack configuration can be represented as harmonic potentials of single layers, more precisely in the form

$$\left.\begin{aligned} G(x,y,z) &= -\frac{1}{4\pi(1-\nu)}\iint f(\xi,\eta)R^{-1}d\xi d\eta \qquad \text{with} \\ R^2 &= (x-\xi)^2 + (y-\eta)^2 + z^2 \ . \end{aligned}\right\} \tag{2.20}$$

The integration is over one of the crack faces. Of the density function $f(\xi,\eta)$ we assume continuity inside C' and furthermore for interior points $(\xi,\eta)$

$$|f(\xi,\eta)| \leq f_o d^{-1/2} \tag{2.21}$$

where $f_o$ is some constant and where d is the distance from the contour C' of $(\xi,\eta)$. Formulas (2.3) lead to

$$\left.\begin{aligned} w = 2(1-\nu)G_z &= f \quad \text{on } C^+ \\ &= -f \quad \text{on } C^- \end{aligned}\right\} \ , \tag{2.22}$$

$$\left.\begin{aligned} &\sigma_z = -\,g(x,y) \text{ on } C^+,\ C^- \quad \text{with} \\ &g(x,y) = -\,\Delta\frac{\mu}{2\pi(1-\nu)}\iint f(\xi,\eta)\cdot\left[(x-\xi)^2 + (y-\eta)^2\right]^{-1/2}\cdot d\xi\ d\eta . \end{aligned}\right\} \tag{2.23}$$

As in (2.19), $\Delta$ stands for the Laplacian operator of the (x,y)-plane. The function $g(x,y)$ represents a pressure distribution within the crack. The stresses $\tau_{zx}$, $\tau_{yz}$ vanish in the (x,y)-plane and in particular on the crack. In the nontrivial case $f(\xi,\eta) \not\equiv 0$ we call G and the associated field fundamental if there are no tractions on $C^+$, $C^-$, i.e. if $g(x,y) \equiv 0$. We call G and the associated field regular if $f(\xi,\eta)$ satisfies a condition more stringent than (2.21), namely

$$|f(\xi,\eta)| \leq f_1 d^{1/2} \tag{2.24}$$

where $f_1$ is a suitable constant. Let s denote the arclength on C', counted from some point of C' in the counterclockwise sense as one looks down at the (x,y)-plane. In the neighborhood of any point s of C' we expect the field of G to show the asymptotic behavior of a field of plane strain for an associated half-plane crack; the latter must have the tangent at s as edge and must follow the inner normal of C' at s. In the case of a regular field the asymptotic behavior will be determined by a stress intensity factor k = k(s). In this context we list in particular the asymptotic relations

$$w = k(s)\,\frac{\kappa+1}{2\mu}\left(\frac{1}{2}d\right)^{1/2} \quad \text{on } C^+ \tag{2.25}$$

$$\sigma_z = k(s)(2d)^{-1/2} \qquad \text{for } z = 0 \text{ and points outside the crack .} \tag{2.26}$$

As for the fundamental field we merely write the analogue of (2.25) in the form

$$w = \beta(s)d^{-1/2} \qquad \text{on } C^+ \tag{2.27}$$

where the intensity function $\beta(s)$ depends on the fundamental field.

In the case of plane strain Betti's theorem of reciprocity was applied to the mixed energy formed by a regular and by a fundamental field. The procedure led to formulas (1.8), (1.8') for k. The same method can be used for the configuration of Figure 2a [2]. This yields the analogue of (1.8') in the form

$$\int_{C'} k(s)\beta(s)ds = \frac{\sqrt{2}}{\pi}\iint_{C^+} M(x,y)g(x,y)dxdy \ . \tag{2.28}$$

$M(x,y)$ is the normal displacement w of the fundamental field on $C^+$. The factor $g(x,y)$ is the pressure within the crack of the regular field. It is obvious that <u>one</u> fundamental field does not permit to determine the function k(s). We need infinitely many or - for practical purposes - a sufficiently large number of linearly independent fundamental fields. In order to find such fields we must

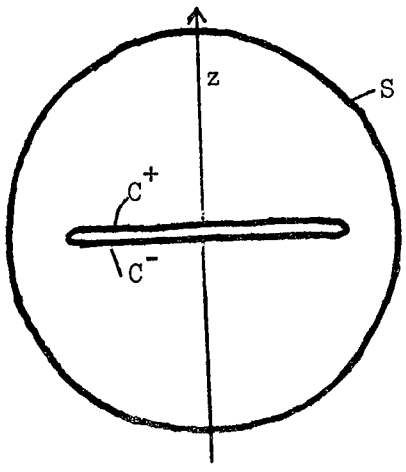



Figure 3

solve the homogeneous integro-differential equation

$$0 = \Delta \iint_{C^+} f(\xi,\eta)\left[(x-\xi)^2 + (y-\eta)^2\right]^{-1/2} d\xi d\eta \tag{2.29}$$

for nontrivial density functions $f(\xi,\eta)$.

Assuming that we possess a fundamental field for the crack configuration of Figure 2a we can construct a fundamental field for a finite elastic body with the same crack. Figure 3 shows a sphere S around the origin. The elastic body is bounded by S and by the faces $C^+$, $C^-$. In analogy to the construction (1.4) we add to our fundamental field for Figure 2a a regular field for Figure 3, such that the modifying regular field has no tractions on the crack while its tractions on S annihilate those of the fundamental field of Figure 2a. The resulting field has no tractions on crack faces and on S; it displays the asymptotic behavior of the initial fundamental field near the edge C' of the crack.

Assume now a fundamental and a regular field for Figure 3, both of mode I. Let the regular field be generated by a distribution of tractions on S. Under these circumstances the analogue of (1.8) is

$$\int_{C'} k(s)\beta(s)ds = \frac{1}{\sqrt{2\pi}} \iint_S (u_f X_r + v_f Y_r + w_f Z_r) dS \tag{2.30}$$

where $u_f, v_f, w_f$ are the displacements of the fundamental field and $X_r, Y_r, Z_r$ the components of traction of the regular one. The fundamental potential G in (2.12) can be used in order to construct an analogue of formula (2.28). Since G has period $2\pi$ one should apply the associated fundamental field to the analysis of $k(y)$ of a regular field with the same period and the same symmetry with respect to y. Moreover it will suffice to consider a slab $0 \leq y \leq \pi$. Further details can be left to the reader.

3. PENNY-SHAPED AND ELLIPTIC CRACK. We return to Figure 2a and interpret C' as an ellipse with half-axes a,b. The ellipse has the equation

$$E(x,y) = 1 - x^2/a^2 - y^2/b^2 = 0 . \tag{3.1}$$

The $\omega$-zeros of the function

$$T(\omega;x,y,z) = 1 - \frac{x^2}{a^2+\omega} - \frac{y^2}{b^2+\omega} - \frac{z^2}{\omega}$$

define associated elliptic coordinates. The largest $\omega$-root of $T = 0$ will play an important role.

At this juncture we turn to the penny-shaped crack by letting b = a. Without essential loss of generality we assume a = 1. Cylindrical coordinates r, $\theta$, z will be useful. We have here x = rcos $\theta$ and y = rsin $\theta$. The function T takes the special form

$$T = 1 - \frac{r^2}{1+\omega} - \frac{z^2}{\omega} \quad . \tag{3.2}$$

The mapping (see also [7])

$$r + iz = \cosh (s + it) \; ; \quad s \geq 0, \; -\tfrac{1}{2}\pi \leq t \leq \tfrac{1}{2}\pi \tag{3.3}$$

permits to represent pairs (r,z) by pairs (s,t) in accord with Figure 4.

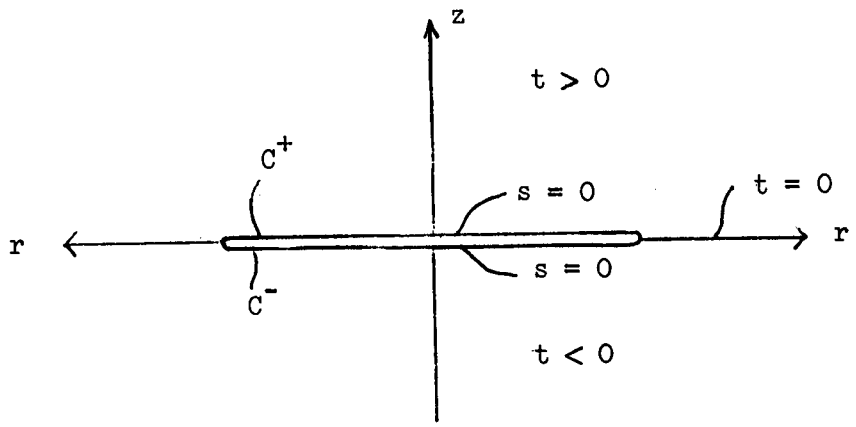



Figure 4

The representation is unique whenever $z \neq 0$ or $r \geq 1$. For points of the crack two different representations appear which permit to distinguish between $C^+$ and $C^-$. From (3.3) it follows that (3.2) has the roots

$$\omega_1 = \sinh^2 s \, , \quad \omega_2 = - \sin^2 t \tag{3.4}$$

The following relations are useful:

$$r = \cosh s \; \cos t \, , \quad z = \sinh s \; \sin t \tag{3.5}$$

$$s_r = t_z = \sinh s \; \cos t \, / N, \quad t_r = - s_z = -\cosh s \; \sin t \, / N \tag{3.6}$$

$$N = \sinh^2 s + \sin^2 t$$

$$\nabla^2 F(s) = \big(F''(s) + \tanh s \, F'(s)\big)/N \quad . \tag{3.7}$$

$$s_z = 1/\sin t = (1 - r^2)^{-1/2} \quad \text{on } C^+$$
$$= -(1 - r^2)^{-1/2} \quad \text{on } C^- \tag{3.8}$$

Furthermore

$$\sinh^2 s \leq r^2 + z^2 \leq \cosh^2 s = 1 + \sinh^2 s \tag{3.9}$$

In what follows we establish an infinite family of fundamental potentials without solving (2.29) directly. We set

$$G_n(x,y,z) = F_n(r,z) \cos n(\theta - \theta') \quad \text{and}$$
$$F_n(r,z) = r^n H_n(s) \quad \text{for } n = 0,1,2,\ldots \tag{3.10}$$

Here $\theta'$ is a constant which may depend on n. We shall try to make $G_n$ a fundamental potential through a suitable choice of $H_n(s)$. Writing altogether

$$G_n = r^n \cos n(\theta - \theta') \cdot H_n(s)$$

we observe that the factor preceding $H_n$ is a harmonic function, i.e.

$$\nabla^2(r^n \cos n(\theta - \theta')) = \Delta\, r^n \cos(\theta - \theta') = 0 \tag{3.11}$$

This in turn together with (3.6), (3.7) yields after steps of an elementary nature

$$\nabla^2 G_n = r^n \cos n(\theta - \theta')\left[H_n''(s) + (2n+1)\tanh s \cdot H_n'(s)\right]/N \tag{3.12}$$

Consequently we must solve

$$H_n''(s) + (2n+1)\tanh s \cdot H_n'(s) = 0 \tag{3.13}$$

We find

$$H_n'(s) = \frac{c}{\cosh^{2n+1} s} \tag{3.14}$$

with some constant c. Integration of $H_n'(s)$ and a special choice of c yield

$$H_n(s) = \alpha_n\left[\frac{1}{2}\pi - \arctan(\sinh s) - \alpha_o \sinh s \sum_{k=1}^{n} \frac{1}{2k\alpha_k \cosh^{2k} s}\right] \tag{3.15}$$

with $\alpha_o = \dfrac{-1}{(1-\nu)\sqrt{2}}$, $\alpha_k = \alpha_o(-1)^k \binom{-1/2}{k}$.

We leave the verification of (3.15) to the reader. Note that the definition of the coefficients $\alpha_k$ is independent of n!

Having established the potential $G_n$ we check on some of its properties. Due to (2.2)

$$w = 2(1-\nu)G_{nz} \quad \text{on } C^+ \tag{3.16}$$

Now (3.14) and (3.8) yield

$$\begin{aligned} G_{nz} &= r^n \cos n(\theta-\theta') H_n'(0) s_z \\ &= c r^n \cos n(\theta-\theta') \cdot (1-r^2)^{-1/2} \quad \text{on } C^+ \end{aligned} \tag{3.17}$$

In the construction of $H_n$ we have chosen the constant c of (3.14) such that

$$\sqrt{2}(1-\nu)c = 1 \tag{3.14'}$$

This choice implies

$$\beta = \cos n(\theta-\theta')$$

for the intensity $\beta(s)$ associated with $G_n$. We still have to verify that $G_n$ does not induce tractions on the crack. The nature of $G_n$ makes it obvious that $\tau_{zx}$, $\tau_{yz}$ vanish on the (x,y)-plane. As for $\sigma_z$ we observe that

$$G_n = r^n \cos n(\theta-\theta') H_n(0) \quad \text{on } C^+, C^- \tag{3.18}$$

Due to (2.19) and (3.11) $\sigma_z$ vanishes. Finally it can be established that the field of $G_n$ has vanishing displacements and stresses at $R = \infty$ where $R = (r^2+z^2)^{1/2}$; moreover the stresses have the order of $R^{-3}$ for large R while the displacements are in the order of $R^{-2}$. In this context we refer to (3.9) with the consequence $R \sim \frac{1}{2} \cdot e^s$ and to (3.14), (3.15) with the consequence

$$H_n(s) = O(e^{-(2n+1)s}) \tag{3.19}$$

for large R. All of this is compatible with the asymptotic behavior for large R of the field of the potential of a single layer. $G_n$ admits a representation (2.20) with the density function

$$f = f_n = \sqrt{2}\, r^n \cos n(\theta-\theta') \cdot (1-r^2)^{-1/2} \; . \tag{3.20}$$

Formula (2.28) takes the special form

$$\int_0^{2\pi} k(\theta)\cos n(\theta-\theta')d\theta = \frac{2}{\pi}\int_0^{2\pi}\int_0^1 r^n(1-r^2)^{-1/2}\cos n(\theta-\theta')g(r,\theta)rdrd\theta \qquad (3.21)$$

for the stress intensity factor $k = k(\theta)$ of the regular field responding to the pressure distribution $g = g(r,\theta)$ within the crack. Since C' is the unit circle we are justified to set $s = \theta$ on C'. It is obvious that the formulas (3.21) for the various n permit to calculate the Fourier components of $k(\theta)$ and thus $k(\theta)$. One can also establish the following formula (see Figure 5)

$$\left.\begin{aligned} k(\theta') &= \frac{\sqrt{2}}{\pi}\iint M(r,\theta,\theta')g(r,\theta)rdrd\theta \quad \text{where} \\ M &= (1-r^2)^{1/2}/d^2 \;;\; d^2 = 1+r^2-2r\cos(\theta-\theta') \end{aligned}\right\} \qquad (3.22)$$

In this case the intensity $\beta(s)$ is a Dirac delta function on C'. An extension of the concept of weight functions in the nature of the case (3.22) has been suggested by Rice [3] in general form. Formula (3.22) appears to be the first concrete example for this idea. Formulas for the penny-shaped crack of type (3.21) are well-known [5,6]. They can be and were indeed derived by direct analysis of the regular field with the aid of Fourier-Hankel transforms. But, as we have just shown, the concepts of fundamental fields and weight functions permit to establish such formulas in a simpler and yet more systematic way.

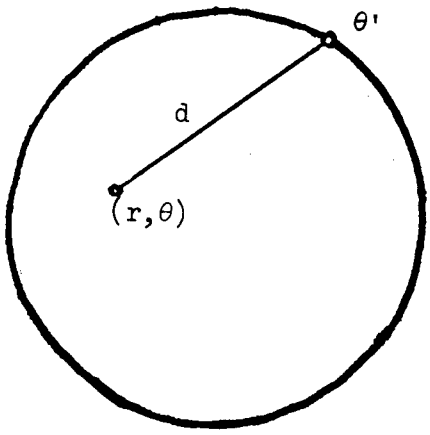



Figure 5

We conclude the discussion of the penny-shaped crack with some formulas for the functions $F_n(r,z)$ in the composition (3.10). We list without proof

$$F_{nr} + \frac{n}{r} F_n = nF_{n-1} + DF_{n-1}; \quad (D-n)F_n = F_{n-1,r} - \frac{n-1}{r} F_{n-1}$$

$$\nabla^2 F_n - \left(\frac{n}{r}\right)^2 F_n = 0; \quad \left[D(D+1)-n(n+1)\right]F_n + F_{n,zz} = 0 \tag{3.23}$$

$$D = r\frac{\partial}{\partial r} + z\frac{\partial}{\partial z} \quad .$$

The operator D preserves harmonicity. The third relation expresses the harmonicity of $G_n$ without reference to the $\theta$ - term. It is possible to establish the formulas of the first line by merely using $\sigma_z = 0$ and the asymptotic behavior of w on the crack in the case of any $G_n$.

Returning to the general elliptic crack we make an extensive use of available literature [7 - 11]. In particular Dyson's formulas [8,9] will be applied. Following Dyson we write the density function f in (2.20) in the form

$$f(x,y) = -4\pi(1-\nu)h(x,y)E^{\lambda-1/2}(x,y) \tag{3.24}$$

where E is the function in (3.1). We are primarily interested in the cases $\lambda = 0$ and $\lambda = 1$; h(x,y) is to be a polynomial in x and y. Under these circumstances the case $\lambda = 1$ will yield a regular potential and the case $\lambda = 0$ a fundamental one for properly chosen h(x,y). Dyson himself admits more general h(x,y). The case $\lambda = 0$ is pertinent to the analysis of an electrically charged disk; so far it has not been applied to elastic analysis. We introduce the following denotations and symbols:

$$Q(s) = s(a^2+s)(b^2+s); \quad \bar{q}(s) = Q^{1/2}(s) \geq 0 \tag{3.25}$$

$$\text{for } s \geq 0$$

$$p = \frac{a^2}{a^2+s}, \quad q = \frac{b^2}{b^2+s}; \quad D = \frac{1}{p}\frac{\partial^2}{\partial x^2} + \frac{1}{q}\frac{\partial^2}{\partial y^2} \tag{3.26}$$

We denote the largest $\omega$-root of $T = 0$ by t; it is nonnegative. These symbols and denotations are unrelated to formerly defined quantities. Dyson and Hobson have shown that the potential G of (2.20) can be rewritten as a single integral,

$$G(x,y,z) = \frac{\pi ab\Gamma(\lambda+\frac{1}{2})}{\Gamma(\frac{1}{2})\lambda\Gamma(\lambda)} \int_t^\infty \frac{T^\lambda}{\bar{q}(s)} M_\lambda h(px,qy) \cdot ds \quad , (\omega \cong s) \tag{3.27}$$

where $M_\lambda$ is the following differential operator:

$$M_\lambda = \sum_{n=0}^{\infty} \frac{s^n T^n D^n}{4^n n!(\lambda+1)(\lambda+2)\ldots(\lambda+n)} . \qquad (3.27')$$

The symbol $\Gamma$ denotes the Gamma function. We replace $\lambda\Gamma(\lambda)$ by unity for $\lambda = 0$. Since $h(x,y)$ is a polynomial only a finite number of terms in (3.27) have to be used. $M_\lambda h(px,qy)$ is therefore a polynomial in x,y whose coefficients are functions of s. Let us now consider G on the crack; in this case t = 0 and thus

$$G(x,y,0) = \frac{\pi ab\Gamma(\lambda+\frac{1}{2})}{\Gamma(\frac{1}{2})\lambda\Gamma(\lambda)} \int_0^{\infty} \frac{T^\lambda}{q(s)} M_\lambda h(px,qy)ds \quad \text{valid for } C^+, C^- . \qquad (3.28)$$

In the cases $\lambda = 0$, $\lambda = 1$ the function $T^\lambda$ is a polynomial in x,y. Altogether we see now that $G(x,y,0)$ is a polynomial in x,y on the crack, and so is $-\sigma_z = \Delta 2\mu G(x,y,0) = g(x,y)$. We can write

$$g = \mathcal{L}_\lambda h \qquad (3.29)$$

where $\mathcal{L}_\lambda$ denotes a linear operator which transform the polynomial h into a polynomial g. The nature of the mapping depends on $\lambda$.

Case $\lambda = 1$

This is the case of ordinary elasticity. $\mathcal{L}_1$ maps the real linear space of all polynomials of degree $\leq m$ (real coefficients) into itself. $\mathcal{L}_1 h = 0$ for some $h \neq 0$ cannot happen. The mapping is therefore 1-1; given g there is a unique h. The mapping does not necessarily transform homogeneous polynomials into homogeneous ones.

Case $\lambda = 0$

If h has degree m then $g = \mathcal{L}_0 h$ has degree not higher than m-2. The case $\mathcal{L}_0 h = 0$ for $h \neq 0$ can happen. We call such an h a fundamental polynomial. It leads to a fundamental field G. Trivial cases are: $h = 1$, $h = x$, $h = y$, $h = xy$.

Here the reader is reminded of the definition of the degree of a polynomial $h(x,y)$. If $h(x,y)$ is a monome, i.e. $h = cx^m y^n$ with $c \neq 0$ then the degree of h is m+n. If h is a combination of monomes we look for the monome of highest degree; that degree is taken as degree of h. A polynomial is homogeneous if all of its monomes have the same degree. The values of h on E = 0 are given by the Fourier polynomial $h(a\cos\theta, b\sin\theta)$; the latter has degree $\leq N$ if h has degree N. Note that $E(a\cos\theta, b\sin\theta) \equiv 0$!

For each degree $m \geq 2$ two fundamental polynomials $h(x,y)$ of degree m can be constructed as follows: Set

$$h(x,y) = x^m + h_1(x,y)E(x,y) \quad \text{where } \mathcal{L}_1 h_1 = -\mathcal{L}_0 x^m \tag{3.30}$$

Since $\mathcal{L}_0 x^m$ has degree $\leq m-2$ the polynomial $h_1$ is of degree $\leq m-2$; consequently $h_1 E$ has degree $\leq m$, and the degree of h cannot exceed m. But $h = x^m = a^m \cos^m\theta$ on $E = 0$; this is a Fourier polynomial of degree m. The degree of $h(x,y)$ cannot be less. Thus h is seen to have exact degree m. Now $h_1 E$ and $\lambda = 0$, $h_1$ and $\lambda = 1$ define the same potential. Hence

$$\mathcal{L}_0(h_1 E) = \mathcal{L}_1 h_1 \quad \text{and } \mathcal{L}_0 h \equiv 0 \ . \tag{3.31}$$

This establishes h as a fundamental polynomial of degree m. In the same vein we construct

$$h(x,y) = x^{m-1}y + h_2(x,y)E(x,y) \quad \text{where } \mathcal{L}_1 h_2 = -\mathcal{L}_0 x^{m-1}y \ . \tag{3.32}$$

With (3.30), (3.32) we have obtained two linearly independent fundamental polynomials of degree m.

The application of the operators $\mathcal{L}_0$, $\mathcal{L}_1$ involves certain elliptic integrals. The following coefficients are needed for the construction of fundamental polynomials:

$$c_{mn}^{(\ell)} = \int_0^\infty s^{-1/2+\ell} p^{m+1/2} q^{n+1/2} ds; \quad c_{mn}^{(0)} = c_{mn} \ ; \tag{3.33}$$

$m,n,\ell$ run through the nonnegative integers. The coefficients satisfy the recursions

$$(2m+1)c_{m+1,n} - (2m+2n+1)c_{mn} + (2n+1)c_{m,n+1} = 0 \tag{3.34}$$

$$c_{mn} = \tau c_{m-1,n} + (1-\tau)c_{m,n-1} \quad \text{with } \tau = a^2/(a^2-b^2)$$

Up to degree 3 fundamental polynomials can be homogeneous. This is no longer so from degree four on. We give some polynomials below:

Fundamental polynomials up to degree 4

$m = 0$: $h = 1$

$m = 1$: $h = x$, $h = y$

$m = 2$: $h = c_{01}x^2 - c_{10}y^2$, $h = xy$

$m = 3$: $h = c_{11}x^3 - 3c_{20}xy^2$, $h = c_{11}y^3 - 3c_{02}x^2y$

$m = 4$: $h = c_{12}x^3y - c_{21}xy^3$, $h = \alpha x^4 + \beta x^2y^2 + \gamma y^4 + \delta x^2 + \epsilon y^2$

with the following coefficients:

$$\alpha = \begin{vmatrix} c_{11} & 3c_{02} \\ c_{21}-c_{12} & -5c_{02}+2c_{12} \end{vmatrix}, \quad \beta = -3\begin{vmatrix} c_{20} & c_{02} \\ 5c_{20}-2c_{21} & -5c_{02}+2c_{12} \end{vmatrix}$$

$$\gamma = \begin{vmatrix} 3c_{20} & c_{11} \\ 5c_{20}-2c_{21} & c_{21}-c_{12} \end{vmatrix}$$

$$\delta = \frac{1}{2}(\tau - 1)\left\{(\alpha - \beta)a^2 + \gamma b^2\right\}; \quad \epsilon = -\frac{1}{2}\tau\left\{\alpha a^2 - (\beta - \gamma)b^2\right\}.$$

For the case $h \equiv 1$, more precisely for the weight function $M = E^{-1/2}$ on the crack the intensity function $\beta$ is

$$\beta(s) = 2^{-1/2}\left(\frac{x^2}{a^4} + \frac{y^2}{b^4}\right)^{-1/4} \tag{3.35}$$

## References


1. H.F. Bueckner, "A Novel Principle for the Computation of Stress Intensity Factors," ZAMM (Zeitschrift fur angewandte Mathematik und Mechanik) Band 46, pp. 529-545, 1970.

2. H.F. Bueckner, "Field Singularities and Related Integral Representations," Chapter 5 in "Methods of Analysis and Solutions of Crack Problems," edited by G.C. Sih, Noordhoff International Publishing, 1973.

3. J.R. Rice, "Some Remarks on Elastic Crack-Tip Stress Fields," Int. J. Solids Structures, Vol. 8, pp. 751-758, 1972; Pergamon Press.

4. R. Labbens, A. Pellissier-Tanot and J. Heliot, "Practical Method for Calculating Stress-Intensity Factors through Weight Functions," in "Mechanics of Crack Growth," ASTM, STP 590.

5. I.N. Sneddon and M. Lowengrub, "Crack Problems in the Classical Theory of Elasticity." The SIAM Series in Applied Mathematics. John Wiley & Sons, Inc., 1969.

6. M.K. Kassir and G.C. Sih, "Three-dimensional crack problems," Noordhoff International Publishing, 1975.

7. H. Bateman, "Partial Differential Equations of Mathematical Physics," pp. 435, 436, New York Dover Publications, 1944.

8. F.W. Dyson, "The Potentials of Ellipsoids of Variable Densities," Quart. J. Math. Oxford Ser. 25 (1891), pp. 259-288.

9. E.W. Hobson, "On some general Formulae for the Potentials of Ellipsoids, Shells and Disks," Proc. London Math. Soc., 27 (1896), pp. 519-544.

10. L.J. Walpole, "Some elastostatic and potential problems for an elliptic disk," Proceedings of the Cambridge Philosophical Society, 67, pp. 225-235 (1970).

11. C.M. Segedin, "A note on geometric discontinuities in elastostatics," Int. J. Engng. Sci., Vol 6, pp. 309-312, (1968).

# THE BUCKLING PRESSURE OF AN ELASTIC PLATE FLOATING ON WATER AND STRESSED UNIFORMLY ALONG THE PERIPHERY OF AN INTERNAL HOLE

Shunsuke Takagi

Corps of Engineers
U.S. Army Cold Regions Research and Engineering Laboratory
Hanover, New Hampshire

## INTRODUCTION

To test the strength of an ice sheet floating on water the following measurement is regularly performed (Zabilanski et al., 1): Dig a hole, place a vertical pile of various shapes and push it breaking through the ice. However, the mechanism of the failure is not yet clarified, and the interpretation of the data is not yet satisfactory. To understand the basic mechanism, an ideally simple case is chosen and analyzed in this paper.

A paper of the same title was presented at the 20th Conference of Army Mathematicians (1974). When numerical work was attempted in the summer of 1975, it was found that the analysis presented in the 20th Conference did not work as expected. A new analysis as reported in this paper was developed, and the numerical computation was carried out.

1. The Problem

Suppose a thin elastic plate floating on water, extending horizontally to infinity, and stressed with uniform horizontal pressure along the periphery of an internal circular hole. We are interested in formulating the buckling pressure and the deformation at the failure.

The vertical deflection $w$ of an elastic plate that rests on a liquid and is subjected to a vertical load $q$ and the horizontal stress of components $N_{xx}$, $N_{yy}$, and $N_{xy}$, is governed by the differential equation,

$$D\nabla^4 w + \gamma w = q + N_{xx}\frac{\partial^2 w}{\partial x^2} + 2N_{xy}\frac{\partial^2 w}{\partial x \partial y} + N_{yy}\frac{\partial^2 w}{\partial y^2} \quad (1.1)$$

where $D$ is the flexural rigidity and $\gamma$ the specific weight of the liquid (Ref. 2). Let $r$ be the radial distance from the center of the hole. In our problem $q$ = o and the deformation is cylindrically symmetric around the center of the hole. Then (1.1) becomes

$$l_o^4\left(\frac{d^2}{dr^2} + \frac{1}{r}\frac{d}{dr}\right)^2 w + w = \frac{1}{\gamma}\left(N_{rr}\frac{d^2 w}{dr^2} + N_{\theta\theta}\frac{1}{r}\frac{dw}{dr}\right) \quad (1.2)$$

where $l_o = (D/\gamma)^{1/4}$ is the characteristic length, and $N_{rr}$ and $N_{\theta\theta}$ are the radial and hoop horizontal stresses in the plate (see Appendix B).

Following the usual treatment (Ref. 2), we assume that the horizontal stress components $N_{xx}$, $N_{xy}$, $N_{yy}$ are in equilibrium by themselves. Then they are derived from a biharmonic function $\phi$ by

$$N_{xx} = \frac{\partial^2 \phi}{\partial y^2}$$

$$N_{yy} = \frac{\partial^2 \phi}{\partial x^2}$$

$$N_{xy} = - \frac{\partial^2 \phi}{\partial x \partial y}$$

In the general polar coordinates they are:

$$N_{rr} = \frac{1}{r} \frac{\partial \phi}{\partial r} + \frac{1}{r^2} \frac{\partial^2 \phi}{\partial \theta^2}$$

$$N_{r\theta} = - \frac{\partial}{\partial r} \left( \frac{1}{r} \frac{\partial \phi}{\partial \theta} \right)$$

$$N_{\theta\theta} = \frac{\partial^2 \phi}{\partial r^2}$$

In our problem $\phi$ is a function of $r$ only and must tend to zero when $r$ becomes infinite. Then they are formulated as

$$N_{rr} = - Ar^{-2}$$

$$N_{r\theta} = 0$$

$$N_{\theta\theta} = Ar^{-2} ,$$

where $A$ is a constant. Constant $A$ is positive because $N_{rr}$ is pressure. Instead of $A$ we introduce nondimensional constant $a$ and express the stress components as

$$\left.\begin{aligned} N_{rr} &= -a\gamma l_o^4 r^{-2} \\ N_{\theta\theta} &= a\gamma l_o^4 r^{-2} \end{aligned}\right\} \quad (1.3)$$

Introduce the nondimensional length $x$,

$$x = r l_o^{-1} \quad (1.4)$$

In this way (1.2) becomes

$$\frac{d^4w}{dx^4} + \frac{2}{x}\frac{d^3w}{dx^3} - \frac{1-a}{x^2}\frac{d^2w}{dx^2} + \frac{1-a}{x^3}\frac{dw}{dx} + w = 0 \qquad (1.5)$$

At $x = \infty$, the condition

$$\left.\begin{aligned} w &= 0 \\ \frac{dw}{dw} &= 0 \end{aligned}\right\} \qquad (1.6)$$

must be satisfied. At $x = x_o$, where $x_o$ is the value of $x$ at the periphery of the internal hole, we consider three conditions: (1) the clamped-edge condition

$$\left.\begin{aligned} w &= 0 \\ \frac{dw}{dx} &= 0 \end{aligned}\right\} \qquad (1.7)$$

2) the simple-edge condition

$$\left.\begin{aligned} w &= 0 \\ \frac{d^2w}{dx^2} + \frac{\nu}{x}\frac{dw}{dx} &= 0 \end{aligned}\right\} \qquad (1.8)$$

and (3) the free-edge condition

$$\left.\begin{aligned} \frac{d^2w}{dx^2} + \frac{\nu}{x}\frac{dw}{dx} &= 0 \\ \frac{d}{dx}\left(\frac{d^2w}{dx^2} + \frac{1}{x}\frac{dw}{dx} + \frac{a}{x^2}\frac{dw}{dx}\right) &= 0 \end{aligned}\right\} \qquad (1.9)$$

where $\nu$ is Poisson's ratio.

The second equation of (1.8) and the first equation of (1.9) are found from $M_r = 0$. The second equation of (1.9) is derived from

$Q_r + (1/r)(\partial M_{r\theta}/\partial\theta) = 0$. The effect of horizontal stress must be counted in $Q_r$. In the rectangular coordinates $x$, $y$, shears $Q_x$ and $Q_y$ are given by

$$\left.\begin{aligned} Q_x &= \frac{\partial M_{xx}}{\partial x} + \frac{\partial M_{xy}}{\partial y} + N_{xx}\frac{\partial w}{\partial x} + N_{xy}\frac{\partial w}{\partial y} \\ Q_y &= \frac{\partial M_{xy}}{\partial x} + \frac{\partial M_{yy}}{\partial y} + N_{xy}\frac{\partial w}{\partial x} + N_{yy}\frac{\partial w}{\partial y} \end{aligned}\right\} \quad (1.10)$$

These equations are found by extending Hitényi's (3) one-dimensional treatment to two-dimensional. In polar coordinates $r$, $\theta$, components of shear $Q_r$ and $Q_\theta$ are given (see Appendix A) by

$$\left.\begin{aligned} Q_r &= \frac{\partial M_r}{\partial r} + \frac{1}{r}\left\{M_{rr} + \frac{\partial M_{r\theta}}{\partial\theta} - M_{\theta\theta}\right\} + \frac{\partial w}{\partial r}N_{rr} + \frac{1}{r}\frac{\partial w}{\partial\theta}N_{r\theta} \\ Q_\theta &= \frac{\partial M_{r\theta}}{\partial r} + \frac{2}{r}M_{r\theta} + \frac{1}{r}\frac{\partial M_{\theta\theta}}{\partial\theta} + \frac{\partial w}{\partial r}N_{r\theta} + \frac{1}{r}\frac{\partial w}{\partial\theta}N_{\theta\theta} \end{aligned}\right\} \quad (1.11)$$

Constant $a$ is the eigenvalue to be determined to satisfy the boundary conditions at $x = x_0$. The first step for the solution of this eigenvalue problem is to discover, given a positive number $a$, two real functions, $w_1(x)$ and $w_2(x)$, that are the solutions of the differential equation (1.5) and meet the boundary conditions at $x = \infty$ in (1.6) but are not restricted at $x = x_0$ in any way. We call them the fundamental solutions. We shall find them later in the following form,

$$w_1 + iw_2 = \int_1^\infty \left\{\left(r^2 + r^4\text{-}1\right)^{-\sqrt{1-a}\,/2} + \left(r^2 + r^4\text{-}1\right)^{\sqrt{1-a}\,/2}\right\} e^{\frac{-1+i}{\sqrt{2}}xr}\left(r^4\text{-}1\right)^{-\frac{1}{2}}dr \quad (1.12)$$

The second step is to express the fundamental solutions as power series of $x$. Let $f_m(x)$ $(m = 0, 1, 2, 3)$ be the Fuchsian type solutions of (1.5) relative to $x = 0$. We shall find linear combinations,

$$w_k(x) = \sum_{m=0}^{3} A_{km} f_m(x) \tag{1.13}$$

by determining constants $A_{km}$ by use of (1.12). The solution $w(x)$ is a linear combination of the fundamental solutions,

$$w(x) = A\, w_1(x) + B\, w_2(x) \tag{1.14}$$

The third step is to solve the simultaneous equations of $A$ and $B$ that are found by substituting (1.14) into the boundary conditions (1.7), (1.8), or (1.9) at $x = x_o$. If a root of the algebraic equation found by letting the determinant of the simultaneous equations equal to zero is positive, the root gives $x_o$. Our problem is then solved.

1a. Abstract of the result.

The main feature of the numerical result is as follows:

1. Buckling takes place under the free-edge condition. Buckling does not take place under the clamped-edge and the simple-edge conditions.

2. Eigenvalue $a$ under the free-edge condition is found in the range $1-\nu^2 \leq a < \infty$, where $\nu$ is Poisson's ratio of the elastic ice plate. When $a = 1-\nu^2$, root $x_o$ is equal to zero. Analysis presented here is complete for the case $1-\nu^2 \leq a \leq 2$, but not complete for the case $2 < a < \infty$. It is believed that the result presented in this paper can practically cover all the cases of our interest.

3. Buckling under the free-edge condition takes the shape as shown in Figure 5 and 6. ($a$ is restricted to $1-\nu^2 \leq a \leq 2$). This shape of

deformation is observed frequently in laboratory experiments and field tests. Therefore we may conclude that buckling is an important mechanism of failure.

## PART I. FUNDAMENTAL SOLUTIONS

### 2. Fuchsian Type Solutions

Equation (1.5) has a regular singularity at $x = 0$. The solutions relative to $x = 0$ are the Fuchsian type power series of $x$. Their indicial numbers are:

$$\nu_0 = 0$$

$$\nu_1 = 2$$

$$\nu_2 = 1 + \mu$$

$$\nu_3 = 1 - \mu$$

where

$$\mu = \sqrt{1-a} \tag{2.1}$$

The four solutions may be expressed with a single formula

$$f_m(x) = \sum_{n=0}^{\infty} a_n^{(m)} x^{\nu_m + 4n} \tag{2.2}$$

where $m = 0, 1, 2, 3$, and

$$a_0^{(m)} = 1$$

and the rest of the coefficients $a_n^{(m)} (n \geq 1)$ are determined by the recurrence formula,

$$a_n^{(m)} = - a_{n-1}^{(m)} \left[ (\nu_m + 4n)(\nu_m + 4n-2)(\nu_m + 4n-1-\mu)(\nu_m + 4n-1+\mu) \right]^{-1} \tag{2.3}$$

Their individual forms are:

$$f_0(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{3n}(2n)!} \frac{\Gamma\left(\frac{1}{4}(3-\mu)\right)\ \Gamma\left(\frac{1}{4}(3+\mu)\right)}{\Gamma\left(n+\frac{1}{4}(3-\mu)\right)\Gamma\left(n+\frac{1}{4}(3+\mu)\right)} x^{4n} \tag{2.4}$$

$$f_1(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{3n}(2n+1)!} \frac{\Gamma\left(\frac{1}{4}(5-\mu)\right)\ \Gamma\left(\frac{1}{4}(5+\mu)\right)}{\Gamma\left(n+\frac{1}{4}(5-\mu)\right)\ \Gamma\left((n+\frac{1}{4}(5+\mu)\right)} x^{4n+2} \tag{2.5}$$

$$f_{k+1}(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{4n}n!} \frac{\Gamma\left(\frac{1}{2}(2\pm\mu)\right)\ \Gamma\left(\frac{1}{4}(3\pm\mu)\right)\ \Gamma\left(\frac{1}{4}(5\pm\mu)\right)}{\Gamma\left(n+\frac{1}{2}\ (2\pm\mu)\right)\ \Gamma\left(n+\frac{1}{4}(3\pm\mu)\right)\ \Gamma\left(n+\frac{1}{4}(5\pm\mu)\right)} x^{4n+1\pm\mu} \tag{2.6}$$

where $k$=1, 2. In (2.6) we have introduced the convention that the upper or lower of the double sign $\pm$ (or $\mp$) should be taken according to $k = 1$ or 2, respectively. This convention is observed throughout the paper. The main objective of PART I is to determine the fundamental solution, i.e. to determine $A_{km}$ in (1.13).

Differential equation (1.5) has an irregular singularity at $x = \infty$. In other words, the solution relative to $x = \infty$, say $f(x)$, can be found in the form

$$f(x) = e^{-\lambda x} \sum_{n=0}^{\infty} p_n\ x^{-\frac{1}{2}-n}$$

where $\lambda$ satisfies $\lambda^4 + 1 = 0$. The series $\sum_{n=0}^{\infty} p_n\ x^{-n}$ in this equation is asymptotic and divergent in this case. Therefore this equation does not provide any means for determining $A_{km}$ in (1.13).

## 3. Countour Integral Solution

In order to find the fundamental solutions, (1.5) must be transformed by means of the contour integral,

$$w(x) = \int_L v(\zeta)\ e^{x\zeta}\ d\zeta \tag{3.1}$$

where L is a contour in the complex plane of $\zeta$ that shall be determined to let a solution of (1.5) satisfy the boundary condition (1.6) at $x = \infty$. Following the usual procedure (Ince (4), pp. 187-188), one arrives at the differential equation of $v(\zeta)$,

$$(1+\zeta^4)\ \frac{d^3v}{d\zeta^3} + 10\zeta^3\ \frac{d^2v}{d\zeta^2} + (23+a)\zeta^2\ \frac{dv}{d\zeta} + (9+3a)\zeta v = 0 \tag{3.2}$$

The contour L selected for this solution is shown in Figure 1.

To find the solution relative to $\zeta = \infty$, let

$$\zeta = \beta r \tag{3.3}$$

where

$$\beta = \exp(3\pi i/4). \tag{3.4}$$

Then the equation (3.2) becomes

$$(r^4-1)\ \frac{d^3v}{dr^3} + 10r^3\ \frac{d^2v}{dr^2} + (23+a)\ r^2\ \frac{dv}{dr} + (9+3a)rv = 0 \tag{3.5}$$

This equation has a regular singular point at $r = \infty$. The indicial numbers $\lambda_m$ $(m = 0, 1, 2)$ at $r = \infty$ are:

$$\lambda_0 = 3$$

$$\lambda_1 = 2 + \mu$$

$$\lambda_2 = 2 - \mu$$

where $\mu$ is given by (2.1). The solution $v_m(r)$ corresponding to the indicial number $\lambda_m$ is:

$$v_m(r) = \sum_{n=0}^{\infty} q_n^{(m)}\ r^{-\lambda_m-4n} \tag{3.6}$$

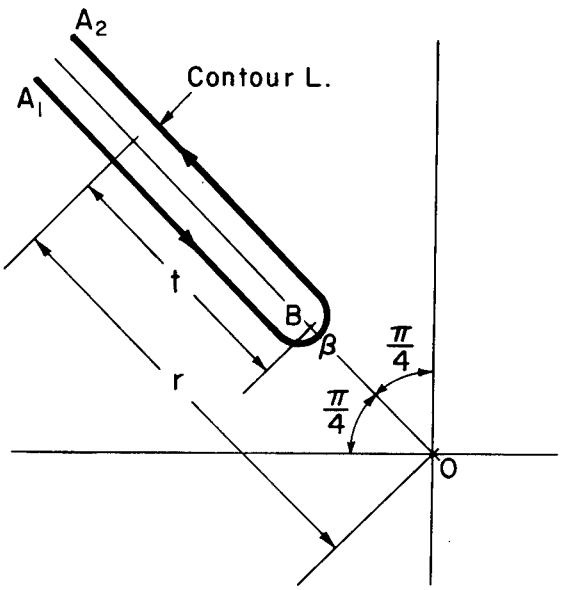



Fig. 1. Contour L on the complex plane $\zeta$. Points $A_1$ and $A_2$ represent points $\infty\beta$ on the respective branch.

where

$$q_n^{(m)} = \prod_{p=0}^{n-1} (4p+\lambda_m)(4p+\lambda_m+2) \left[(4p+\lambda_m+2-\mu)(4p+\lambda_m+2+\mu)\right]^{-1} \quad (3.7)$$

The contour L must be such that the point $\zeta = \beta$ is a branch point of $v(\zeta)$. This condition is satisfied by the series $v_1(r)$ and $v_2(r)$, as shown by (3.10) below. These two series can be expressed by means of hypergeometric series F ( , , ; ;) as

$$v_k(r) = r^{-2\mp\mu} \; F\left((2\pm\mu)/4,\ (4\pm\mu)/4;\ (2\pm\mu)/2;\ r^{-4}\right) . \quad (3.8)$$

The hypergeometric series are summed up by use of the formula

$$F(\alpha, \tfrac{1}{2}+\alpha;\ 2\alpha;\ z) = 2^{2\alpha-1}\,(1-z)^{-\frac{1}{2}} \left[1+\sqrt{1-z}\right]^{1-2\alpha} \quad (3.9)$$

[Handbook (Ref. 5) p. 556, Formula (15.1.14)]. Thus $v_k(r)$, where $k = 1$ or 2, reduces to

$$v_k(r) = (r^4-1)^{-\frac{1}{2}} \left[(1/2)(r^2+\sqrt{r^4-1})\right]^{\mp\mu/2} \quad (3.10)$$

This equation shows that $\zeta = \beta$ is a branch point. Formula (3.9) can be proved by showing that the one on the right-hand side satisfies the hypergeometric differential equation of the one on the left-hand side and also that they satisfy the same initial conditions at $z = 0$.

Suppose that $v_k(r)$ on one of the branch $A_1B$ of $L$ in Figure 1 is given by (3.10). Then, $v_k(r)$ on the other branch $A_2B$ of $L$ is given by

$$-(r^4-1)^{-\frac{1}{2}} \left[(1/2)(r^2+\sqrt{r^4-1})\right]^{\pm\mu/2}$$

Thus one finds the integral solution,

$$F(x) = \int_1^\infty \left\{\left(r^2+\sqrt{r^4-1}\right)^{-\frac{\mu}{2}} + \left(r^2+\sqrt{r^4-1}\right)^{\frac{\mu}{2}}\right\} e^{\beta x r}\,(r^4-1)^{-\frac{1}{2}} dr \quad (3.11)$$

The fundamental solutions $w_1(x)$ and $w_2(x)$ are given by the real and imaginary parts of $F(x)$,

$$F(x) = w_1(x) + i\, w_2(x) \tag{3.12}$$

4. Integration of the Integral Solution

We shall integrate (3.11) to a linear combination of $f_o(x)$, $f_1(x)$, $f_2(x)$ and $f_3(x)$. The first step for this goal is to change the range of integration in (3.11). Introduce a complex variable $z = \beta r$, where $\beta$ is given by (3.4). Use of $z$ transforms (3.11) to

$$F(x) = \int_{\beta}^{\infty\beta} \left\{ \left(\beta^{-2}z^2 + \sqrt{-z^4-1}\right)^{-\frac{\mu}{2}} + \left(\beta^{-2}z^2 + \sqrt{-z^4-1}\right)^{\frac{\mu}{2}} \right\} e^{zx}\ \beta^{-1}(-z^4-1)^{-\frac{1}{2}}\, dz \tag{4.1}$$

The range of integration $\beta \sim \infty\ \beta$ in (4.1) shall be changed to $\beta \sim 0$ and $0 \sim -\infty$. Thus (4.1) becomes

$$F(x) = \left\{ \int_{\beta}^{0} + \int_{0}^{-\infty} \right\} \left\{ \left(\beta^{-2}z^2 + i\sqrt{z^4+1}\right)^{-\frac{\mu}{2}} + \left(\beta^{-2}z^2 + i\sqrt{z^4+1}\right)^{\frac{\mu}{2}} \right\} e^{zx}\ i^{-1}\beta^{-1}(z^4+1)^{-\frac{1}{2}}\, dz \tag{4.2}$$

In the above equations, quantities inside the square roots are chosen to be positive in order to insure correct forms in the respective ranges. Letting $z = \beta r$ in the first integral and $z = -r$ in the second integral, (4.2) transforms to a summation of normal forms of integration,

$$F(x) = ih_1(x) + ih_2(x) - \infty\exp(-\mu\pi i/4)\, g_1(x) - \beta\exp(\mu\pi i/4)\, g_2(x) \tag{4.3}$$

where

$$h_k(x) = \int_0^1 \left(r^2 + i\sqrt{1-r^4}\right)^{\mp\frac{\mu}{2}} e^{\beta x r}(1-r^4)^{-\frac{1}{2}}\, dr \tag{4.4}$$

and

$$g_k(x) = \int_0^\infty \left(r^2 + \sqrt{r^4+1}\right)^{\mp \frac{\mu}{2}} e^{-xr} \quad (r^4+1)^{-\frac{1}{2}} dr \tag{4.5}$$

Expansion of $\exp(\beta xr)$ transforms (4.4) to power series of $x$,

$$h_k(x) = h_o^{(k)} + \beta x\, h_1^{(k)} + \frac{1}{2!}(\beta x)^2 h_2^{(k)} + \ldots \tag{4.6}$$

where

$$h_n(k) = \int_0^1 \left(r^2 + i\sqrt{1-r^4}\right)^{\mp \frac{\mu}{2}} r^n (1-r^4)^{-\frac{1}{2}} dr \tag{4.7}$$

Integration of (4.7) will be carried out later. Integral (4.5) transforms to power series of $x$

$$g_1(x) = g_o^{(1)} + g_1^{(1)}x + g_\mu^{(1)}x^{1+\mu} + g_2^{(1)}x^2 + \ldots \tag{4.8}$$

and

$$g_2(x) = g_o^{(2)} + g_\mu^{(2)}x^{1-\mu} + g_1^{(2)}x + g_2^{(2)}x^2 + \ldots \tag{4.9}$$

as explained in the next section. Thus one finds $F(x)$ in the following form.

$$F(x) = B_o + B_\mu^{(2)}x^{1-\mu} + B_1 x + B_\mu^{(1)}x^{1+\mu} + B_2 x^2 + \ldots \tag{4.10}$$

where

$$B_o = i\left(h_o^{(1)} + h_o^{(2)}\right) - \beta\left(e^{-\frac{\mu\pi i}{4}} g_o^{(1)} + e^{\frac{\mu\pi i}{4}} g_o^{(2)}\right) \tag{4.11}$$

$$B_1 = i\beta\left(h_1^{(1)} + h_1^{(2)}\right) - \beta\left(e^{-\frac{\mu\pi i}{4}} g_1^{(1)} + e^{\frac{\mu\pi i}{4}} g_1^{(2)}\right) \tag{4.12}$$

$$B_2 = \frac{1}{2}\left(h_2^{(1)} + h_2^{(2)}\right) - \beta\left(e^{-\frac{\mu\pi i}{4}} g_2^{(1)} + e^{\frac{\mu\pi i}{4}} g_2^{(2)}\right) \tag{4.13}$$

$$B_\mu^{(k)} = -\beta e^{\mp \frac{\mu\pi i}{4}} g_\mu^{(k)} \tag{4.14}$$

In this calculation we tentatively assume that $0 < \alpha < 1$. Series are arranged in the ascending order on this assumption. The formulas for the

values of $a$ outside the range $0 < a < 1$ will be derived from the formulas in this range. Note that the entries in (4.10), except $B_1 x$, are the first terms of $f_m(x)$, where $m = 0, 1, 2, 3$. The first-order term, $B_1 x$, is not contained in any of $f_m(x)$; it is proved later that $B_1 = 0$. The entries in (4.10) are sufficient to express $w_k(x)$ as a linear combinations of $f_m(x)$.

4a. Formulas of $g_n^{(k)}$

We shall give the integral forms of $g_n^{(k)}$ by successively developing (4.5) into series of $x$. Integration of these formulas will be carried out in the next section.

Letting $x = 0$ in (4.5), one finds $g_o^{(k)}$:

$$g_o^{(k)} = \int_0^\infty (r^2 + \sqrt{r^4+1})^{\frac{\mp\mu}{2}} (r^4+1)^{-\frac{1}{2}} dr \qquad (4a.1)$$

To find $g_\mu^{(2)}$, the formula

$$g_2(x) - g_o^{(2)} = \int_0^\infty (r^2 + \sqrt{r^4+1})^{\frac{\mu}{2}} (e^{-rx} - 1) (r^4+1)^{-\frac{1}{2}} dr$$

shall be transformed by introducing $\xi = rx$ to

$$= x^{1-\mu} \int_0^\infty (\xi^2 + \sqrt{\xi^4+x^4})^{\frac{\mu}{2}} (e^{-\xi} - 1) (\xi^4+x^4)^{-\frac{1}{2}} d\xi \qquad (4a.2)$$

Letting $x = 0$ inside the integral, one finds

$$g_\mu^{(2)} = 2^{\frac{\mu}{2}} \int_0^\infty \xi^{\mu-2} (e^{-\xi} - 1) \, d\xi \qquad (4a.3)$$

To find $g_1^{(2)}$, multiply $x^{1-\mu}$ on (4a.3) and subtract it from (4a.2). Thus one finds

$$g_2(x) - g_o^{(2)} - g_\mu^{(2)} x^{1-\mu}$$

$$= x^{1-\mu} \int_o^\infty \left\{ (\xi^4+x^4)^{-\frac{1}{2}} (\xi^2+\sqrt{\xi^4+x^4})^{\frac{1}{2^\mu}} - 2^{\frac{\mu}{2}} \xi^{\mu-2} \right\} (e^{-\xi}-1)\, d\xi$$

Letting $\xi = rx$, this transforms to

$$= x \int_o^\infty \phi_1(r) r \frac{e^{-rx}-1}{rx} dr \tag{4a.4}$$

where

$$\phi_1(r) = (r^2+\sqrt{r^4+1})^{\frac{1}{2^\mu}} (r^4+1)^{-\frac{1}{2}} - 2^{\frac{\mu}{2}} r^{\mu-2} \tag{4a.5}$$

Because of the inequality $1 \geq (1 - e^{-u})/u \geq 1 - u/2$ the integrand of (4a.4) is uniformly bounded. One can, therefore, let $x \to 0$ inside the integral. Thus

$$g_1^{(2)} = -\int_o^\infty \phi_1(r) \; r \, dr \tag{4a.6}$$

To find $g_2^{(2)}$, multiply $x$ on (4a.6) and subtract it from (4a.4). Thus one finds

$$g_2(x) - g_o^{(2)} - g_\mu^{(2)} x^{1-\mu} - g_1^{(2)} x$$

$$= x^2 \int_1^\infty \phi_1(r) \; r^2 \frac{e^{-xr} - 1 + rx}{r^2x^2} dr$$

Because of the inequality $0 \geq (1-u-e^{-u})/u^2 \geq -\frac{1}{2}$ the integrand of the last integral is uniformly bounded. One can, therefore, let $x \to 0$ inside the integral. Thus

$$g_2^{(2)} = \frac{1}{2} \int_o^\infty \phi_1(r) \; r^2 \, dr \tag{4a.7}$$

To find $g_1^{(1)}$, one may simply differentiate $g_1(x)$ in (4.5) with regard to $x$ and let $x = 0$ in the result. Thus one finds

$$g_1^{(1)} = -\int_0^\infty (r^2 + \sqrt{r^4 + 1})^{-\frac{\mu}{2}} [r/(r^4+1)^{\frac{1}{2}}] \, dr \qquad (4a.8)$$

To find $g_\mu^{(1)}$, use (4a.1) and (4a.8) to derive the formula

$$g_1(x) - g_0^{(1)} - g_1^{(1)}x = \int_0^\infty (r^2 + \sqrt{r^4 + 1})^{-\frac{\mu}{2}} (e^{-rx} - 1 + rx)(r^4+1)^{-\frac{1}{2}} \, dr$$

Letting $\xi = rx$, this becomes

$$= x^{1+\mu} \int_0^\infty (\xi^4+x^4)^{-\frac{1}{2}} (\xi^2+\sqrt{\xi^4+x^4})^{-\frac{\mu}{2}} (e^{-\xi}-1+\xi) \, d\xi \qquad (4a.9)$$

Letting $x = 0$ inside the integral, one finds

$$g_\mu^{(1)} = 2^{-\frac{\mu}{2}} \int_0^\infty \xi^{-\mu-2} (e^{-\xi} - 1 + \xi) \, d\xi \qquad (4a.10)$$

To find $g_2^{(1)}$, use (4a.9) and (4a.10) to derive the formula

$$g_1(x) - g_0^{(1)} - g_1^{(1)}x - g_\mu^{(1)} x^{1+\mu}$$

$$= x^{1+\mu} \int_0^\infty \left\{ (\xi^4+x^4)^{-\frac{1}{2}} (\xi^2+\sqrt{\xi^4+x^4})^{-\frac{\mu}{2}} - 2^{-\frac{\mu}{2}} \xi^{-\mu-2} \right\} (e^{-\xi}-1+\xi) \, d\xi$$

Letting $\xi = rx$, this transforms to

$$= x^2 \int_0^\infty \phi_2(r) \, r^2 \, \frac{e^{-rx} - 1 + rx}{r^2x^2} \, dr \qquad (4a.11)$$

where

$$\phi_2(r) = (r^4+1)^{-\frac{1}{2}} (r^2+\sqrt{r^4+1})^{-\frac{\mu}{2}-2^{-\frac{\mu}{2}}} r^{-\mu-2}$$

Because the integrand of (4a.11) is uniformly bounded, one can let $x \to 0$ inside the integral. Thus one finds

$$g_2^{(1)} = \frac{1}{2}\int_0^\infty \phi_2(r)\ r^2\ dr \tag{4a.12}$$

4b. Evaluation of $g_n^{(k)}$

The independent variable $\eta$ introduced by

$$r^2 + \sqrt{r^4+1} = \eta^{-1/2} \tag{4b.1}$$

is useful for the following integrations. This transforms to

$$r^2 = (1-\eta)/(2\eta^{1/2})$$

$$\sqrt{1+r^4} = (1+\eta)/(2\eta^{1/2})$$

and

$$r\ dr = -[(1+\eta)/(4\eta^{3/2})]\ d\eta$$

Substituting these in (4a.1), one finds

$$g_o^{(k)} = (2\sqrt{2})^{-1}\ B\left(\frac{1}{2}, \frac{1\mp\mu}{4}\right)$$

$$= (2\sqrt{2\pi})^{-1}\ \Gamma\left(\frac{1+\mu}{4}\right)\Gamma\left(\frac{1-\mu}{4}\right) \sin\frac{\pi(1\mp\mu)}{4} \tag{4b.2}$$

Eq. (4a.8) is similarly integrated to

$$g_1^{(1)} = -\mu^{-1} \tag{4b.3}$$

Use of $\eta$ transforms (4a.6) to

$$g_1^{(2)} = -\frac{1}{4}\int_0^1 \left[1-(1+\eta)(1-\eta)^{\frac{\mu}{2}-1}\right]\eta^{-\frac{\mu}{4}-1}\ d\eta$$

By letting $1+\eta = 2-(1-\eta)$, the last integral is divided in two portions; thus $g_1^{(2)}$ becomes

$$= -\frac{1}{4} J_1 - \frac{1}{2} J_2$$

where

$$J_1 = \int_0^1 \left[1 - (1-\eta)^{\frac{\mu}{2}}\right] \eta^{-\frac{\mu}{4}-1} \, d\eta$$

and the remaining integral $J_2$ simply integrates to

$$J_2 = -B(\mu/2, 1 - \mu/4)$$

After partial integration of $\eta^{-\frac{\mu}{4}-1}$, $J_1$ integrates to

$$J_1 = -4/\mu + 2B(\mu/2, 1 - \mu/4)$$

Thus one finds

$$g_1^{(2)} = \mu^{-1} \tag{4b.4}$$

Combining (4b.3) and (4b.4) the result may be shown with a single formula,

$$g_1^{(k)} = \mp \mu^{-1} \tag{4b.5}$$

After partial integration of $\zeta^{\mu-2}$, (4a.3) integrates; after two times of partial integrations of $\zeta^{-\mu-2}$, (4a.10) integrates; the two results are shown here with a single formula:

$$g_\mu^{(k)} = \pm 2^{\mp\mu/2} [\mu(1\pm\mu)]^{-1} \Gamma(1\mp\mu) \tag{4b.6}$$

We express $g_2^{(2)}$ in (4a.7) and $g_2^{(1)}$ in (4a.12) with a single formula,

$$g_2^{(k)} = \frac{1}{2}\int_0^\infty \left\{(r^4+1)^{-\frac{1}{2}} (r^2 + r^4+1)^{\mp\mu/2} - 2^{\mp\mu/2} r^{-2\mp\mu}\right\} r^2 \, dr$$

Use of $\eta$ changes this to

$$g_2^{(k)} = \frac{1}{8\sqrt{2}}\int_0^1 \left\{1 - (1+\eta)(1-\eta)^{-1\mp\frac{\mu}{2}}\right\} \eta^{-\frac{5}{4}\pm\frac{\mu}{4}} (1-\eta)^{1/2} \, d\eta$$

Letting $1 + \eta = 2 - (1 - \eta)$, this can be divided in two integrals,

$$g_2^{(k)} = (8\sqrt{2})^{-1} K_1 + (4\sqrt{2})^{-1} K_2$$

where

$$K_1 = \int_0^1 1 - \left((1-\eta)^{\mp \frac{\mu}{2}}\right) \eta^{-\frac{5}{4} \pm \frac{\mu}{4}} (1-\eta)^{1/2} \, d\eta$$

and the remaining integral $K_2$ simply integrates to

$$K_2 = -B\left(\frac{1 \mp \mu}{2}, \frac{3 \pm \mu}{4}\right)$$

After the partial integration of $\eta^{-\frac{5}{4} \pm \frac{\mu}{4}}$, $K_1$ integrates to

$$K_1 = \frac{2}{-1 \pm \mu} \left\{ B\left(\frac{1}{2}, \frac{3 \pm \mu}{4}\right) - (1 \mp \mu) \, B\left(\frac{1 \mp \mu}{2}, \frac{3 \pm \mu}{4}\right) \right\}$$

Thus one finds

$$g_2^{(k)} = \left[4\sqrt{2}(-1 \pm \mu)\right]^{-1} B\left(\frac{1}{2}, \frac{3 \pm \mu}{4}\right)$$

$$= -\left[\sqrt{2\pi}(1-\mu^2)\right]^{-1} \Gamma\left(\frac{3+\mu}{4}\right) \Gamma\left(\frac{3-\mu}{4}\right) \sin\frac{\pi(1 \pm \mu)}{4} \qquad (4b.7)$$

4c. Integration of $h_n^{(k)}$

Let

$$r^2 + i\sqrt{1 - r^4} = \zeta \qquad (4c.1)$$

Then

$$r^2 = (1 + \zeta^2)/(2\zeta)$$

$$\sqrt{1 - r^4} = i\,(1 - \zeta)^2/(2\zeta)$$

and

$$r\,dr = \left[(\zeta^2 - 1)/(4\zeta^2)\right] d\zeta$$

Substituting these in (4.7), one gets

$$h_n^{(k)} = i2^{-\frac{n+1}{2}} \int_i^1 \zeta^{-\frac{n+1}{2} \mp \frac{\mu}{2}} (1+\zeta^2)^{\frac{n-1}{2}} \, d\zeta \qquad (4c.2)$$

For $n = 1$, this integrates to

$$h_1^{(k)} = \mp (i/\mu) \left[1 - \exp(\mp\mu\pi i/4)\right] \qquad (4c.3)$$

To integrate (4c.2) for $n=0$ and 2, it is noted that $h_n^{(1)} + h_n^{(2)}$ is real. To show this, let $r^2 = \cos\theta$ in (4.7) to get

$$h_n^{(1)} + h_n^{(2)} = \int_0^{\frac{\pi}{2}} (\cos\theta)^{\frac{n-1}{2}} \cos\mu\theta \, d\theta$$

which is real. Divide the contour of (4c.2) in two parts,

$$h_n^{(k)} = i2^{-\frac{n+1}{2}} \left\{ I_n^{(k)} - J_n^{(k)} \right\}$$

where

$$I_n^{(k)} = \int_0^1 \zeta^{\mp\frac{\mu}{2} - \frac{n+1}{2}} (1+\zeta^2)^{\frac{n-1}{2}} \, d\zeta$$

and

$$J_n^{(k)} = \int_0^i \zeta^{\mp\frac{\mu}{2} - \frac{n+1}{2}} (1+\zeta^2)^{\frac{n-1}{2}} \, d\zeta$$

Letting $\zeta = i\sqrt{x}$, the latter integrates to

$$J_n^{(k)} = \frac{1}{2} \exp\left\{\frac{\pi i}{4}(\mp\mu+1-n)\right\} \Gamma\left(\frac{\mp\mu+1-n}{4}\right) \Gamma\left(\frac{n+1}{2}\right) \Big/ \Gamma\left(\frac{\mp\mu+3+n}{4}\right)$$

Transforming the Gama-functions, this yields

$$J_o^{(k)} = (2\sqrt{\pi})^{-1} \Gamma\left(\frac{1+\mu}{4}\right) \Gamma\left(\frac{1-\mu}{4}\right) \exp\left(\frac{\pi i(1\mp\mu)}{4}\right) \sin\frac{\pi(1\mp\mu)}{4}$$

and

$$J_2^{(k)} = -4\left[(1-\mu^2)\sqrt{\pi}\right]^{-1}\Gamma\left(\frac{3+\mu}{4}\right)\Gamma\left(\frac{3-\mu}{4}\right)\exp\left(\frac{\pi i(-1\mp\mu)}{4}\right)\sin\frac{\pi(1\mp\mu)}{4}.$$

Thus one finds

$$h_o^{(1)}+h_o^{(2)} = i2^{-\frac{1}{2}}\left(I_o^{(1)}+I_o^{(2)}\right)+(2\sqrt{2\pi})^{-1}\Gamma\left(\frac{1+\mu}{4}\right)\Gamma\left(\frac{1-\mu}{4}\right)\left(\cos\frac{\mu\pi}{2}-i\right)$$

and

$$h_2^{(1)}+h_2^{(2)} = i2^{-\frac{3}{2}}\left(I_2^{(1)}+I_2^{(2)}\right)+\sqrt{2}\left[\sqrt{\pi}(1-\mu^2)\right]^{-1}\Gamma\left(\frac{3+\mu}{4}\right)\Gamma\left(\frac{3-\mu}{4}\right)\left(\cos\frac{\mu\pi}{2}+i\right)$$

Taking the real parts, one finds

$$h_o^{(1)} + h_o^{(2)} = (2\sqrt{2\pi})^{-1}\Gamma\left(\frac{1+\mu}{4}\right)\Gamma\left(\frac{1-\mu}{4}\right)\cos\frac{\mu\pi}{2} \tag{4c.4}$$

and

$$h_2^{(1)} + h_2^{(2)} = (1-\mu^2)^{-1}\sqrt{\frac{2}{\pi}}\,\Gamma\left(\frac{3+\mu}{4}\right)\Gamma\left(\frac{3-\mu}{4}\right)\cos\frac{\mu\pi}{2} \tag{4c.5}$$

4d. Fundamental solutions for $0 < a < 1$.

Substituting (4b.2) and (4c.4) into (4.11), one gets

$$B_o = (2\sqrt{2\pi})^{-1}\Gamma\left(\frac{1+\mu}{4}\right)\Gamma\left(\frac{1-\mu}{4}\right) \tag{4d.1}$$

Substituting (4b.5) and (4c.3) into (4.12), one gets

$$B_1 = 0 \tag{4d.2}$$

Substituting (4b.7) and (4c.5) into (4.13), one gets

$$B_2 = i\left[\sqrt{2\pi}\,(1-\mu^2)\right]^{-1}\Gamma\left(\frac{3+\mu}{4}\right)\Gamma\left(\frac{3-\mu}{4}\right) \tag{4d.3}$$

Substituting (4b.6) into (4.14), one gets

$$B_\mu^{(k)} = \mp\left[\mu(1-\mu^2)\right]^{-1}\,2^{\mp\frac{\mu}{2}}\,\Gamma(2\mp\mu)\exp\frac{(3\mp\mu)\pi i}{4} \tag{4d.4}$$

Thus one finds

$$F(x) = B_o f_o(x) + B_2 f_2(x) + B_\mu^{(1)} f_2(x) + B_\mu^{(2)} f_3(x) \tag{4d.5}$$

When $0 < a < 1$, functions $f_m(x)$ $(m = 0, 1, 2, 3)$ are real. Therefore, fundamental solutions $w_1(x)$ and $w_2(x)$ are found by decomposing the coefficients into real and imaginary parts. Thus

$$w_1(x) = p_o f_o(x) + p_1 f_2(x) + p_2 f_3(x) \qquad (4d.6)$$

and

$$w_2(x) = q_o f_2(x) + q_1 f_2(x) + q_2 f_3(x) \qquad (4d.7)$$

where

$$p_o = (2\sqrt{2\pi})^{-1} \Gamma\left(\frac{1+\mu}{4}\right) \Gamma\left(\frac{1-\mu}{4}\right) \qquad (4d.8)$$

$$q_o = \left[\sqrt{2\pi}\,(1-\mu^2)\right]^{-1} \Gamma\left(\frac{3+\mu}{4}\right) \Gamma\left(\frac{3-\mu}{4}\right) \qquad (4d.9)$$

$$p_k = \mp \left[\mu(1-\mu^2)\right]^{-1} 2^{\mp\frac{\mu}{2}} \Gamma(2\mp\mu) \cos\frac{(3\mp\mu)\pi}{4} \qquad (4d.10)$$

$$q_k = \mp \left[\mu(1-\mu^2)\right]^{-1} 2^{\mp\frac{\mu}{2}} \Gamma(2\mp\mu) \sin\frac{(3\mp\mu)\pi}{4} \qquad (4d.11)$$

5. Fundamental Solutions for $a = 1$

When $a = 1$, (1.5) reduces to

$$\frac{d^4w}{dx^4} + \frac{2}{x}\frac{d^3w}{dx^3} + w = 0$$

Nevel (1961) gave the Fuchsian type solutions of this equation with the notations,

$$\mathrm{nev}_o(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{3n}(2n)!} \left[\frac{\Gamma(\frac{3}{4})}{\Gamma(n+\frac{3}{4})}\right]^2 x^{4n}$$

$$\mathrm{nev}_1(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{4n}(n!)^2} \frac{\Gamma(\frac{3}{4})\,\Gamma(\frac{5}{4})}{\Gamma(n+\frac{3}{4})\,\Gamma(n+\frac{5}{4})} x^{4n+1}$$

$$\mathrm{nev}_2(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{3n}(2n+1)!} \left[\frac{\Gamma(\frac{5}{4})}{\Gamma(n+\frac{5}{4})}\right]^2 x^{4n+2}$$

and

$$nel_1(x) = nev_1(x)\ \log x - \sum_{n=1}^{\infty} \frac{(-1)^n}{4^{4n}(n!)^2} \frac{\Gamma(\frac{3}{4})\ \Gamma(\frac{5}{4})}{\Gamma(n+\frac{3}{4})\ \Gamma(n+\frac{5}{4})} \sum_{k=1}^{n} \left(\frac{1}{4k-1} + \frac{1}{2k} + \frac{1}{4k+1}\right) x^{4n+1}$$

To find the fundamental solutions, the unknown function $w(x)$ must be transformed to the unknown function $v(\zeta)$ defined by the contour integral (3.1). The transformed differential equation is

$$(1 + \zeta^4)\ \frac{dv}{d\zeta} + 2\zeta^3 v = 0$$

and one finds

$$v(\zeta) = (1+\zeta^4)^{-1/2}$$

Therefore the complex solution $w(x)$ for $a = 1$ is found in the integral form

$$w(x) = \int_L \frac{e^{\zeta x}}{\sqrt{1+\zeta^4}}\ d\zeta \qquad (5.1)$$

The contour L is the one shown in Figure 1. To find the fundamental solutions in the form of the linear combination of the *nev* functions, J. Dieudonné (1958), as explained in Nevel (1968), expanded (5.1) into power series in the neighborhood of $x = 0$, and determined the first few coefficients. The fundamental solutions thus found are denoted here by $w_1^N$ and $w_2^N$:

$$w_1^N(x) = (4\sqrt{2\pi})^{-1}\ \Gamma^2(\tfrac{1}{4})\ nev_0(x) - \pi(2\sqrt{2})^{-1}\ nev_1(x) + (2\sqrt{2\pi})^{-1}\ \Gamma^2(\tfrac{3}{4})\ nev_2(x)$$

and

$$w_2^N(x) = (8\sqrt{\pi})^{-1}\ \Gamma^2(\tfrac{1}{4})\ nev_0(x) - (4\sqrt{\pi})^{-1}\ \Gamma^2(\tfrac{3}{4})\ nev_2(x) + nel_1(x) - (1 - \gamma + \log\sqrt{2})\ nev_1(x)$$

where $\gamma$ is Enler's constant 0.5772156.

We shall show in the following that $w_1(x)$ and $w_2(x)$ in (4d.6) and (4d.7), respecitvely, gives

$$\lim_{a\to 1} \; w_1(x) \;=\; w_1^N(x) \;+\; \sqrt{2}\, w_2^N(x)$$

$$\lim_{a\to 1} \; w_2(x) \;=\; w_1^N(x) \;-\; \sqrt{2}\, w_2^N(x)$$

To show this, note that

$$\lim_{a\to 1} \; p_o\, f_o\,(x) \;=\; (2\sqrt{2\pi})^{-1}\; \Gamma^2\!\left(\tfrac{1}{4}\right)\; nev_o(x)$$

$$\lim_{a\to 1} \; q_o\, f_2(x) \;=\; (2\sqrt{2\pi})^{-1}\; \Gamma^2\!\left(\tfrac{3}{4}\right)\; nev_2(x)$$

$$\lim_{a\to 1} \; f_{k+1}(x) \;=\; nev_1(x)$$

We shall prove, therefore, that

$$\lim_{a\to 1} \left(p_1\, f_2(x) \;+\; p_2\, f_3(x)\right) \;=\; -\sqrt{2}\,\left(1 - \gamma + \log\sqrt{2} \;+\; \tfrac{\pi}{4}\right)\, nev_1(x) \;+\; \sqrt{2}\, nel_1(x) \tag{5.2}$$

$$\lim_{a\to 1} \left(q_1\, f_2(x) \;+\; q_2\, f_3(x)\right) \;=\; \sqrt{2}\,\left(1 - \gamma + \log\sqrt{2} - \tfrac{\pi}{4}\right)\, nev_1(x) \;-\; \sqrt{2}\, nel_1(x) \tag{5.3}$$

The left-hand side of (5.2) becomes

$$\lim_{a\to 1} \left(p_1\, f_2(x) + p_2\, f_3(x)\right) \;=\; \sum_{n=0}^{\infty} \frac{(-1)^n}{4^{4n} n!}\, x^{4n+1} \lim_{\mu\to o} \frac{1}{\mu} \left(-\, C_1(n) + C_2(n)\right)$$

where

$$C_k(n) \;=\; 2^{\mp\frac{\mu}{2}}\, \Gamma(2\mp\mu)\, \frac{\Gamma\left(\frac{1}{2}(2\pm\mu)\right)\; \Gamma\left(\frac{1}{4}(3\pm\mu)\right)\; \Gamma\left(\frac{1}{4}(5\pm\mu)\right)}{\Gamma\left(n+\frac{1}{2}(2\pm\mu)\right)\; \Gamma\left(n+\frac{1}{4}(3\pm\mu)\right)\; \Gamma\left(n+\frac{1}{4}(5\pm\mu)\right)}\; x^{\pm\mu}\, \cos\frac{(3\mp\mu)\pi}{4}$$

Taking the limit, one finds that

$$\lim_{\mu\to 0}\frac{1}{\mu}\left(-C_1(n)+C_2(n)\right)=$$

$$-\sqrt{2}\left[n!\left(\tfrac{5}{4}\right)_n\left(\tfrac{3}{4}\right)_n\right]^{-1}\left\{1-\gamma-\log\frac{x}{\sqrt{2}}+\frac{\pi}{4}+\sum_{p=1}^{n}\left(\frac{1}{4p+1}+\frac{1}{2p}+\frac{1}{4p-1}\right)\right\}$$

where we have introduced the convention that the summation $\sum_{p=1}^{n}$ disappears when $n = 0$. This equation proves (5.2).

The left-hand side of (5.3) becomes

$$\lim_{a\to 1}\left(q_1 f_2(x)+q_2 f_3(x)\right)=\sum_{n=0}^{\infty}\frac{(-1)^n}{4^{4n}n!}x^{4n+1}\lim_{\mu\to 0}\frac{1}{\mu}\left(-S_1(n)+S_2(n)\right)$$

where $S_k(n)$ can be given from $C_k(n)$ by replacing $\cos\left((3\mp\mu)\pi/4\right)$ with $\sin\left((3\mp\mu)\pi/4\right)$. Taking the limit, one finds that

$$\lim_{\mu\to 0}\frac{1}{\mu}\left(-S_1(n)+S_2(n)\right)=$$

$$\sqrt{2}\left[n!\left(\tfrac{5}{4}\right)_n\left(\tfrac{3}{4}\right)_n\right]^{-1}\left\{1-\gamma-\log\frac{x}{\sqrt{2}}-\frac{\pi}{4}+\sum_{p=1}^{n}\left(\frac{1}{4p+1}+\frac{1}{2p}+\frac{1}{4p-1}\right)\right\}$$

where, by convention, the summation $\sum_{p=1}^{n}$ disappears when $n = 0$. This proves (5.3).

6. Fundamental Solutions for $a = 0$.

When $a = 0$, (1.5) reduces to

$$\left(\frac{d^2}{dx^2}+\frac{1}{x}\frac{d}{dx}\right)^2 w+w=0$$

as may be derived by putting $N_{rr} = 0$ and $N_{\theta\theta} = 0$ in (1.2). This equation can be decomposed in two equations,

$$\left(\frac{d^2}{dx^2}+\frac{1}{x}\frac{d}{dx}-i\right)w_1=0 \qquad (6.1)$$

and

$$\left(\frac{d^2}{dx^2} + \frac{1}{x}\frac{d}{dx} + i\right)w_2 = 0 \qquad (6.2)$$

The solutions of the two equations satisfying the boundary condition (1.6) at $x = \infty$ are

$$w_1 = \text{ker}x + i\ \text{kei}x \qquad (6.3)$$

and

$$w_2 = \text{ker}x - i\ \text{kei}x \qquad (6.4)$$

giving the fundamental solutions ker$x$ and ker$x$. Thus

$$w = A\ \text{ker}x + B\ \text{kei}x \qquad (6.5)$$

We shall prove that $w_1(x)$ in (4d.6) and $w_2(x)$ in (4d.7) satisfy

$$\lim_{a\to o} w_1(x) = \sqrt{2}\ \text{ker}x \qquad (6.6)$$

and

$$\lim_{a\to o} w_2(x) = \sqrt{2}\ \text{kei}x \qquad (6.7)$$

First we note that

$$\lim_{\mu\to 1} f_o(x) = \lim_{\mu\to 1} f_3(x) = \text{ber}x$$

and

$$\lim_{\mu\to 1} f_1(x) = \lim_{\mu\to 1} f_2(x) = 4\ \text{bei}x$$

Letting

$$\lambda = 1 - \mu$$

We transform (4d.6) to

$$w_1 = +\frac{1}{\mu(1+\mu)}\,2^{-\frac{\mu}{2}}\,\Gamma(2-\mu)\cdot\frac{1}{\lambda}\sin\frac{\lambda\pi}{4}\cdot f_2 +$$

$$+\frac{1}{\lambda}\left\{\frac{2}{\pi}\,\Gamma\left(\frac{2-\lambda}{4}\right)\Gamma\left(\frac{4+\lambda}{4}\right)\cdot f_o + \frac{1}{(1-\lambda)(2-\lambda)}\,2^{\frac{1-\lambda}{2}}\,\Gamma(3-\lambda)\cos\frac{\lambda\pi}{4}\cdot f_3\right\}$$

Letting $\lambda\to o$,

$$\lim_{\lambda\to o} w_1 = \frac{\pi}{2\sqrt{2}} \operatorname{bei} x + \sqrt{2}(\log 2-\alpha) \operatorname{ber} x + \sqrt{2} \lim_{\lambda\to o}\left(\frac{\partial f_o}{\partial\lambda} - \frac{\partial f_3}{\partial\lambda}\right)$$

Substituting from (2.4) and (2.6), one finds that

$$\lim_{\lambda\to o} \left(\frac{\partial f_o}{\partial\lambda} - \frac{\partial f_3}{\partial\lambda}\right) = -\log x \operatorname{ber} x + \sum_{n=1}^{\infty} \frac{(-1)^n x^{4n}}{4^{2n}\left((2n)!\right)^2} \sum_{p=1}^{2n} \frac{1}{p}$$

which proves (6.6).

We transform (4d.7) to

$$w_2 = \frac{1}{\mu(1+\mu)} 2^{\frac{\mu}{2}} \Gamma(2+\mu)\cdot \frac{1}{\lambda} \sin\frac{\lambda\pi}{4} \cdot f_3 +$$

$$+ \frac{1}{\lambda}\left\{ \frac{1}{\sqrt{2\pi}(2-\lambda)} \Gamma\left(\frac{2+\lambda}{4}\right) \Gamma\left(\frac{4-\lambda}{4}\right)\cdot f_1 - \frac{1}{(1-\lambda)(2-\lambda)} 2^{-\frac{1-\lambda}{2}} \Gamma(1+\lambda) \cos\frac{\lambda\pi}{4} \cdot f_2 \right\}$$

Letting $\lambda\to o$,

$$\lim_{\lambda\to o} w_2 = \frac{\pi}{2\sqrt{2}} \operatorname{ber} x - \sqrt{2}(1-\gamma+\log\sqrt{2}) \operatorname{bei} x + \frac{1}{2\sqrt{2}} \lim_{\lambda\to o} \left(\frac{\partial f_1}{\partial\lambda} - \frac{\partial f_2}{\partial\lambda}\right)$$

Substituting from (2.5) and (2.6), one finds that

$$\lim_{\lambda\to o} \left(\frac{\partial f_1}{\partial\lambda} - \frac{\partial f_2}{\partial\lambda}\right) = 4\log x \operatorname{bei} x - 4 \sum_{n=1}^{\infty} \frac{(-1)^n x^{4n+2}}{4^{2n+2}\left((2n+1)!\right)^2} \sum_{p=2}^{2n+1} \frac{1}{p}$$

which proves (6.7).

6a. Eigenvalues for $a = 0$.

When $a = 0$, no horizontal pressure works on the plate, and buckling should not take place under any boundary conditions. We shall prove below that this is true under the boundary conditions (1.7), (1.8), and (1.9).

The following formulas are needed for the proof. Substituting either (6.3) into (6.1) or (6.4) into (6.2), one finds the relations,

$$\ker''x + x^{-1}\ker'x + \mathrm{kei}x = 0 \tag{6a.1}$$

and

$$\mathrm{kei}''x + x^{-1}\mathrm{kei}'x - \ker x = 0 \tag{6a.2}$$

We shall prove that no positive number $x_o$ can satisfy the clamped-edge condition (1.7). The determinant of (1.7) is given by

$$D_1 = \begin{vmatrix} \ker x & \ker'x \\ \mathrm{kei}x & \mathrm{kei}'x \end{vmatrix}$$

when (6.5) is used. Differentiating $D_1$, one finds the differential equation

$$\frac{dD_1}{dx} + x^{-1}D_1 = \ker^2 x + \mathrm{kei}^2 x$$

Solving this equation under the boundary condition that $D_1 = 0$ at $x = \infty$, one finds

$$D_1 = -\frac{1}{x}\int_x^\infty \xi(\ker^2\xi + \ker^2\xi)\, d\xi$$

which is negative for any positive $x$, proving our contention.

We shall prove that no positive number $x_o$ can satisfy the simple-edge condition (1.8). The determinant of (1.8) transforms to

$$D_2 = \begin{vmatrix} \ker x & (1-\nu)x^{-1}\ker'x + \mathrm{kei}x \\ \mathrm{kei}x & (1-\nu)x^{-1}\mathrm{kei}'x - \ker x \end{vmatrix}$$

This equation transforms to

$$D_2 = -\frac{1-\nu}{x}\int_x^\infty \xi(\ker^2\xi + \mathrm{kei}^2\xi)\, d\xi - (\ker^2 x + \mathrm{kei}^2 x),$$

which is negative for any positive $x$, proving our contention.

We shall prove that no positive number $x_o$ can satisfy the free-edge condition (1.9). The determinant of (1.9) transforms to

$$D_3 = \begin{vmatrix} (1-\nu)x^{-1}\ \mathrm{ker}'x + \mathrm{kei}x & -\mathrm{kei}'x \\ (1-\nu)x^{-1}\ \mathrm{kei}'x - \mathrm{ker}x & \mathrm{ker}'x \end{vmatrix}$$

This equation transforms to

$$D_3 = \frac{1-\nu}{x}(\mathrm{ker}'^2x + \mathrm{kei}'^2x) + \frac{1}{x}\int_x^\infty \xi(\mathrm{ker}^2\xi + \mathrm{kei}^2\xi)\,d\xi$$

which is positive for any positive $x$, proving our contention.

7. Fundamental Solutions for $a > 1$.

For $a > 1$, $\mu$ defined in (2.1) must be replaced with $\mu = ik$, where

$$\kappa = \sqrt{a-1} \tag{7.1}$$

To compute $\Gamma(x + iy)$, we use the formulas

$$\left|\Gamma(x+iy)/\Gamma(x)\right|^2 = \prod_{n=0}^{\infty}\left[1 + y^2/(x+n)^2\right]^{-1} \tag{7.2}$$

and

$$\mathrm{Arg}\ \Gamma(x+iy) = y\psi(x) + \sum_{n=0}^{\infty}\left\{y/(x+n) - \tan^{-1}[y/(x+n)]\right\} \tag{7.3}$$

[Handbook (ref. 5), p. 256]. These formulas can be proved by use of Euler's formula for the Gamma function (ref. (9), p. 237).

Using these formulas, coefficients of $F(x)$ in (4d.5) become

$$B_o = \frac{1}{2}(2\pi)^{-\frac{1}{2}}\,\Gamma^2(\tfrac{1}{4})\prod_{p=0}^{\infty}\left[1 + \kappa^2(4p+1)^{-2}\right]^{-1} \tag{7.4}$$

$$B_2 = \frac{i}{a}(2\pi)^{-\frac{1}{2}}\,\Gamma^2(\tfrac{3}{4})\prod_{p=0}^{\infty}\left[1 + \kappa^2(4p+3)^{-2}\right]^{-1} \tag{7.5}$$

$$B_\mu^{(k)} = \mp(a\kappa)^{-1}\ R\ \exp(\pm\tfrac{\kappa\pi}{4} + \tfrac{\pi i}{4} \mp i\kappa\log\sqrt{2} \mp i\Theta) \tag{7.6}$$

where $R$ and $\Theta$ are defined by

$$\Gamma(2\mp i\kappa) = R\exp(\mp i\Theta)$$

They are given by

$$R = \prod_{n=0}^{\infty} (n+2)\left[(n+2)^2 + \kappa^2\right]^{-\frac{1}{2}}$$

and

$$\Theta = \kappa(1-\gamma) + \sum_{n=0}^{\infty}\left\{[\kappa/(n+2)] - \tan^{-1}[\kappa/(n+2)]\right\}$$

Functions $f_o(x)$ and $f_1(x)$ are real. To decompose the complex function $f_{k+1}(x)$ into the real and imaginary parts, the denominator of (2.3),

$$(4n+1\pm\mu)(4n-1+\mu)(4n)(4n\pm 2\mu)$$

is transformed to

$$= 8n[2n(16n^2 + 4 - 5a) \pm i\kappa(32n^2 - a)]$$

Thus one finds

$$f_{k+1}(x) = \sum_{n=0}^{\infty} (-1)^n P_n^{(o)} \exp(\mp i\rho_n)\, x^{4n+1\pm i\kappa}$$

where

$$P_o^{(o)} = 1$$

$$\rho_o = 0$$

$$P_n^{(o)} = 8^{-2n}(n!)^{-1}\left[\prod_{p=1}^{n}\left\{p^2(4p^2 + 1 - \frac{5}{4}a)^2 + (a-1)(4p^2 - \frac{a}{8})^2\right\}\right]^{-\frac{1}{2}}$$

$$\rho_n = \sum_{p=1}^{n}\tan^{-1}\left[\kappa(4p^2 - \frac{a}{8})\Big/\left\{p(4p^2 + 1 - \frac{5}{4}a)\right\}\right]$$

for $n \geq 1$.

Fundamental solutions $w_1(x)$, $w_2(x)$, and their derivatives are found by decomposing $F(x)$ and its derivatives into real and imaginary parts. We formulated them (up to the third derivatives) as follows:

$$\frac{d^m w_1}{dx^m} = \frac{\Gamma^2 \frac{1}{4}}{2\sqrt{2\pi}} \prod_{p=0}^{\infty} \left[1 + \frac{a-1}{(4p+1)^2}\right]^{-1} \frac{d^m f_o}{dx^m} -$$

$$- \frac{R \exp(\kappa\pi/4)}{a\kappa} \sum_{n=0}^{\infty} (-1)^n P_n^{(m)} x^{4n+1-m} \cos(\theta_1 - \rho_n + \phi_n^{(m)})$$

$$+ \frac{R \exp(-\kappa\pi/4)}{a\kappa} \sum_{n=0}^{\infty} (-1)^n P_n^{m} x^{4n+1-m} \cos(\theta_2 + \rho_n - \phi_n^{(m)}) \qquad (7.7)$$

$$\frac{d^m w_2}{dx^m} = \frac{\Gamma^2 \frac{3}{4}}{\sqrt{2\pi a}} \prod_{p=0}^{\infty} 1 + \frac{a-1}{(4p+3)^2}^{-1} \frac{d^m f_1}{dx^m} -$$

$$- \frac{R \exp(\kappa\tau/4)}{a\kappa} \sum_{n=0}^{\infty} (-1)^n P_n^{(m)} x^{4n+1-m} \sin(\theta_1 - \rho_n + \phi_n^{(m)})$$

$$+ \frac{R \exp(\kappa\pi/4)}{a\kappa} \sum_{n=0}^{\infty} (-1)^n P_n^{(m)} x^{4n+1-m} \sin(\theta_2 + \rho_n - \phi_n^{(m)}) \qquad (7.8)$$

where

$$\theta_k = \frac{\pi}{4} \pm \kappa \log\frac{x}{\sqrt{2}} \mp (1-\gamma)\kappa \mp \sum_{p=0}^{\infty} \left(\frac{\kappa}{p+2} - \tan^{-1}\frac{\kappa}{p+2}\right)$$

$$P_n^{(1)} = P_n^{(o)} \left[(4n+1)^2 + \kappa^2\right]^{-\frac{1}{2}}$$

$$P_n^{(2)} = P_n^{(1)} \left[(4n)^2 + \kappa^2\right]^{-\frac{1}{2}}$$

$$P_n^{(3)} = P_n^{(2)} \left[(4n-1)^2 + \kappa^2\right]^{-\frac{1}{2}}$$

$$\phi_n^{(0)} = 0$$

$$\phi_n^{(1)} = \tan^{-1}\left[\kappa/(4n+1)\right]$$

$$\begin{cases} \phi_o^{(2)} = \phi_o^{(1)} + \dfrac{\pi}{2} \\ \phi_n^{(2)} = \phi_n^{(1)} + \tan^{-1}[\kappa/(4n)] \qquad \text{for } n \geq 1 \end{cases}$$

$$\begin{cases} \phi_o^{(3)} = \phi_o^{(2)} + \pi - \tan^{-1}\kappa \\ \phi_n^{(3)} = \phi_n^{(2)} + \tan^{-1}[\kappa/(4n-1)] \qquad \text{for } n \geq 1 \end{cases}$$

## PART II. ASYMPTOTIC EXPANSIONS

Values of a series solution developed in PART I must overlap on a certain range of $x$ with the values of an asymptotic expansion determined corresponding to the respective series solution. The series may be used for any $x$ less than the overlapping range, and the asymptotic expansion may be used for any $x$ larger than the overlapping range.

8. Asymptotic Expansion for $0 < a \leq 1$.

Using analytical continuation of the hypergeometric function in (3.8) from the range $1 < r < \infty$ into the neighborhood of $r = 1$ (more exactly in the range $|1-r^4| < 1$), one finds that $v_k(r)$ in the contour integral solution (3.11) defined in the range $1 < r < \infty$ is analytically continued to

$$v_k(r) = v_1^{(k)} \; F\left(\tfrac{1}{4}(2+\mu), \tfrac{1}{4}(2-\mu); \tfrac{3}{2}; 1-r^4\right) + \\ + v_2^{(k)} \; r^2(r^4-1)^{-\frac{1}{2}} \; F\left(\tfrac{1}{4}(2+\mu), \tfrac{1}{4}(2-\mu); \tfrac{1}{2}; 1-r^4\right) \qquad (8.1)$$

defined in the neighborhood of $r = 1$, where

$$v_1^{(k)} = -2\sqrt{\pi}\,\Gamma\left(\tfrac{1}{2}(2\pm\mu)\right)\left[\Gamma\left(\tfrac{1}{4}(2\pm\mu)\right)\Gamma\left(\pm\tfrac{1}{4}\mu\right)\right]^{-1} \qquad (8.2)$$

$$v_2^{(k)} = \frac{\sqrt{\pi}}{2}\,\Gamma\left(\tfrac{1}{2}(2\pm\mu)\right)\left[\Gamma\left(\tfrac{1}{4}(2\pm\mu)\right)\Gamma\left(\tfrac{1}{4}(4\pm\mu)\right)\right]^{-1} \qquad (8.3)$$

Double signs may not appear in the hypergeometric functions on the right-hand side of (8.1) because of their symmetric properties with regard to the first and second parameters.

Letting $r = 1 + t$ and developing the hypergeometric functions on the right-hand side of (8.1) into power series, one can integrate (8.1) to a complex-form asymptotic expansion for $0 \leq a \leq 1$,

$$F(x) \sim$$

$$(v_1^{(1)} + v_1^{(2)})\, e^{-x/\sqrt{2}} \left\{ \frac{\pi}{x} \exp\left(\frac{ix}{\sqrt{2}} + \frac{\pi i}{8}\right) + \frac{P_1}{2} \frac{\pi}{x^3} \exp\left(\frac{ix}{\sqrt{2}} + \frac{3\pi i}{8}\right) + \frac{3P_2}{4} \frac{\pi}{x^5} \exp\left(\frac{ix}{\sqrt{2}} + \frac{5\pi i}{8}\right) + \ldots \right\}$$

$$+ (v_2^{(1)} + v_2^{(2)})\, e^{-x/\sqrt{2}} \left\{ \frac{1}{x} \exp\left(\frac{ix}{\sqrt{2}} + \frac{\pi i}{8}\right) + \frac{B_1}{x^2} \exp\left(\frac{ix}{\sqrt{2}} + \frac{\pi i}{2}\right) + \frac{2B_2}{x^3} \exp\left(\frac{ix}{\sqrt{2}} + \frac{3\pi i}{4}\right) + \ldots \right\} \tag{8.4}$$

where

$$P_1 = -(1 + 2a)/4$$

$$P_2 = (9 + 20a + 4a^2)/96$$

$$B_1 = -(3 + a)/6$$

$$B_2 = -(3 + a)(5 + a)/120$$

Asymptotic expansions for $w_1(x)$ and $w_2(x)$ are given by the real and imaginary parts of (8.4), respectively.

9. Asymptotic Expansion for $1 \leq a \leq 2$.

A form of asymptotic expansion for $a \geq 1$ is found by letting $\mu = i\kappa$ in the coefficients of $v_k^{(1)} + v_k^{(2)}$ $(k = 1,2)$. In this case formulas (7.2) and (7.3) need to be modified to include the case $x = 0$. The modified formulas are:

$$\left|\Gamma(iy)\right|^2 = y^{-2} \prod_{n=1}^{\infty} [1 + (y/n)^2]^{-1} \tag{9.1}$$

and

$$\text{Arg } \Gamma(iy) = -\frac{\pi}{2}\,\text{sign}(y) - y\gamma + \sum_{n=1}^{\infty}\left(\frac{y}{n} - \tan^{-1}\frac{y}{n}\right), \qquad (9.2)$$

where

$$\text{sign}(y) = 1 \quad \text{for } y > 0$$

$$= -1 \quad \text{for } y < 0$$

The result of the transformation becomes extremely simple:

$$v_1^{(1)} + v_1^{(2)} = \cos(\kappa \log 2) \qquad (9.3)$$

and

$$v_0^{(1)} + v_0^{(2)} = \kappa \sin[\kappa(\tfrac{1}{4}\gamma + \log\sqrt{2})] \qquad (9.4)$$

Substituting these into (8.4) the complex form asymptotic expansion for $a \geq 1$ is found.

Our numerical computation shows that this asymptotic expansion is effective only for $a$ close to 1. We used this formula for $2 \geq a \geq 1$.

10. Asymtotic expansion for $a \geq 2$.

Letting $\mu = i\kappa$, the integral solution (3.11) transforms to

$$\frac{1}{2}\,F(x) = \int_1^{\infty} e^{\beta x r} \cos\left[\frac{\kappa}{2}\log\left(r^2 + \sqrt{r^4-1}\right)\right] (r^4-1)^{-\frac{1}{2}}\, dx \qquad (10.1)$$

Expanding the integrand in the neighborhood of $r = 1$ by letting $r = 1 + t$, and using the approximations,

$$\log(r^2 + \sqrt{r^4-1}) = 2\sqrt{t} + O(t^{3/2})$$

and

$$\sqrt{r^4-1} = 2\sqrt{t} + O(t)$$

one finds the integral asymptotic solution,

$$F(x) \sim \int_0^\infty e^{\beta x(1+t)} \cos(\kappa\sqrt{t}) \frac{dt}{\sqrt{t}} \tag{10.2}$$

To evaluate this integral, define the function,

$$G_k(x) = \frac{1}{2} \int_0^\infty e^{\beta x t \mp i\kappa\sqrt{t}} \frac{dt}{\sqrt{t}} \tag{10.3}$$

Then (10.2) becomes

$$F(x) \sim e^{\beta x} \left( G_1(x) + G_2(x) \right) \tag{10.4}$$

Letting $t = \xi^2$, (10.3) transforms to

$$G_k(x) = \int_0^\infty \exp\left[ \beta x (\mu \mp \frac{\beta\kappa}{2x})^2 + \frac{\kappa^2}{4\beta x} \right] d\xi$$

Define $z$ by

$$\beta x \, (\xi \mp \frac{\beta\kappa}{2x})^2 = -z^2$$

The root $z$ of this equation satisfying the condition that the real part of $z$ must approach positive infinity as $\xi \to \infty$ is

$$z = \exp(-\frac{\pi i}{8}) \, x^{\frac{1}{2}} \, \xi \mp \exp(\frac{5\pi i}{8}) \, \frac{1}{2}\kappa x^{-\frac{1}{2}}$$

Use of $z$ thus defined transforms (10.3) to

$$G_k(x) = x^{-\frac{1}{2}} \exp(\frac{\kappa^2}{4\beta x} + \frac{\pi i}{8}) \int_{\alpha_k}^{\infty \exp(5\pi i/8)} \exp(-z^2) \, dz$$

where

$$\alpha_k = \mp \exp(\frac{5}{8}\pi i) \, \frac{1}{2}\kappa x^{-\frac{1}{2}}$$

Transforming the contour of integration to the sum of two contours, $\alpha_k \sim o$ and $o \sim +\infty$, one finds

$$G_1(x) + G_2(x) = \sqrt{\pi/x} \exp\left( \kappa^2/(8\beta x) + \pi i/8 \right) \tag{10.5}$$

Thus one gets

$$F(x) \sim \sqrt{\pi/x}\, \exp(\beta x + \kappa^2/(8\beta x) + \pi i/8) \tag{10.6}$$

We use this equation for $a \geq 2$.

11. Fundamental Solutions for Large $a$ and Small $x$.

Our numerical computation shows that the overlapping range of the series solution and the asymptotic expansion moves to small values of $x$ as the values of $a$ increases. When $a = 2$, the series solution and the asymptotic expansion overlap in the neighborhood of $x = 6$. When $a = 6$, they overlap in the neighborhood of $x = 1$, showing that the fundamental solutions at this value of $a$ is ineffective. For larger $a$, fundamental solutions must be transformed to a more effective form.

Following formulas were used for the transformation. For large values of $y$

$$\Gamma(x \pm iy) \sim \sqrt{2\pi}\; e^{-\frac{1}{2}\pi y}\; y^{x-\frac{1}{2}}\; \left(\frac{y}{e}\right)^{\pm iy}\; e^{\pm \frac{1}{2}\pi i(x-\frac{1}{2})} \tag{11.1}$$

and

$$\Gamma(x \pm iy) \left[\Gamma(n+x \pm iy)\right]^{-1} \sim e^{\mp \frac{1}{2}n\pi i}\; y^{-n} \tag{11.2}$$

where $x$ and $y$ are real. These formulas can be derived by transforming the asymptotic expansions of the Gamma-functions by using the assumption that $y$ is large.

When $x$ is small, the number of terms needed for the summation of series $f_m(x)$ $(m = 0, 1, 2, 3)$ in (2.4) $\sim$ (2.6) are fairly small. Letting $\kappa$ be large under this condition, formulas (11.1) and (11.2) may be applied to transform series $f_m(x)$. Thus one finds

$$f_0(x) \sim \cos[(2\kappa)^{-1}x^2] \tag{11.3}$$

$$f_1(x) \sim 2\kappa \sin[(2\kappa)^{-1}x^2] \tag{11.4}$$

and

$$f_{k+1}(x) \quad x^{1\mp i\kappa} \exp[\mp i(4\kappa^3)^{-1}x^4] \tag{11.5}$$

Also one finds

$$\Gamma\left(\tfrac{1}{4}(1+i\kappa)\right)\ \Gamma\left(\tfrac{1}{4}(1-i\kappa)\right) \sim 4\pi\,\kappa^{-\frac{1}{2}}\,e^{-\frac{1}{4}\kappa\pi} \tag{11.6}$$

$$\Gamma\left(\tfrac{1}{4}(3+i\kappa)\right)\ \Gamma\left(\tfrac{1}{4}(3-i\kappa)\right) \sim \pi\,\kappa^{\frac{1}{2}}\,e^{-\frac{1}{4}\kappa\pi} \tag{11.7}$$

and

$$\Gamma(2\mp i\kappa) \sim (2\pi\kappa^3)^{\frac{1}{2}}\,e^{-\frac{1}{2}\kappa\pi}\,(\kappa/e)^{\mp i\kappa}\,e^{\pm\frac{3}{4}\pi i} \tag{11.8}$$

Thus for extremely large $\kappa$ and small $x$, one finds the complex expression,

$$F(x) \sim A(1+i(2\kappa)^{-1}x^2) + Bx^{1+i\kappa} + Cx^{1-i\kappa}, \tag{11.9}$$

where

$$A = (2\pi/\kappa)^{-\frac{1}{2}}\,e^{-\frac{1}{4}\kappa\pi} \tag{11.10}$$

$$B = i2^{-\frac{1}{2}i\kappa}\,(2\pi)^{\frac{1}{2}}\,\kappa^{-\frac{3}{2}}\,e^{-\frac{1}{4}\pi\kappa+i\kappa}\,\kappa^{-i\kappa} \tag{11.11}$$

$$C = -\,2^{\frac{1}{2}i\kappa}\,(2\pi)^{\frac{1}{2}}\,\kappa^{-\frac{3}{2}}\,e^{-\frac{3}{4}\pi\kappa-i\kappa}\,\kappa^{i\kappa}. \tag{11.12}$$

## PART III EIGENVALUES

### 12. Computation of $x_o$

Our numerical computation shows that the clamped-edge condition (1.7) and the simple-edge condition (1.8) do not yield any positive number $x_o$ as a root of the respective determinent equations. The free-edge condition (1.9) always yields roots or a root. We shall discuss below only the free-edge condition.

Define operators

$$L = \frac{d^2}{dx^2} + \frac{\nu}{x}\frac{d}{dx} \tag{12.1}$$

and

$$M = \frac{d^3}{dx^3} + \frac{1}{x}\frac{d^2}{dx^2} - \frac{1-a}{x^2}\frac{d}{dx} \tag{12.2}$$

Then the determinant $D_3$ found by substituting (1.14) into (1.9) is given by

$$D_3 = \begin{vmatrix} L(w_1) & M(w_1) \\ L(w_2) & M(w_2) \end{vmatrix} \tag{12.3}$$

Root $x_o$ thus found in the range $0 < a \leq 2$ are shown in Figure 2 and 3.

To discuss the neighborhood of $x_o = 0$ in these figures, take the first term of the series $f_m(x)$ $(m = 0, 1, 2, 3)$, and approximate $w_1(x)$ and $w_2(x)$ in (4d.6) and (4d.7) with

$$w_1(x) = p_o + p_1 x^{1+\mu} + p_2 x^{1-\mu} + 0(x^2) \tag{12.4}$$

and

$$w_2(x) = q_o x^2 + q_1 x^{1+\mu} + q_2 x^{1-\mu} + 0(x^{5-\mu}) \tag{12.5}$$

Because $M(x^{1+\mu}) = 0$, $M(w_1)$ is negligible against $M(w_2)$. Therefore the root of (12.3) is given by $L(w_1) = 0$, which yields

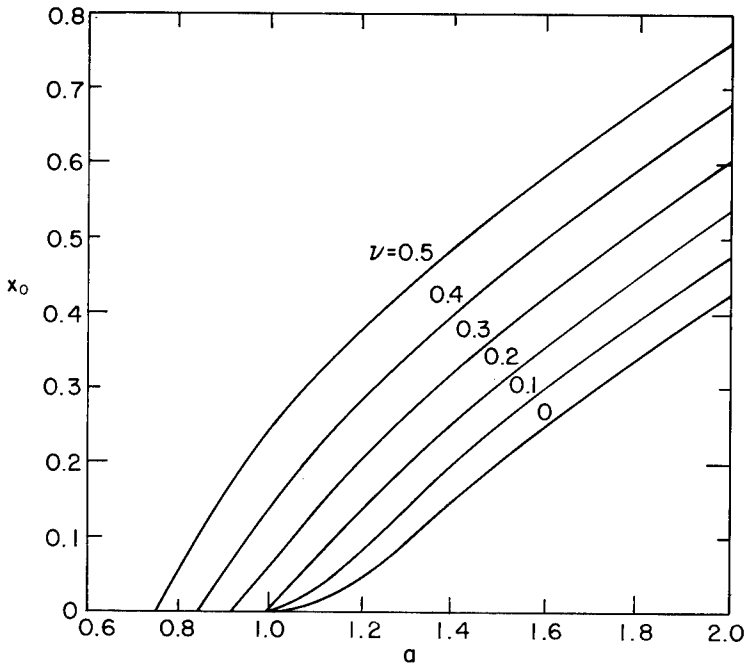



Fig. 2. Values of $x_o$ in the range $0 < a \leq 2$ using $\nu$ (Poisson's ratio) as parameters.

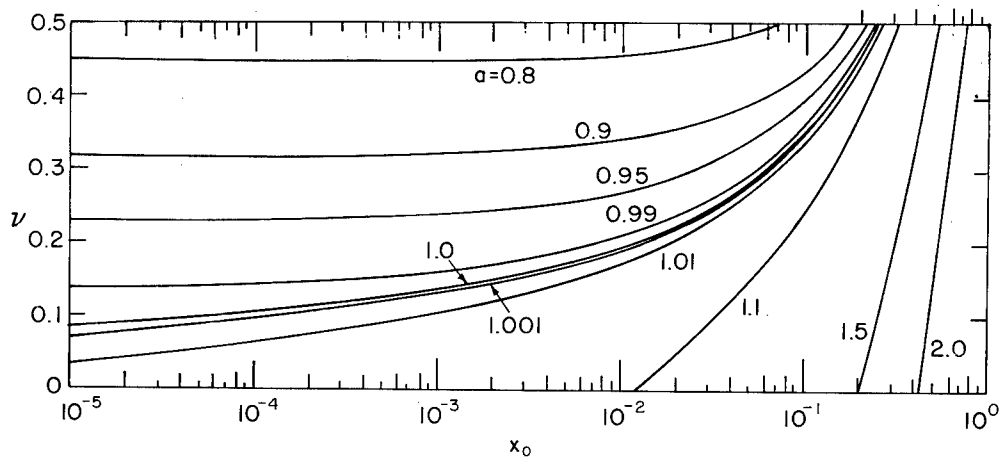



Fig. 3. Values of $x_o$ expressed with logarithmic scale in the range $0 < a \leq 2$ using $a$ as parameter.

$$\left(\frac{1}{\sqrt{2}}\,x_o\right)^{2\mu} = \frac{(\nu-\mu)\ \Gamma(1+\mu)\ \cos[(3+\mu)\pi/4]}{(\nu+\mu)\ \Gamma(1-\mu)\ \cos[(3-\mu)\pi/4]} \tag{12.6}$$

This equation shows that the condition $\mu \leq \nu$, i.e. $a \geq -\nu^2$, must be met. When $\mu = \nu$, $x_o$ becomes equal to zero. Each curve in Figure 2, therefore, terminates at the intersection with the axis of abscissa, whose coordinate is $a = 1 - \nu^2$.

For small $\mu$, (12.6) becomes

$$\ln\left(\frac{1}{\sqrt{2}}\,x_o\right) = \frac{\pi}{4} - \gamma - \frac{1}{\nu} + \frac{1}{3}\ \mu^2\left(-\zeta(3) - \nu^{-3} + \frac{1}{32}\ \pi^3\right) \tag{12.7}$$

where $\zeta(3)$ is Rieman's Zeta-function. Our numerical computation shows that (12.7) gives close approximation over the entire lengths of the curves in the neighborhood of $a = 1\text{-}0$ in Figure 3.

To discuss the neighborhood of $x_o = 0$ for the case $1 \leq a \leq 2$, we used the complex form $F(x)$ in (4d.5) with coefficients given by (7.4) ~ (7.6). Taking the first terms of $f_m(x)$, one finds that

$$M(F) = 2B_2\ (1+\kappa^2)x^{-1}$$

Because $B_2$ is a pure imaginary, the real part of M(F), i.e. $M(\omega_1)$, is negligible against the imaginary part of M(F), i.e. $M(\omega_2)$. Therefore $D_3 = 0$ is equivalent to $L(\omega_1) = 0$. Equating the real part of L(F) equal to zero, one finds that $x_o$ is approximated by the root of

$$\tan(\alpha+\kappa\ln x) = \frac{[(\nu-\kappa^2)\ \exp(\frac{1}{4}\kappa\pi) - \kappa(1+\nu)\ \exp(-\frac{1}{4}\kappa\pi)] \cdot [\kappa(1+\nu)\ \exp(\frac{1}{4}\kappa\pi) + (\nu-\kappa^2)\ \exp(-\frac{1}{4}\kappa\pi)]}{} \tag{12.8}$$

where

$$\alpha = \frac{\pi}{4} - \kappa\log\sqrt{2} - (1-\alpha)\kappa - \sum_{n=0}^{\infty}{}'\left(\frac{\kappa}{n+2} - \tan^{-1}\frac{\kappa}{n+2}\right) \tag{12.9}$$

For small $\kappa$ (12.8) reduces to

$$\nu H(x) + 1 + \nu - \frac{\nu\pi}{4} =$$

$$= \kappa^2 \left[\frac{\pi}{4}\left\{1 + \frac{(1+\nu)\pi}{8} - \frac{\nu\pi^2}{96}\right\} - \left\{1 - \frac{(1+\nu)\pi}{4} + \frac{\nu\pi^2}{32}\right\} H(x)\right] \qquad (12.10)$$

where

$$H(x) = \log\left[(\sqrt{2})^{-1} x\right] - 1 + \gamma \qquad (12.11)$$

Our numerical computation shows that this equation gives close approximation over the entire lengths of the curves in the neighborhood of $a = 1 + 0$ in Figure 3.

Equation (12.8) shows that, if $x_n$ is a root, then $x_{n+1}$ given by

$$x_{n+1} = x_n \exp(-\pi/\kappa) \qquad (12.12)$$

is also a root. Therefore infinitely many roots exist in the neighborhood of $x = 0$. Roots $x_0$, $x_1$, $x_2$ and $x_3$ are shown in Figure 4 where $\kappa = \sqrt{a-1}$ is used as the ordinate. The solid line covers the values we actually computed. They may be extended to the left of the solid lines by means of (12.10) and (12.12).

The asymptotic behavior of the large roots can be found by using $F(x)$ in (10.6) to compute L(F) and M(F). Assuming that $\zeta$ defined by

$$\zeta = \kappa^2/(4x^2) \qquad (12.13)$$

is of the ordinary magnitude for large $x$, one finds

$$L(F) \sim F(x)\ (\beta^2 - 2\zeta + \beta^{-2}\zeta^2) \qquad (12.14)$$

and

$$M(F) \sim F(x)\ (1+\zeta^2)\ [\exp(\tfrac{\pi i}{4}) - \zeta \exp(-\tfrac{\pi i}{4})] \qquad (12.15)$$

Thus one discovers that there are two asymptotic roots,

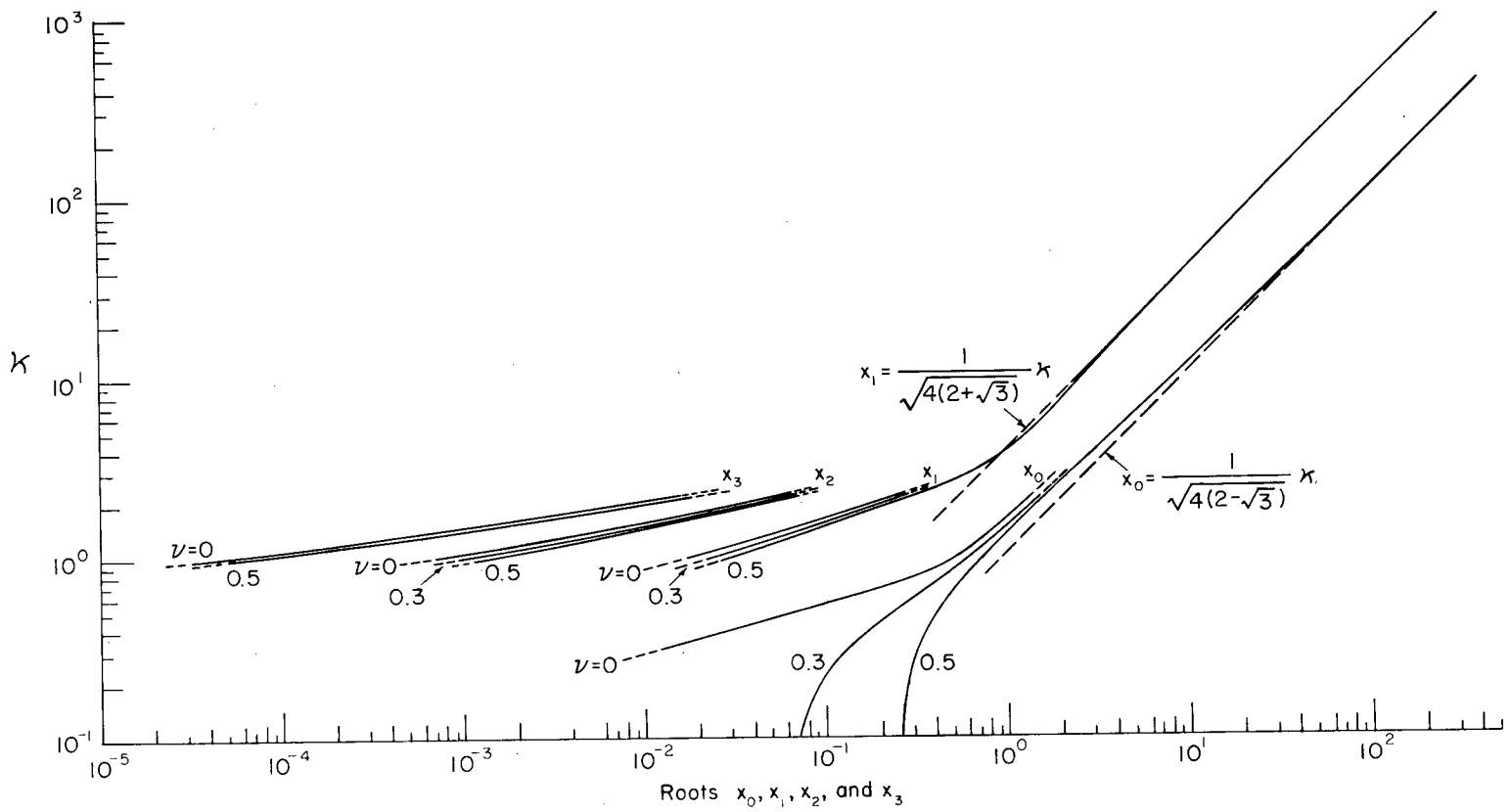



Fig. 4. Values of roots $x_0$, $x_1$, $x_2$, and $x_3$ as functions of $\kappa = \sqrt{a-1}$ in the range $0 \leq \kappa < \infty$ using $\nu$ (Poisson's ratio) as parameter.

$$x_h = (4\zeta_h)^{-\frac{1}{2}} \kappa \tag{12.16}$$

where

$$\zeta_h = 2 \mp \sqrt{3} \tag{12.17}$$

In these two equations, suffix $h$ is defined by

$$h = k - 1, \tag{12.18}$$

where the old convention for suffix $k$ is still observed. The two lines in Figure 4 expressing the two equations in (12.16) are shown by the broken lines.

Asymptotic roots were also computed retaining all the terms in L(F) and M(F) that were found by letting F($x$) be (10.6). Carrying out the computation of $D_3$ as given by (12.3), one finds that the equation $D_3 = 0$ reduces to

$$\sum_{n=0}^{4} N_n x^{-n} = 0 \tag{12.19}$$

where

$$N_o = (1+\zeta)(1+\zeta^2)(1-4\zeta+\zeta^2) \tag{12.20}$$

$$N_1 = \sqrt{2}\,[(1-\nu) + 4\zeta - 2\zeta^2 + 8\zeta^3 - (3-\nu)\zeta^4]$$

$$N_2 = -\frac{1}{2} + 5(2-\nu)\zeta - (6-\nu)\zeta^2 + (12-7\nu)\zeta^3$$

$$N_3 = \frac{1}{4}\sqrt{2}\,[-(3+\nu) + 2(9-14\nu)\zeta - (15-13\nu)\zeta^2]$$

$$N_4 = \frac{1}{16}\,[(-3+8\nu) + 3\,(15-16\nu)\zeta]$$

The positive roots of $N_o$ are $\zeta_o$ and $\zeta_1$ in (12.17). The solid lines running close to the broken lines in Figure 4 cover the values of

$x_o$ and $x_1$ computed for $\nu = 0.5$ by use of (12.19). The values of $x_o$ and $x_1$ in the range $a \leq 2$ were computed by using the series (7.7) and (7.8).

The asymptotic behavior of the small roots can be found by using (11.9) to compute L(F) and M(F). One finds that

$$M(F) = i A (\kappa x)^{-1} (1+\kappa^2) \tag{12.21}$$

Because $A$ is real, $M(w_1)$ is negligible against $M(w_2)$, and $D_3 = 0$ is equivalent to $L(w_1) = 0$. For large $\kappa$, this yields

$$x = \sqrt{2}\, e^{-1}\kappa$$

which, however, is not small. Therefore small roots do not accumulate at point $x = 0$, when $\kappa$ is large.

This conclusion does not yet exclude the possible existence of roots that are too small to be found with the asymptotic expansion (10.6) but too large to be found with the approximation (11.9). It is probably true, however, that roots $x_n$ $(n \geq 2)$ become equal to zero at certain values of $\kappa$ and do not extend indefinitely to large values of the ordinate.

Extension of the curves expressing $x_n$ $(n \geq 2)$ beyond the ordinate $\kappa \geq \sqrt{3}$ was not attempted. Our interest was originally in small $a$, and moreover we did not have enough time to have series improve for the case $a > 1$. However, we believe that small roots are not important for engineering purposes and need not be known in detail.

13. Deformation

Forms of deformation corresponding to the roots $x_n$ were calculated in the range $1-\nu^2 \leq a \leq 2$ by assuming the normalization,

$$w(x_n) = 1 \tag{13.1}$$

Two cases, ($a = 1$, $\nu = 0.3$) and ($a = 4$, $\nu = 0.3$), are shown in Figure 5 and 6, respectively.

These forms of deformation have often been observed in laboratories and fields when floating ice plates are compressed. We are now convinced that buckling is frequently taking place.

Forms of deformation other than shown in Figure 5 and 6 can be guessed by use of Figure 7 and 8, where the values of $w_{min}$ (minimum depression) and $x_{min}$ (defined by $w_{min} = w(x_{min})$) determined for $x_o$ are shown. (See Figure 5 for the definition of $x_{min}$ on a curve of deformation). Values $a$ in these figures are restricted to $1-\nu^2 \leqq a \leqq 2$. We did not compute them for the case $a > 2$, nor for $x_n$ ($n \geqq 1$) except for the cases shown in Figure 6. The broken lines in these figures are determined by the terminal condition $\nu = \mu$.

The deformation at fracture shall be determined by assuming that the stress at $x_{min}$ reaches the fracture stress $\sigma_f$. In the general polar coordinates, stress components $\sigma_{rr}$, $\sigma_{\theta\theta}$, $\sigma_{r\theta}$ are given (see Appendix C) by

$$\sigma_{rr} = -\frac{Eh}{2(1-\nu)}\left[\frac{\partial^2 w}{\partial r^2} + \frac{\nu}{r}\frac{\partial w}{\partial r} + \frac{\nu}{r^2}\frac{\partial^2 w}{\partial \theta^2}\right]$$

$$\sigma_{\theta\theta} = -\frac{Eh}{2(1-\nu)}\left[\frac{1}{r}\frac{\partial w}{\partial r} + \frac{1}{r^2}\frac{\partial^2 w}{\partial \theta^2} + \nu\frac{\partial^2 w}{\partial r^2}\right] \tag{13.2}$$

and

$$\sigma_{r\theta} = -\frac{Eh}{2(1+\nu)}\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial w}{\partial \theta}\right)$$

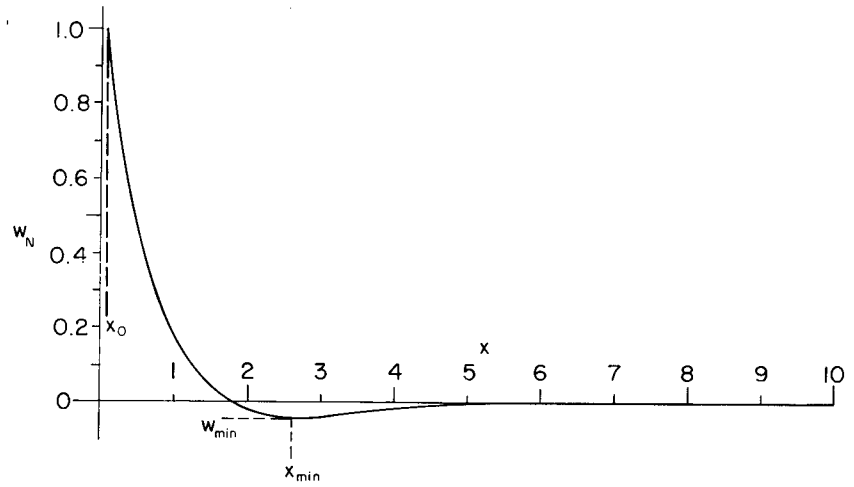



Fig. 5. Normalized deformation for the case ($a$ = 1, $\nu$ = 0.3).

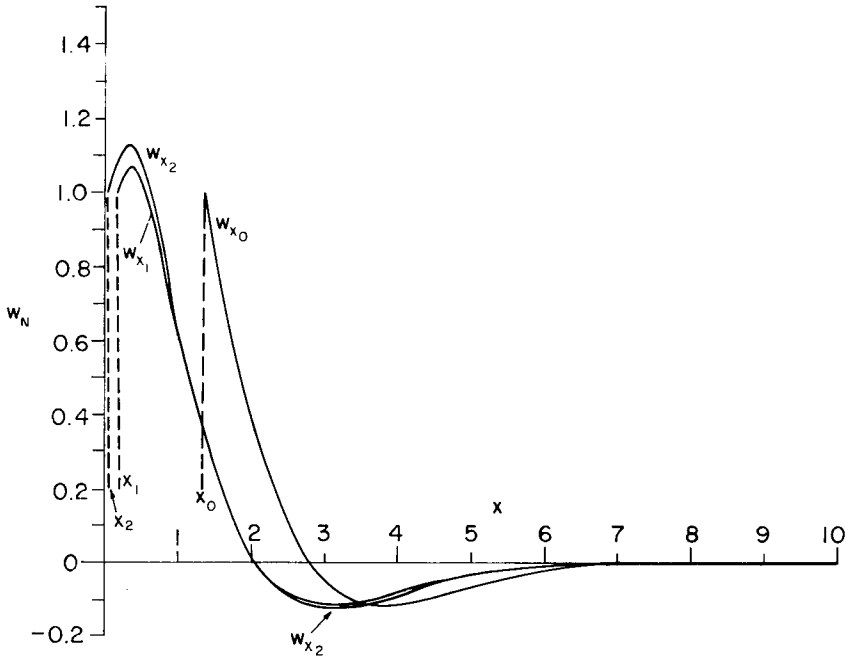



Fig. 6. Normalized deformation for the case ($a$ = 4, $\nu$ = 0.5).

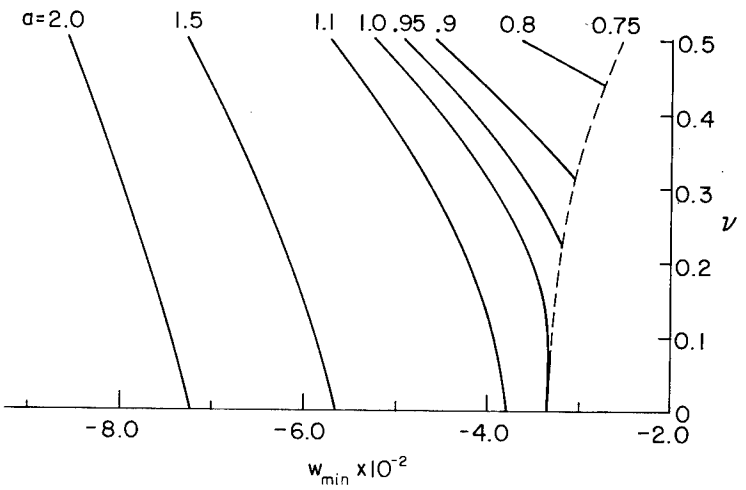



Fig. 7. Values of $w_{min}$ determined corresponding to $x_o$.

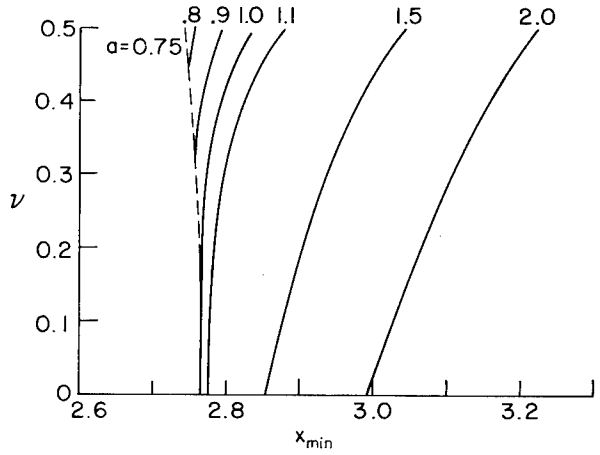



Fig. 8. Values of $w_{min}$ determined corresponding to $x_o$.

where $h$ is the thickness of the plate. In our case of axisymmetry, introducing the nondimensional length $x$ defined by (1.4), the above formulas become

$$\sigma_{rr} = -\frac{Eh}{2(1-\nu)\ell_o^2}\left[\frac{d^2w}{dx^2} + \frac{\nu}{x}\frac{dw}{dx}\right] \tag{13.3}$$

$$\sigma_{\theta\theta} = -\frac{Eh}{2(1-\nu)\ell_o^2}\left[\frac{1}{x}\frac{dw}{dx} + \nu\frac{d^2w}{dx^2}\right]$$

and

$$\sigma_{r\theta} = 0$$

At point $x_{min}$, where $dw/dx = 0$, therefore,

$$|\sigma_{rr}| > |\sigma_{\theta\theta}|$$

Let $w_N(x)$ be the normalization of $w(x)$ at $x = x_o$. Then the depression is given by

$$w(x) = K\, w_N(x) \tag{13.4}$$

where $K$ shall be determined by applying the condition that

$$|\sigma_{rr}| = \sigma_f \quad \text{at} \quad x = x_{min} \tag{13.5}$$

where $\sigma_f$ is the fracture strength. Summing up the above results, $K$ is found:

$$K = 2\ell_o^2\, \sigma_f\, (Eh)^{-1}\, H(\nu,a) \tag{13.6}$$

where

$$H(\nu,a) = (1-\nu^2)/\left(\frac{d^2w}{dx^2}\right)_{x_{min}} \tag{13.7}$$

Values of $H(\nu,a)$ are shown in Figure 9 for the case $1-\nu^2 \leq a \leq 2$.

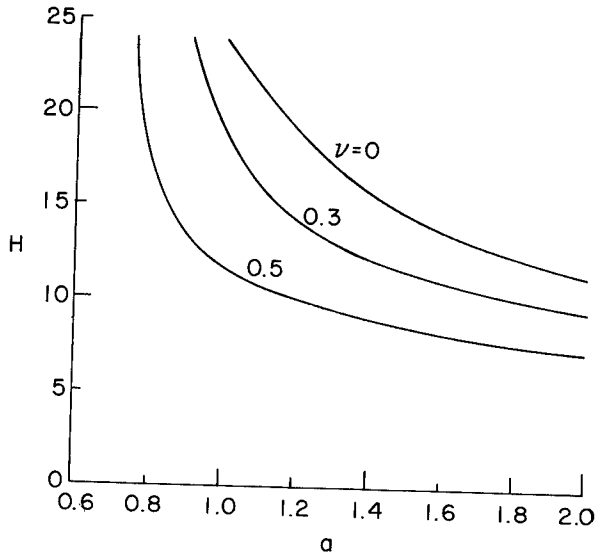



Fig. 9. Values of $H(\nu,a)$.

ACKNOWLEDGEMENT

Professor Jean Dieudonné (Maître de Conférence à la Faculté des Science de Nice) solved the boundary value problem of (5.1) during his visit at Northwestern University in 1958 at the request of the U.S. Army Snow, Ice and Permafrost Research Establishment, then at Evanston, Ill., as recorded in references (5) and (9). His method inspired the author for developing the method of transforming the integral solution (1.12) to the linear combination (1.13).

Numerical computation was performed with the help of Duncan M. Fitchet, student at Dartmouth College.

## REFERENCES


(1) Zabilansky, L.J., D.E. Nevel, and F.D. Haynes: Ice forces on model structures. Canadian Journal of Civil Engineering, 2(1975), 400 - 417.

(2) Mansfield, E.H. (1964) The Bending and Stretching of Plates. The MacMillan Company, New York.

(3) Hetényi, M. (1946) Beams on Elastic Foundations. The University of Michigan (Ann Arbor) Press.

(4) Ince, E.L. (1972) Ordinary Differential Equations, Republished by Dover Publications, Incorporated.

(5) Handbook of Mathematical Functions. National Bureau of Standards (1964).

(6) Nevel, D. (1961) The narrow free infinite wedges on an elastic foundation. U.S. Army Cold Regions Research and Engineering Laboratory Research Report 79.

(7) Dieudonné, J. (1958) Study of the equation $y'''' + 2xy'''' + w = 0$. Unpublished manuscript.

(8) Nevel, D. (1968) The general solution of a wedge on an elastic foundation. U.S. Army Cold Regions Research and Engineering Laboratory Research Report 247.

(9) Whittaker, E.T. and G.N. Watson (1952) A course of Modern Analysis. Cambridge at the University Press.

## APPENDICES

In the following Appendices A, B, and C, transformation of tensor components utilized in this paper are derived by use of the tensor notation where tensors are expressed in combinations of components and base vectors. This tensor expression yields simpler and more enjoyable analysis of component transformations in Euclidean space than the conventional tensorial expressions where base vectors are omitted, because geometric and mechanical quantities are explicitly shown in the former and therefore the meaning of the step by step computation is clear.

In the Appendix D, the deformation for the case $a = \infty$ is derived. In the Appendix E, the buckling of the semi-infinite plate is discussed. It is interesting to note that both cases pertains to the case of $x_o = \infty$ but they are substantially different.

### A. Transformation of (1.10) to (1.11).

Shear's $Q_x$ and $Q_y$ in rectangular coordinates are the magnitude per unit length of the shears acting on a side normal to the $x$-axis and $y$-axis, respectively, (see Figure 10). We shall begin with expressing $Q_x$ and $Q_y$ as components of a vector. Let $\mathbf{c}_x$ and $\mathbf{c}_y$ be unit vectors in the $x$- and $y$-directions. In Figure 11, let $\mathbf{c}_n$ be a unit vector normal to the hypotenuse $AB$ (Fig. 11). Vector $\mathbf{c}_n$ is given by

$$\mathbf{c}_n \, ds = \mathbf{c}_x \, dy + \mathbf{c}_y \, dx \tag{A.1}$$

because $\mathbf{c}_n$ thus defined satisfies the condition

$$\mathbf{c}_n \cdot \mathbf{c}_x = dy/ds$$

and

$$\mathbf{c}_n \cdot \mathbf{c}_y = dx/ds$$

where the dot (·) between two vectors means the scalar product of the two vectors. Let $Q_s$ be the shear per unit length of the hypotenuse *AB*. It is given by

$$Q_s \, ds = Q_x \, dy + Q_y \, dx \tag{A.2}$$

We can now prove that the equation

$$\mathbf{Q} = Q_x \mathbf{c}_x + Q_y \mathbf{c}_y \tag{A.3}$$

is the desired vector combination of $Q_x$ and $Q_y$, because the relation

$$\mathbf{Q} \cdot \mathbf{c}_n = Q_s$$

is satisfied.

Substitute (1.10) into (A.3) and transform the result to a tensor-invariant form:

$$Q = \nabla \cdot \mathbf{M} + \nabla w \cdot N \tag{A.4}$$

where

$$\nabla = \mathbf{c}_x \frac{\partial}{\partial x} + \mathbf{c}_y \frac{\partial}{\partial y} \tag{A.5}$$

$$\mathbf{M} = M_{xx} \mathbf{c}_x \mathbf{c}_x + M_{xy}(\mathbf{c}_x \mathbf{c}_y + \mathbf{c}_y \mathbf{c}_x) + M_{yy} \mathbf{c}_y \mathbf{c}_y \tag{A.6}$$

$$\mathbf{N} = N_{xx} \mathbf{c}_x \mathbf{c}_x + N_{xy}(\mathbf{c}_x \mathbf{c}_y + \mathbf{c}_y \mathbf{c}_x) + N_{yy} \mathbf{c}_y \mathbf{c}_y \tag{A.7}$$

In (A.4), a convention is made that **a.bc** means **(a.b)c.**

Let $\mathbf{u}_r$ and $\mathbf{u}_\theta$ be the unit vectors in the $r$- and $\theta$-directions. They are given by

$$\mathbf{u}_r = \mathbf{c}_x \cos\theta + \mathbf{c}_y \sin\theta$$

and (A.8)

$$\mathbf{u}_\theta = -\mathbf{c}_x \sin\theta + \mathbf{c}_y \cos\theta$$

These equations yield

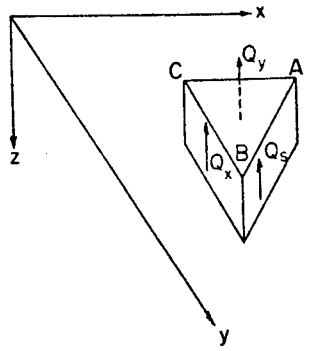



Fig. 10. Definition of $Q_x$ and $Q_y$.

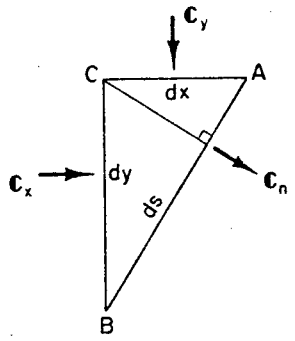



Fig. 11. Definition of $C_n$.

$$\left.\begin{aligned} \frac{\partial u_r}{\partial r} &= 0 \\ \frac{\partial u_\theta}{\partial r} &= 0 \\ \frac{\partial u_r}{\partial \theta} &= u_\theta \end{aligned}\right\} \qquad (A.9)$$

and

$$\frac{\partial u_\theta}{\partial \theta} = -u_r$$

In the polar coordinates, (A.5) becomes

$$\nabla = u_r \frac{\partial}{\partial r} + u_\theta \frac{\partial}{r\partial\theta} \qquad (A.10)$$

$$Q = \left(u_r \frac{\partial}{\partial r} + u_\theta \frac{\partial}{r\partial\theta}\right).M + \left(u_r \frac{\partial w}{\partial r} + u_\theta \frac{\partial w}{\partial \theta}\right).N, \qquad (A.11)$$

where

$$M = M_{rr}u_r u_r + M_{r\theta}(u_r u_\theta + u_\theta u_r) + M_{\theta\theta}\, u_\theta u_\theta \qquad (A.12)$$

and

$$N = N_{rr}u_r u_r + N_{r\theta}\,(u_r u_\theta + u_\theta u_r) + N_{\theta\theta}\, u_\theta u_\theta \qquad (A.13)$$

In the polar coordinates, (A3) becomes

$$Q = Q_\gamma u_r + Q_\theta u_\theta \qquad (A.14)$$

Carry out the differentiation in (A.11) by use of (A.9) and the scalar products indicated by dot (·) and identify the components with those of (A.14), then one finds (1.11).

B. Transformation of (1.1) to (1.2)

We shall prove the formula in the general polar coordinates,

$$N_{xx}\frac{\partial^2 w}{\partial x^2} + 2N_{xy}\frac{\partial^2 w}{\partial x \partial y} + N_{yy}\frac{\partial^2 w}{\partial y^2}$$

$$= N_{rr}\frac{\partial^2 w}{\partial r^2} + 2N_{r\theta}\frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial w}{\partial \theta}\right) + N_{\theta\theta}\left(\frac{1}{r^2}\frac{\partial^2 w}{\partial \theta^2} + \frac{1}{r}\frac{\partial w}{\partial r}\right) \qquad (B.1)$$

The left-hand side of (B.1) is a tensor-invarient form $\mathbf{N} \cdot\cdot \nabla\nabla w$ where N is given by (A.7), and $\nabla\nabla w$ is a dyadic,

$$\nabla\nabla w = \left(c_x \frac{\partial}{\partial x} + c_y \frac{\partial}{\partial y}\right)\left(c_x \frac{\partial w}{\partial x} + c_y \frac{\partial w}{\partial y}\right)$$

The double dot $(\cdot\cdot)$ of $ab\cdot\cdot cd$ means $(b.c)(a.d)$. In polar coordinates,

$$\nabla\nabla w = \left(u_r \frac{\partial}{\partial r} + u_\theta \frac{\partial}{r\partial\theta}\right)\left(u_\gamma \frac{\partial}{\partial r} + u_\theta \frac{\partial w}{r\partial\theta}\right)$$

Carrying out the differentiation given by (A.9), one finds

$$\nabla\nabla w = \frac{\partial^2 w}{\partial r^2} u_r u_r + \frac{\partial}{\partial r}\left(\frac{1}{r}\frac{\partial w}{\partial\theta}\right)(u_r u_\theta + u_\theta u_r) + \left(\frac{1}{r^2}\frac{\partial^2 w}{\partial\theta^2} + \frac{1}{r}\frac{\partial w}{\partial r}\right) u_\theta u_\theta \qquad (B.2)$$

Carrying out the double dot products by use of N in (A7) and $\nabla\nabla w$ in (B.2), one finds that $\mathbf{N}\cdot\cdot\nabla\nabla w$ becomes the right-hand side of (B.1).

C. Proof of (13.2)

Substituting Equation (1.4) of Mansfield (1), one can transform the tensor equation

$$\sigma = \sigma_x c_x c_x + \sigma_y c_y c_y + \tau_{xy}(c_x c_y + c_y c_x)$$

to an tensor-invariant form

$$\sigma = -[EZ/(1-\nu^2)]\left(\nabla\nabla w + \Delta\Delta w\right) \qquad (C.1)$$

The tensor-invariant operator

$$\Delta = c_x \frac{\partial}{\partial y} - c_y \frac{\partial}{\partial x}$$

becomes

$$\Delta = u_r \frac{\partial}{r\partial\theta} - u_\theta \frac{\partial}{\partial r} \qquad (C.2)$$

in the polar coordinates.

Substituting (A.10) and (C.2), and carrying out the differentiation as given by (A.9), one finds that (C.1) becomes

$$\sigma = \sigma_{rr} c_r c_r + \sigma_{r\theta}(u_r u_\theta + u_\theta u_r) + \sigma_{\theta\theta}\, u_\theta u_\theta$$

when $\sigma_{rr}$, $\sigma_{r\theta}$, and $\sigma_{\theta\theta}$ are given by (13.2).

d. Deformation for $a = \infty$.

Let $w_1(x)$ and $w_2(x)$ be defined with the real and imaginary parts of the right-hand side of (10.6):

$$w_1(x) = R \cos I$$

and

$$w_2(x) = R \sin I$$

where

$$R(x) = (\pi/x)^{\frac{1}{2}} \exp\left[-x/\sqrt{2} - \kappa^2/(4\sqrt{2}x)\right]$$

and

$$I(x) = \pi/8 + x/\sqrt{2} - \kappa^2/(4\sqrt{2}x)$$

The depression is given by

$$w(x) = A\, w_1(x) + B\, w_2(x)$$

When $\kappa = \infty$, there are two positive roots given by (12.16). The ratio

$$\frac{A}{B} = -\frac{L(w_2)}{L(w_1)} = -\frac{M(w_2)}{M(w_1)}$$

can be computed by using (12.14) and (12.15). Thus one finds

$$\frac{A}{B} = \frac{(1+\zeta_h)\cos I_h + (1-\zeta h)\sin I_h}{(1+\zeta_h)\sin I_h - (1-\zeta_h)\cos I_h}$$

where

$$I_h = I(x_h)$$

and $h$ is defined by (12.18). Normalizing $w(x)$ at $x = x_h$, one finds

$$A = (R_h)^{-1}\left(\cos I_h \pm (\sqrt{3})^{-1} \sin I_h\right)$$

and

$$B = (R_h)^{-1}\left(\sin I_h \mp (\sqrt{3})^{-1} \cos I_h\right)$$

Thus the normalized deformation $w_N(x)$ is given by

$$w_N(x) = \left[R(x)/R_h\right]\left[\cos(I(x) - I_h) \mp (\sqrt{3})^{-1} \sin(I(x) - I_h)\right] \quad \text{(D.1)}$$

Letting

$$x - x_h = \xi$$

and assuming $\xi$ to be finite, one may let $x \to \infty$ in (D.1). Thus one finds

$$w_N(x) = \exp(-\xi/\sqrt{2})\left[\cos(\xi/\sqrt{2}) \mp (\sqrt{3})^{-1} \sin(\xi/\sqrt{2})\right] \quad \text{(D.2)}$$

The maximum of $w_N(x)$ occurs at

$$\tan(\xi/\sqrt{2}) = -2 + \sqrt{3}$$

which is negative. Therefore $w_N(x)$ is always decreasing for $\xi \geq 0$. The deformation at $a = \infty$, therefore, does not take a minimum, as those (shown in Figure 5 and 6) of case $1-\nu^2 \leq a \leq 2$ do.

E. Buckling of semi-infinite plate

We shall show that the deformation discovered in the preceding section is different from the buckling deformation of a rectangular semi-infinite floating plate.

We assume that uniform pressure $N_{xx}$ is applied on the axis $y$, the axis $x$ extending from $x = 0$ to $x = \infty$. Then from (1.1) one gets

$$\frac{d^4w}{dx^4} + \gamma w = N_{xx}\frac{d^2w}{dx^2} \quad \text{(E.1)}$$

where we have put $q = 0$. Defining new $x$ by the quotient of old $x$ divided by the characteristic length $\ell_o = (D/\gamma)^{1/4}$, (E.1) becomes

$$\frac{d^4w}{dx^4} + 2a\frac{d^2w}{dx^2} + w = 0 \quad \text{(E.2)}$$

where we have put

$$N_{xx} = -2a\gamma\ell_o^2 \tag{E.3}$$

Because $N_{xx} \leqq 0$, the relation

$$a \geqq 0$$

must be satisfied. The solution of the differential equation (E.2) is

$$w_k = \exp(\lambda_k x) \tag{E.4}$$

where

$$\lambda_k^2 = -a \pm \sqrt{a^2 - 1} \tag{E.5}$$

The real part of $\lambda_k$ must be chosen to be negative. The convention with regard to suffix $k$ is still observed.

When $a = 1$, the general solution is given by

$$w = A \cos x + B \sin x$$

which we do not accept, because the boundary condition at $x = \infty$ cannot be met.

When $0 \leqq a < 1$, letting

$$a = \cos 2\eta \tag{E.6}$$

the general solution is given by

$$w = e^{-\beta x} [A \cos\alpha x + B \sin\alpha x] \tag{E.7}$$

where

$$\beta = \sin\eta \tag{E.8}$$

and

$$\alpha = \cos\eta \tag{E.9}$$

The condition

$$0 < \eta \leqq \pi/4 \tag{E.10}$$

must be met to satisfy the conditions with regard to $a$ and $\lambda_k$.

When $a > 1$, letting

$$a = \cosh 2\eta$$

one finds four fundamental solutions

$$\cos\nu x,\ \sin\nu x,\ \cos(x/\nu),\ \text{and}\ \sin(x/\nu)$$

where

$$\nu = \exp(\eta)$$

We shall discuss below only the case $0 \leq a < 1$, because the boundary condition at $x = \infty$ cannot be satisfied in the other cases.

The free-edge condition for this case is

$$\left.\begin{aligned} &\frac{d^2w}{dx^2} = 0 \\ &\text{and} \\ &\frac{d^3w}{dx^3} + 2a\frac{dw}{dx} = 0 \end{aligned}\right\} \qquad (E.11)$$

The second equation of (E.11) is derived from the first equation of (1.10). Substituting (E.7), one finds that the eigenvalue is given by

$$\eta = \pi/6$$

i.e.

$$a = 1/2 \qquad (E.12)$$

The deformation for this case is

$$w(x) = A \exp(-x/2) \cos[(\sqrt{3}\pi/2) + (\pi/6)] \qquad (E.13)$$

where $A$ is arbitrary. The maximum of $w(x)$ occurs at

$$x = 4\pi/(3\sqrt{3})$$

Therefore the deformation in this section is different from the deformation in the preceding section.

For the simple-edge condition, the eigenvalue is given by

$$a = 1$$

which we do not accept.

# Nonlinear Theory of the Response of Pavements to Vibratory Loads

Richard A. Weiss
Pavement Investigations Division
Soils and Pavements Laboratory
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi

ABSTRACT. A nonlinear model of the pavement response to a dynamic load is presented which has applications to the vibratory nondestructive method of testing pavements. The parameters of the model have been determined by comparison with actual dynamic load-deflection curves. The model gives a quantitative description of the dependence of the measured dynamic load-deflection curves on the strength of the pavement, static load of the vibrator, and the frequency of operation of the vibrator. The model determines the elastic modulus of the subgrade from the measured load-deflection curves. The nonlinear dynamical model is applied to the laboratory determination of the resilient modulus with the result that the resilient modulus is expressed analytically in terms of the static confining pressure, dynamic deviator stress, and material parameters which describe the linear and nonlinear behavior of soil under dynamic and static force loading.

## I. INTRODUCTION

The Waterways Experiment Station (WES) has for many years used the

method of nondestructive testing of airfield pavements.[1-5] This method of testing pavements is relatively quick accurate, reproducible, and inexpensive. When the nondestructive test method is used an airfield runway need not be shut down for long periods of time as is the case for the destructive testing of pavements.

The instrument used for the vibratory nondestructive testing of pavements is a mechanical vibrator whose force payload to the pavement surface is generated either by a hydraulic system or a mechanism of counter-rotating weights. The WES 16-kip vibrator applies a static load of 16 kips to the pavement surface and a dynamic load to the pavement surface which can be varied from 0 to 15 kips. Both static and dynamic loads are applied to the pavement surface through a circular 18-in. diameter baseplate.

Four types of nondestructive tests are generally performed on pavements, and these consist of the following measurements:

a. Dynamic load-deflection curves giving the dynamic amplitude as a function of the dynamic load.

b. Frequency response spectrum giving the dynamic amplitude as a function of frequency for a fixed dynamic load.

c. Deflection basin measurements.

d. Rayleigh wave dispersion curves giving phase velocity versus wavelength.

Only the dynamic load-deflection curves and the frequency response spectrum measurements will be considered in this paper.

A typical measured frequency response curve appears in Fig. 1, and a typical measured dynamic load-deflection curve appears in Fig. 2. Most of the WES measurements of the dynamic load-deflection curves were done at a frequency of 15 Hz. Experience has shown that the dynamic load-deflection curves are relatively smooth for this frequency. The frequency response spectrum may contain multiple resonance peaks.

Two basic theoretical approaches have been taken to describe the experimental data:

1. a linear theory of the frequency response spectrum
2. a nonlinear theory of the dynamic load-deflection curves

The two types of dynamic pavement response models that have been considered are shown in Fig. 3. Single-mass and multiple-mass models have been developed in the linear theory, while only a single mass model was developed with a nonlinear spring constant. It was found that multiple-mass pavement response models are somewhat intractable because they contain many parameters. Only the single-mass pavement response models are considered in this paper. The elements of the spring-mass-dashpot model must be determined in terms of the characteristic forms of the measured frequency response spectrum and the measured dynamic load-deflection curves.

## II. DYNAMIC FREQUENCY RESPONSE THEORY

The dynamic frequency response spectrum measured at the pavement surface is often quite complex and difficult to interpret. Many factors probably contribute to produce its characteristic shape. In order to extract some information about pavement and subgrade structure from the measured dynamic frequency response spectrum it is necessary to use a simple

dynamic pavement response model to fit the measured frequency response spectrum with the theoretically predicted frequency response spectrum. This fit will yield the parameters of the dynamic model from which the pavement and subgrade structure can be determined. The frequency response spectrum of the single-mass model has one resonance peak, and this predicted resonance peak is fit to the second resonance peak of the measured response spectrum. The second peak is chosen because an examination of many frequency response spectra has shown this peak to be more consistent and less affected by electronic equipment than the other peaks. Generally the second peak is the most pronounced.

The second resonance peak is associated with a resonance frequency and a resonance amplitude as indicated in Fig. 4. The resonance amplitude and frequency was used to calculate the elements of the spring model — effective mass, effective spring constant, and effective damping constant. The elements of the single-mass spring model can be simply related to the resonance peak.

## DETERMINATION OF ELEMENTS OF THE SPRING MODEL

Within the framework of the single-mass spring model[6-9] the dynamic amplitude of the pavement surface response to a sinusoidal dynamic load can be written as

$$A = F_D/S \tag{1}$$

$$S = \sqrt{(k-m\omega^2)^2 + C^2\omega^2} \tag{2}$$

where A = amplitude of the dynamic displacement of the pavement surface as represented by a linear spring model, $F_D$ = dynamic load applied to the pave-

ment surface, S = dynamic stiffness, k = linear spring constant, m = effective mass of the pavement-subgrade system, $\omega$ = angular frequency, and C = damping coefficient. The resonance frequency and amplitude[6] can be obtained from (1) and (2) to be

$$f_R = \frac{1}{2\pi}\sqrt{\frac{k}{m}}\sqrt{1 - 2D^2} \tag{3}$$

$$A_R = \frac{F_D}{2kD\sqrt{1 - D^2}} \tag{4}$$

$$D = \frac{C}{2\sqrt{km}} \tag{5}$$

where $f_R$ = resonance frequency, $A_R$ = responance amplitude, and D = damping ratio. The three elements of the linear spring model that are to be obtained are k , m and C . In order to determine these three parameters another piece of information, in addition to $f_R$ and $A_R$ , is necessary. This is given by

$$J(\omega) = A_R/A \tag{6}$$

where $J(\omega)$ = ratio of the resonance amplitude to the amplitude at some nearby frequency. The theoretical value of this ratio is given by

$$J(k,m,C,\omega) = \frac{\sqrt{(k - m\omega^2)^2 + C^2\omega^2}}{\sqrt{(k - m\omega_R^2)^2 + C^2\omega_R^2}} \tag{7}$$

The three measured quantities which are extracted from the frequency response curve are $f_R$ , $A_R$ and $J(\omega)$.

The spring model elements $k$ , $m$ and $C$ must now be obtained in terms of $f_R$ , $A_R$ and $J(\omega)$. The equations in (3) - (6) can be inverted to determine $k$ and $D$ in the following manner

$$k = 4\pi^2 m f_R^2 \sqrt{1 + \left( \frac{F_D}{4\pi^2 m f_R^2 A_R} \right)^2} \tag{8}$$

$$D^2 = \frac{1}{2} - \frac{1}{2} \left[ 1 + \left( \frac{F_D}{4\pi^2 m f_R^2 A_R} \right)^2 \right]^{-1/2} \tag{9}$$

The $k$ and $D$ terms have now been expressed in terms of the effective mass. Using (8) and (9) it is now possible to express $J(k,m,C,\omega)$ in terms of the effective mass as the only unknown parameter as follows

$$J(m,\omega) = \frac{\sqrt{\left| 1 - \frac{m\omega^2}{k} \right|^2 + \frac{4mD^2\omega^2}{k}}}{\sqrt{\left| 1 - \frac{m\omega_R^2}{k} \right|^2 + \frac{4mD^2\omega_R^2}{k}}} \tag{10}$$

The only unknown independent variable in $J(m,\omega)$ is now the effective mass. By sweeping through a series of values of $m$ and calculating numerical values of $J(m,\omega)$ it is possible to determine the specific value of $m$ for which $J(m,\omega)$ is equal to the experimental value of the J-ratio, i.e.,

$J(m,\omega) = J(\omega)$. This condition determines the value of the effective mass required by the spring-mass-dashpot model to fit the experimentally measured dynamic frequency response curve. Placing this calculated value of the effective mass into (5), (8) and (9) gives the proper values of $k$ and $C$ required to fit the experimental frequency response data. The necessary computer programs to accomplish this work on a digital computer have been developed and will be referred to as the WES Dynamic Frequency Response Program.

## DETERMINATION OF SUBGRADE MODULUS BY FREQUENCY RESPONSE METHOD

The value of the spring constant that is determined from the measured frequency response spectrum will be used to determine the subgrade modulus. The theory of the linear elastic layered half-space predicts a theoretical value of the static spring constant $k_T$ which depends on the radius of the loaded area and on the elastic constants of the subgrade and the pavement layers. Computer programs are available which calculate the value of $k_T$ if the Young's modulus and Poisson's ratio of each layer of the half-space is known. A well known computer program of this kind is the Chevron Program. The procedure for determining the Young's modulus $E_S$ of the subgrade is shown in Fig. 5. The measured values of $f_R$, $A_R$ and $J(\omega)$ are inserted into the WES Dynamic Frequency Response Program and values of $k$, $m$ and $C$ are determined. The Young's modulus and Poisson's ratio of the layers of the pavement are selected and entered into the Chevron Program. The subgrade modulus $E_S$ is then iterated in the Chevron Program and a series of values of $k_T$ are determined. The proper value of $E_S$ is determined by the condition

$$k = k_T \tag{11}$$

The predicted value of $E_S$ will depend on the values of the elastic moduli selected for the pavement layers.

NUMERICAL RESULTS OF FREQUENCY RESPONSE METHOD

Values of $k$, $m$, $C$ and $E_S$ have been obtained for several airport pavement sites and are listed in Table I. This table has listed the sites according to increasing values of the Dynamic Stiffness Modulus (DSM), which is the slope of the dynamic load-deflection curves at a dynamic load of 14 kips. It is seen that the measured spring constant $k$ increases with increasing pavement strength and that $k$ is not equal to the DSM value. The effective mass is presented as a ratio to the above-surface (vibrator) mass, and increases with the strength of the pavement. The effective mass is not equal to the above-surface mass and any theory which apriori assumes that $m = m_v$ cannot be used to fit the experimental frequency response data. The value of the damping constant also increases with increasing pavement strength. The predicted values of $E_S$ are compared to those modulus values that are predicted by the CBR method ($E_S = 1500$ CBR). The values of $E_S$ predicted by the combined WES Frequency Response Program and the Chevron Program are 3 to 5 times larger than those predicted by the CBR method.

There are several possible reasons for the discrepancy in the values of $E_S$ predicted by these two methods:

a. the pavement-subgrade system is nonlinear under dynamic and static loading

b. the subgrade is not uniform and the theoretical layered

elastic half-space model requires a rigid boundary below the subgrade

c. reflections from a lower boundary layer add to the motion of the pavement surface

When a rigid boundary such as bedrock is present relatively close to the pavement surface it is possible that the effects listed in b. and c. may be of importance for determining the motion of a pavement surface that is subjected to a sinusoidal dynamic loading. However, the discrepancy between the values of $E_S$ predicted by the CBR method and that predicted by the frequency response spectra method also occurs in cases where the subgrade is relatively uniform and contains no obvious discontinuities. Therefore only the fact that the response of pavements and subgrades to dynamic and static loads is nonlinear remains as a possible explanation for the discrepancy in the values of $E_S$ determined by these two methods.

## III. NONLINEAR THEORY OF PAVEMENT RESPONSE TO DYNAMIC SURFACE LOADINGS

An alternative method for determining the subgrade modulus from vibratory nondestructive test data is the use of the dynamic load-deflection curves measured at the pavement surface for a fixed frequency and a fixed static load. These dynamic load-deflection curves are generally nonlinear for weak pavements and become more linear for stronger pavements. Over the years the WES has collected an extensive set of dynamic load-deflection curves that have been obtained on many airfield pavements throughout the country.

The nonlinear dynamic load-deflection curves were measured at a frequency of 15 Hz and at a static surface loading of 16 kips. The nonlinear

dynamic theory must account for the frequency and static load conditions under which the dynamic load-deflection curves were measured. The predicted subgrade modulus should be free of the particular loading characteristics of the vibrator. Therefore, in addition to the static Young's modulus some other parameters have to be introduced which will account for the observed nonlinearity of the dynamic load-deflection curves. These nonlinear parameters must also account for the nonlinear behavior of the static load-deflection curves. The predicted subgrade modulus value will be independent of the particular loading characteristics of the vibrator — frequency, static load, and dynamic load. Only the natural overburden pressure will be reflected in the subgrade modulus value.

The determination of the elastic constants and the static and dynamic nonlinear parameters of the pavements and subgrades from measured dynamic load-deflection data requires a nonlinear dynamic theory of pavement response[4].

## EQUATION OF MOTION OF A NONLINEAR OSCILLATOR

The nonlinear theory of pavement response to a vibratory load assumes that the pavement-subgrade system can be described by a lumped mass nonlinear oscillator whose equation of motion is written as

$$m\ddot{x} + C\dot{x} + k_{00}x + bx^3 + ex^5 = F_D + F_S \qquad (12)$$

where $m$ = effective mass of the pavement-subgrade system, $x$ = total displacement of the pavement surface beneath the vibrator baseplate, $C$ = damping constant, $k_{00}$ = linear spring constant, $b$ = third order nonlinear pavement parameter, $e$ = fifth order nonlinear pavement parameter,

$F_D$ = dynamic load applied to the pavement surface, and $F_S$ = static load applied to the pavement surface. The total displacement of the pavement surface is decomposed into a static and a dynamic part as follows

$$x = x_e + \xi \tag{13}$$

where $x_e$ = static elastic displacement of the pavement surface, and $\xi$ = dynamic elastic displacement of the pavement surface. Placing (13) into (12) gives the following equation of motion

$$m\ddot{\xi} + C\dot{\xi} + \left(k_{00} + 3bx_e^2 + 5ex_e^4\right)\xi + b\xi^3 + e\xi^5 + \xi g(x_e,\xi) = F_D \tag{14}$$

where

$$g(x_e,\xi) = 3bx_e\xi + 10ex_e^3\xi + 10ex_e^2\xi^2 + 5ex_e\xi^3 \tag{15}$$

For convenience in manipulating (14) it is necessary to use a time averaged expression for (15)

$$g(x_e,\xi) = 3a_1bx_e^2 + 5a_2ex_e^4 + a_3b\xi^2 + a_4e\xi^4 \tag{16}$$

where $a_1$ , $a_2$ , $a_3$ and $a_4$ are coefficients to be determined from the measured dynamic load-deflection data. Combining (16) and (14) gives the motion equation as

$$m\ddot{\xi} + C\dot{\xi} + k_0\xi + b\theta\xi^3 + e\eta\xi^5 = F_D \tag{17}$$

where

$$k_0 = k_{00} + 3b\varepsilon_2x_e^2 + 5e\varepsilon_4x_e^4 \tag{18}$$

$$\theta = 1 + a_3 \tag{19}$$

$$\eta = 1 + a_4 \tag{20}$$

$$\varepsilon_2 = 1 + a_1 \tag{21}$$

$$\varepsilon_4 = 1 + a_2 \tag{22}$$

The parameters $\theta$ , $\eta$ , $\varepsilon_2$ and $\varepsilon_4$ depend on the pavement strength and are determined by requiring (17) to adequately describe the dynamic load-deflection curves. The nonlinear parameters b and e determine the static load-deflection curves, as can be seen from (12)

$$F_S = k_{00} x_e + b x_e^3 + e x_e^5 \tag{23}$$

In general it is found that $b < 0$ and $e > 0$ for pavements and most subgrades.

THEORY OF DYNAMIC LOAD-DEFLECTION CURVES

The problem remains to solve the nonlinear equation (17). This can be done by casting (17) into an equivalent linear form for which the dynamic amplitude is given by

$$\xi = F_D/S \tag{24}$$

where

$$S = \sqrt{(k - m\omega^2)^2 + C^2\omega^2} \tag{25}$$

where S = dynamic stiffness, k = dynamic spring constant, m = effective mass, $\omega$ = angular frequency and C = damping constant. The requirement that (24) and (25) be a solution of (17) is that the spring constant in (25) is given by[4]

$$k = k_0 + \frac{3}{4} b\theta\xi^2 + \frac{5}{8} e\eta\xi^4 \tag{26}$$

Therefore the spring constant for a nonlinear system depends on the dynamic and static displacements of the pavement surface.

Placing (26) into (25) and (24), and solving for the dynamic amplitude yields the result[4]

$$\xi = \frac{F_D}{S_0} (1 + \alpha_1\psi + \alpha_2\psi^2 + \cdots) \tag{27}$$

where

$$S_0 = \sqrt{(k_0 - m\omega^2)^2 + C^2\omega^2} \tag{28}$$

$$\psi = F_D^2/S_0^4 \tag{29}$$

$$\alpha_1 = -\frac{3}{4} b\theta(k_0 - m\omega^2) \tag{30}$$

$$\alpha_2 = \frac{7}{2}\left(\frac{3}{4}\right)^2 b^2\theta^2(k_0 - m\omega^2)^2 - S_0^2\left[\frac{5}{8}\eta e(k_0 - m\omega^2) + \frac{1}{2}\left(\frac{3}{4}\right)^2 b^2\theta^2\right] \tag{31}$$

As seen from (27) - (30) the degree of nonlinearity of a dynamic load-deflection curve depends on the strength of the pavement and the frequency of operation of the vibrator. The strength of the pavement affects the degree of nonlinearity of the dynamic load-deflection curves through the term $S_0^{-4}$ that appears in (27) and (29). The $S_0^{-4}$ term shows that strong pavements tend to be more linear than weak pavements. From (30) it is clear that there is a critical frequency for which the first order nonlinear term vanishes and this frequency is given by

$$f_c = \frac{1}{2\pi}\sqrt{\frac{k_0}{m}} \tag{32}$$

At this frequency the dynamic load-deflection curves should become especially linear in the regions of low dynamic force if the second order nonlinear term is comparatively small. The straightening effect at the critical frequency will not be strongly evident if the second order nonlinear term is comparatively large.

DYNAMIC NATURE OF THE SPRING CONSTANT

The measurement of the dynamic load-deflection curves determine the linear and nonlinear parameters of a pavement system — $k_{00}$ , b , e , $\theta$ , $\eta$ , $\varepsilon_2$ , $\varepsilon_4$ . Equation (26) shows that the spring constant k that is determined from a dynamic analysis of the nonlinear properties of a pavement-subgrade system is dependent on the dynamic and static displacements of the pavement surface as well as on the elastic constants of the pavement-subgrade system. Therefore the spring constant k that is determined from the dynamic response of a nonlinear pavement system is a dynamic quantity that is not analogous to an ordinary static spring constant. The theoretical static spring constant determined from a static linear elastic program such as the Chevron Program will depend only on the elastic constants of the pavement. Therefore the value of k determined from the dynamic response data of a nonlinear pavement cannot logically be compared to the static $k_T$ value determined from static layered elastic computer programs. Static plate bearing tests will result in a spring constant which will also not be directly comparable to the spring constant determined from an analysis

of dynamic data.

FINITE DEPTH OF INFLUENCE

The static linear and nonlinear parameters $k_{00}$, b and e respectively can be related to the elastic parameters of the pavement layers and to the depth of influence of the static stress-strain field[4]. The finite depth of influence is written in terms of the static deflection of the pavement surface as

$$\ell = \ell_0 + \ell_2 x_e^2 + \ell_4 x_e^4 \tag{33}$$

For the simplest case of a vibrator placed on the surface of a subgrade, the static parameters are

$$k_{00} = \frac{2\pi a^2 \psi (1 - \nu) G}{\ell_0 (1 - 2\nu)} \tag{34}$$

$$b = - \frac{4\pi a^2 \psi \ell_2 (1 - \nu) G}{\ell_0^2 (1 - 2\nu)} \tag{35}$$

$$e = \frac{6\pi a^2 \psi \delta (1 - \nu) G}{\ell_0 (1 - 2\nu)} \tag{36}$$

where

$$\delta = \left( \frac{\ell_2}{\ell_0} \right)^2 - \frac{\ell_4}{\ell_0} \tag{37}$$

and $\psi$ = volume factor for the frustum of the cone of stress and strain. It is through equations similar to (34) - (37) that the connection is made between the elastic parameters of the pavement system and the theoretical

expression for the dynamic stiffness as given by (25) and (26).

MODEL PARAMETERS

The model parameters $k$ , $m$ , $C$ , $k_{00}$ , $b$ , $e$ , $\ell_0$ , $\ell_2$ , $\ell_4$ , $\theta$ , $\eta$ , $\varepsilon_2$ and $\varepsilon_4$ depend on vibrator characteristics and on the structure of the pavement and subgrade. This dependence is in general very complicated and difficult to determine theoretically. The simplest way to attach the model parameters to the strength of a pavement-subgrade system is to determine these parameters in terms of the measured dynamic stiffness modulus (DSM) of a pavement. The DSM is the slope of the load-deflection curve measured by the WES 16-kip vibrator in the region of large dynamic load; it is in fact the tangent modulus of the dynamic load-deflection curves for $F_D \sim 15$ kips. The DSM value is a suitable choice for a parameter in terms of which to describe the model parameters because it is a measure of the bulk strength of the pavement and subgrade. The model parameters expressed in terms of the measured DSM correspond to the WES 16-kip vibrator. The vibrator characteristics appear in these parameters because the subgrade modulus to be determined is intended to be independent of the dynamic characteristics of the vibrator. A corresponding set of vibrator parameters will have to be developed for any other vibrator that is to be used for nondestructive testing of pavements.

The model parameters are presented as a function of the measured DSM in Figs. 6 through 15. From these figures it is seen that $k$ , $m$ , $C$ and $\ell_0$ are increasing functions of the strength of the pavement. The dynamic spring constant presented in Fig. 6 corresponds to a dynamic load of 15 kips. The depth of influence of the static stress-strain field increases with increasing

pavement strength while the static deflection of the pavement surface under a fixed static load decreases with increasing strength. As seen from Fig. 7 the effective mass is generally much larger than the above-surface mass, and it would be incorrect to assume that the only lumped-mass of the vibrator-pavement-subgrade system is the vibrator mass itself. The effective mass of the dynamic model includes the inertial effects of the mechanical radiation field in the pavement and subgrade. In all cases of the pavements investigated it was found that $b < 0$ and $e > 0$.

## DETERMINATION OF SUBGRADE MODULUS FROM DYNAMIC LOAD-DEFLECTION CURVES

The nonlinear dynamic response model that has been outlined in the preceding section can be used in conjunction with a dynamic load-deflection curve measured at the pavement surface to determine the modulus of the subgrade beneath the pavement. A computer program has been developed which calculates the theoretical dynamic response of a pavement in terms of the elastic moduli of the pavement layers and subgrade and in terms of the empirically determined parameters $\theta$ , $\eta$ , $\varepsilon_2$ , $\varepsilon_4$ , m and C which have been expressed in terms of the measured DSM values of the pavement. A typical example of the vibratory nondestructive input data to the computer program is shown in Table II. The computer program calculates a theoretical load-deflection curve in terms of the b and e coefficients that are determined from measured load-deflection curves and in terms of the elastic moduli of the pavement layers and the subgrade. The elastic moduli of the pavement layers are selected from laboratory tests and CBR measurements. The subgrade modulus is then determined by requiring that the theoretically predicted dynamic load-deflection curve agree with the measured dynamic load-deflection curve. This procedure for determining the subgrade modulus is shown in Fig. 16. The numerical results of this procedure for a few pavement sites are presented in Table III. The values of the subgrade modulus predicted by the nonlinear dynamic response theory are in general agreement with those predicted by the empirical relation $E_S = 1500$ CBR.

## IV. LABORATORY CONFIRMATION OF VIBRATORY NONDESTRUCTIVE FIELD TEST DATA

It is of interest to be able to correlate the laboratory value of the resilient modulus $M_r$ of a soil sample taken from the subgrade at a pavement

or soil site for which the subgrade modulus has been predicted by the vibratory nondestructive testing method. Such a correlation is difficult to achieve because the loading conditions on the soil sample for the laboratory tests are different from the loading conditions on the subgrade during vibratory nondestructive testing. The loading conditions differ in terms of the magnitude of the static and dynamic stresses and in terms of the frequency of application of the dynamic stress.

In its natural state, an element of soil in the subgrade is subjected only to the overburden pressure. When a vibrator is operated on the surface of a pavement or subgrade, an additional static and dynamic stress is applied to an element of soil in the subgrade. For the WES 16-kip vibrator the static load applied to the surface is 16 kips, while the dynamic load can be varied up to 15 kips and is applied sinusoidally with a frequency of 15 Hz. The stress field in the subgrade is nonuniform and can be calculated by standard elasticity theory.

The laboratory sample for resilient modulus testing is cylindrical in shape with a typical diameter of 3 inches and a length of 6 inches. The cylindrical sample is subjected to a static confining pressure and then a dynamic load is applied in the axial direction. The stress is uniform along the axis of the laboratory sample. The total stress along the axis of the laboratory sample is written as

$$\sigma = \sigma_D + \sigma_S \tag{38}$$

where $\sigma_D$ = dynamic stress in axial direction of sample, and $\sigma_S$ = confining

pressure. The axial dynamic stress is also called the deviator stress and is written as $\sigma_D = \sigma - \sigma_S$, where $\sigma$ = total stress along the axis of the specimen. The resilient modulus has been measured for a number of soil and pavement materials, and $M_r$ has been found to depend on $\sigma_S$ and $\sigma_D$. The dependence of $M_r$ on the dynamic deviator stress is such that $M_r$ at first decreases with increasing values of $\sigma_D$, attains a minimum value, and then increases with further increase of the deviator stress[10].

The dynamic stress acting along the axial direction of the soil specimen during the laboratory resilient modulus test is applied as a series of pulses in the form of haversines with a pulse of 1 second duration being applied every 3 seconds. The characteristic frequency of the dynamic loading on the sample will therefore be in the range of 0.3 - 1.0 Hz, and this is much lower than the frequency of 15 Hz at which the vibratory nondestructive field tests are conducted. The large difference in the frequencies used for these two types of tests requires that an adequate account of frequency effects be included in the theoretical analysis of both laboratory and field vibratory tests.

## NONLINEAR DYNAMICAL ANALYSIS OF THE RESILIENT MODULUS TEST

A dynamical theory of the resilient modulus test has been developed which is similar in form to the analysis developed for the vibratory nondestructive field tests. The basic result of this theory is that the dynamic displacement of the test specimen can be written as

$$\xi = F_D/S = A_C\sigma_D/S \tag{39}$$

$$S = \sqrt{(k - m\omega^2)^2 + C^2\omega^2} \tag{40}$$

where $\xi$ , $F_D$ and $\sigma_D$ = resilient dynamic displacement, dynamic load, and dynamic stress on the cylinder end in the axial direction; $S$ , $k$ , $m$ , $C$ and $A_C$ = dynamic stiffness, spring constant, effective mass, damping constant, and area of loaded end of the cylinder respectively; $\omega$ = effective angular frequency component of the dynamic load applied to the soil sample.

The nonlinear theory of vibrations that was outlined earlier in this paper for the vibratory nondestructive field tests can also be used to calculate the quantities in (39) and (40). This nonlinear theory shows that the spring constant is given by

$$k = k_0 + \frac{3}{4}b\theta\xi^2 + \frac{5}{8}e\eta\xi^4 \tag{41}$$

$$k_0 = k_{00} + 3b\varepsilon_2 x_e^2 + 5e\varepsilon_4 x_e^4 \tag{42}$$

where $b$ , $e$ , $\theta$ , $\eta$ , $\varepsilon_2$ and $\varepsilon_4$ = parameters which characterize the soil sample, and $x_e$ = resilient static displacement of the soil sample in the axial direction. The coefficients $k_{00}$ , $b$ and $e$ could be determined from the resilient static stress-strain curve if such a curve could be measured. The resilient static stress-strain curve of the soil sample is determined by

$$F_S = \sigma_S A_C = k_{00} x_e + b x_e^3 + e x_e^5 \tag{43}$$

where $\sigma_S$ = static confining pressure, and $F_S$ = total static force applied to the cylinder end.

The solution of (39) - (42) can be written as[4]

$$\xi = \frac{F_D}{S_0}\left|1 + \alpha_1\psi + \alpha_2\psi^2\right| \tag{44}$$

where

$$S_0 = \sqrt{(k_0 - m\omega^2)^2 + C^2\omega^2} \tag{45}$$

$$\psi = F_D^2/S_0^4 = A_C^2\sigma_D^2/S_0^4 \tag{46}$$

$$\alpha_1 = -\frac{3}{4}b\theta(k_0 - m\omega^2) \tag{47}$$

$$\alpha_2 = \frac{7}{2}\left(\frac{3}{4}\right)^2 b^2\theta^2(k_0 - m\omega^2)^2 - S_0^2\left|\frac{5}{8}\eta e(k_0 - m\omega^2) + \frac{1}{2}\left(\frac{3}{4}\right)^2 b^2\theta^2\right| \tag{48}$$

The dynamic stiffness of the soil sample can be obtained from (39) and (44) to be

$$S = S_0(1 + \beta_1\psi + \beta_2\psi^2) \tag{49}$$

$$\beta_1 = -\alpha_1 \tag{50}$$

$$\beta_2 = \alpha_1^2 - \alpha_2 \tag{51}$$

The quantities necessary for the calculation of the resilient modulus have now been determined.

## CALCULATION OF THE RESILIENT MODULUS

The resilient modulus is defined as the slope of the unloading portion of the dynamic stress-strain curve of the soil sample, and is given by

$$M_r = \frac{d\sigma_D}{d\varepsilon_D} = \frac{L}{A_C}\left(\frac{d\xi}{dF_D}\right)^{-1} \tag{52}$$

where $\varepsilon_D$ = dynamic strain in axial direction, L = length of the soil sample, and $A_C$ = area of end of the cylindrical sample. In (52) $\xi$ is assumed to describe the unloading portion of the resilient dynamic load-deflection curve of the soil sample. Combining (44) and (52) gives

$$M_r = M_{r0}\left(1 + \delta_1\psi + \delta_2\psi^2\right) \tag{53}$$

where

$$\delta_1 = -3\alpha_1 \tag{54}$$

$$\delta_2 = 9\alpha_1^2 - 5\alpha_2 \tag{55}$$

$$M_{r0} = \frac{L}{A_C} S_0 \tag{56}$$

For the low frequency and small mass with which the resilient modulus tests are conducted, the inertial and damping terms in (40) and (45) can be neglected and the following approximations can be made

$$S \sim k \tag{57}$$

$$S_0 \sim k_0 \tag{58}$$

The same approximations can be made in (47) and (48). Combining (42), (56) and (58) gives the following approximation

$$M_{r0} \sim E_0 + E_2 x_e^2 + E_4 x_e^4 \tag{59}$$

where

$$E_0 = \frac{L}{A_C} k_{00} \tag{60}$$

$$E_2 = \frac{L}{A_C} 3b\epsilon_2 \tag{61}$$

$$E_4 = \frac{L}{A_C} 5e\epsilon_4 \tag{62}$$

The quantities $E_0$ , $E_2$ and $E_4$ are soil parameters which are independent of the size of the soil sample and machine characteristics. The calculation of the resilient Poisson's ratio requires further study.

The expression for $M_r$ given by (53) - (56) characterizes the resilient modulus in terms of $\sigma_D$ , $\sigma_S$ and $\omega$ . The parameters required to describe the resilient modulus are $k_{00}$ , b , e , $\theta$ , $\eta$ , $\epsilon_2$ , $\epsilon_4$ , m , L , $A_C$ . These parameters will depend on the type of testing machine, size of soil sample, and the type of soil constituting the soil sample; and therefore the parameters will have to be determined for each type of testing machine. Typical values of the parameters describing a resilient modulus test as described by (39) - (62) are given for lean clay in Table IV. It is possible to determine resilient modulus parameters which are independent of the size of the soil sample and independent of the type of testing apparatus. The parameters $E_0$ , $E_2$ and $E_4$ that occur in (60) - (62) are soil parameters and are independent of the sample size or loading conditions. It is the quantity $E_0$ that must be compared with

the value of $E_S$ determined from the vibratory load-deflection curves that were measured directly on the subgrade. The value of $E_S$ was determined in a manner such that its value is independent of the static and dynamic loads exerted by the vibrator.

The preceding analysis shows that the characteristic shape of the nonlinear dynamic load-deflection curves measured in the field by the WES 16-kip vibrator is due in part to the basic nonlinear response of the material in the subgrade to dynamic loads. The signs of the coefficients describing the resilient modulus test: $\alpha_1 > 0$ , $\alpha_2 > 0$ , $\delta_1 < 0$ , $\delta_2 > 0$ , $b < 0$ , and $e > 0$ , determine to a large extent the signs of the corresponding coefficients determined from the vibratory nondestructive tests conducted on pavements and subgrades. However, inertial, damping and frequency effects will affect the values of $\alpha_1$ and $\alpha_2$ that are determined by vibratory nondestructive testing. For the vibratory nondestructive tests done on pavements and subgrades at 15 Hz, it is generally found that $\alpha_1 > 0$ and $\alpha_2 > 0$ which is in agreement with the signs of the corresponding coefficients describing the resilient modulus laboratory test. For frequencies different from 15 Hz and for exceptional pavement cases it is found that $\alpha_1 > 0$ and $\alpha_2 < 0$ or $\alpha_1 < 0$ and $\alpha_2 > 0$ . Therefore, the combination of the large effective mass associated with a pavement and subgrade, and the relatively high frequency of operation of the WES 16-kip vibrator can produce a dynamic load-deflection curve which has a shape which is considerably different from the shape of the dynamic load-deflection curve measured in the laboratory during a resilient modulus test.

Because of the finite size of the soil sample for the resilient modulus test, the effective mass of the soil sample is, to a good approximation, equal

to the actual mass of the sample. The effective mass that enters the dynamical calculations for the vibratory nondestructive field tests is generally quite large compared to the moving mass of the vibrator because of the large inertial effects associated with the pavement and subgrade. The large effective mass and high frequency of the vibratory nondestructive field tests indicate that the inertial and damping terms are comparable or larger than the elastic effects, $m\omega^2 \sim k$ and $C\omega \sim k$. The relatively small mass of the soil sample used for the laboratory resilient modulus tests and the low frequency at which these tests are conducted suggest that for this case, $m\omega^2 << k$ and $C\omega << k$, and the linear and nonlinear elastic properties are measured directly in this test.

The resilient modulus tests combined with the nonlinear dynamical theory of these tests indicate that the static nonlinear elastic coefficients $b$ and $e$ have the signs $b < 0$ and $e > 0$. It is this basic property of soils that is responsible for making the corresponding coefficients determined from field tests exhibit the same signs. It is the nonzero values of $b$ and $e$ as determined from the resilient modulus that are responsible for the finite depth of influence of the static stress-strain field in the subgrade beneath a static load placed on the pavement surface. The intrinsic nonlinearity exhibited by the soil during the resilient modulus tests is responsible for the finite depth of influence of the static stress-strain field in an actual soil formation.

## V. CONCLUSION

The nonlinear dynamic pavement response model that is presented in this paper gives a quantitative description of the dynamic response of a pavement surface under the action of the dynamic and static load applied to the

pavement surface by the WES 16-kip vibrator. The model parameters - spring constants, effective mass, damping constant and finite depth of influence of the static load have been determined as a function of pavement strength as represented by the measured DSM. The nonlinear pavement response model gives a theoretical expression for the pavement response in terms of these parameters and in terms of the elastic constants of the pavement and subgrade. For a suitable choice of the elastic moduli of the pavement layers, it is possible to predict the value of the subgrade modulus from the dynamic load-deflection curve measured at the pavement surface.

Of much importance to pavement engineers is an estimation of the strength and condition of a subgrade as measured by its subgrade modulus. The nonlinear elastic response model of the dynamic load-deflection curve combined with measured values of this curve is sufficient to determine the subgrade elastic modulus quickly and accurately. This work was funded by the Federal Aviation Administration.

## REFERENCES


1. Hall, J. W., Jr., "Nondestructive Testing of Pavements: Tests on Mulitple-Wheel Heavy Gear Load Sections at Eglin and Hurlburt Airfields," Technical Report No. AFWL-TR-71-64, Mar 1972, Air Force Weapons Laboratory, Kirtland Air Force Base, N. Mex.

2. _____, "Nondestructive Testing of Pavements: Final Test Results and Evaluation Procedure," Technical Report No. AFWL-TR-72-151, Jun 1973, Air Force Weapons Laboratory, Kirtland Air Force Base, N. Mex.

3. Green, J. L. and Hall, J. W., Jr., "Nondestructive Vibratory Testing of Airport Pavements," Vol. I, Evaluation Methodology and Experimental Test Results, Report No. FAA-RD-73-205-I, Sept. 1975, Department of Transportation, Federal Aviation Administration.

4. Weiss, R. A., "Nondestructive Vibratory Testing of Airport Pavements," Vol. II, Theoretical Study of the Dynamic Stiffness and its Application to the Vibratory Nondestructive Method of Testing Pavements, Report No. FAA-RD-73-2-5-II, April 1975, Department of Transportation, Federal Aviation Administration.

5. Tomita, H., "Field NDE of Airport Pavements," Materials Evaluation, Vol. XXXIII, No. 7, July 1975.

6. Richart, F. E., Jr., Hall, J. R. and Woods, R. D., Vibrations of Soils and Foundations, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1970.

7. Jacobsen, L. S. and Ayre, R. S., Engineering Vibrations, McGraw-Hill, New York, 1958.

8. Meirovitch, L., Elements of Vibration Analysis, McGraw-Hill, New York, 1975.

9. Snowden, J. C., Vibration and Shock in Damped Mechanical Systems, John Wiley, New York, 1968.

10. Fredlund, D. G., Bergan, A. T. and Sauer, E. K., "Deformation Characterization of Subgrade Soils for Highways and Runways in Northern Environments," Canadian Geotechnical Journal, Vol. 12, May 1975, pg. 213.

TABLE I

NUMERICAL RESULTS FOR FREQUENCY RESPONSE METHOD

| LOCATION | DSM kips/in | $m/m_v$ | C $10^4$lb·sec/in | k kips/in | $E_S$(CHEVRON) $10^3$psi | $E_S$(CBR) $10^3$psi |
|---|---|---|---|---|---|---|
| B2 | 700 | 1.7 | 1.0 | 2137 | 65 | 21 |
| N18 | 770 | 2.0 | 0.8 | 1500 | 58 | 27 |
| W1 | 860 | 1.8 | 0.4 | 2620 | 136 | 30 |
| B3 | 1630 | 2.0 | 1.1 | 2140 | 35 | 25 |
| W2 | 1940 | 2.4 | 1.3 | 2470 | 69 | 30 |
| P14 | 2120 | 2.5 | 1.5 | 2610 | 139 | 30 |
| P13 | 2780 | 4.4 | 2.0 | 3500 | 153 | 30 |
| B1 | 3120 | 10.0 | 2.8 | 4270 | 140 | 21 |

TABLE II

INPUT OF WES NONLINEAR DYNAMIC PROGRAM

SITE B2A

DSM = 700 kips/in

| $F_D$ kips | A in. |
|---|---|
| 0 | 0.0 |
| 2 | 0.003 |
| 4 | 0.007 |
| 6 | 0.011 |
| 8 | 0.015 |
| 10 | 0.020 |
| 12 | 0.025 |
| 14 | 0.030 |

TABLE III

RESULTS OF WES NONLINEAR DYNAMIC PROGRAM

| SITE | DSM kips/in | SUBGRADE CBR | $E_S$(CBR) $10^3$psi | $E_S$(WES NONLINEAR) $10^3$psi |
|---|---|---|---|---|
| TETS | 450 | 8 | 12.0 | 13.0 |
| B2A | 700 | 14 | 21.0 | 22.8 |
| N18 | 770 | 18 | 27.0 | 25.9 |
| WES-AC | 780 | 4 | 6.0 | 6.7 |
| W1 | 860 | 20 | 30.0 | 18.8 |
| N23A | 980 | 18 | 27.0 | 28.1 |
| B3 | 1680 | 17 | 25.5 | 11.1 |
| W2C | 1940 | 20 | 30.0 | 35.5 |
| P14A | 2120 | 20 | 30.0 | 13.7 |
| P13 | 2780 | 20 | 30.0 | 17.7 |
| B1 | 3120 | 14 | 21.0 | 9.0 |
| WES-PCC | 3500 | 4 | 6.0 | 6.8 |

TABLE IV

PARAMETERS DESCRIBING THE DYNAMIC CHARACTERISTICS
OF THE RESILIENT MODULUS LABORATORY TEST

| | | | | | |
|---|---|---|---|---|---|
| $A_C$ | $in^2$ | 6.16 | $\delta_1$ | $lb^2/in^4$ | $-5.66 \times 10^{13}$ |
| L | in | 6.0 | $\delta_2$ | $lb^4/in^8$ | $1.45 \times 10^{27}$ |
| W | lb | 5.0 | $\alpha_1$ | $lb^2/in^4$ | $1.89 \times 10^{13}$ |
| m | lb $sec^2/in$ | 0.013 | $\alpha_2$ | $lb^4/in^8$ | $3.5 \times 10^{26}$ |
| $\omega$ | $sec^{-1}$ | 6.0 | $\beta_1$ | $lb^2/in^4$ | $-1.89 \times 10^{13}$ |
| $m\omega^2$ | lb/in | 0.468 | $\beta_2$ | $lb^4/in^8$ | $7.2 \times 10^{24}$ |
| C | lb sec/in | 30.0 | $\theta$ | dimensionless | 30.0 |
| $C\omega$ | lb/in | 180.0 | $\eta$ | dimensionless | 50.0 |
| $k_{00}$ | lb/in | $1.5 \times 10^4$ | $\varepsilon_2$ | dimensionless | 31.0 |
| $k_0$ | lb/in | $4.0 \times 10^4$ | $\varepsilon_4$ | dimensionless | 54.0 |
| k | lb/in | $4.0 \times 10^4$ | $E_0$ | $lb/in^2$ | $1.5 \times 10^4$ |
| b | $lb/in^3$ | $-2.0 \times 10^7$ | $E_2$ | $lb/in^4$ | $-2.0 \times 10^9$ |
| e | $lb/in^5$ | $3.6 \times 10^{11}$ | $E_4$ | $lb/in^6$ | $9.7 \times 10^{13}$ |

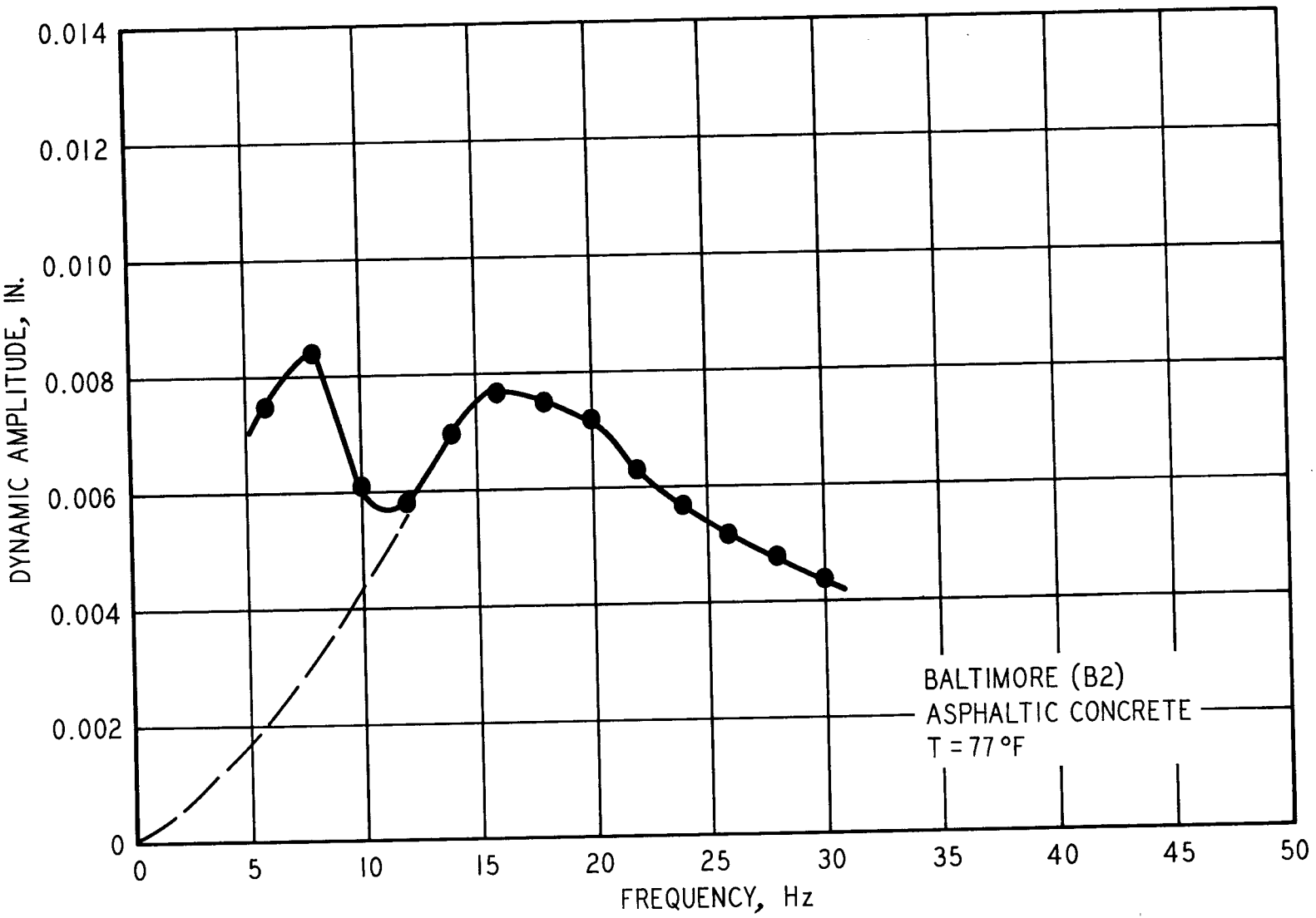



Figure 1. Typical frequency response curve

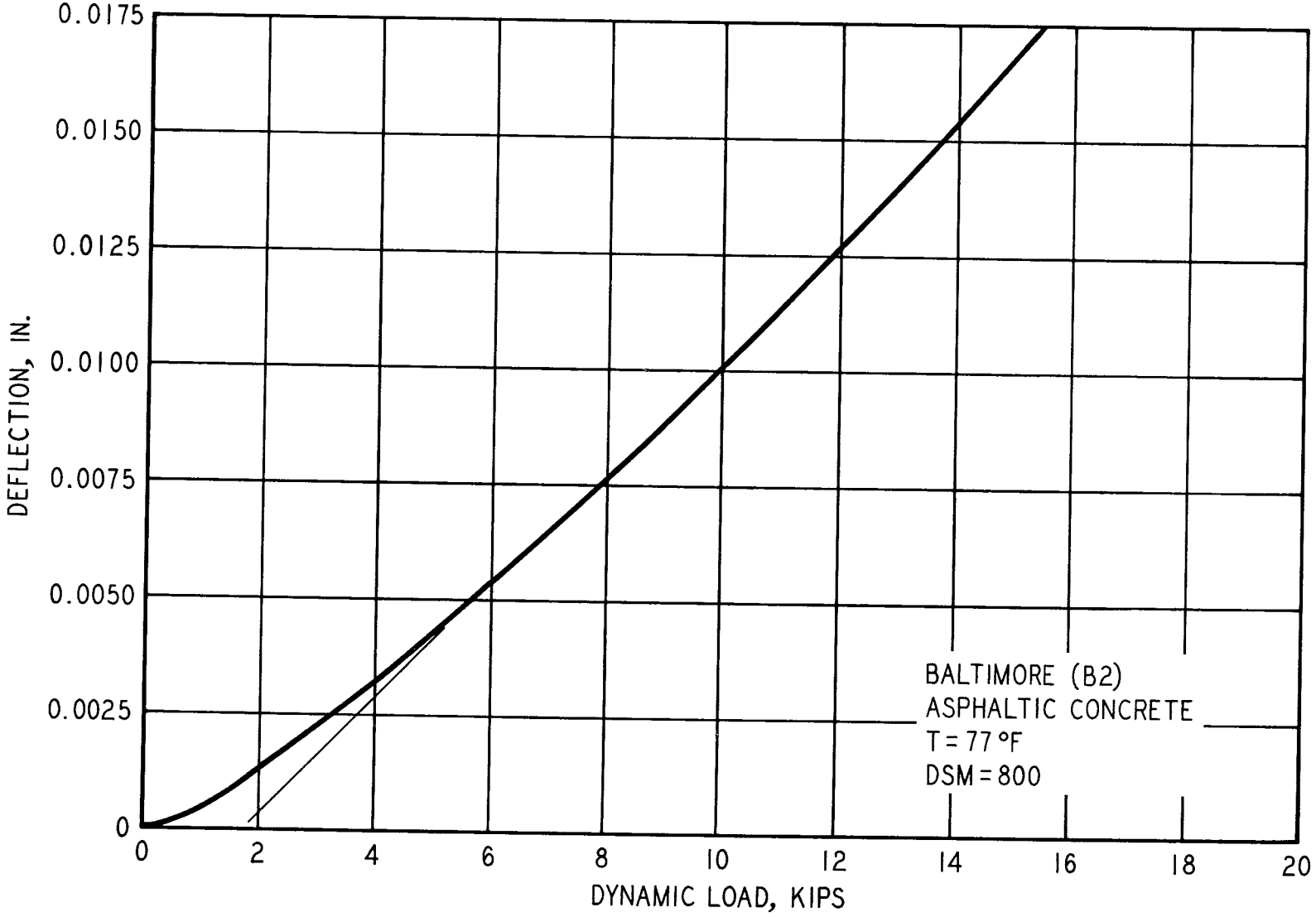



Figure 2. Typical dynamic load-deflection curve

## DYNAMIC PAVEMENT RESPONSE MODELS

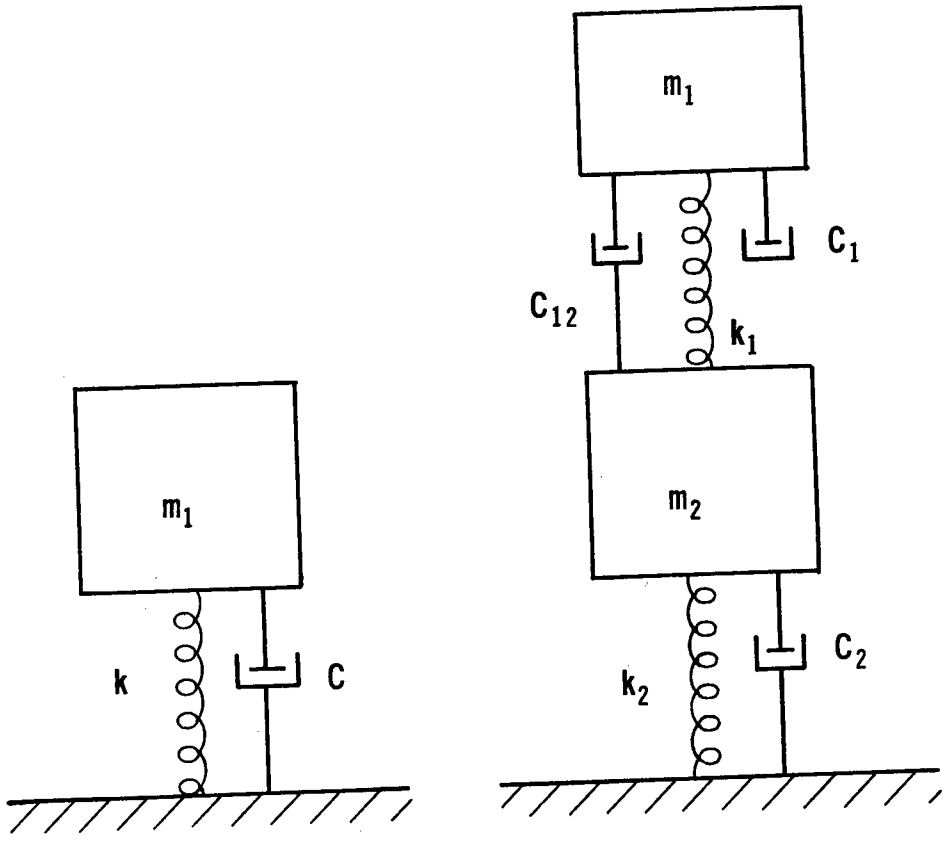
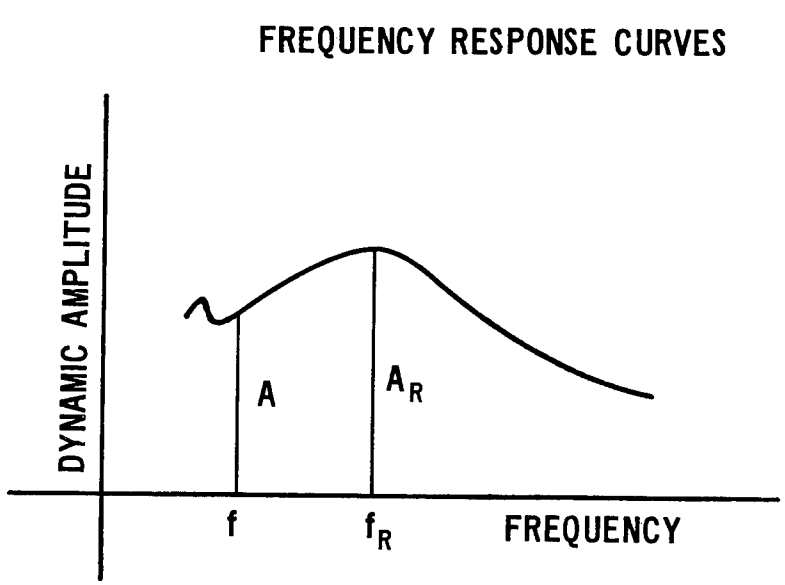



Figure 3. Single and double mass dynamic pavement response models




MEASURED QUANTITIES: $f_R$ = RESONANCE FREQUENCY

$A_R$ = DYNAMIC AMPLITUDE AT RESONANCE

$J(f) = A_R/A$ = RATIO OF AMPLITUDE AT RESONANCE TO AMPLITUDE AT ARBITRARY FREQUENCY.

Figure 4. Measured quantities obtained from frequency response curves

# FREQUENCY RESPONSE METHOD

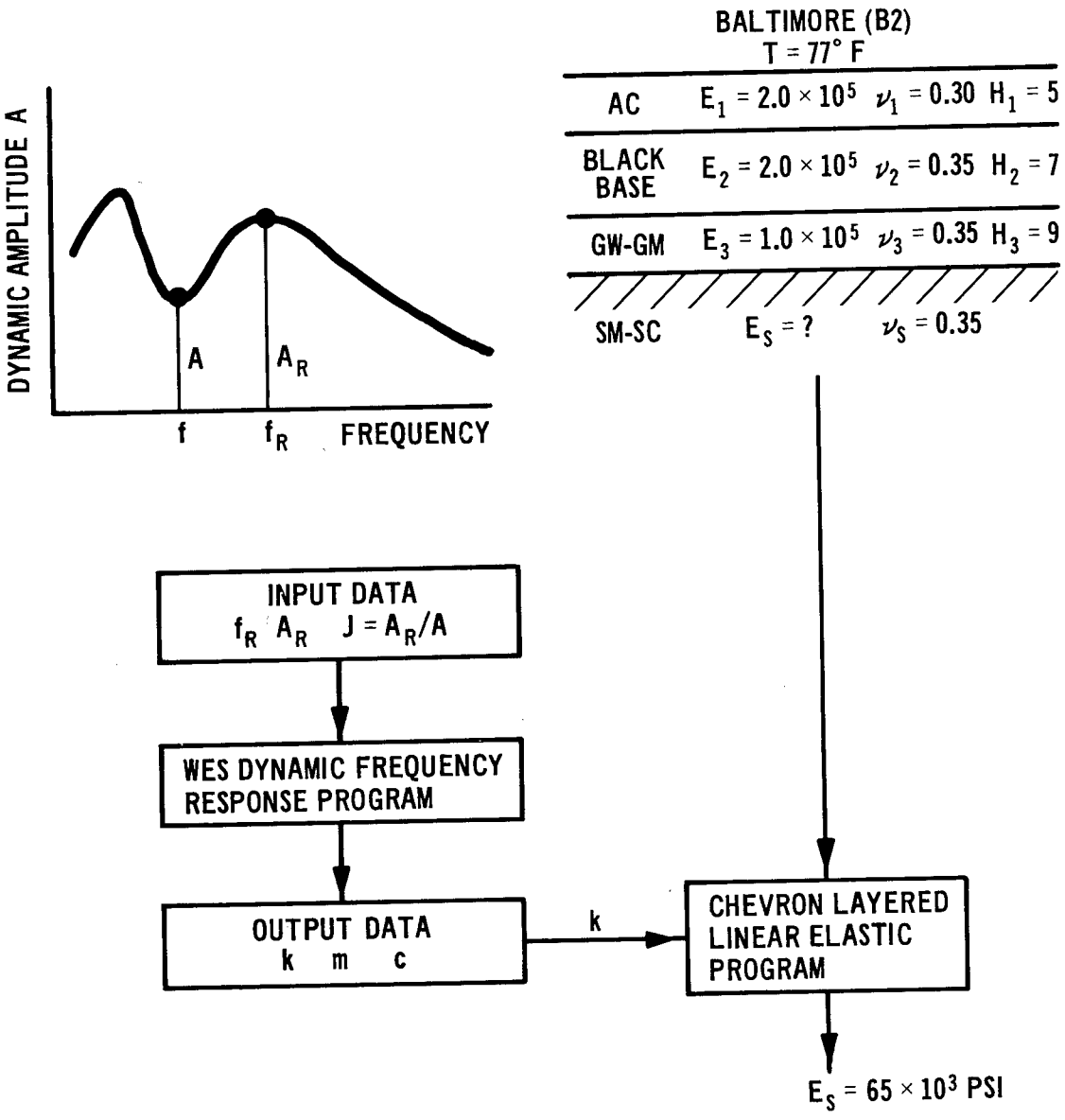



Figure 5. Method of calculating subgrade modulus from measured frequency response curves

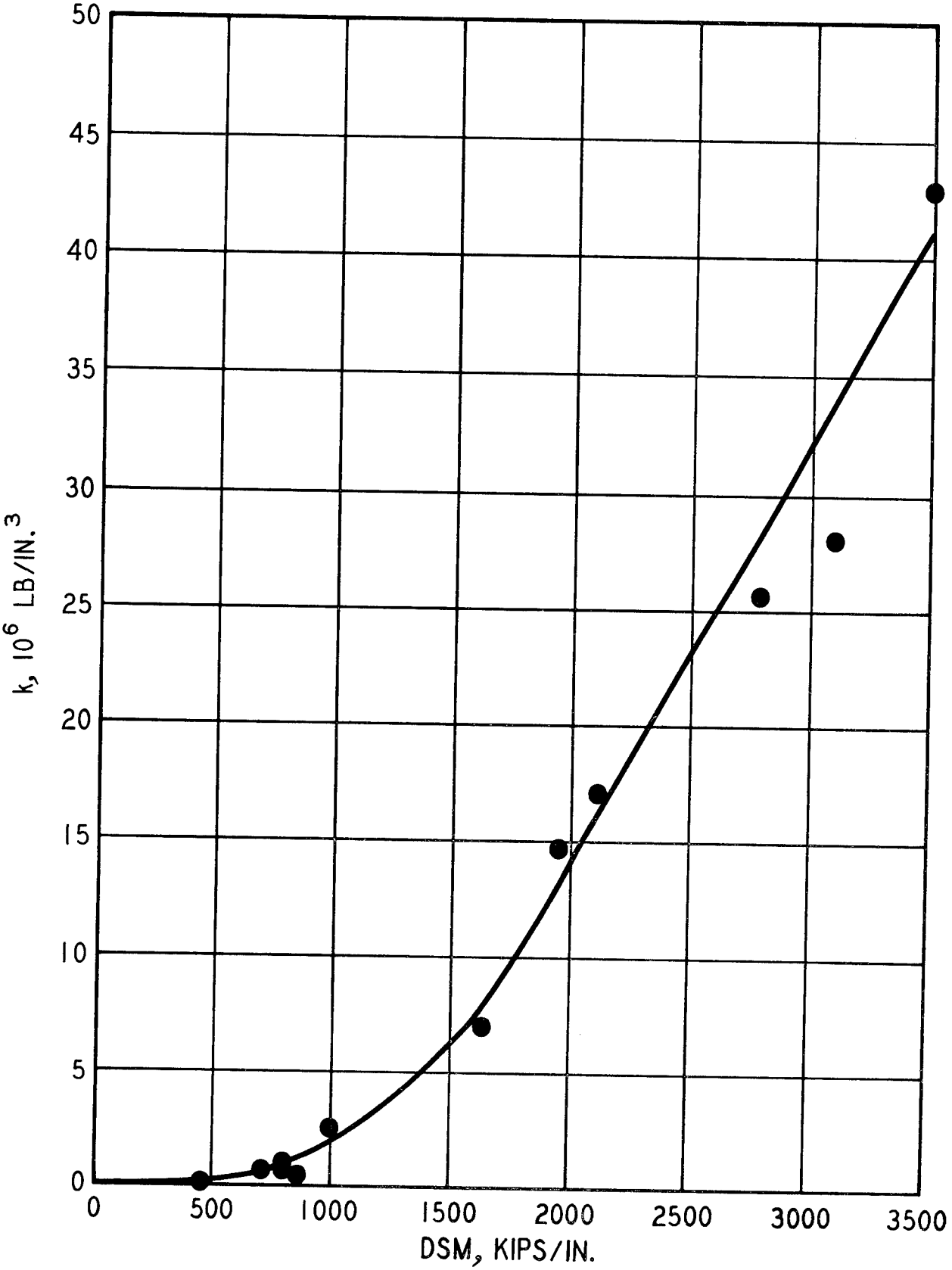



Figure 6. Dynamic spring constant versus measured DSM

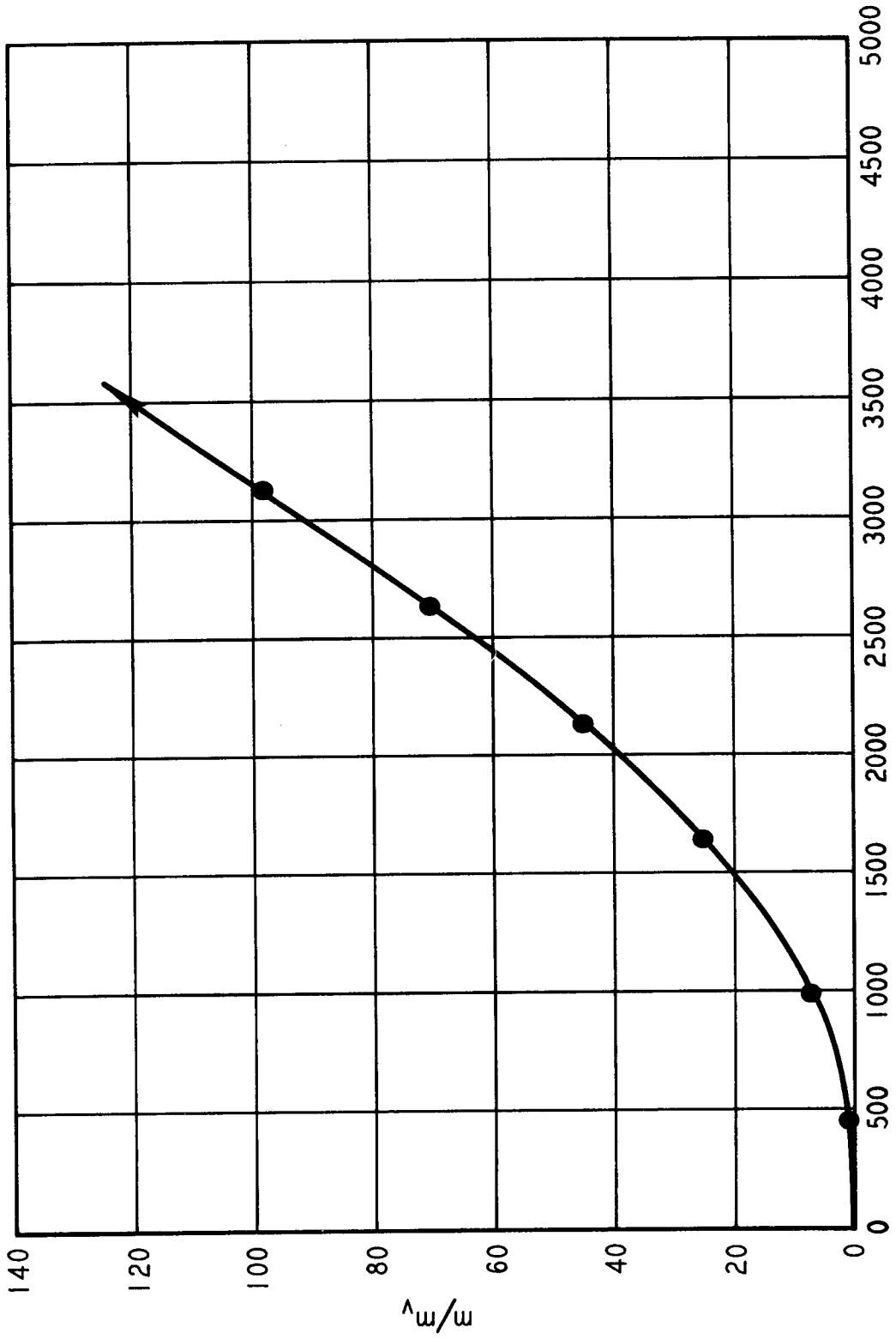



Figure 7. Effective mass ratio versus measured dynamic stiffness

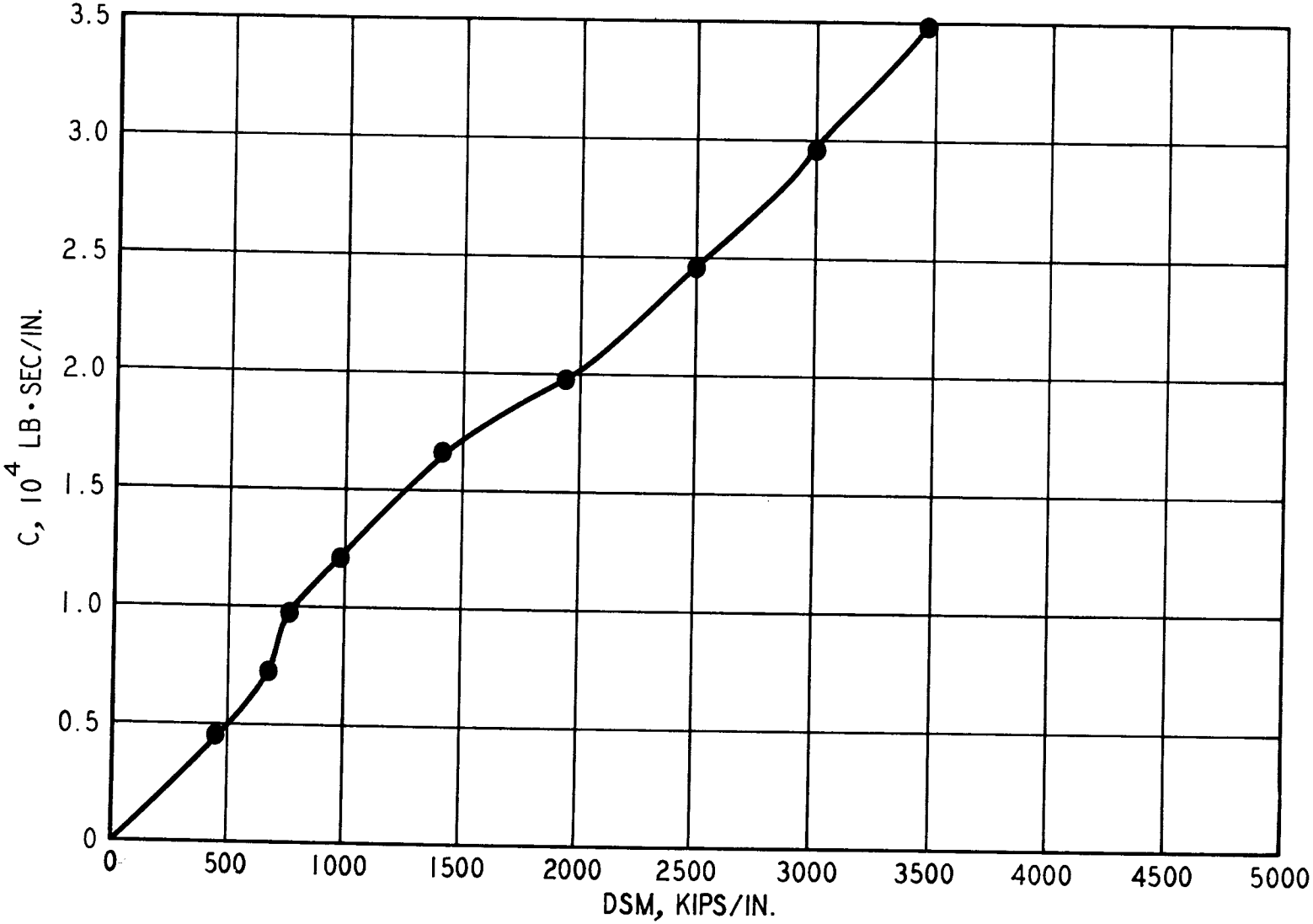



Figure 8. Damping constant versus measured values of DSM

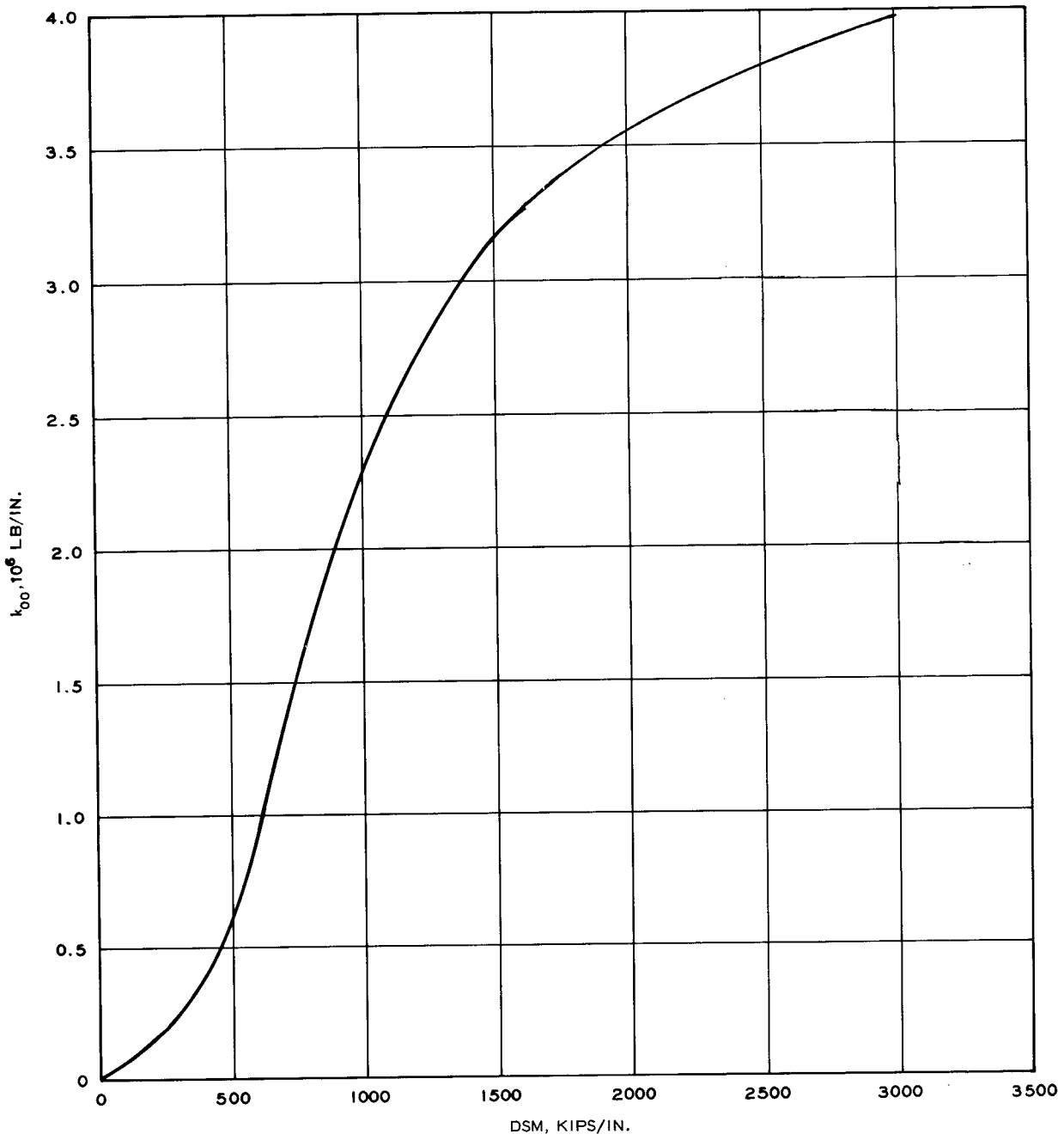



Figure 9. Linear static spring constant versus DSM value

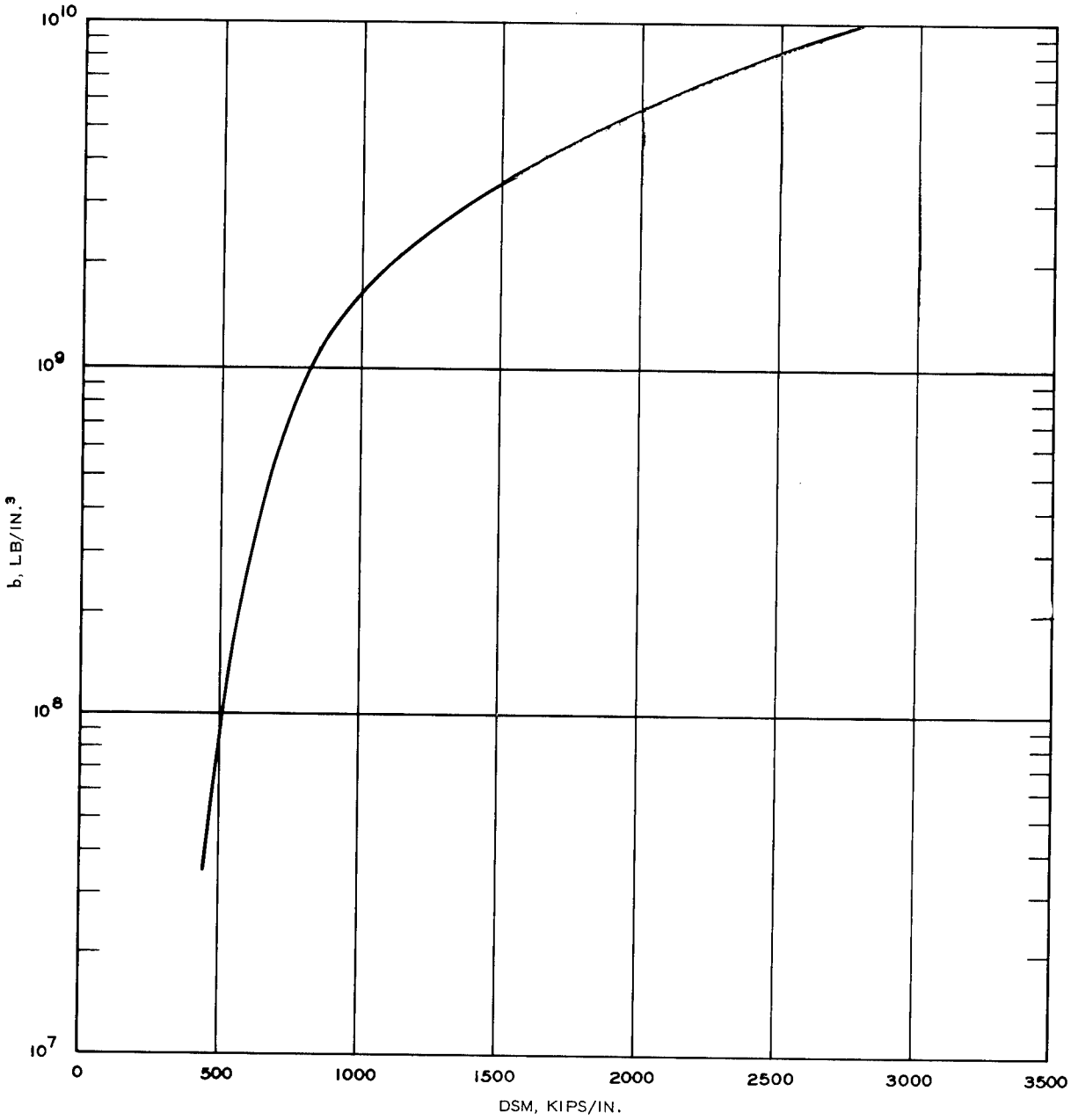



Figure 10. Third order static nonlinear parameter versus DSM value

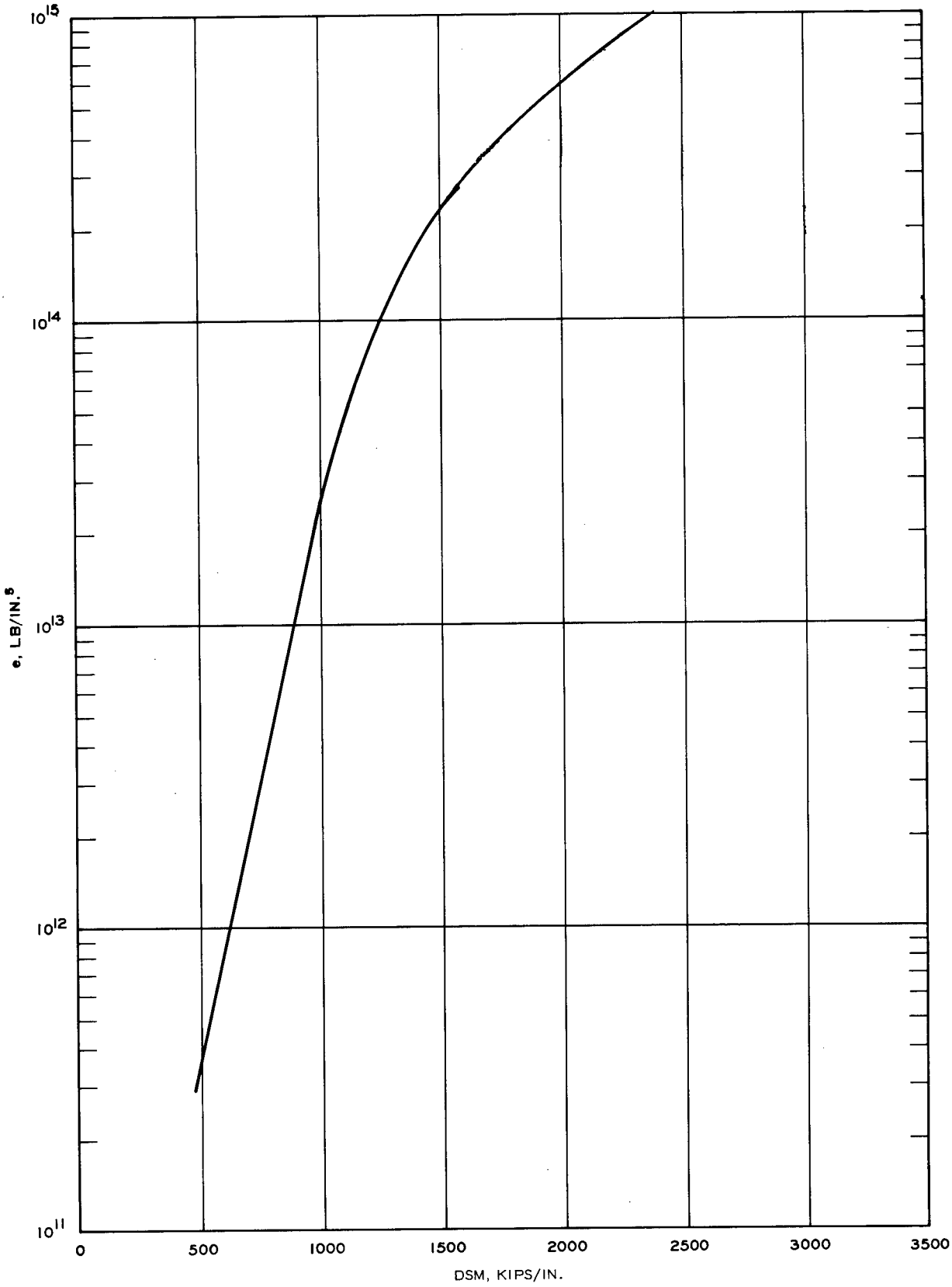



Figure 11. Fifth order static nonlinear parameter versus DSM value

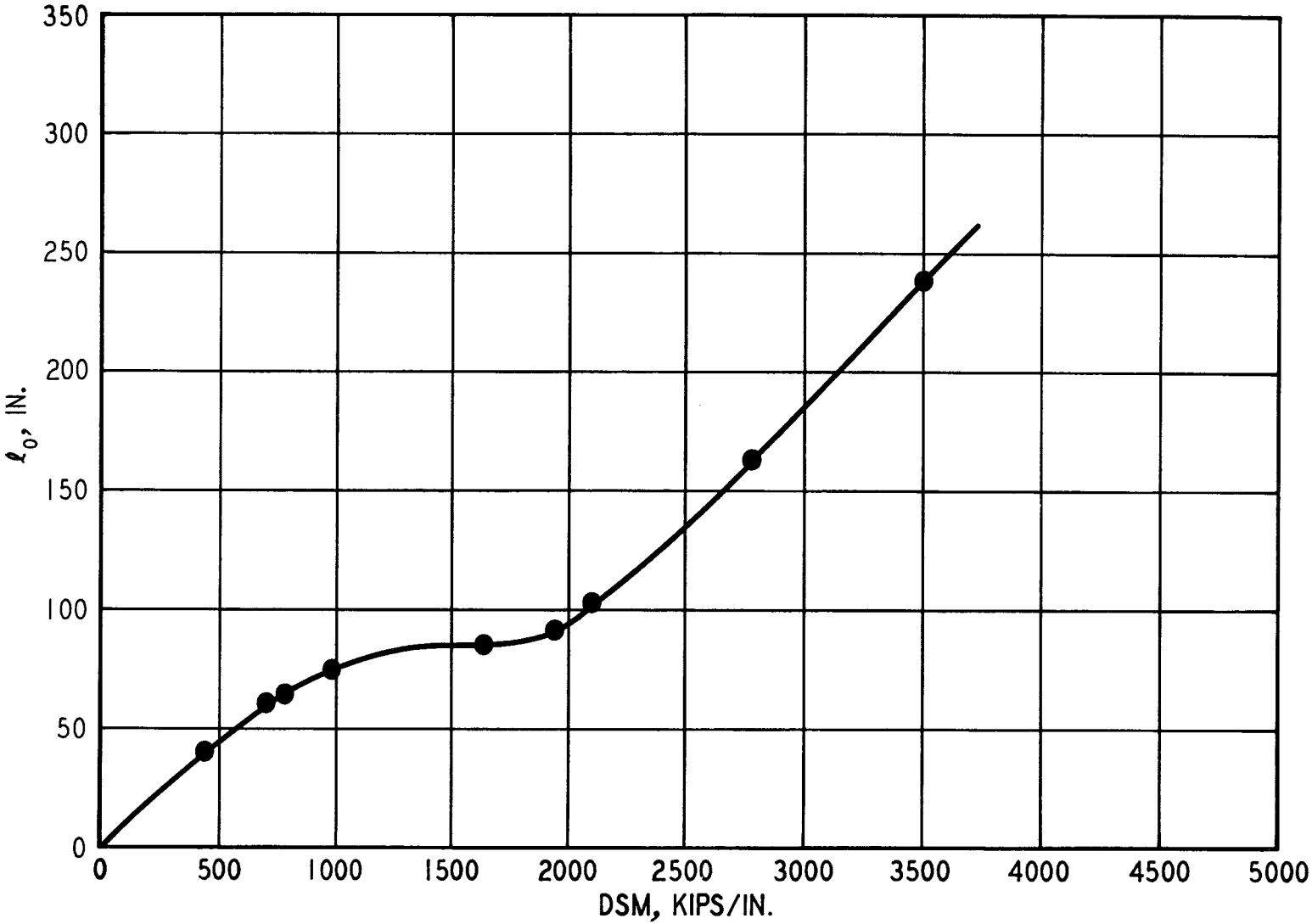



Figure 12. Leading term of the finite depth of influence versus measured values of DSM

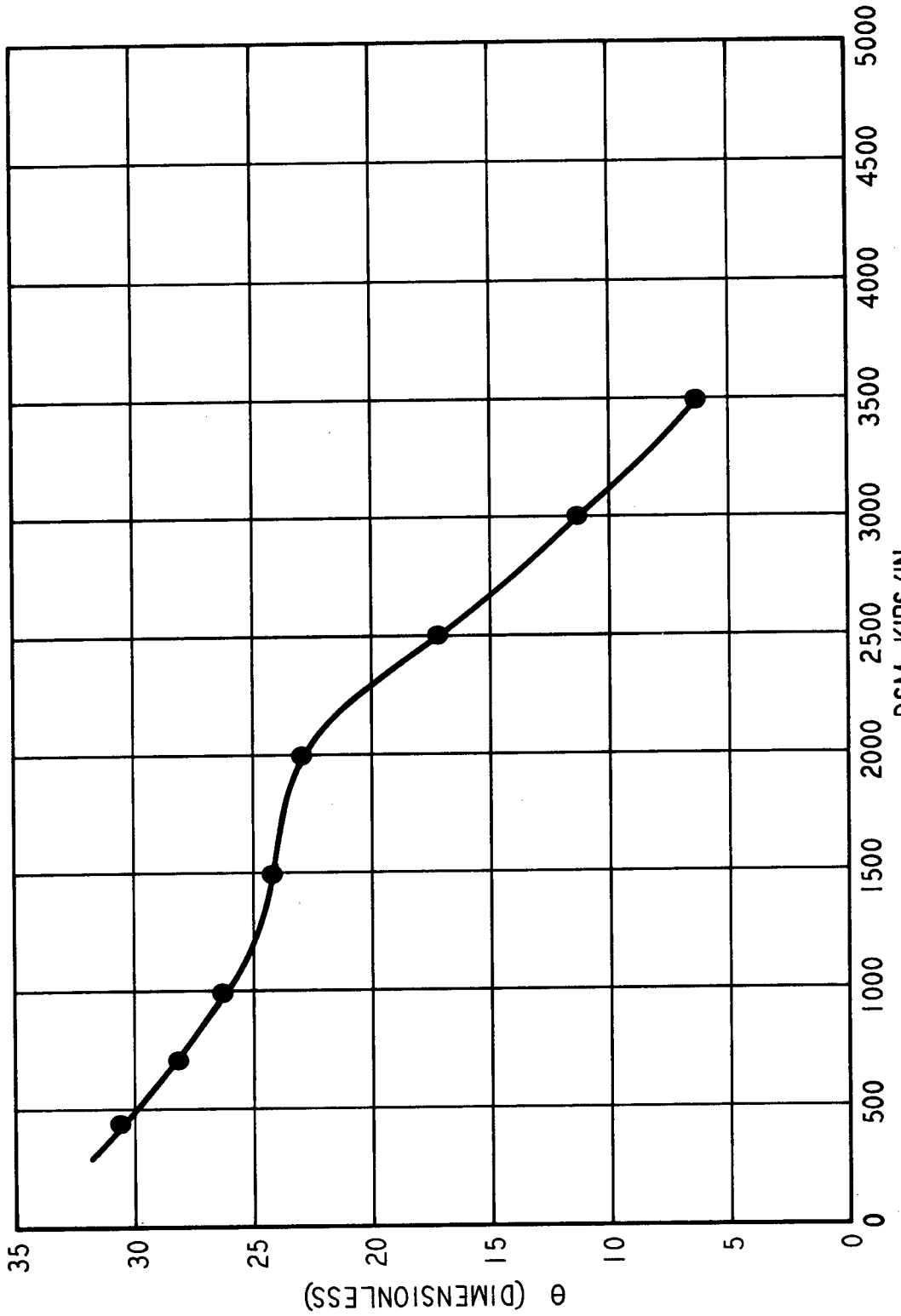



Figure 13. Dimensionless parameter θ versus measured values of values of DSM

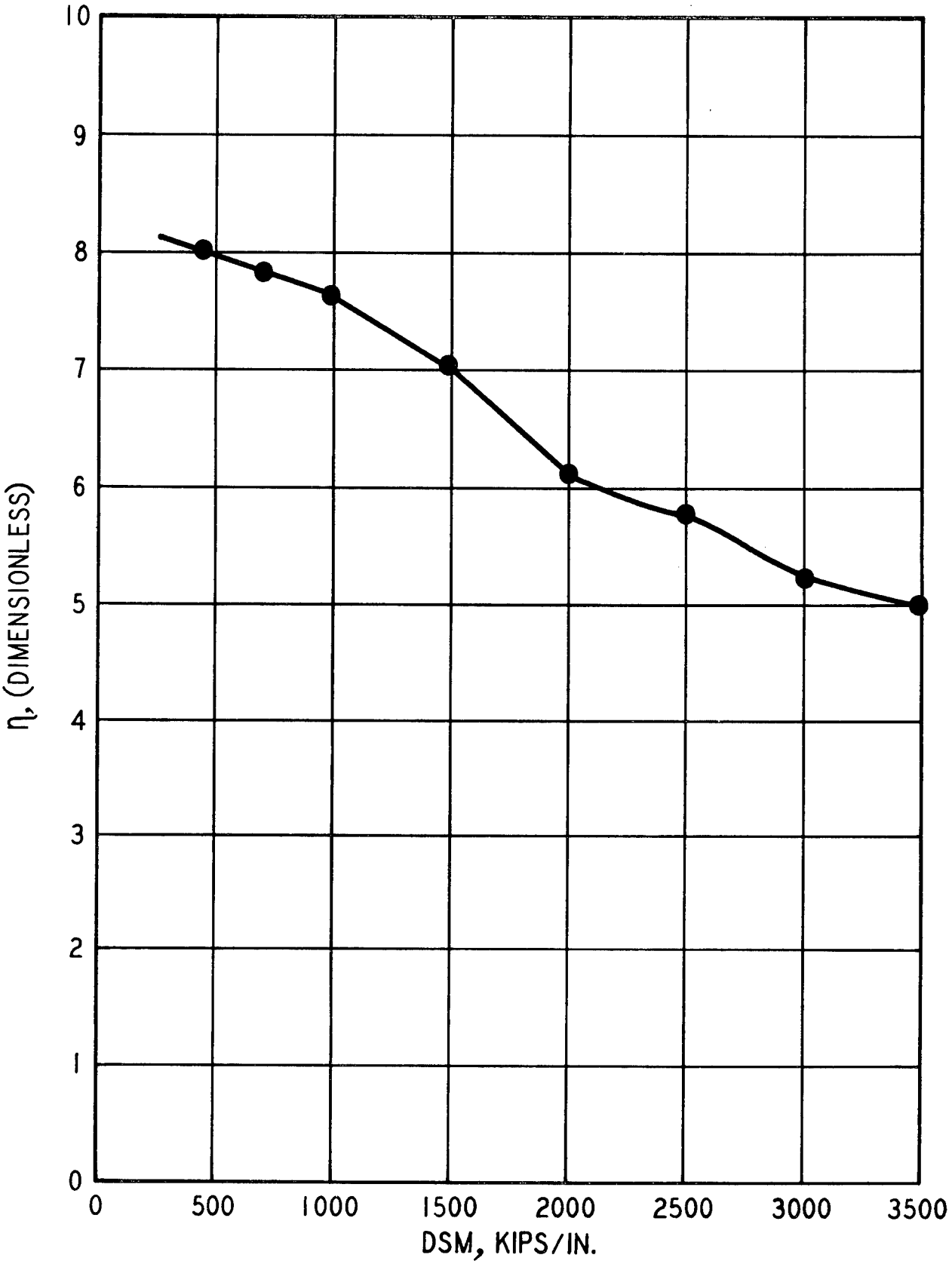



Figure 14. Dimensionless parameter $\eta$ versus measured DSM values

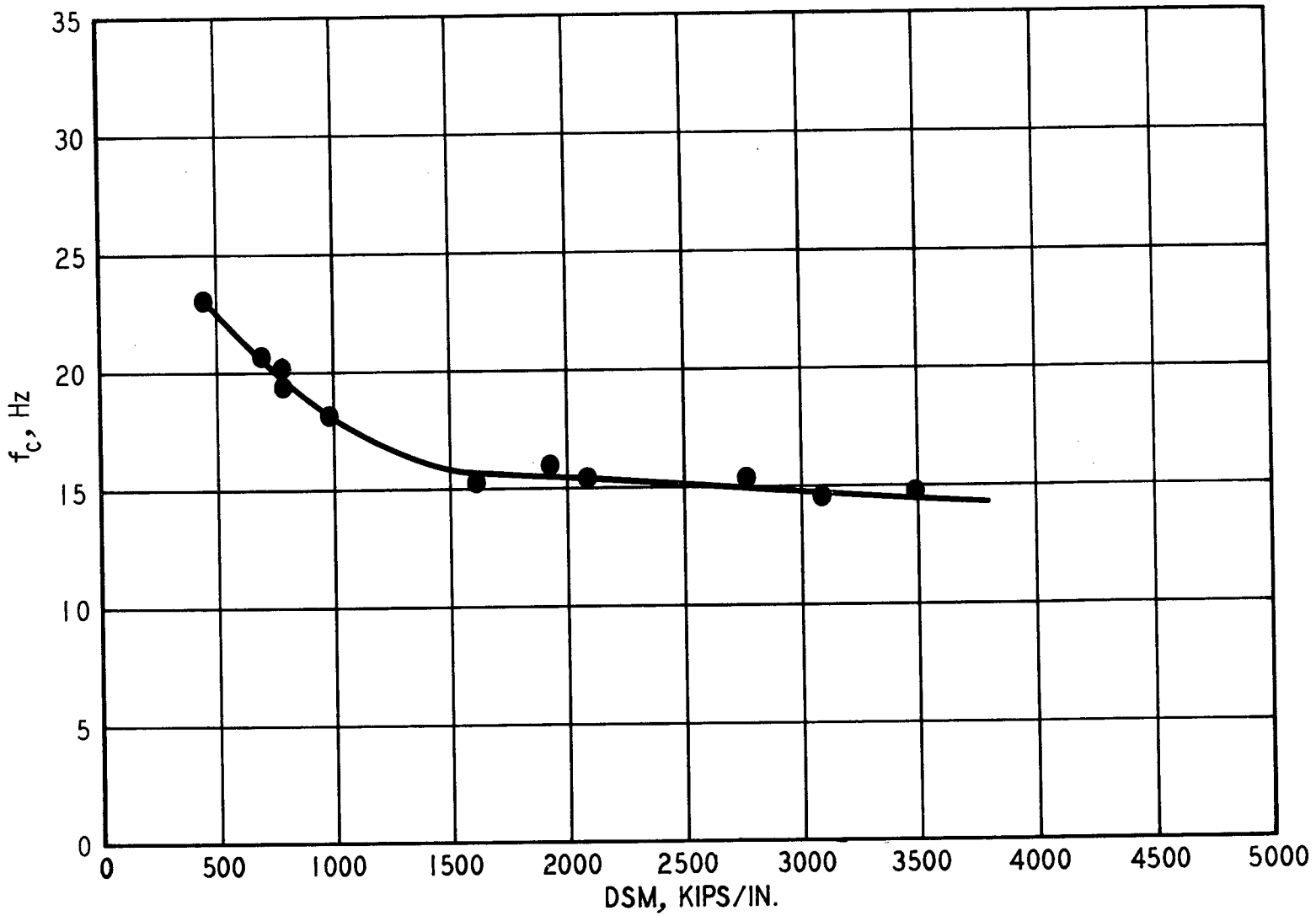



Figure 15. Critical frequency versus measured DSM

# DYNAMIC LOAD-DEFLECTION METHOD

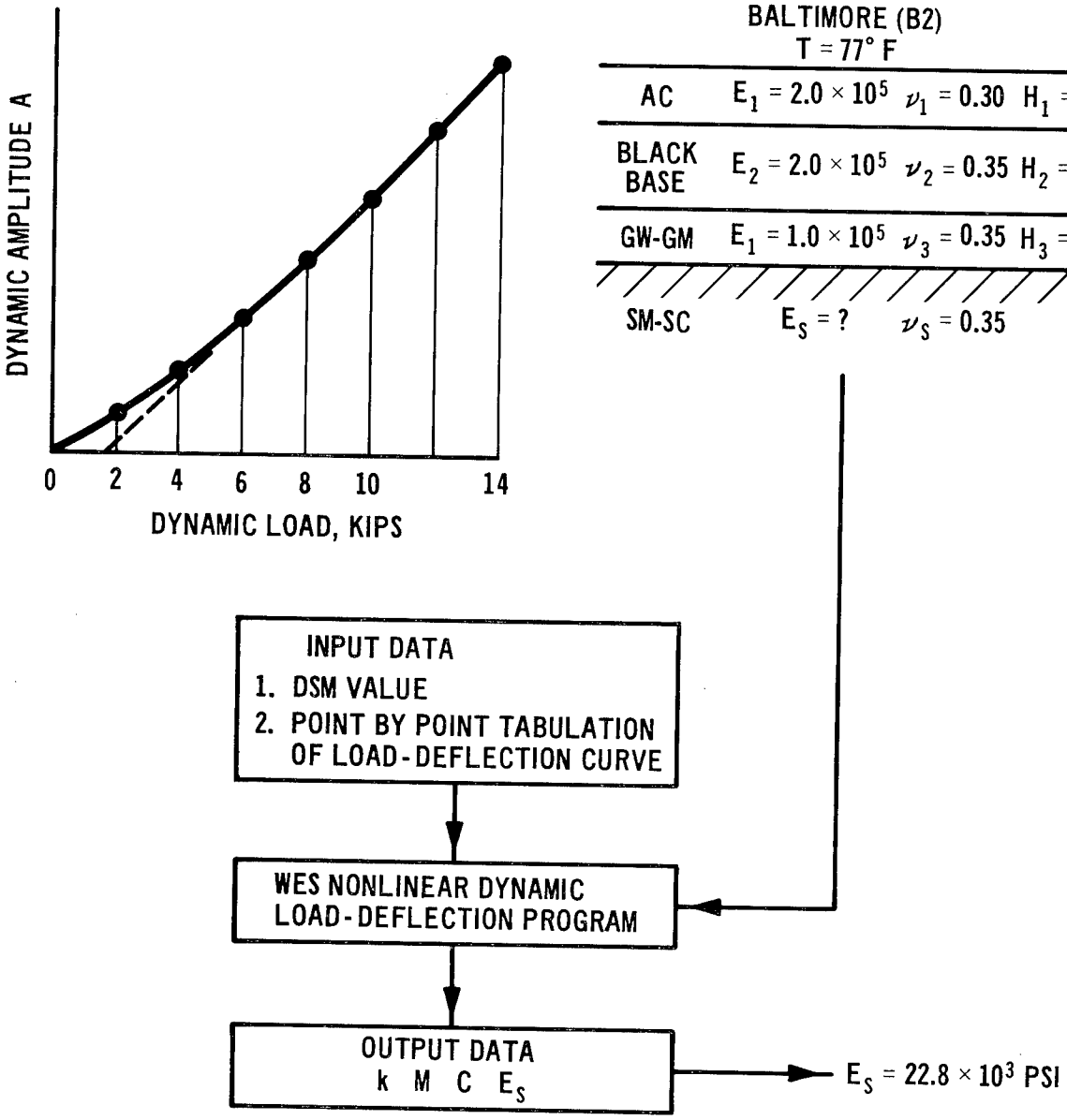



Figure 16. Determination of subgrade modulus from measured dynamic load-deflection curves

# CHARACTERIZATION OF BEHIND ARMOR EFFECTS FOR LONG ROD PENETRATORS

Victor D. Maki
Engineering Branch
Ballistic Modeling Division
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

ABSTRACT. This study was needed to provide information on the behind armor effects essential to armored vehicle analysis and in the design of future kinetic energy penetrators. Both spall and rod penetrator fragment data was examined for gross characteristic statistical trends. Use of least squares was employed to ascertain causes for similarities in the data base. A linear function relating fragment mass to velocity was employed to study effects of variation in projectile materials, initial projectile weights, striking velocities, length to diameter ratios and plate thicknesses. Kolmogorov-Smirnov type test statistics were used to determine whether or not a unique parent weight distribution existed between various firings. The Weibull, Poisson and Truncated Normal cumulative distribution functions were also compared with empirical weight distributions for several selected firings. This paper summarizes the characteristics found.

1. INTRODUCTION. Whenever armored vehicles of any kind are attacked by metal rod penetrators, fragments are sprayed inside the vehicle which damage components and personnel. To facilitate a greater understanding of those mechanical processes involved, a gross characterization was done that includes fragment numbers, mass distributions, and spatial locations behind 6.35 and 12.7 millimeter rolled homogenous steel targets. The fragment data base used for this analysis is comprised of 140 test firings completed at the BRL in 1970. In the data base projectile weights, length to diameter ratios, projectile material types,target plate thickness, fragment masses, fragment locations, (see Figure 1), and velocities were found recorded in a BRL Memorandum Report[1]. This data was transcribed onto IBM punched cards for computer reduction and analysis. Initially, the natural log of fragment mass and velocity were fitted with a first degree polynomial of the form, $y = a_0 + a_1 x$ where y denotes the ln velocity parameter and x the ln of fragment mass. This polynomial fitting and plotting technique was later published by the author as a BRL Systems

[1] L. Herr and C. Gabarek, "Ballistic Performance and Beyond Armor Data for Rods Impacting Steel Armor Plates," US Army Ballistic Research Laboratories Memorandum Report #2575, January 1976.

Programming Bulletin[2]. The volume of linear equations and plots produced was found to be valuable as a convenient index in a search for trends from firing to firing which later led to a zone analysis of the data. A zone definition and the results of the zone analysis are on the following page.

Zone number 1 is represented by the innermost circle on the recovery media surface and is measured by an angle of ten degrees with respect to the shotline. Zones 2-thru-5 are defined by an angle increase of ten degrees per zone. For all firings the shotline was orthogonal to the target surface plane. Projectile striking velocities were in the 900 to 1500 meters/sec range. The straight line function, ln (fragment velocity) = $a_0 + a_1$ (ln fragment mass) when fitted on a zone per zone basis revealed a distinct trend. As zone angle increased, the slope values, $a_1$'s were more negative in value. This agrees with the basic conservation of energy law of physics.

2. The Weibull Distribution Function. In a testing of the Poisson, Truncated Normal, and Weibull distribution functions the latter provided the best fit to the fragment mass parameter. A detailed report of how the Weibull distribution function parameters were estimated can be found in Reference 3. A two sided Kolmogorov-Smirnov type test was employed as a criteria for best fit. The empirical cumulative distribution function, (Equation 1) was computed for the fragment mass parameters for several selected firings. A graphical and numerical comparison with the Weibull cumulative distribution function, (Equation 2) was then performed.

$$F_N(x) = \begin{cases} 0, & x < x_{(1)} \\ K/N, & x_{(K)} \leq x < x_{(K+1)} \\ 1, & x_{(N)} \leq x \end{cases} \qquad (1)$$

$$F(x) = 1.0 - \exp - \left(\frac{x}{\hat{\theta}}\right)^{\hat{\gamma}} \qquad (2)$$

x, is the fragment mass parameter.

[2]Victor D. Maki, "PDT Plot Subroutine with Bi-variate Analysis," US Army Ballistic Research Laboratories Systems Programming Bulletin #SPB-G-74, 17 July 1974.

[3]Victor D. Maki, "Three Probability Density Function FORTRAN Subroutines," US Army Ballistic Research Laboratories Interim Memorandum Report #396, June 1975.

The computed maximum absolute difference was numerically compared with, $1.36/\sqrt{N}$ which is fully described in Reference 4. If the computed maximum absolute difference was found to be less than the above statistic, a decision was made to accept the Weibull distribution function for describing fragment mass. Included in this paper is a plot of this type test for round number 5 (see Figure 2). For "good" fitting of the Weibull distribution function to a large number of firings fragment masses greater than 100 grains should be ignored. Because rod penetrator fragment mass distributions are characteristically bi-modal, more than 90 percent of the fragments can be found in the first mode and therefore, for this data set, ignoring the second mode caused no significant loss in accuracy.

## 3. ZONE DEFINITION.

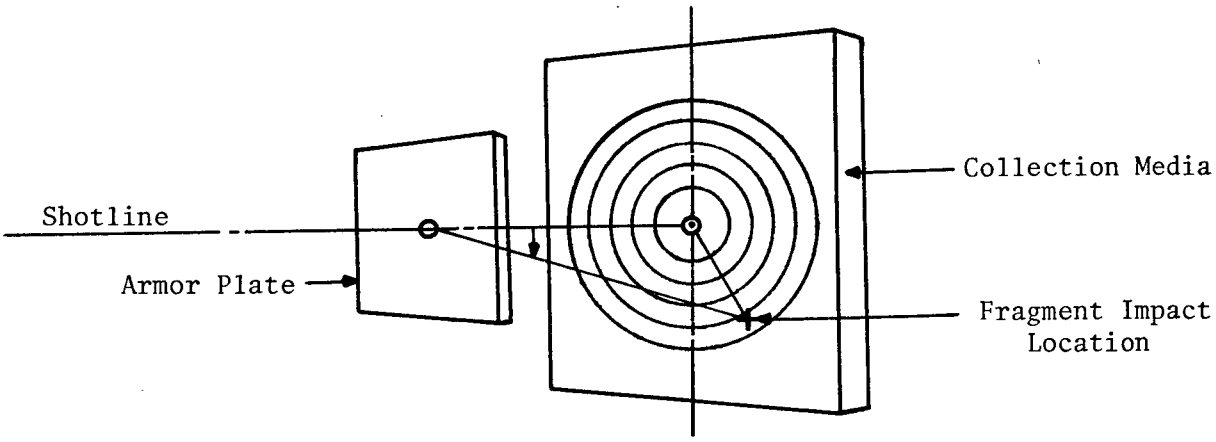



Figure 1

Zone 1 is represented by the innermost circle and forms an angle of ten degrees measured from the shotline. Zones 2-thru-5 are defined by an angle increase of ten degrees per zone.

[4] W. J. Conover, Practical Non-Parametric Statistics, published by John Wiley, 1971, New York, NY.

4. TWO-SIDED KOLMOGOROV-SMIRNOV TYPE TEST

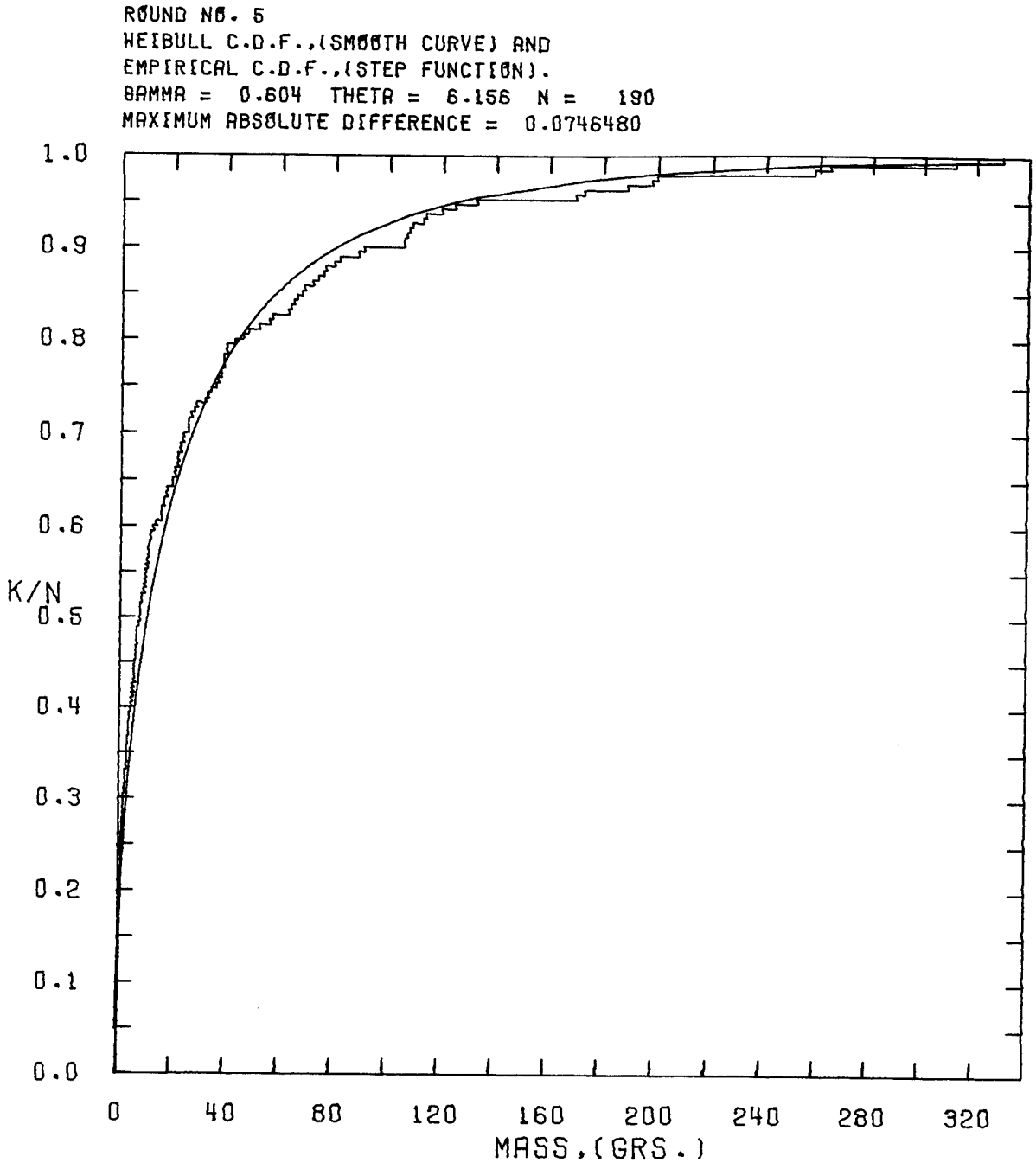



Figure 2

$1.36/\sqrt{N}$ = .09866477

Since the maximum absolute difference is smaller than the above value, the Weibull distribution is accepted for this firing.

The fitting of the Weibull distribution to spall fragment mass is documented in Reference 5. Further interest on the readers part on this topic should be directed to Mr. John Misey, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland 21005.

## 5. CONCLUSIONS OF ANALYSIS.

a. As zone angle increases, average fragment velocity decreases, numbers of fragments decrease and average mass increases.

b. A two-sided Kolmogorov-Smirnov type test shows the Weibull distribution function is a good choice for describing fragment mass less than 100 grains.

c. Rod penetrator fragment mass distributions are characteristically bi-modal.

---

[5] "Behind Armor Data for Long Rod Penetrators," paper presented by Mr. John Misey at the Second Annual Automatic Cannon Caliber Munitions Symposium, 25 September 1975 at Frankfort Arsenal.

## REFERENCES


1. L. Herr and C. Grabarek, "Ballistic Performance and Beyond Armor Data for Rods Impacting Steel Armor Plates," US Army Ballistic Research Laboratories Memorandum Report #2575, January 1976.

2. Victor D. Maki, "PDT Plot Subroutine with Bi-variate Analysis," US Army Ballistic Research Laboratories Systems Programming Bulletin #SPB-G-74, 17 July 1974.

3. Victor D. Maki, "Three Probability Density Function FORTRAN Subroutines," US Army Ballistic Research Laboratories Interim Memorandum Report #396, June 1975.

4. W. J. Conover, Practical Non-Parametric Statistics, published by John Wiley, 1971, New York, NY.

5. "Behind Armor Data for Long Rod Penetrators," paper presented by Mr. John Misey at the Second Annual Automatic Cannon Caliber Munitions Symposium," 25 September 1975 at Frankfort Arsenal.


## BIBLIOGRAPHY


1. A. Clifford Cohen, Maximum Likelihood Estimation in the Weibull Distribution Based on Complete and on Censored Samples, Technometrics Magazine, Vol. 7, No. 4, 1865, pp. 579-588.

2. John E. Freund, Mathematical Statistics, published by Prentice-Hall, Incorporated, Englewood Cliffs, NJ, 1962.

3. W. T. Eadie, D. Drijard, F. E. James, M. Roos, and B. Sadoulet, Statistical Methods in Experimental Physics, published by North-Holland Publishing Company, Amsterdam and London, 1971.

4. Paul R. Rider, An Introduction to Modern Statistical Methods, published by John Wiley and Sons, Incorporated and London, Chapman and Hall, Limited, 1939.

5. Herbert Arkin and Raymond R. Colton, Statistical Methods, fifth edition, published by Barnes and Noble Books, A Division of Harper and Row, Publishers New York, Evanston, San Francisco, London, 1970.

# MATHEMATICAL MODELS OF SYSTEMS AND TACTICS IN LAND COMBAT

Roger F. Willis

US Army TRADOC Systems Analysis Activity
White Sands Missile Range, New Mexico 88002

ABSTRACT. This paper covers a variety of mathematical models that have recently been developed, tailored to specific decision problems in tactics and alternative system tradeoffs. These models emphasize rapid and flexible variation of assumptions, investigation of alternative tactics, tradeoffs between system parameters, tradeoffs between the elements of a force and various optimizations. Alternative mathematical formulations include linear versus non-linear, constant versus time-varying coefficients and stochastic versus deterministic.

## 1. INTRODUCTION

Flexible and efficient mathematical models are required for use in different phases of a particular force evaluation or combat developments study. In an early phase these models can be used to compare and screen alternatives -- alternative systems, alternative tactics or alternative mixes. In late phases the same models can be used for sensitivity analysis, to give approximate answers to "what if" questions -- i.e., to determine how study results might change if certain assumptions are varied. In most studies the major analytical tool will be a large, relatively slow and expensive computer model or simulation or computer-assisted wargame (e.g., DIVWAG). The mathematical models presented in this paper are intended to supplement the large models, to provide additional insights and to enrich the study results. These models can also be used to develop hypotheses (e.g., about the relative merits of alternative tactics) that can then be tested with high resolution stochastic simulations.

2. We consider the class of models consisting of sets of ordinary differential equations in which each equation represents the time rate of change of the number remaining of a particular type of weapon. The equations can be deterministic or stochastic, linear or non-linear, with constant coefficients or variable coefficients. We consider ten specific decision problems and the mathematical categories of models as follows:

a. deterministic, linear, constant coefficients

(1) tradeoff between ground forces and aircraft

(2) remotely piloted vehicles

(3) air defense suppression

(4) optimum artillery mix

b. deterministic, linear, variable coefficients

(1) antiarmor target priorities

(2) optimum disengagement time

(3) electronic warfare

c. deterministic, non-linear

(1) weapon effectiveness

(2) force required

d. stochastic

(1) time to achieve goal

3. More complete statements of the decision problems are:

a. To what extent can tanks be traded off for close support aircraft?

b. How many remotely piloted vehicles are required to support one maneuver battalion?

c. How should artillery fire be allocated between counterbattery fire and suppression of air defense?

d. What is the optimum mix of artillery types and numbers for supporting a mechanized infantry division?

e. How should tank fire be allocated between three or four types of anti-armor weapons?

f. What is the optimum time for a defending anti-armor force to disengage?

g. What is the payoff from detecting enemy surveillance systems and attacking or jamming them?

h. What weapon effectiveness is required against a given enemy force:

(1) if replacements are available?

(2) if no replacements are available?

i. What force size is required against a given enemy?

j. With a given force available, how much time would be required to reduce an enemy force to a specified level?

4. In this paper we will present models for only four of these decision problems: a, e, i and j. These particular examples were selected to illustrate the four categories of models and several different measures of effectiveness. In the first problem, involving tradeoffs between ground and air, we are interested in the broader question of what mix of ground forces and air forces do we need in NATO? What factors should be incorporated in a simple model designed to give gross, order-of-magnitude answers to this question? Some of them are: aircraft availability rate and sortie rate, attrition of aircraft, allocation of aircraft against alternative target types (e.g., tanks or artillery), lethality of air-delivered weapons, tank effectiveness, tank vulnerability and artillery effectiveness. The model is presented in Figure 1, with variables and factors defined as follows:

$X_1(t)$ = Red tanks

$X_2(t)$ = Red artillery

$Y_1(t)$ = Blue tanks

$Y_2(t)$ = Blue artillery

$Y_3(t)$ = Blue aircraft

J = rate at which a Blue tank can kill Red tanks

K = rate at which a Red tank can kill Blue tanks

P = Blue aircraft attrition rate per sortie flown

b = rate at which a Blue artillery weapon can kill Red tanks

k = average number of Red tanks killed per aircraft sortie

s = sortie rate, per available aircraft

V = aircraft availability rate, taking into account NORM, NORS, etc.

r = replacement rate for Red tanks

L = rate at which a Blue artillery weapon can kill Red artillery

M = rate at which a Red artillery weapon can kill Blue artillery

N = average number of Red artillery weapons killed per aircraft sortie

f = fraction of Blue aircraft sorties employed against Red tanks (the rest are used against artillery)

g = fraction of Blue artillery employed against Red tanks (the rest are used against artillery)

| | |
|---|---|
| RED TANKS | $\frac{dX_1}{dt} = - JY_1 - bgY_2 - VskfY_3 + r$ |
| RED ARTILLERY | $\frac{dX_2}{dt} = - L(1-g)Y_2 - VsN(1-f)Y_3$ |
| BLUE TANKS | $\frac{dY_1}{dt} = - KX_1$ |
| BLUE ARTILLERY | $\frac{dY_2}{dt} = - MX_2$ |
| BLUE AIRCRAFT | $\frac{dY_3}{dt} = - VsPY_3$ |

Figure 1

Rates at which Committed Strengths Change

From the first differential equation in Figure 1 we see that Red tanks are killed by Blue tanks ($Y_1$), Blue artillery ($Y_2$) and Blue aircraft ($Y_3$). To some extent Red tank losses are compensated for by replacement tanks, at a rate of r per minute. The Blue commander has two weapon allocation problems: allocation of available aircraft against tanks (f) and against artillery (1-f); allocation of available artillery against tanks (g) and against artillery (1-g).

5. The solutions of this model express the numbers of weapons of each type surviving (and committed) as functions of time. This model can be used to investigate tradeoffs between tanks and close air support aircraft in the following way. We set a tactical goal and calculate the various combinations of "number of tanks" and "number of aircraft",each of which will achieve the goal. An example of a goal is: "Reduce the Red tank strength by 100 within 2 hours. The tradeoff curves (tanks versus aircraft) will usually depend on the values assumed by all the other factors in the model, such as Red tank effectiveness, Blue tank effectiveness, Blue aircraft attrition rate, number of Blue artillery tubes available, etc. The tradeoff curves also vary with the type of goal required. "Reduce the Red to Blue tank force ratio by 50% in 6 hours" would give different curves.

6. For example, if we leave out Red and Blue artillery to simplify the calculations and make the following assumptions

| | |
|---|---|
| J = .003 | S = .004 |
| K = .001 | P = .05 |
| V = 0.70 | k = 2 |

we get the following results, for the Blue goal of killing the required number of Red tanks within 16-2/3 hours:

| Number of Red tanks killed | Number of Blue aircraft | Number of Blue tanks |
|---|---|---|
| 400 | 100 | 133 |
| | 75 | 200 |
| | 50 | 267 |
| 300 | 100 | 33 |
| | 75 | 100 |
| | 50 | 167 |

7. The tradeoff between Blue tanks and Blue aircraft depend on two major uncertainties: the duration of combat and the ratio of Blue tank effectiveness (J) to Red tank effectiveness (K). We see this directly in the following results, based on the assumptions: V = 0.70, S = 0.004, k = 2, p = 0.05.

| Ratio of Blue tank effectiveness to Red tank effectiveness | Combat Time (minutes) | Number of Blue tanks equivalent to one Blue aircraft |
|---|---|---|
| 1 to 1 | 2000 | 8.5 |
| | 1000 | 6.2 |
| | 500 | 5.8 |
| 3 to 1 | 2000 | 5.3 |
| | 1000 | 2.7 |
| | 500 | 2.2 |

8. For the next decision problem (3e) the question is: How should Red tank fire be allocated between three or four types of Blue anti-armor weapons (targets)? The Red side makes tactical judgments about allocation of fire. Here we let $f_1$ be the fraction of Red fire directed against type 1 Blue weapons, $f_2$ the fraction of Red fire directed against type 2 Blue weapons, etc. We could let the $f_i$ factors vary with time during the battle, but in the examples given here we assume that for a given battle each $f_i$ is given a fixed value, with the sum of $f_i$ equal to one.

9. The ability of individual weapons to kill targets (detect, hit, kill) is assumed to vary with time during the battle (comparable to variations with range as intervisibility, detection and weapon accuracy change). The model represents the dynamics of combat as the battle progresses, the rates at which the numbers of weapons surviving changes due to attrition. The model, a set of N + 1 differential equations, is given below in para 10. The factors and variables are defined as follows:

$X$ = number of Red tanks

$Y_1$ = number of type 1 Blue weapons (e.g., M60A1E3)

$Y_2$ = number of type 2 Blue weapons (e.g., TOW on M113)

$Y_3$ = number of type 3 Blue weapons (e.g., TOW on jeep)

$Y_4$ = number of type 4 Blue weapons (e.g., DRAGON)

$\vdots$

$Y_N$ = number of type N Blue weapons.

$f_i$ = fraction of Red tank fire allocated against type i Blue weapons (This could include target opportunities as well as target priorities.)

$K_j$ (t) = average rate at which a type i Blue weapon can kill Red tanks (This includes engagement opportunities, hit probability, rate of fire and kill probability given a hit.)

$J_i$ (t) = average rate at which a Red tank can kill type i Blue weapons

We assume that each $K_i$ and $J_i$ is a linear function of time. In particular,

$$K_i(t) = a_i + b_i t$$

$$J_i(t) = c_i + d_i t$$

10. The model is:

Red tanks

$$\frac{dX}{dt} = - K_1(t)Y_1 - K_2(t)Y_2 - \ldots - K_N(t)Y_N$$

Blue weapons

$$\frac{dY_1}{dt} = - J_1(t)\ f_1\ X(t)$$

$$\frac{dY_2}{dt} = - J_2(t)\ f_2\ X(t)$$

$$\vdots$$

$$\frac{dY_N}{dt} = - J_N(t)\ f_N\ X(t)$$

11. We consider five alternative tactical allocation schemes for the Red tanks, as follows:

a. Initial Blue strength.

$$f_i = \frac{Y_i(o)}{\sum_{j=1}^{N} Y_j(0)}$$

b. Equal priorities by target type.

$$f_i = \frac{1}{N}$$

c, Initial threat to Red tanks.

$$f_i = \frac{a_i}{\sum_{j=1}^{N} a_j}$$

d. Initial ease of killing.

$$f_i = \frac{c_i}{\sum_{j=1}^{N} c_j}$$

e. Later threat to Red tanks (time $\bar{t}$, e.g., $\bar{t}$ = 5 or 10).

$$f_i = \frac{a_i + b_i \bar{t}}{\sum_{j=1}^{N} (a_j + b_j \bar{t})}$$

12. In order to compare these Red alternatives we must make assumptions about the initial force sizes on both sides and the coefficients representing weapon kill capabilities. In a number of runs we used the following values:

| Blue weapons | Blue versus Red | | Red versus Blue | |
|---|---|---|---|---|
| | $a_i$ | $b_i$ | $c_i$ | $d_i$ |
| type 1(tanks) | .152 | .163 | .013 | .015 |
| type 2 (long range ATGM) | .053 | .270 | .008 | .040 |
| type 3 (short range ATGM) | .000 | .600 | .000 | .060 |

With overall Red to Blue initial force ratios on the order to 3 to 1 or 4 to 1 the order of preference for the Red alternatives in para 11 turned out as follows:

Best: e. later threat to Red tanks
a. initial Blue strength
b. equal priorities by type
d. initial ease of killing
Worst: c. initial threat to Red tanks

13. In the next decision problem (3i) we consider the tactical question of how a given initial Blue force should be broken up into smaller units for employment against the enemy. If the effectiveness of the defending Blue force does not depend on the absolute scale of the battles fought by the units (but only on the force ratio) then it might not matter how the initial force is broken up into units. We have investigated many alternative types of models with respect to this question. Here we present results for four of them:

Model A

$$\frac{dR}{dt} = -KB(t) \qquad \frac{dB}{dt} = -JR(t)$$

Model B

$$\frac{dR}{dt} = -K[B(t)]^M \qquad \frac{dB}{dt} = -J[R(t)]^N$$

Model C

$$\frac{dR}{dt} = - LB(t)R(t) + aR(t)$$

$$\frac{dB}{dt} = - JR(t)B(t) + bB(t)$$

Model D

$$\frac{dR}{dt} = - (p + qt)B(t)$$

$$\frac{dB}{dt} = - (a + bt)R(t)$$

14. In evaluating combat of maneuver units the single most meaningful measure of effectiveness is the cumulative loss ratio--i.e., the ratio of total Red losses to total Blue losses. This loss ratio will depend on many factors, including (in most cases) the initial force ratio--the ratio of the initial number of Red weapons to the initial number of Blue weapons. If we calculate the cumulative loss ratio at the particular time $\bar{t}$ at which Blue has a fraction "A" of his force surviving (e.g., A = 0.70) the result for Model A is:

$$L(\bar{t}) = \frac{F_o}{1-A}\left\{1 - \sqrt{1 - \frac{c(1-A^2)}{F_o^2}}\right\}$$

where $F_o$ is the initial force ratio ($\frac{R_o}{B_o}$) and $c = \frac{K}{J}$, the ratio of individual weapon effectiveness coefficients. It is clear that, for Model A, L does not depend on the scale of the battle ($B_o$ or $R_o$) but only on the initial ratio of forces.

15. For Model B, the loss ratio is:

$$L(\bar{t}) = \frac{F_o}{1-A}\left\{1 - \sqrt[N+1]{1 - \frac{c\,(1 - A^{M+1})R_o^{M-N}}{F_o^{M+1}}}\right\}$$

If M does not equal N then L does depend on the scale ($R_o$) but if M = N then it does not.

16. The cumulative loss ratio L satisfies the following equation when B(t) equals $AB_o$ for Model C:

$$b \log\{F_o - (1 - A)L\} + J(1 - A)\, B_o\, L =$$

$$b \log F_o + a \log A + K(1 - A)\, B_o$$

Since $B_o$ appears explicitly the loss ratio does depend on scale for Model C.

17. Based on the Taylor series solutions of Model D, the cumulative loss ratio, at any time t, is:

$$L(t) = \frac{pt - \frac{t^2}{2}(paF_o - q) + \frac{t^3}{6}[p^2a - (pb+2qa)F_o] - \cdots}{aF_o t - \frac{t^2}{2}(ap - bF_o) + \frac{t^3}{6}[a^2pF_o - (aq + 2bp)] - \cdots}$$

This expression depends on $F_o$ but not on $B_o$ or $R_o$ (and hence not on the scale of battle).

18. The final decision problem to be illustrated in this paper is 3j: with a given force available, how much time would be required to reduce an enemy force to a specified level? For example, Red has six tank platoons and Blue has three tank platoons. How long would it take Red to reduce Blue to 65% of his initial strength (e.g., with about 2 platoons left)? These numbers are too small for stable results from a deterministic model. Thus, we consider the following stochastic model, developed by Isbell and Marlow. At time t, the probability that exactly R Red units and exactly B Blue units are surviving is:

$$P(R, B; R_o, B_o, t)$$

where $R_o$ and $B_o$ are the initial strengths. If f and g are transition probabilities, in small increments of time, for Red and Blue respectively, then it is assumed that the function P satisfies the following differential equations:

$$\frac{dP(R,B)}{dt} = f(R+1, B)\,P(R+1,B) + g(R,B+1)P(R,B+1) - \left[f(R, B) + g(R, B)\right] P(R, B)$$

*where* $P(R, B, t) = 0$ *if* $R > R_o$ *or* $B > B_o$

*and* $P(R_o, B_o, 0) = 1$.

If we assume that f and g are linear functions of weapon characteristics and that Red and Blue weapon are equally capable, then the solutions are as given in Figure 2, where the functions F satisfy the following relations:

$$F(R, B; R_o, B_o) = \frac{B}{R + B + 1} F(R + 1, B; R_o, B_o) + \frac{R}{R + B + 1} F(R, B + 1; R_o, B_o)$$

# PROBABILITY THAT EXACTLY R AND B ARE SURVIVING AT TIME t IS:

$$P(R, B; R_0, B_0, t) =$$

$$F(R, B; R_0, B_0) \left| \begin{matrix} R_0+B_0 \\ R+B \end{matrix} \right| (e^{bt}-1)^{R_0+B_0-R-B} e^{-bt(R_0+B_0)}$$

FIGURE 2

19. Examples of specific solutions are the following:

a. Initial Red force: 6 platoons
Initial Blue force: 3 platoons

| Probability | Time required to reduce Blue force by 65% (t minutes or less) |
|---|---|
| .63 | 7 |
| .56 | 5 |
| .42 | 3 |
| .16 | 1 |

b. Initial Red force: 4 platoons
Initial Blue force: 2 platoons

| Probability | Time required to reduce Blue force by 65% (t minutes or less) |
|---|---|
| .42 | 7 |
| .32 | 5 |
| .19 | 3 |
| .06 | 1 |

# EVALUATION OF SEVERAL 'BEST FIT' METHODS AS THEY PERTAIN TO THE SUPERPOSITION OF SOLUTIONS IN A MULTIPOINT BOUNDARY VALUE PROGRAM

John H. Walker

U.S. Army Test and Evaluation Command
White Sands Missile Range
White Sands, New Mexico

S. Bart Childs, Ph.D.

Department of Industrial Engineering
Texas A&M University
Texarkana, Texas

ABSTRACT. A shooting method is the superposition of initial value solutions of ordinary differential equations such that boundary are "met" or a performance index is minimized.

The results of meeting noisy boundary conditions in least squares and minimax norms are presented. The example problem is a damped, forced harmonic oscillator.

The procedures are basic to system identification problems.

1. INTRODUCTION. The linear boundary value problem is governed by the ordinary differential equation

$$\dot{y} = Ly + f \tag{1}$$

where $y$ and $\dot{y}$ are, respectively, the vector of $n$ state variables and its derivative, $L$ and $f$ are matrix and vector functions of the independent variable $t$, time. The solution of this differential equation is subject to a set of boundary conditions

$$q_i(y(t_i)) = b_i \qquad i = 1,2,\ldots,m \geq n \tag{2}$$

where $q_i$ is the boundary condition operator that specifies a linear combination of the state variables equal to the boundary value, $b_i$, at time, $t_i$.

A shooting method is to superimpose appropriately independent solutions of (1). This can be written as

$$y = \sum_{j=0}^{n} p^{(j)} a_j \tag{3}$$

where $p^{(j)}$ is a particular solution of (1) and $a_j$ is the corresponding superposition constant.

The independence properties can be assured by the following strategy. Assume $p^{(0)}(0) = \alpha$. Then take

$$p_i^{(j)}(0) = \alpha_i + \delta_{ij}\beta_j \qquad i, j = 1,2,\ldots,n \tag{4}$$

where $\delta$ is the Kronecker delta and all $\beta_j \neq 0$. The above strategy gives a determinant (Wronskian) of the associated homogeneous differential equations, at $t=0$, of the product of the $\beta$'s.

The superposition of particular solutions also requires that

$$\sum_{j=0}^{n} a_j = 1. \tag{5}$$

Childs et al. [1970] give more details on the strategy and a proof of (5). If $\alpha$ is the initial value vector that makes (1) satisfy (2) then it is obvious that $a_0=1$ and $a_1=a_2=\ldots=a_n=0$. How close the actual superposition constants come to these values is an indication of the merit of the numerical method.

The superposition (3) is substituted into the boundary conditions (2). If the boundary conditions are linear in $y$, then the operators $q$ and $\Sigma$ may be interchanged giving

$$\sum_{j=0}^{n} [q_i(p^{(j)}(t_i))]\, a_j = b_i \qquad i = 1,2,\ldots,m. \tag{6}$$

The use of shooting procedures results in the particular solutions, $p^{(j)}$, being known and we observe that the bracketed terms in (6) are simply coefficients of an algebraic equation in the unknown superposition constants. We write (5) and (6) as the matrix equation

$$Sa = d \tag{7}$$

where $S_{0,j} = 1, \qquad d_0 = 1$

and $$S_{i,j} = q_i(p^{(j)}(t_i)), \qquad d_i = b_i \qquad \begin{matrix} i = 1,2,\ldots,m \\ j = 0,1,\ldots,n \end{matrix} . \tag{8}$$

2. NONLINEARITIES. If the differential equation is nonlinear then we may write it as

$$\dot{y} = g(y,t). \tag{9}$$

Equation (9) is linearized via a Taylor series expansion in order to obtain an equation linear in $z$;

$$\dot{z} = g(w,t) + \left[\frac{\partial g}{\partial y}\right]_{y=w} (z - w) \tag{10}$$

where $w$ is a reference solution or a previous approximation to $y$. Equation (10) may be rewritten as

$$\dot{z} = Jz + g' \tag{11}$$

where
$$J = \frac{\partial g}{\partial y}_{y=w} \quad \text{and} \quad g' = g(w,t) - Jw \ . \tag{12}$$

Therefore, if we are given a nonlinear differential equation subject to boundary conditions, we may approximate it by the linear equation

$$\dot{z} = Jz + g' \tag{13}$$

subject to the boundary conditions

$$q_i(z(t_i)) = b_i \qquad i = 1,2,\ldots,m. \tag{14}$$

It is obvious that (13) and (14) are analagous to (1) and (2) and so we proceed in the same fashion. The only difference is that now the solution is obtained iteratively.

We again superimpose $n+1$ particular solutions of (13) that meet the boundary conditions (14):

$$z = \sum_{j=0}^{n} p^{(j)} a_j \tag{15}$$

where each $p^{(j)}$ satisfies

$$\dot{p}^{(j)} = Jp^{(j)} + g'. \tag{16}$$

The superposition constants are determined by

$$\sum_{j=0}^{n} a_j = 1 \tag{17}$$

and
$$q_i\left(\sum_{j=0}^{n} (p^{(j)}(t_i))a_j\right) = b_i \qquad i = 1,2,\ldots,m \geq n \ . \tag{18}$$

If the operators, $q_i$, are linear then (17) and (18) form a set of linear equations analagous to (7).

If any of the boundary condition operators, $q_i$, are nonlinear, then they must also be linearized by a Taylor series expansion. The linearization is done with respect to the superposition constants with the initial reference values of the vector, a, as

$$\begin{aligned} a_0 &= 1 \\ a_j &= 0 \qquad j = 1,2,\ldots,n \ . \end{aligned} \tag{19}$$

The reader can refer to Childs et al. [1970] and to Roberts and Shipman [5] for more details on these linearization procedures.

3. OVERDETERMINED BOUNDARY CONDITIONS. If the number of boundary conditions, $m$, is greater than the order of the differential equation, $n$, then (7) constitutes an overdetermined set of linear equations with the $a_j$'s unknown. Not all of the equations can be met exactly, some will have to be met in a "best fit" sense. Let's assume that $\rho$ of the $m$ boundary conditions are to be met exactly, then $\rho+1$ of the $m+1$ equations (the superposition condition is included) must be met exactly. Equation (7) may be partitioned as follows;

$$\begin{bmatrix} S_1 & S_2 \\ S_3 & S_4 \end{bmatrix} \begin{bmatrix} a_e \\ a_o \end{bmatrix} = \begin{bmatrix} d_e \\ d_o \end{bmatrix} \tag{20}$$

The components $S_1$, $S_2$ and $d_e$ correspond with the equations to be met exactly. By suitable matrix operations, (20) can be transformed into

$$\begin{bmatrix} I & S_2' \\ 0 & S_4' \end{bmatrix} \begin{bmatrix} a_e \\ a_o \end{bmatrix} = \begin{bmatrix} d_e' \\ d_o' \end{bmatrix} \tag{21}$$

Two matrix equations results from (21):

$$S_4' a_o = d_o' \tag{22}$$

$$a_e = d_e' - S_2' a_o \tag{23}$$

Equation (22) is solved in a "best fit" sense for $a_o$, which is then substituted into (23) for $a_e$.

Once the superposition constants, $a_j$, are found, they are multiplied by their appropriate particular solutions at $t=0$, that is, $p^{(j)}(0)$, which yields an estimate of $y(0)$, i.e.

$$y(0) = \sum_{j=0}^{n} p^{(j)}(0) a_j \,. \tag{24}$$

If the problem is nonlinear, the LHS of (24) is taken as the unperturbed particular solution at $t=0$, $p^{(o)}(o)$. Independent perturbed solutions are generated by the strategy described in (4) and a new set of superposition constants found. The method is repeated until convergence of consecutive $p^{(0)}(0)$ vectors are observed (i.e., $a_o$ will approach unity and all other $a_j$'s will approach zero).

There are two principal methods of solving overdetermined systems of linear equations. They are:

1) least squares solution
2) minimax or Chebyshev solution.

a. Least-Squares Solution: Given

$$S_4' a_o = d_o' \tag{25}$$

the residual vector can be written,

$$R = S_4' a_o - d_o' \quad . \tag{26}$$

The least-squares solution is the vector, $a_o$, that minimizes the sum of the squares of the components of the residual vector, $R$, is:

$$a_o = (S_4')^T S_4'^{-1} (S_4')^T d_o' \quad . \tag{27}$$

This is substituted back into (23) to find $a_e$ .

b. Minimax Solution: The minimax solution is the vector $a_o$ which minimizes the largest absolute value of the components of the residual vector (26). That is, we want to minimize $max(r_1, r_2, \ldots, r_{m-p})$ . The method advanced by Powell is used in the program. See [3] and [4] for more specifics on the minimax method.

4. RESULTS. We considered the following problem

$$\ddot{x} + \mu\dot{x} + \xi x = sin(t) \tag{28}$$

which is the equation of motion for a forced, damped harmonic oscillator. By the change of variables

$$y_1 = x, \; y_2 = \dot{x}, \; y_3 = \mu, \; y_4 = \xi \tag{29}$$

(28) may be replaced by

$$\begin{aligned} \dot{y}_1 &= y_2 \\ \dot{y}_2 &= -y_4 y_1 - y_3 y_2 + sin(t) \\ \dot{y}_3 &= 0 \\ \dot{y}_4 &= 0 \; . \end{aligned} \tag{30}$$

Initial values were selected for the state variables and solutions for $y$ and $\dot{y}$ were generated on the interval $0 \le t \le 15$. At times, $t = 1, 2, 3, \ldots, 15$, the value of $\dot{y}_2$ was observed. These values were taken as the exact boundary values (to 8 significant figures). Six sets of "noisy" data were produced by two techniques. The first was to round off the exact boundary values to 1,2, and 3 decimal places to the right of the decimal point. The second technique was to add a random variable that was normally distributed with a mean of *zero* and a standard deviation, $\sigma$. Three different values of $\sigma$ were used: *.01*, *.1*, and *.5*. See Table 1 for the sets of boundary values. The program was run using each set of the corrupted boundary values as data. The errors between the originally selected initial conditions and those that the computer estimated from the noisy data were computed. Both least-squares and minimax were employed to solve the overdetermined system, for each data set. See Tables 2, 3, 4 and 5. Two criteria were chosen as the basis for evaluating the "closeness" of fit: (1) the sum of the absolute values of the errors and (2) the sum of the squares of the errors. In all cases, the least-squares solution proved the better fit, as expected. The accuracy of the parameter estimates is impressive, even with the noisiest data.

| $t_i$ | Boundary values rounded to 3 decimal places | Boundary values rounded to 2 decimal places | Boundary values rounded to 1 decimal place |
|---|---|---|---|
| 1 | -0.220 | -0.22 | -0.2 |
| 2 | 0.35D-01 | 0.3D-01 | 0.0 |
| 3 | -0.474 | -0.47 | -0.5 |
| 4 | -0.589 | -0.59 | -0.6 |
| 5 | 0.393 | 0.39 | 0.4 |
| 6 | 1.597 | 1.60 | 1.6 |
| 7 | 1.452 | 1.45 | 1.5 |
| 8 | -0.388 | -0.39 | -0.4 |
| 9 | -2.324 | -2.32 | -2.3 |
| 10 | -2.274 | -2.27 | -2.3 |
| 11 | 0.88D-01 | 0.9D-01 | 0.1D-00 |
| 12 | 2.711 | 2.71 | 2.7 |
| 13 | 2.997 | 3.00 | 3.0 |
| 14 | 0.401 | 0.40 | 0.4 |
| 15 | -2.816 | -2.82 | -2.8 |

| $t_i$ | Boundary values with N(0,.01) r.v. added | Boundary values with N(0,.1) r.v. added | Boundary values with N(0,.5) r.v. added |
|---|---|---|---|
| 1 | -0.21869550 | -0.20667165 | -0.15323232 |
| 2 | 0.37798268D-01 | 0.66064361D-01 | 0.19169151 |
| 3 | -0.46666120 | -0.40497946 | -0.13083809 |
| 4 | -0.59655859 | -0.66370844 | -0.96215247 |
| 5 | 0.39299204 | 0.39030898 | 0.37838423 |
| 6 | 1.5960593 | 1.5846639 | 1.5340173 |
| 7 | 1.4712304 | 1.6443249 | 2.4136342 |
| 8 | -0.38313658 | -0.33635166 | -0.12841859 |
| 9 | -2.3169182 | -2.2529945 | -1.9688892 |
| 10 | -2.2631763 | -2.1656611 | -1.7322599 |
| 11 | 0.91826552D-01 | 0.12353162 | 0.26444312 |
| 12 | 2.6876393 | 2.4736753 | 1.5227239 |
| 13 | 3.0007328 | 3.0323136 | 3.1726727 |
| 14 | 0.39058453 | 0.29953824 | -0.10511233 |
| 15 | -2.8133056 | -2.7851188 | -2.6598437 |

TABLE 1

| | True value of initial conditions | I.C. estimates using Least Squares. B.V.'s input with 8 significant figures | I.C. estimates using Minimax. B.V.'s input with 8 significant figures |
|---|---|---|---|
| $X(0)$ | 1.0 | 0.99999999 | 0.99999997 |
| $\dot{X}(0)$ | 0.5 | 0.49999999 | 0.49999998 |
| $\mu$ | 0.2 | 0.20000000 | 0.20000000 |
| $\xi$ | 1.0 | 1.00000000 | 1.00000000 |
| | | $\sum\|e_i\|$ = 2.0D-08 | $\sum\|e_i\|$ = 4.0D-08 |
| | | $\sum e_i^2$ = 2.0D-16 | $\sum e_i^2$ = 1.0D-15 |

TABLE 2

| | I. C. estimates using Least Squares. B.V.'s are rounded to 3 decimal places | I. C. estimates using Minimax. B.V.'s are rounded to 3 decimal places |
|---|---|---|
| $X(0)$ | 1.0000524 | 0.99976973 |
| $\dot{X}(0)$ | 0.49978310 | 0.49976451 |
| $\mu$ | 0.20004417 | 0.20009774 |
| $\xi$ | 0.99996730 | 0.99998927 |
| | $\sum \lvert e_i \rvert$ = 3.4617D-04 | $\sum \lvert e_i \rvert$ = 5.7423D-04 |
| | $\sum e_i^2$ = 5.2812D-08 | $\sum e_i^2$ =11.8148D-08 |

| | I. C. estimates using Least Squares. B.V.'s are rounded to 2 decimal places | I. C. estimates using Minimax. B.V.'s are rounded to 2 decimal places |
|---|---|---|
| $X(0)$ | 1.0011527 | 1.0008828 |
| $\dot{X}(0)$ | 0.50086499 | 0.50155621 |
| $\mu$ | 0.19989372 | 0.19996940 |
| $\xi$ | 1.0001414 | 0.99988326 |
| | $\sum \lvert e_i \rvert$ = 2.2654D-03 | $\sum \lvert e_i \rvert$ = 2.5864D-03 |
| | $\sum e_i^2$ = 2.1082D-06 | $\sum e_i^2$ = 3.2157D-06 |

TABLE 3

| | I.C. estimates using Least Squares. B.V.'s are rounded to 1 decimal place | I.C. estimates using Minimax. B.V.'s are rounded to 1 decimal place |
|---|---|---|
| $X(0)$ | 0.95849170 | 0.93474852 |
| $\dot{X}(0)$ | 0.50061069 | 0.48097522 |
| $\mu$ | 0.20243731 | 0.20584127 |
| $\xi$ | 0.99890736 | 0.99848235 |
| | $\sum\|e_i\|$ = 4.5649D-02 | $\sum\|e_i\|$ = 9.1635D-02 |
| | $\sum e_i^2$ = 1.7304D-03 | $\sum e_i^2$ = 4.6561D-03 |

| | I.C. estimates using Least Squares. N(0,.01) r.v. added to the B.V.'s | I.C. estimates using Minimax. N(0,.01) r.v. added to the B.V.'s |
|---|---|---|
| $X(0)$ | 1.0009066 | 0.99134015 |
| $\dot{X}(0)$ | 0.49193995 | 0.48700959 |
| $\mu$ | 0.20106603 | 0.20179481 |
| $\xi$ | 1.0003406 | 0.99973471 |
| | $\sum\|e_i\|$ = 9.7755D-03 | $\sum\|e_i\|$ = 23.7104D-03 |
| | $\sum e_i^2$ = 6.7049D-05 | $\sum e_i^2$ = 24.7035D-05 |

TABLE 4

| | I.C. estimates using Least Squares. N(0,.1) r.v. added to the B.V.'s | I.C. estimates using Minimax. N(0,.1) r.v. added to the B.V.'s |
|---|---|---|
| $X(0)$ | 1.0112150 | 0.91127941 |
| $\dot{X}(0)$ | 0.41910446 | 0.36854682 |
| $\mu$ | 0.21071785 | 0.21777510 |
| $\xi$ | 1.0033087 | 0.99682169 |
| | $\sum \lvert e_i \rvert$ = .1061 | $\sum \lvert e_i \rvert$ = .2411 |
| | $\sum e_i^2$ = 6.7957D-03 | $\sum e_i^2$ = 2.5477D-02 |

| | I.C. estimates using Least Squares. N(0,.5) r.v. added to the B.V.'s | I.C. estimates using Minimax. N(0,.5) r.v. added to the B.V.'s |
|---|---|---|
| $X(0)$ | 1.1020264 | 0.49280864 |
| $\dot{X}(0)$ | 0.90461984D-01 | -0.18505768 |
| $\mu$ | 0.25470622 | 0.28377993 |
| $\xi$ | 1.0143196 | 0.97314194 |
| | $\sum \lvert e_i \rvert$ = .5809 | $\sum \lvert e_i \rvert$ = 1.3029 |
| | $\sum e_i^2$ = .1813 | $\sum e_i^2$ = .7343 |

TABLE 5

## References


1. Childs, B., Doiron, H., Holloway, C., "Numerical Solutions of Multipoint Boundary Value Problems in Non-Linear Systems", Int. J. Systems Science, Vol. 2, No. 1, pp. 59-66, 1971.

2. Childs, B., Luckinbill, D., Bryan, J., Boyd, J.H., Jr., "Numerical Solutions of Multipoint Boundary Value Problems in Linear Systems", Int. J. Systems Science, Vol. 2, No. 1, pp. 49-57, 1971.

3. Madsen, K. and Powell, M.J.D., "A Fortran Subroutine That Calculates the Minimax Solution of Linear Equations Subject to Bounds on the Variables", Harwell, Oxfordshire, England, Feb. 1975.

4. Powell, M.J.D., "The Minimax Solution of Linear Equations Subject to Bounds on the Variables", Harwell, Oxfordshire, England, Dec. 1974.

5. Roberts, S.M. and Shipman, J.S., Two-Point Boundary Value Problems: Shooting Methods, Elsevier, 1972.

6. Zupp, G.A. and Childs, Bart, "Applications of Quasilinearization Theory to Systems Identification", NASA TN D-5300, Manned Spacecraft Center, Houston, Texas, July, 1969.

# A STATISTICAL STUDY OF NUMERICAL ANALYSIS

## APPLIED TO THE REGRESSION OF nTH ORDER DIFFERENTIAL EQUATIONS

Craig D. Hunter

DRCPM-PBM-T-PA
Picatinny Arsenal
Dover, New Jersey
Formerly, Intern Training Center-DARCOM

Bart Childs

Department of Industrial Engineering
Texas A&M University
Texarkana, Texas

ABSTRACT. An extension of regression analysis from the usual algebraic models to differential equation models is given. A shooting method, superposition of appropriately independent initial value solutions of differential equations, is used. The shooting method used is based on particular solutions of the governing differential equations. Nonlinear differential equations and/or boundary conditions can be accommodated.

The statistics of linear regression are generated through a straightforward analysis of variance. These provide the basis of "acceptance" or "rejection" of the regression.

The statistics generated include an (uncorrected) ANOVA tables, general F-test on the regression, $R^2$ value, the coefficient of variation, covariance matrix of the superposition constants, estimate of the variance about the regression, estimate of the variance of the parameters, and the confidence intervals of these estimates.

The procedures are basic to system identification problems.

1. INTRODUCTION. The linear boundary value problem is governed by the ordinary differential equation

$$\dot{y} = Ly + f \tag{1}$$

where $y$ and $\dot{y}$ are, respectively, the vector of $n$ state variables and its derivative, $L$ and $f$ are matrix and vector functions of the independent variable $t$, time. The solution of this differential equation is subject to a set of boundary conditions

$$q_i(y(t_i)) = b_i \qquad i = 1, 2, \ldots, m \geq n \tag{2}$$

where $q_i$ is the boundary condition operator that specifies a linear combination of the state variables equal to the boundary value, $b_i$, at time, $t_i$. We are concerned only with those cases where $m>n$ and the boundary conditions are to be met in a least squares sense.

A shooting method is to superimpose appropriately independent solutions of equation (1). This can be written as

$$y = \sum_{j=0}^{n} p^{(j)} a_j \tag{3}$$

where $p^{(j)}$ is a particular solution of (1) and $a_j$ is the corresponding superimposition constant.

The independence properties can be assured by the following strategy. Assume $p^{(0)}(0) = \alpha$. Then take

$$p_i^{(j)}(0) = \alpha_i + \delta_{ij}\beta_j \qquad i, j = 1, 2, \ldots, n \tag{4}$$

where $\delta$ is the Kronecker delta and all $\beta_j \neq 0$. The above strategy gives a determinant (Wronskian) of the associated homogeneous differential equations, at $t=0$, of the product of the $\beta$'s.

The superposition of particular solutions also requires that

$$\sum_{j=0}^{n} a_j = 1. \tag{5}$$

Childs et al. [1970] give more details on the strategy and proof of (5). If $\alpha$ is the initial value vector that makes (1) satisfy (2) then it is obvious that $a_0 = 1$ and $a_1 = a_2 = \ldots = a_n = 0$. How close the actual superposition constants come to these values is an indication of the merit of the numerical method.

The superposition (3) is substituted into the boundary conditions (2). If the boundary conditions are linear in $y$, then the operators $q$ and $\Sigma$ may be interchanged giving

$$\sum_{j=0}^{n} q_i(p^{(j)}(t_i))\ a_j = b_i \qquad i = 1, 2, \ldots, m. \tag{6}$$

The use of shooting procedures results in the particular solutions, $p^{(j)}$, being known and we observe that the bracketed terms in (6) are simply coefficients of an algebraic equation in the unknown superposition constants. We write (5) and (6) as the matrix equation

$$Sa = d \tag{7}$$

where $S_{0,j} = 1, \quad d_0 = 1$

and $S_{i,j} = q_i(p^{(j)}(t_i)), \quad d_i = b_i \quad i = 1, 2, \ldots, m \quad j = 0, 1, \ldots, n.$

2. THE OVER-DETERMINED SYSTEM. The solution of the system (7) is easily obtained for the data sets we have considered. We rewrite (7) partitioning (and rearranging if necessary) the elements of the vectors and matrices

$$\left[\begin{array}{c|c} S_1 & S_2 \\ \hline S_3 & S_4 \end{array}\right] \left[\begin{array}{c} a_e \\ \hline a_l \end{array}\right] = \left[\begin{array}{c} b_1 \\ \hline b_m \end{array}\right] = \left[\begin{array}{c} d_e \\ \hline d_l \end{array}\right] \tag{8}$$

The equality sign is understood to mean "equality" (as much as our numerical procedures allow) for the upper portion of (8) and "least squares" fit for the lower portion. The equality conditions come from the superposition constraint (5) and any boundary conditions that may exist which should be met "exactly." Weighting of the rest of the boundary conditions can be done but is not shown.

A straightforward method of solution is by elementary operations (Gaussian reduction with maximum pivot selection) to transform (8) into the equality (9) and the "least square" fit (10)

$$I\, a_e + {}'S_2 a_l = {}'d_e \tag{9}$$

$${}'S_4 a_l = {}'d_l \tag{10}$$

The $'()$ denotes the values have been affected by the reduction process. Note that ${}'S_1 = I$ and ${}'S_3 = 0$.

The least square solution for $a_l$ in (10) is obtained in the usual manner, the normal equations result from premultiplying by the transpose of ${}'S_4$. The result is substituted into (9) to obtain the rest of the $a$ vector.

The "correct" initial value vector can be calculated from

$$y(0) = p^{(0)}(0) + Ba \tag{11}$$

where $B$ is a diagonal matrix with indices varying from $0$ through $n$. $B_{00} = 0$, $B_{ii} = \beta_i$ for $i = 1, 2, \ldots, n$. Our computational procedure is to repeat the process with $p^{(0)}(0) = y(0)$ from (11) such that we should have

$$a_j = 0 \qquad j = 1, 2, \ldots, n \tag{12}$$

This will aid in construction of confidence limits of parameter estimates. This also gives a convenient quantity

$$\hat{q}_i = q_i(p^{(0)}(t_i)) = b_i - \varepsilon_i \tag{13}$$

which is the "predicted" boundary value.

3. ANALYSIS OF VARIANCE. An uncorrected ANOVA table is presented below in terms of the nomenclature introduced. The significant calculations are presented in terms of vector products. These products are over the $\hat{q}$ and $b$ vectors and *are formed over the elements associated with least squares boundary conditions only, any exact boundary conditions are ignored in these products.*

TABLE 1

ANOVA TABLE

| Source | Sum of Squares | Degrees of Freedom | Mean Square |
|---|---|---|---|
| Due to regression | $\hat{q}^T b$ | n | SS/n |
| About the regression (residual) | $b^T b - \hat{q}^T b$ | m-k-n | $s^2$ |
| Total (uncorrected) | $b^T b$ | m-k | |

Notice in the degrees of freedom column that $m$ is the total number of boundary conditions and $k$ of those are to be met exactly.

We define $\bar{b}$ to be the mean of the least square boundary values

$$\bar{b} = (\Sigma\, b_i)/(m-k) \tag{14}$$

The following formulae are used to calculate the usual statistics:

$$R^2 = \frac{\Sigma\, (\hat{q}_i b_i - \bar{b}\,)^2}{b^T b - \bar{b}^2 (m-k)} \tag{15}$$

$$s^2 = MS\ (residual) = estimated\ variance\ of\ system \tag{16}$$

$$F_{cal} = MS\ (regression)/s^2 \tag{17}$$

In (15) the summation and product are over the least square boundary conditions.

The $F_{cal}$ value must exceed a Fischer's $F$ with:

Probability of $1 - \alpha$ ($\alpha$ is the producers risk)

Numerator degrees of freedom $n$

Denominator degrees of freedom $m-k-r$

for the regression to be *accepted.*

The estimated variances of the boundary values are:

$$est.\ var.\ \ (\hat{q}_i) = ('S_{4i})\ [('S_4)('S_4)^T]^{-1}\ ('S_{4i})^T\ s^2 \tag{18}$$

where the $i$ subscript denotes the $i$th row of the $'S_4$ matrix. The resulting confidence limits in terms of the $t$ statistic are:

$$\hat{q}_i \pm t(\gamma,\ 1 - \alpha/2)\ \sqrt{[est.\ var\ (q_i)]} \tag{19}$$

where $\gamma = m-k-n$ and $\alpha$ is the producer's risk.

We have stated these procedures are basic to system identification procedures. We are most interested in $y(0)$ and its covariance in those cases. Recall equations (9) and (10) and we denote

$$r^2 = ('d_l)^T('d_l)/(m-n) \tag{20}$$

The covariance matrix of $a_l$ is

$$C_l = [('S_4)^T('S_4)]^{-1} r^2 \tag{21}$$

Likewise, the covariance matrix of $a_e$ is

$$C_e = ('S_2)\ C_l\ ('S_2)^T \tag{22}$$

The final covariance matrix is formed by appropriate multiplying by the perturbation matrix B giving

$$B \begin{bmatrix} C_e & 0 \\ 0 & C_l \end{bmatrix} \quad B^T = \begin{bmatrix} 0 & 0 - - - 0 \\ 0 & \\ \vdots & C_y \\ 0 & \end{bmatrix} \tag{23}$$

The *zero*th row and column of the result are null reflecting the variance of the superposition constraint (5). The $i$th diagonal element is the estimated variance of $y_i(0)$. Its square root is the estimated standard deviation of $y_i(0)$. The confidence limits (which Draper and Smith point out should be viewed with caution) are

$$y_i(0) \pm t(m-n,\ 1-\alpha/2)\ [est.\ std.\ dev.]_i \tag{24}$$

A more stringent confidence limit would be a hyperellipsoid like

$$e^T (C_y)^{-1} e \leq n\, r^2\, F(n,\ m-n,\ 1-\alpha) \tag{25}$$

where the vector $e$ is within the confidence limits about $y(0)$. This kind of statement is difficult to use if $n>2$.

4. AN EXAMPLE. Consider the problem of determining the coasting dynamics of an automobile. The three force elements of a model of the phenomenon are [5]

a. Rolling friction due to tire flexing and Coulomb friction on gear train.

b. Aerodynamic resistance proportional to the square of velocity.

c. Product of mass and deceleration.

The differential equation may be written as

$$\ddot{x} + \frac{C_d A_f \rho}{M} \frac{(\dot{x})^2}{2} + \mu_f g = 0 \tag{26}$$

where

$\rho$ = air density (slugs/ft$^3$)

$A_f$ = frontal area of vehicle (ft$^2$)

$M$ = mass of vehicle (slugs)

$g$ = acceleration of gravity (ft/sec$^2$)

$C_d$ = coefficient of drag

$\mu_f$ = rolling friction coefficient

$\dot{x}$ = velocity (ft/sec)

$\ddot{x}$ = acceleration (ft/sec$^2$)

Since the displacement, $x$, does not appear in (26), we can write this as a single first order ordinary differential equation (substituting $y_1$ for $\dot{x}$)

$$\dot{y}_1 = -\frac{y_2 A_f \rho}{M} \frac{y_1^2}{2} + y_3 g \tag{27}$$

The most economical procedure to obtain sufficient measurements would be to coast an automobile in neutral from some speed like say 80 miles/hour and record the speed at intervals of say 5 seconds. The boundary values in Figure 1 are velocities in ft/sec at 5 second intervals. Since (27) is nonlinear, the usual Newton type linearization procedures are employed, see Walker [4] in this proceedings or Childs [1], [2] for more details.

Figure 2 is the output of our program which is based on (27), the data in Figure 1, parameter values common in engineering use, and automobile parameters from the Road and Track Test Annual for 1966 for a Sunbeam Alpine. The coefficient of drag $C_d = y_2 = 0.5025 \pm 0.0690$ and coefficient of rolling friction $\mu_f = y_3 = 0.0169 \pm 0.0031$ resulted.

5. CONCLUSIONS. Regression analysis with differential equation models is feasible. It could significantly affect design of experiments when such models are relevant. Determination of the parameters in the simple example would be expensive if one had to use wind tunnels or treadmills.

## References


1. Childs, B., et al., "Numerical Solution of Multipoint Boundary Value Problems in Linear Systems," Int. J. Systems Science, Vol. 2, No. 1, pp. 49-57, 1971.

2. Childs, B., et al., "Numerical Solution of Multipoint Boundary Value Problems in Non-Linear Systems," Int. J. Systems Science, Vol. 2, No. 1, pp. 59-66, 1971.

3. Draper, N.R. and Smith H., Applied Regression Analysis, New York: John Wiley & Sons, Inc., 1966.

4. Walker, John H. and Childs, B., "Evaluation of Several 'Best Fit' Methods as They Pertain to the Superposition of Solutions in a Multi-point Boundary Value Program," this proceedings.

5. Zupp, G.A. and Childs, B., "Applications of Quasilinearization Theory to Systems Identification," NASA TN D-5300, Manned Spacecraft Center, Houston, Texas, July, 1969.

```
**********************************************************************************************************************
**********************************************************************************************************************
     FOLLOWING IS THE OUTPUT OF INPUT

     LINEAR=    0                                        STAT=    1          ALPHA= 0.500000000-01
         LINEAR= ZERO IMLIES NON-LINEAR                      STAT= ONE IMPLIES POST STATISTICAL STUDY TO BE EXECUTED
         LINEAR= NON-ZERO IMPLIES LINEAR                     STAT= ZERO IMPLIES NO POST SATISTICAL STUDY

     ACCURACY                          ACC= 0.10000D-05
     FOR GJWRLS                        CHECK= 0.10000D-19     DET=  1.0000
     FUDGE FACTOR FOR DTFIND           RF= 0.50000

     NUMBER OF DIFFERENTIAL EQ         NDE=       3
     NUMBER OF BOUNDARY CONDITIONS     NBV=       9
     NUMBER OF TERMS IN PWR SERIES     NTERMS=   20
     NO. EXACT BOUNDARY CONDITIONS     NEMAX=     1
     NO. OF NON-TRIVIAL DIFF EQ        NEQ=       1

     BOUNDARY VALUES     APPLIED ON(IBV)        IQBV      TIME OF BV
      117.30000                  1              1        0.0
      104.90000                  1              1         5.0000000
      95.300000                  1              1         10.000000
      87.300000                  1              1         15.000000
      79.200000                  1              1         20.000000
      71.900000                  1              1         25.000000
      65.700000                  1              1         30.000000
      59.800000                  1              1         35.000000
      54.000000                  1              1         40.000000

     INITIAL TIME OF INTEGRATION       TO= 0.0

     INITIAL GUESSES OF Y                          ICEX      LOWER LIMIT          UPPER LIMIT
          Y 1  YI( 1)=  110.00000                   1         0.0                  0.0
          Y 2  YI( 2)= 0.40000000                   1         0.0                  0.0
          Y 3  YI( 3)= 0.30000000D-01               1         0.0                  0.0

     PERTURBATION SCALER          PTBS= 0.100000000D 00
         ABSOLUTE VALUE LIMITS    PTMIN= 0.0
                                  PTMAX=  2.0000

     IOUT=    0
     QUIT=   15

     CONCLUDES FORMAL OUTPUT OF THE INPUT
**********************************************************************************************************************
**********************************************************************************************************************
```



FIGURE 1

FOLLOWING IS THE OUTPUT OF THE POST STATISTICAL STUDY

| SOURCE | SUM OF SQUARES | DEGREES OF FREEDOM | MEAN SQUARE | |
|---|---|---|---|---|
| REGRESSION | 63716.493 | 3 | 21238.831 | |
| RESIDUAL | 0.96692256 | 6 | 0.16115376 | = S**2 |
| TOTAL(UNC) | 63717.460 | 9 | | |

PER-CENT VARIATION R**2= 99.973342 COEFFICIENT OF VARIATION CV= 0.49129134D-02

TEST OF OVERALL REGRESSION FCAL= 131792.34 F(ALPHA)= 4.7571592 P(F(ALPHA).GT.FCAL)= 0.83446503D-06
FOR RISK OF ALPHA= 0.50000000D-01
*****ACCEPT REGRESSION*****

SPECIFICS OF THE BOUNDARY CONDITIONS T(ALPHA)= 2.4469080 ALPHA= 0.50000000D-01

| OBSERVED VALUES | ESTIMATED VALUE | RESIDUAL | ESTIMATED S.E. | LOWER LIMIT | UPPER LIMIT |
|---|---|---|---|---|---|
| 117.30000 | 116.89493 | 0.40507283 | 0.35231109 | 116.03285 | 117.75700 |
| 104.90000 | 105.52164 | -0.62164188 | 0.20683878 | 105.01553 | 106.02776 |
| 95.300000 | 95.609284 | -0.30928430 | 0.19612970 | 95.129373 | 96.089196 |
| 87.300000 | 86.856482 | 0.44351834 | 0.20538341 | 86.353927 | 87.359036 |
| 79.200000 | 79.037830 | 0.16217005 | 0.19521417 | 78.560159 | 79.515501 |
| 71.900000 | 71.985065 | -0.85064950D-01 | 0.17431129 | 71.558541 | 72.411589 |
| 65.700000 | 65.568615 | 0.13138498 | 0.17133949 | 65.149363 | 65.987867 |
| 59.800000 | 59.682176 | 0.11782363 | 0.21489984 | 59.156336 | 60.208016 |
| 54.000000 | 54.240515 | -0.24051461 | 0.30186196 | 53.501886 | 54.979143 |

MEAN OF OBSERVATIONS= 81.711111
SUM OF THE RESIDUALS= 0.346409600-02

SPECIFICS OF THE YI SOLUTIONS T(ALPHA)= 2.4469080 ALPHA= 0.50000000D-01

| ESTIMATED VALUE | EST VARIANCE | CONFIDENCE INTERVAL | | |
|---|---|---|---|---|
| 116.89493 | 0.12412310 | 116.89493 | (+-) | 0.86207283 |
| 0.50251918 | 0.79454589D-03 | 0.50251918 | (+-) | 0.68972685D-01 |
| 0.16924118D-01 | 0.15727466D-05 | 0.16924118D-01 | (+-) | 0.30686476D-02 |

COVARIANCE MATRIX OF THE SUPERPOSITION CONSTANTS

| | | |
|---|---|---|
| 0.31031D-01 | 0.61974D-01 | -0.65375D-01 |
| 0.61974D-01 | 0.31464 | -0.40651 |
| -0.65375D-01 | -0.40651 | 0.54909 |

CONCLUDES STAT PACKAGE



FIGURE 2

# MULTI-LEVEL ADAPTIVE SOLUTIONS TO BOUNDARY-VALUE PROBLEMS*

Achi Brandt
Weizmann Institute of Science, Rehovot, Israel
IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598

ABSTRACT. The boundary-value problem is discretized on several grids (or finite-element spaces) of widely different mesh sizes. Interactions between these levels enable us (i) to solve the possibly nonlinear system of n discrete equations in O(n) operations (40n additions and shifts for Poisson problems). (ii) To conveniently adapt the discretization (the local mesh size, local order of approximation, etc.) to the evolving solution in a nearly optimal way, obtaining "∞-order" approximations and low n, even when singularities are present. General theoretical analysis of the numerical process. Numerical experiments with linear and nonlinear, elliptic and mixed-type (transonic flow) problems - confirm theoretical predictions.

## 1. INTRODUCTION.

In most numerical procedures for solving partial differential equations, the analyst first discretizes the problem, choosing approximating algebraic equations on a finite dimensional approximation space, and then devises a numerical process to (nearly) solve this huge system of discrete equations. Usually, no real interplay is allowed between discretization and solution processes. This results in enormous waste: The discretization process, being unable to predict the proper resolution and the proper order of approximation at each location, produces a mesh which is too fine. The algebraic system thus becomes unnecessarily large in size, while accuracy usually remains rather low, since local smoothness of the solution is not being properly exploited. On the other hand, the solution process fails to take advantage of the fact that the algebraic system to be solved does not stand by itself, but is actually an approximation to continuous equations, and therefore can itself be approximated by other (much simpler) algebraic systems.

The purpose of the work reported here is to study how to intermix discretization and solution processes, thereby making both of them orders-of-magnitude more effective. The method to be proposed is not "saturated", that is, accuracy grows indefinitely as computations proceed. The rate of convergence (overall error E as function of computational work W) is in principle of "infinite order", e.g., $E \sim \exp(-\beta^d W)$ for a d-dimensional problem which has a solution with scale-ratios $\geq\beta>0$; or $E \sim \exp(-W^{1/2})$, for problems with arbitrary thin layers (see Sec. 9).

* The research reported here was partly supported by the Israel Commission for Basic Research. Part of the research was conducted at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia.

This paper will appear in the journal Mathematics of Computation. Permission of the editor of that journal to publish it in this technical manual is appreciated.

The basic idea of the Multi Level Adaptive Techniques (MLAT) is to work not with a single grid, but with sequence of grids ("levels") of increasing fineness, which may be introduced and changed in the process, and which constantly interact with each other. For description purposes, it is convenient to regard this technique as composed of two main concepts:

(1) The Multi Grid (MG) method for solving discrete equations. This method iteratively solves a system of discrete (finite-difference or finite-element) equations on a given grid, by constant interactions with a hierarchy of coarser grids, taking advantage of the relation between different discretizations of the same continuous problem. This method can be viewed in two complimentary ways: One is to view the coarser grids as correction grids, accelerating convergence of a relaxation scheme on the finest grid by efficiently liquidating smooth error components. (See general description in Sec. 2 and algorithm in Sec. 4.) Another point of view is to regard finer grids as the correction grids, improving accuracy on coarser grids by correcting their forcing terms. The latter is a very useful point of view, making it possible to manipulate accurate solutions on coarser grids, with only infrequent "visits" to pieces of finer levels. (This is the basis for the multi-grid treatment of non-uniform grids; cf. Secs. 7.2 and 7.5. The FAS mode for nonlinear problems and the adaptive procedures stem from this viewpoint.) The two seemingly different approaches actually amount to the same algorithm (in the simple case of "coextensive" levels).

The multi-grid process is very efficient: A discrete system of n equations (n points in the finest grid) is solved, to the desired accuracy, in O(n) computer operations. If $\bar{P}$ parallel processors are available, the required number of computer steps is $O(n/\bar{P} + \log n)$. For example, only 40n additions and shifts are required for solving the 5-point Poisson equation on a grid with n points (see Sec. 6.3). This efficiency does not depend on the shape of the domain, the form of the boundary conditions, or the mesh-size, and is not sensitive to choice of parameters. The memory area required is essentially only the minimal one, that is, the storage of the problem and the solution. In fact, if the amount of numerical data is small and only few functionals of the solution are wanted, the required memory is only O(log n), with no need for external memory (see Sec. 7.5).

Multi-grid algorithms are not difficult to program, if the various grids are suitably organized. We give an example (Appendix B) of a FORTRAN program, showing the typical structure, together with its computer output, showing the typical efficiency. With such an approach, the programming of any new multi-grid problem is basically reduced to the programming of a usual relaxation routine. The same is true for nonlinear problems where no linearization is needed, due to the FAS (Full Approximation Storage) method introduced in Sec. 5.

Multi-grid solution times can be predicted in advance, - a recipe is given and compared with numerical tests (Sec. 6). The basic tool is the local mode (Fourier) analysis, applied to the locally linearized-freezed difference equations, ignoring far boundaries. Such an analysis yields a very good approximation to the behavior of the high-frequency error modes, which are exactly the only significant modes in the multi-grid process, since the low-frequency error modes are

liquidated in the coarse-grids processing, with negligible amounts of computational work. Thus, mode analysis gives a very realistic prediction of convergence rates per unit work. (For model problems, the analysis can be made rigorous; see Appendix C.) The mode analysis can, therefore, be used to choose suitable relaxation schemes (Sec. 3) and suitable criteria for switching and interpolating between the grids (Appendix A). Our numerical tests ranged from simple elliptic problems to non-linear mixed-type (transonic flow) problems, which included hyperbolic regions and discontinuities (shocks). The results show that, as predicted by the mode analysis, errors are reduced by an order of magnitude (factor 10) expending computational work equivalent to 4 to 5 relaxation sweeps on the finest grid.

(2) Adaptive discretization. Mesh-sizes, orders of approximation and other discretization parameters are treated as spatial variables. Using certain general internal criteria, these variables are controlled in a sub-optimal way, adapting themselves to the computed solution. The criteria are devised to obtain maximum overall accuracy for a given amount of calculations; or, equivalently, minimum of calculations for given accuracy. (In practice only near-optimality should of course be attempted, otherwise the required control would become more costly than the actual computations. See Sec. 8.) The resulting discretization will automatically resolve thin layers (when required), refine meshes near singlular points (that otherwise may "contaminate" the whole solution), exploit local smoothness of solutions (in proper scale), etc. (see Sec. 9).

Multi-grid processing and adaptive discretization can be used independently of each other, but their combination is very fruitful: MG is the only fast (and convenient) method to solve discrete equations on the non-uniform grids typically produced by the adaptive procedure. Its iterative character fits well into the adaptive process. The two ideas use and relate similar concepts, similar data structure, etc. In particular, an efficient and very flexible way to construct any adaptive grid is as a sequence of uniform subgrids, the same sequence used in the multi-grid process, where finer levels may be confined to increasingly narrower subdomains to produce the desired local refinement. In this structure, the difference equations can be defined separately on each of the uniform subgrids, which interact with each other through the multi-grid process. Thus, difference equations should only be constructed on equidistant points, which makes it easy to employ high and adaptive orders of approximation. Moreover, the finer, localized subgrids may be defined in terms of suitable local coordinates, facilitating, for example, the use of high-order approximation near a certain piece of the boundary, with all these pieces naturally patched together by the multi-grid process (Sec. 7).

The presentation in this article is mainly in terms of finite-difference solutions to partial-differential boundary-value problems. The basic ideas, however, are more general, applicable to integro-differential problems, functional minimization problems, etc., and to finite-elements discretization. The latter is briefly discussed in Secs. A.5 and 7.3.

Contents of the article:

1. Introduction
2. Multi-grid philosophy
3. Relaxation and its smoothing rate
   3.1. An example
   3.2. General results
   3.3. Acceleration by weighting
4. A multi-grid algorithm (Cycle C) for linear problems
   4.1. Indefinite problems and the size of the coarsest grid
5. The FAS (Full Approximation Storage) algorithm
6. Performance estimates and numerical tests
   6.1. Predictability
   6.2. Multi-grid rates of convergence
   6.3. Overall multi-grid computational work
   6.4. Numerical experiments: Elliptic problems
   6.5. Numerical experiments: Transonic flow problems
7. Non-uniform grids
   7.1. Organizing non-uniform grids
   7.2. The multi-grid algorithm on non-uniform grids
   7.3. Finite-elements generalization
   7.4. Local transformations
   7.5. Segmental refinement
8. Adaptive discretization techniques
   8.1. Basic principles
   8.2. Continuation methods
   8.3. Practice of discretization control
   8.4. Generalizations
9. Adaptive discretization: Case studies
   9.1. Uniform-scale problems
   9.2. One-dimensional case
   9.3. Singular perturbation: Boundary layer resolution
   9.4. Singular perturbation without boundary layer resolution
   9.5. Boundary corners
   9.6. Singularities

10. Historical notes and acknowledgements

References

Appendices:

A. Interpolation and stopping criteria: Analysis and rules

A.1. Coarse-grid amplification factors

A.2. The coarse-to-fine interpolation $I_{k-1}^{k}$

A.3. Effective smoothing rate

A.4. The fine-to-coarse weighting of residuals $I_{k}^{k-1}$ and the coarse grid operator $L^{k-1}$

A.5. Finite-elements procedures

A.6. Criteria for slow convergence rates

A.7. Convergence criteria on coarser grids

A.8. Convergence on the finest grid

A.9. Partial relaxation sweeps

A.10. Convergence criteria on non-uniform grids

B. Sample multi-grid program and output

C. Rigorous bound to model-problem convergence rate

Figures

## 2. MULTI-GRID PHILOSOPHY.

Suppose we have a set of grids $G^0, G^1 \ldots, G^M$, all approximating the same domain $\Omega$ with corresponding meshsizes $h_0 > h_1 > \ldots > h_M$. For simplicity one can think of the familiar uniform square grids, with the mesh-size ratio $h_{k+1}:h_k = 1:2$. Suppose further that a differential problem of the form

$$(2.1) \qquad LU(x) = F(x) \quad \text{in } \Omega, \qquad \Lambda U(x) = \Phi(x) \text{ on the boundary } \partial\Omega,$$

is given. On each grid $G^k$ this problem can be approximated by difference equations of the form

$$(2.2) \qquad L^k U^k(x) = F^k(x) \quad \text{for } x \in G^k, \qquad \Lambda^k U^k(x) = \Phi^k(x) \quad \text{for } x \in \partial G^k.$$

(See example in Sec. 3.1). We are interested in solving this discrete problem on the finest grid, $G^M$. The main idea is to exploit the fact that the discrete problem on a coarser grid, $G^k$ say, approximates the same differential problem and hence can be used as a certain approximation to the $G^M$ problem. A simple use of this fact has long been made by various authors (e.g., [14]); namely, they first solved (approximately) the $G^k$ problem, which involves an algebraic system much smaller and thus much easier to solve than the given $G^M$ problem, and then they iterpolated

their solution from $G^k$ to $G^M$, using the result as a first approximation in some iterative process for solving the $G^M$ problem. A more advanced technique was to use a still coarser grid in a similar manner when solving the $G^k$ problem, and so on. The next natural step is to ask whether we can exploit the proximity between the $G^k$ and $G^M$ problems not only in generating a good first approximation on $G^M$, but also in the process of improving the first approximation.

More specifically let $u^M$ be an approximate solution of the $G^M$ problem, and let

$$(2.3) \qquad L^M u^M = F^M - f^M, \qquad \Lambda^M u^M = \Phi^M - \phi^M.$$

The discrepancies $f^M$ and $\phi^M$ are called the residual functions, or residuals. Assuming for simplicity that L and $\Lambda$ are linear (cf. Sec. 5 for the non-linear case), the exact discrete solution is $U^M = u^M + V^M$, where the correction $V^M$ satisfies the residual equations

$$(2.4) \qquad L^M V^M = f^M, \qquad \Lambda^M V^M = \phi^M.$$

Can we solve this equation, to a good first approximation, again by interpolation from solutions on coarser grids? As it is, the answer is generally negative. Not every $G^M$-problem has meaningful approximation on a coarser grid $G^k$. For instance, if the right-hand-side $f^M$ fluctuates rapidly on $G^M$, with wavelength less than $4h_M$, these fluctuations are not visible on coarser grids. Such rapidly-fluctuating residuals $f^M$ are exactly what we get when the approximation $u^M$ has itself been obtained as an interpolation from a coarser-grid solution.

An effective way to damp rapid fluctuations in residuals is by usual relaxation procedures, e.g., the Gauss-Seidel relaxation (see Sec. 3). At the first few iterations such procedures usually seem to have fast convergence, with residuals (or corrections) rapidly decreasing from one iteration to the next, but soon after the convergence rate levels off and becomes very slow. Closer examination (see Sec. 3 below) shows that the convergence is fast as long as the residuals have strong fluctuations on the scale of the grid. As soon as the residuals are smoothed out, convergence slows down.

This is then exactly the point where relaxation sweeps should be discontinued and approximate solution of the (smoothed out) residual equations by coarser grids should be employed.

The Multi-Grid (MG) methods are systematic methods of mixing relaxation sweeps with approximate solution of residual equations on coarser grids. The residual equations are in turn also solved by combining relaxation sweeps with corrections through still coarser grids, etc. The coarsest grid $G^0$ is coarse enough to make the solution of its algebraic system inexpensive compared with, say, one relaxation sweep over the finest grid.

The following sections further explain these ideas. Sec. 3.1 explains, through a simple example, what is a relaxation sweep and shows that it indeed smooths out the residuals very efficiently. The smoothing rates of general difference systems are summarized in Sec. 3.2. A full multi-grid algorithm, composed of relaxation sweeps over the various grids with suitable interpolations in between, is then presented in Sec. 4. An important modification for nonlinear problems is described in Sec. 5 (and used later as the basic algorithm for non-uniform grids and adaptive procedures). Appendix A supplements these with suitable stopping criteria, details of the interpolation procedures and special techniques (partial relaxation).

## 3. RELAXATION AND ITS SMOOTHING RATE.

3.1 An Example. Suppose, for example, we are interested in solving the partial differential equation.

$$(3.1) \qquad LU(x,y) \equiv a \frac{\partial^2 U(x,y)}{\partial x^2} + b \frac{\partial^2 U(x,y)}{\partial y^2} = F(x,y)$$

with some suitable boundary conditions. Denoting by $U^k$ and $F^k$ approximations of U and F, respectively, on the grid $G^k$, the usual second-order discretization of (3.1) is

$$(3.2) \qquad L^k U^k_{\alpha,\beta} \equiv a \frac{U^k_{\alpha+1,\beta} - 2U^k_{\alpha,\beta} + U^k_{\alpha-1,\beta}}{h_k^2} + b \frac{U^k_{\alpha,\beta+1} - 2U^k_{\alpha,\beta} + U^k_{\alpha,\beta-1}}{h_k^2} = F^k_{\alpha,\beta}$$

where

$$U^k_{\alpha,\beta} = U^k(\alpha h_k, \beta h_k), \quad F^k_{\alpha,\beta} = F^k(\alpha h_k, \beta h_k); \ \alpha,\beta \text{ integers.}$$

(In the multi-grid context it is important to define the difference equations in this divided form, without, for example, multiplying throughout by $h_k^2$, in order to get the proper relative scale at the different levels.) Given an approximation u to $U^k$, a simple example of a relaxation scheme to improve it is the following.

Gauss-Seidel Relaxation: The points $(\alpha,\beta)$ of $G^k$ are scanned one by one in some prescribed order; e.g., lexicographic order. At each point the value $u_{\alpha,\beta}$ is replaced by a new value, $\bar{u}_{\alpha,\beta}$, such that equation (3.2) at that point is satisfied. That is, $\bar{u}_{\alpha,\beta}$ satisfies

$$(3.3) \qquad a \frac{u_{\alpha+1,\beta} - 2\bar{u}_{\alpha,\beta} + \bar{u}_{\alpha-1,\beta}}{h_k^2} + b \frac{u_{\alpha,\beta+1} - 2\bar{u}_{\alpha,\beta} + \bar{u}_{\alpha,\beta-1}}{h_k^2} = F^k_{\alpha,\beta},$$

where the new values $\bar{u}_{\alpha-1,\beta}$, $\bar{u}_{\alpha,\beta-1}$ are used since, in the lexicographic order, by the time $(\alpha,\beta)$ is scanned new values have already replaced old values at $(\alpha-1,\beta)$ and $(\alpha,\beta-1)$.

A complete pass, scanning in this manner all the points of $G^k$, is called a (Gauss-Seidel lexicographic) $G^k$ relaxation sweep. The new approximation $\bar{u}$ does not satisfy (3.2), and further relaxation sweeps may be required to improve it. An important quantity therefore is the rate of convergence, $\mu$ say, which may be defined by

$$(3.4)\qquad \mu = \frac{||\bar{v}||}{||v||}, \qquad \text{where } v = U^k - u, \quad \bar{v} = U^k - \bar{u},$$

$||\cdot||$ being any suitable discrete norm.

The rate of convergence of the above relaxation scheme is asymptotically very slow. That is, except for the first few relaxation sweeps we have $\mu = 1 - O(h_k^2)$. This means that we have to perform $O(h_k^{-2})$ relaxation sweeps to reduce the error order of magnitude.

In the multi-grid method, however, the role of relaxation is not to reduce the error, but to smooth it out; i.e., to reduce the high-frequency components of the error (the lower frequencies being reduced by relaxation sweeps on coarser grids). In fact, since smoothing is basically a local process (high frequencies have short coupling range), we can analyze it in the interior of $G^k$ by (locally) expanding the error in Fourier series. This will allow us to study separately the convergence rate of each Fourier component, and, in particular, the convergence rate of high-frequency components, which is the rate of smoothing.

Thus to study the $\theta=(\theta_1,\theta_2)$ Fourier component of the error functions $v$ and $\bar{v}$ before and after the relaxation sweep, we put

$$(3.5)\qquad v_{\alpha,\beta} = A_\theta e^{i(\theta_1\alpha+\theta_2\beta)} \quad \text{and} \quad \bar{v}_{\alpha,\beta} = \bar{A}_\theta e^{i(\theta_1\alpha+\theta_2\beta)}.$$

Subtracting (3.2) from (3.3), we get the relation

$$(3.6)\qquad a\,(v_{\alpha+1,\beta} - 2\bar{v}_{\alpha,\beta} + \bar{v}_{\alpha-1,\beta}) + c\,(v_{\alpha,\beta+1} - 2\bar{v}_{\alpha,\beta} + \bar{v}_{\alpha,\beta-1}) = 0,$$

from which, by (3.5),

$$(ae^{i\theta_1} + ce^{i\theta_2})\,A_\theta + (ae^{-i\theta_1} + ce^{-i\theta_2} - 2a - 2c)\,\bar{A}_\theta = 0.$$

Hence the convergence rate of the $\theta$ component is

$$(3.7)\qquad \mu(\theta) = \left|\frac{\bar{A}_\theta}{A_\theta}\right| = \left|\frac{ae^{i\theta_1} + ce^{i\theta_2}}{2a+2c - ae^{-i\theta_1} - ce^{-i\theta_2}}\right|.$$

Define $|\theta| = \max(|\theta_1|,|\theta_2|)$. In domains of diameter $O(1)$ the lowest Fourier components have $|\theta| = O(h_k)$, and their convergence rate therefore is, $\mu(\theta) = 1 - O(h_k^2)$. Here, however, we are interested in the <u>rate of smoothing</u>, which is defined by

$$(3.8)\qquad \bar{\mu} = \max_{\hat{\rho}\pi \le |\theta| \le \pi} \mu(\theta),$$

where $\hat{\rho}$ is the mesh-size ratio and the range $\hat{\rho}\pi \leq |\theta| \leq \pi$ is the suitable range of high-frequency components, i.e., the range of components that cannot be approximated on the coarser grid, because its mesh-size is $h_{k-1} = h_k / \hat{\rho}$. We will assume here that $\hat{\rho} = \frac{1}{2}$, which is the usual ratio (cf. Sec. 6.2).

Consider first the case a=c (Poisson equation). A simple calculation shows that $\bar{\mu} = \mu(\frac{\pi}{2}, \arccos\frac{4}{5}) = .5$. This is a very satisfactory rate; it implies that 3 relaxation sweeps reduce the high-frequency error-components by almost an order of magnitude. Similar rates are obtained for general a and c, provided a/c is of moderate size.

The rate of smoothing is less remarkable in the degenerate case a<<c (or c<<a). For instance,

$$\mu(\tfrac{\pi}{2}, 0) = \left[\frac{a^2 + c^2}{a^2 + (c+2a)^2}\right]^{1/2} \tag{3.9}$$

which approaches 1 as $a \to 0$. Thus, for problems with such a degeneracy, Gauss-Seidel relaxation is not a suitable smoothing scheme. But other schemes exist. For example,

Line Relaxation: Instead of treating each point $(\alpha,\beta)$ of $G^k$ separately, one takes simultaneously a line of points at a time, where a line is the set of all points $(\alpha,\beta)$ in $G^k$ with the same $\alpha$ (a vertical line). All the values $u_{\alpha\beta}$ on such a line are simultaneously replaced by new values $\bar{u}_{\alpha\beta}$ which simultaneously satisfy all the equations (3.2) on that line. (This is easy and inexpensive to do, since the system of equations to be solved for each such line is a tridiagonal, diagonally dominant system. See, e.g., in [17].) As a result, we get the same relation as (3.3) above, except that $u_{\alpha,\beta+1}$ is replaced by $\bar{u}_{\alpha,\beta+1}$. Hence, instead of (3.7) we will get:

$$\mu(\theta) = \left|\frac{a}{2(a+c - c\cos\theta_2) - ae^{-i\theta_1}}\right| \tag{3.10}$$

from which one can derive the smoothing rate

$$\bar{\mu} = \max\left\{5^{-1/2}, \frac{a}{a+2c}\right\}, \tag{3.11}$$

which is very satisfactory, even in the degenerate case a<<c.

3.2. General results. The above situation is very general (see [4] and Chapter 3 of [3]): For any uniformly elliptic system of difference equations, it can be shown that few relaxation sweeps are enough to reduce the high-frequency error components by an order of magnitude.

The same holds for degenerate elliptic systems, provided a suitable relaxation scheme is selected. A scheme of line-relaxation which alternately use all line directions and all sweeping directions is suitable for any degenerate case. Moreover, such a scheme is suitable even for non-elliptic systems, provided it is used "selectively"; i.e., the entire domain is swept in all directions, but new values are not introduced at points where a local test shows the equation to be non-elliptic and the forward characteristic direction to conflict with the current sweeping direction.

By employing local mode analysis (analysis of Fourier components) similar to the example above, one can explicitly calculate the smoothing rate $\bar{\mu}$ for any given difference equation with any given relaxation scheme. (Usually $\bar{\mu}$ should be calculated numerically; an efficient FORTRAN subroutine exists; typical values are given in Table 1, in Sec. 6.2). In this way, one can select the best relaxation scheme from a given set of possibilities. The selection of the difference equation itself may also take this aspect into account. This analysis can also be done for non-linear problems (or linear problems with non-constant coefficients), by local linearization and coefficients freeze. Such localization is fully justified here, since we are interested only in a local property (the property of smoothing. By contrast, one cannot make similar mode analysis to predict the overall convergence rate $\mu$ of a given relaxation scheme, since this is not a local property).

An important feature of the smoothing rate $\bar{\mu}$ is its insensitivity. In the above example no relaxation parameters were assumed. We could introduce the usual relaxation parameter $\omega$; i.e., replace at each point the old value $u_{\alpha,\beta}$ not with the calculated $\bar{u}_{\alpha,\beta}$, but with $u_{\alpha,\beta} + \omega(\bar{u}_{\alpha,\beta}-u_{\alpha,\beta})$. The mode analysis shows, however, that no $\omega \neq 1$ provides a smoothing rate better than $\omega=1$. In other cases, $\omega=1$ is not optimal, but its $\bar{\mu}$ is not significantly larger than the minimal $\bar{\mu}$. In delayed-displacement relaxation schemes a value $\omega < \omega_{critical} < 1$ should often be used to obtain $\bar{\mu} < 1$, but there is no sensitive dependence on the precise value of $\omega$, and suitable values are easily obtained from the local mode analysis. Generally, the smoothing rate of delayed-displacement schemes is somewhat worse than that of immediate-displacement schemes, and the latter should, therefore, be preferred, except when parallel processing is used.

3.3. Acceleration by weighting. The rate of smoothing $\bar{\mu}$ may sometimes be further improved by various parameters introduced into the scheme. Since $\bar{\mu}$ is reliably obtained from the local mode analysis, we can optimize these parameters to minimize $\bar{\mu}$. For linear problems, such optimal parameters can be determined once and for all, since they do not depend on the shape of the domain. For nonlinear problems precise optimization is expensive, and one should prefer the simpler, more robust relaxation schemes, such as SOR.

One general way of parametrization is the weighting of corrections. We first calculate, in any relaxation scheme, the required correction $\delta u_\nu = \bar{u}_\nu - u_\nu$ (where $\nu = (\alpha,\beta)$ or, for a general dimension d,

$\nu = (\nu_1, \nu_2, \ldots, \nu_d)$, $\nu_j$ integers). Then, instead of introducing these corrections, we actually introduce corrections which are some linear combination of $\delta u_\nu$ at neighboring points. That is, the actual new values are

$$\tilde{u}_\nu = u_\nu + \sum_{\gamma \in \Gamma} \omega_\gamma \, \delta u_{\nu+\gamma} \tag{3.12}$$

where the weights $\omega_\gamma$ are the parameters, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_d)$, $\gamma_j$ integers and $\Gamma$ a small set near $(o,o,\ldots,o)$. For any fixed $\Gamma$ we can optimize the weights. In case $\Gamma = \{o\}$, $\omega_o$ is the familiar relaxation parameter. Weighting larger $\Gamma$ is useful in delayed displacement relaxation schemes. For immediate-displacement line relaxation, weighting along the line may be useful.

Examples. In case of simultaneous displacement (Jacobi) relaxation for the 5-points Poisson equation, the optimal weights for $\Gamma=\{o\}$ is $\omega_{oo} = .8$, for which the smoothing rate is $\bar{\mu} = .60$. For the set $\Gamma = \{(\gamma_1,\gamma_2): |\gamma_1|+|\gamma_2| \leq 1\}$ the optimal weights are $\omega_{oo} = 6\omega_{o,\pm 1} = 6\omega_{\pm 1,0} = 48/41$, yielding $\bar{\mu} = 9/41$. This rate seems very attractive; the smoothing obtained in one sweep equals that obtained by $(\log \frac{9}{41}) \,/\, (\log \frac{1}{2}) = 2.2$ sweeps of Gauss-Seidel relaxation. Actually, however, each sweep of this weighted-Jacobi relaxation requires 9 additions and 3 multiplications per grid point, whereas each Gauss-Seidel sweep requires only 4 additions and 1 multiplication per point, so that the two methods have almost the same convergence rate per operation, Gauss-Seidel being slightly faster. The weighted Jacobi scheme is considerably more efficient than any other simultaneous-displacement scheme, but like any carefully weighted scheme, it is considerably more sensitive to various changes.

The acceleration by weighting can be more significant for higher-order equations. For the 13-points biharmonic operator, Gauss-Seidel relaxation requires 12 additions and 3 multiplications per grid point and gives $\bar{\mu} = .802$, while weighted Jacobi (with weights $\omega_{oo} = 1.552$, $\omega_{o,\pm 1} = \omega_{\pm 1,o} = .353$) requires 17 additions and 5 multiplications per point and gives $\bar{\mu} = .549$, which is 2.7 times faster. (The best relaxation sweep for the biharmonic equation $\Delta^2 U = F$ is to write it as the system $\Delta V = F$, $\Delta U = V$ and sweep Gauss-Seidel, alternatively on U and V. Such a double sweep costs 8 additions and 2 multiplications per grid point, and yields $\bar{\mu}=.5$. But a similar procedure is not possible for general 4-th order equations.)

## 4. A MULTI-GRID ALGORITHM (CYCLE C) FOR LINEAR PROBLEMS

There are several actual algorithms for carrying out the basic multi-grid idea, each with several possible variations. We present here an algorithm (called "Cycle C" in [3]) which is easy to program, generally applicable and never significantly less efficient than the others ("Cycle A" and "Cycle B"). The operation of the algorithm for linear problems is easier to learn, and is therefore described first. In the next section the FAS (Full-Approximation-Storage) mode of operation, suitable for nonlinear problems and other important generalizations, will be described. A flow-chart of the algorithm is given in Figure 1. (For completion, we also flowchart, in Fig. 2, Cycles A and B.) A sample FORTRAN program of this cycle, together with a computer output, is given in Appendix B.

Cycle C starts with some approximation $u_o^M$ being given on the finest grid $G^M$. In the linear case one can start with any approximation, but a major part of the computations is saved if $u_o^M$ has smooth residuals (e.g., if $u_o^M$ satisfies the boundary conditions and $L^M u_o^M - F^M$ is smooth. As explained in Sec. 6, smoothing the residuals involves most of the computational effort). In the nonlinear case, one may have to use a continuation procedure, usually performed on coarser grids (cf. Sec. 8.2). Even for linear problems, the most efficient algorithm is to obtain $u_o^M$ by interpolating from an approximate solution $u^{M-1}$ calculated on $G^{M-1}$ by a similar algorithm. (Hence the denomination "cycle" for our present algorithm, which would generally serve as the basic step in processes of continuation, refinement and grid adaptation, or as a time step in evolution problems). For highest efficiency, the interpolation from $u^{M-1}$ to $u_o^M$ should be of sufficiently high order, to exploit all smoothness in $u^{M-1}$. (Cf. (A.7) in Sec. A.2, and see also Sec. 6.3.)

The basic rule in Cycle C is that each $v^k$ (the function defined on the grid $G^k$; k=0, 1, .., M-1) is designed to serve as a correction for the approximation $v^{k+1}$ previously obtained on the next finer grid $G^{k+1}$, if and when that approximation actually requires a coarse-grid correction, i.e., if and when relaxation over $G^{k+1}$ exhibits slow rate of convergence. Thus, the equations to be (approximately) satisfied by $v^k$ are [1]

$$L^k v^k = f^k, \qquad \Lambda^k v^k = \phi^k, \tag{4.1}$$

where $f^k$ and $\phi^k$ are the residuals (to the interior equation and the boundary condition, respectively) left by $v^{k+1}$, that is,

$$f^k = I_{k+1}^k \left(f^{k+1} - L^{k+1} v^{k+1}\right), \qquad \phi^k = I_{k+1}^k \left(\phi^{k+1} - \Lambda^{k+1} v^{k+1}\right). \tag{4.2}$$

---

[1] We denote by $V^k$ the functions in the equations, to distinguish from their computed approximations $v^k$. When $v^k$ is changing in the algorithm (causing $V^{k-1}$ to change), $V^k$ remains fixed.

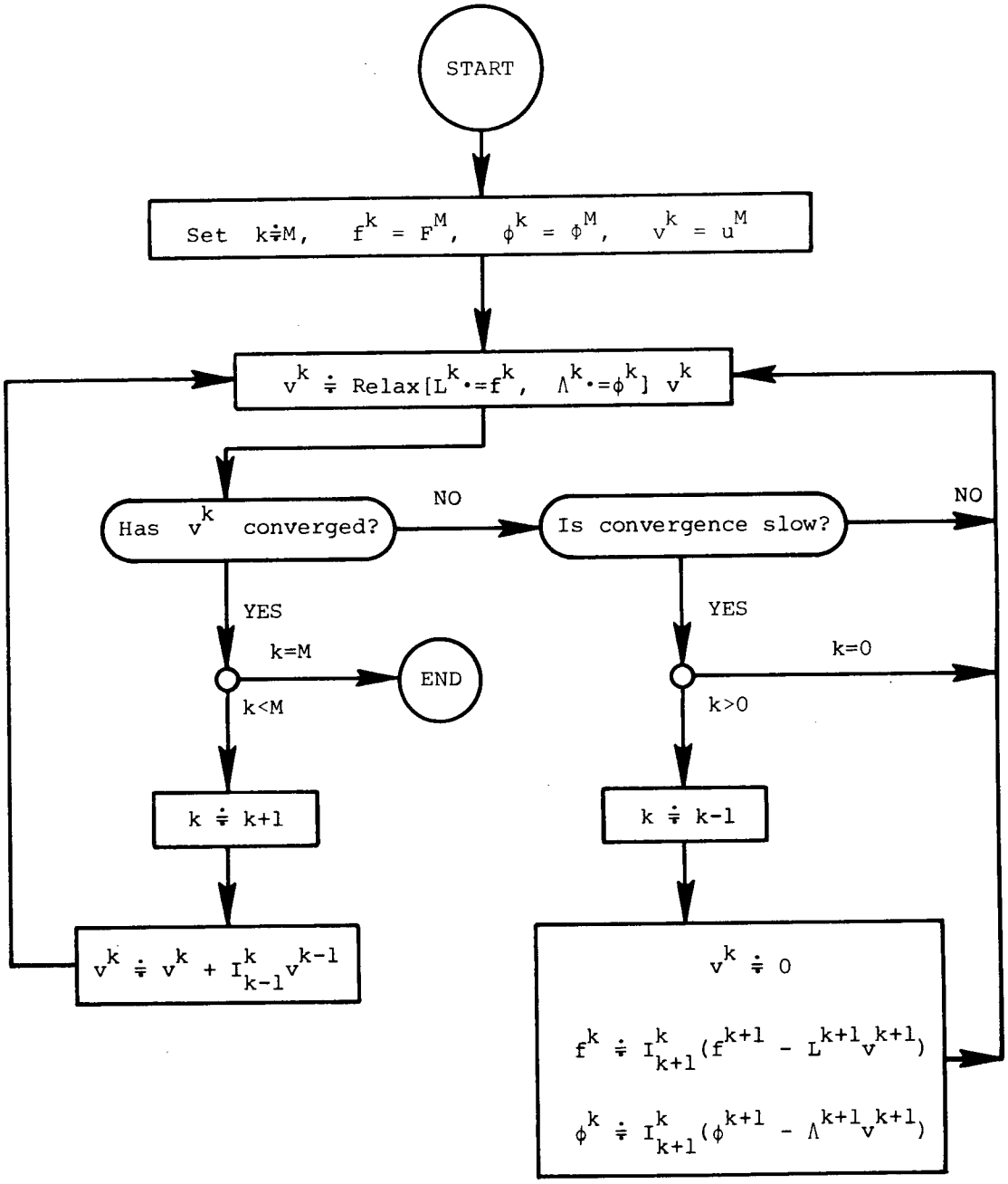



Figure 1. Cycle C, Linear Problems.

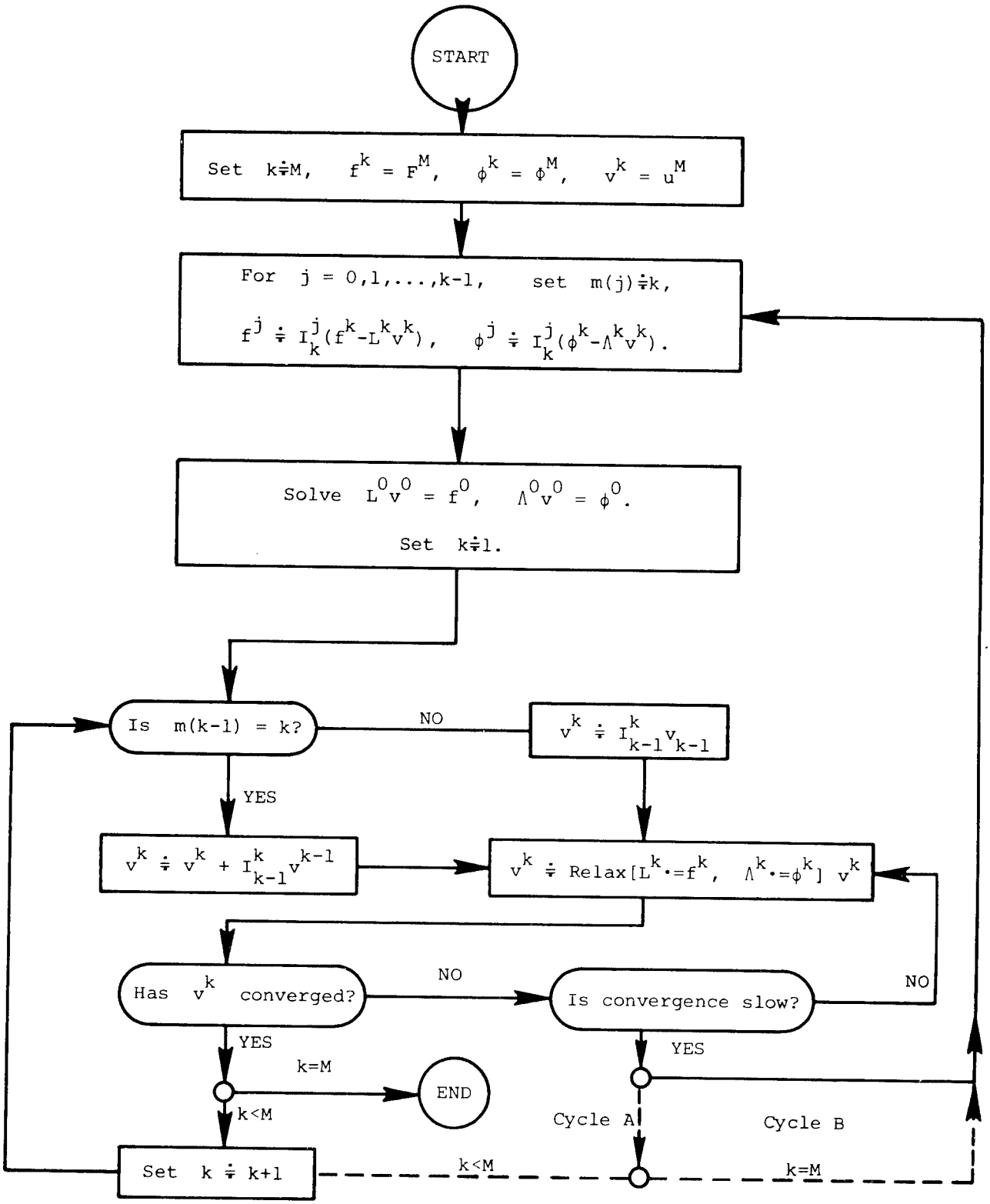



Figure 2. Cycles A and B, Linear Problems.

We use the notation $I_m^k$ to represent interpolation from $G^m$ to $G^k$. In case $m>k$, $I_m^k$ may represent a simple transfer of values to the coarser grid $G^k$ from the corresponding points in the finer grid $G^m$; or instead, it may represent transfer of some weighted averages. In case $k>m$, as in step (e) below, $I_m^k$ is usually a polynominal interpolation of a suitable order (at least the order of the differential equation. See Secs. A.2 and A.4 for more details).

The equations on $G^k$ are thus defined in terms of the approximate solution on $G^{k+1}$. On the finest grid $G^M$, the equations are the original ones; namely

$$f^M = F^M, \qquad \phi^M = \Phi^M, \qquad v^M = u^M. \tag{4.3}$$

The steps of the algorithm are the following:

(a) Set $k \leftarrow M$ ($k$ is the working level, and we start at the finest level), and introduce the given approximation $v^M \leftarrow u_o^M$.

(b) Improve $v^k$ by one relaxation sweep for the difference equations (4.1). Symbolically, we write such a sweep as

$$v^k \leftarrow \text{Relax}\left[L^k \cdot = f^k,\ \Lambda^k \cdot = \phi^k\right] v^k \tag{4.4}$$

(c) If relaxation has sufficiently converged (the precise criterion is described in Secs. A.7 and A.8), go to Step (f). If not, and if the convergence rate is still fast (by a criterion given in Sec. A.6) go back to Step (b). If convergence is not obtained and the rate is slow, go to Step (d).

(d) If $k=o$ (the slow convergence has taken place at the coarsest grid $G^o$), go back to Step (b) (to continue relaxation nevertheless, since on $G^o$ relaxation is very inexpensive. If, however, the problem is indefinite, then slow rate of divergence may occur, in which case the $G^o$ problem should be solved directly. This is as inexpensive as relaxation, but requires additional programming. See Sec. 4.1 below). If $k>o$, lower $k$ by 1 (to compute correction on the next, coarser level). Compute $f^k$ and $\phi^k$ on this new level, using (4.2), put $v^k=o$ as the starting approximation, and go to Step (b).

(e) If $k=M$ (convergence has been obtained on the finest level), the algorithm is terminated. If $k<M$ ($v^k$ has converged and is ready to serve as a correction to $v^{k+1}$), put

$$v^{k+1} \leftarrow v^{k+1} + I_k^{k+1} v^k. \tag{4.5}$$

Then advance $k$ by 1 (to resume computations in the finer level) and go to Step (b).

The storage required for this algorithm is only a fraction more than the number of locations, $2n$ say, required to store $u^M$ and $F^M$ on the finest grid. Indeed, for a $d$-dimensional problem, a storage of roughly $2n/2^d$ locations is required to store $v^{M-1}$ and $f^{M-1}$, the next level requires $2n/2^{2d}$, etc. The total for all levels is

(4.6) $$2n(1+2^{-d}+2^{-2d}+\dots) \le 2n\frac{2^d}{2^d-1} .$$

(In the FAS version below, a major reduction of storage area is possible through segmental refinement. See Sec. 7.5.)

4.1. Indefinite Problems and the Size of the Coarsest Grid. If, on any grid $G^k$, the boundary value-problem (4.1) is a non-definite elliptic problem, with eigenvalues

(4.7) $$\lambda_1^k \le \lambda_2^k \le \dots \le \lambda_\ell^k < 0 < \lambda_{\ell+1}^k \le \lambda_{\ell+2}^k \le \dots ,$$

and with the corresponding eigenfunctions $v_1^k, v_2^k, \dots, v_\ell^k, v_{\ell+1}^k, \dots,$

then it cannot be solved by straight relaxation. Any relaxation sweep will reduce the error components in the space spanned by $v_{\ell+1}^k, v_{\ell+2}^k, \dots,$ but will magnify all components in the span of $v_1^k, v_2^k, \dots, v_\ell^k$. A multi-grid solution, however, is not seriously affected by this magnification, provided the magnified components are suitably reduced by the coarse-grid corrections. This will usually be the case, since these components are basically of low frequency and are well approximated on coarser grids. But care should be taken regarding the coarsest grid:

On the coarsest grid, an indefinite problem should be solved directly, (i.e., not by relaxation of any kind. Semi-iterative solutions, like Newton iterations for non-linear problems, are, of course, permissible). Furthermore, this grid should be fine enough to provide rough approximation to $v_1^k, v_2^k, \dots, v_\ell^k$ for any k, hence also for the corresponding differential eigenfunctions. This means that $G^0$ should contain at least $O(\ell)$, probably $2\ell$, points. Or, in other words, $G^0$ should be just fine enough to still have smoothing capability at any finer level $G^k$. For example, if SOR relaxations with $\omega<\omega_c$ are used, $h_0$ should satisfy (see [4] or Sec. 3 in [3])

(4.8) $$\mathrm{Re}\{B(\theta,h) / b_0(h)\} > 0, \qquad (0 \le h \le h_0),$$

where $B(\theta,h)$ is the symbol of $L^h$ (see (A.3) in Appendix A) and $b_0(h)$ is its central coefficient.

Usually, $G^0$ can still be coarse enough to have the direct solution of its equations still far less expensive than, say, one relaxation sweep over the finest grid, so that the indefinite problem is solved with the same overall efficiency as definite problems.

## 5. THE FAS (FULL APPROXIMATION-STORAGE) ALGORITHM.

In the FAS mode of the multi-grid algorithms, instead of storing a correction $v^k$ (designed to correct the finer-level approximation $u^{k+1}$), the idea is to store the full current approximation $u^k$, which is the sum of the correction $v^k$ and its base approximation $u^{k+1}$:

$$u^k = I_{k+1}^k u^{k+1} + v^k, \quad (k=0,1,\ldots,M-1). \tag{5.1}$$

In terms of these full-approximation functions, we can rewrite the correction equations (4.1-3) as [2]

$$L^k U^k = \bar{F}^k, \qquad \Lambda^k U^k = \bar{\Phi}^k, \tag{5.2}$$

where

$$\bar{F}^k = L^k(I_{k+1}^k u^{k+1}) + I_{k+1}^k (\bar{F}^{k+1} - L^{k+1} u^{k+1}),$$

$$\bar{\Phi}^k = \Lambda^k(I_{k+1}^k u^{k+1}) + I_{k+1}^k (\bar{\Phi}^{k+1} - \Lambda^{k+1} u^{k+1}), \tag{5.3}$$

$$(k=0,1,\ldots,M-1),$$

and where for k=M we have the original problem, i.e.,

$$\bar{F}^M = F^M, \qquad \bar{\Phi}^M = \Phi^M. \tag{5.4}$$

For linear problems, equations (5.2-4) are exactly equivalent to (4.1-3). The advantage of the FAS mode is that equations (5.2-4) apply equally well to nonlinear problems. To see this, consider for instance the nonlinear equation $L^M U^M = F^M$ given on the finest grid. Given an approximate solution $u^M$ we can still improve it by relaxation sweeps, with smoothing rates $\bar{\mu}$ (varying over the domain, but still reliably estimated by mode analyses, applied locally to the linearized-freezed equation). As in the linear case, the smoothed-out functions are the residual

$$f^M = F^M - L^M u^M$$

and the correction $U^M - u^M$. Therefore, the equation that can be approximated on coarser grids is the residual equation

$$L^M U^M - L^M u^M = f^M .$$

Its coarser-grid approximation is

$$L^{M-1} U^{M-1} - L^{M-1} I_M^{M-1} u^M = I_M^{M-1} f^M, \tag{5.5}$$

which is the same as (5.2) for k=M-1. In interpolating $U^{M-1}$ (or a computed approximation $u^{M-1}$) back to $G^M$, we should actually interpolate $U^{M-1} - I_M^{M-1} u^M$, because this is the coarse-grid approximation to the

---

[2] Again we distinguish between the notation $U^k$ used to write the equations and the computed approximation $u^k$. Equation (5.2), for k<M, is not equivalent to (2.2), although they both use the notation $U^k$.

smoothed-out function $U^M - u^M$. Similarly, in interpolating an (approximate) solution $U^k$ of (5.2) to the finer grid $G^{k+1}$, the polynominal interpolation should operate on the correction. Thus the interpolation is

$$u^{k+1} \leftarrow u^{k+1} + I_k^{k+1} \left(u^k - I_{k+1}^k u^{k+1}\right) , \tag{5.6}$$

which is equivalent to (4.5). Note that generally,

$$I_k^{k+1} I_{k+1}^k u^{k+1} \neq u^{k+1} .$$

The FAS (Cycle C) algorithm is the same algorithm as in Sec. 4, with the FAS equations (5.2-4) replacing (4.1-3), and with (5.6) replacing (4.5). It is flowcharted in Fig. 3.

The FAS mode has several important advantages: It is suitable for general nonlinear problems, with the same procedures (relaxation and interpolation routines) used at all levels. Thus, for example, only one relaxation routine should be written. Moreover, this mode is suitable for composite grids (non-uniform grids created by increasingly finer levels being defined on increasingly smaller subdomains; see Sec. 7.2), which is the basis for grid adaptation on one hand, and segmental refinement (see Sec. 7.5) on the other hand. Generally speaking, the basic feature of the FAS mode is that the function stored on a coarse grid $G^k$ coincides there with the fine-grid solution: $u^k = I_M^k u^M$. This enables us to manipulate accurate solutions on coarse grids.

The storage required for the FAS algorithm is again given by (4.6). With segmental refinement (Sec. 7.5) it can be reduced far below that, even to O(log n).

An important by-product of the FAS mode is a good estimate for the truncation error, which is useful in defining natural stopping criteria (see Sec. A.8) and grid-adaptation criteria (Sec. 8.3). Indeed, for any $k<m\leq M$ it can easily be shown (by induction on m, using (5.2-3)) that

$$\begin{aligned} \bar{F}^k - I_m^k \bar{F}^m &= L^k(I_m^k u^m) - I_m^k L^m u^m , \\ \bar{\Phi}^k - I_m^k \bar{\Phi}^m &= \Lambda^k(I_m^k u^m) - I_m^k \Lambda^m u^m , \end{aligned} \tag{5.7}$$

which are exactly the $G^m$ approximations to the $G^k$ truncation errors.

A slight disadvantage of the FAS mode is the longer calculation required in computing $\bar{F}^k$, which is almost twice as long as the calculation of $f^k$ in the former (Correction-Storage) mode. This extra calculation is equivalent to one extra relaxation sweep on $G^k$, but only for $k<M$, and is about 5% to 10% of the total amount of calculations. Hence, for linear problems on uniform grids, the CS mode is slightly preferable.

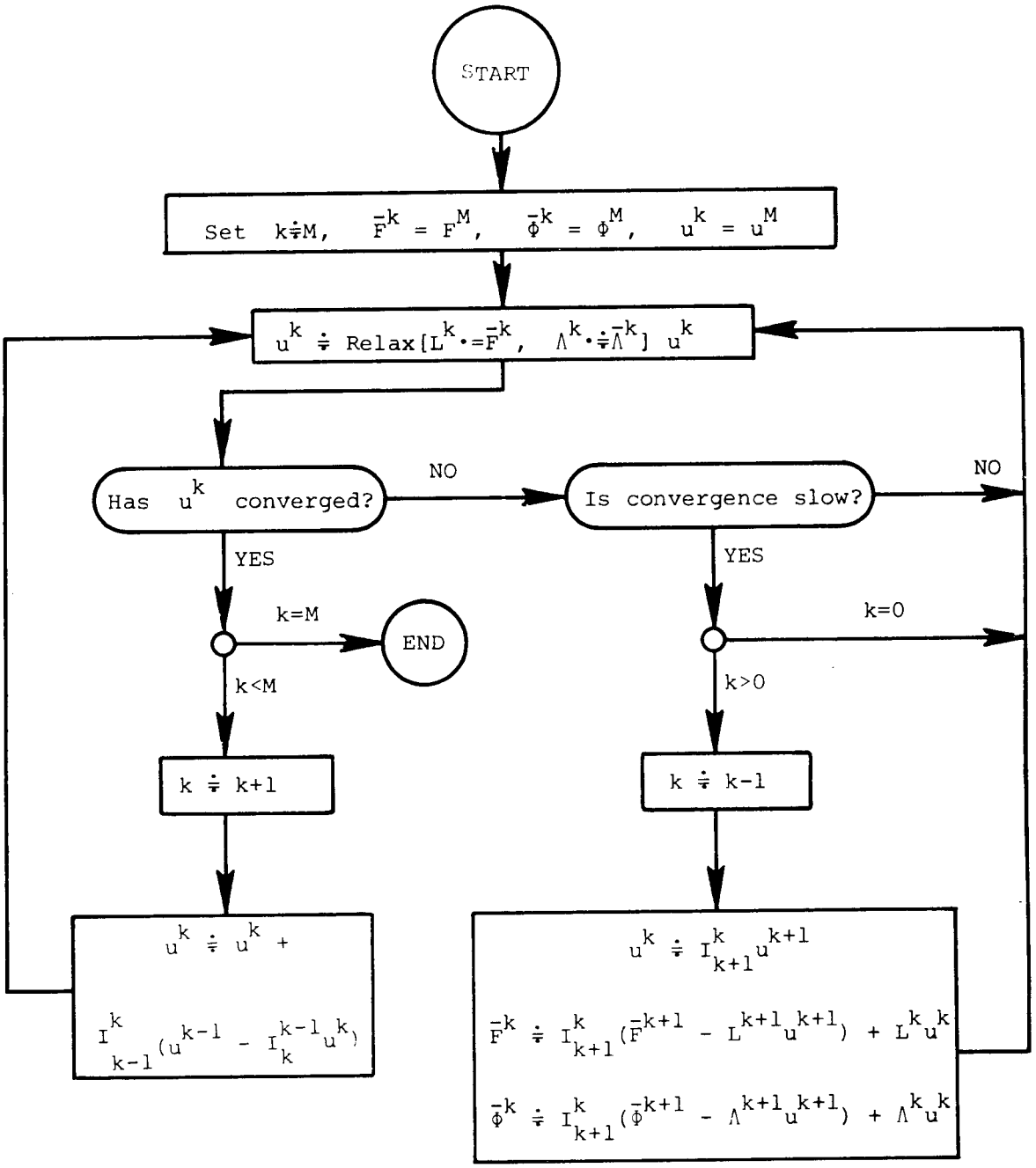



Figure 3. Cycle C, Full-Approximation-Storage.

## 6. PERFORMANCE ESTIMATES AND NUMERICAL TESTS

6.1. Predictability. An important feature of the multi-grid method is that, although iterative, its total computational work can be predicted in advance by local mode (Fourier) analysis. Such an analysis, which linearizes and freezes the equations and ignores distant boundaries, gives a very good approximation to the behavior of high-frequency components (since they have short coupling range), but usually fails to approximate the behavior of the lowest frequencies (which interact at long distances). The main point here, however, is that these lowest frequencies may indeed be ignored in the multi-grid work estimates, since their convergence is obtained on coarser grids, where the computational work is negligible. The purpose of the work on the finer grids is only to converge the high frequencies. Thus, the mode-analysis predictions, although not rigorous, are likely to be very realistic. In fact, these predictions are in full agreement with the results of our computational tests. (For rigorous bounds - see App. C).

6.2. Multi-Grid Rates of Convergence. To get a convenient measure of convergence per unit work, we define as our Work Unit (WU) the computational work in one relaxation sweep over the finest grid $G^M$. The number of computer operations in such a unit is roughly $wn$, where $n$ is the number of points in $G^M$ and $w$ is the number of operations required to compute the residual at each point. (In parallel processing the count should, of course, be different. Also, the work unit should be further specified when comparing different discretization and relaxation schemes.) If the mesh-size ratio is $\hat{\rho} = h_{k+1}/h_k$ and the problem's domain is d-dimensional, then a relaxation sweep over $G^{m-j}$ costs approximately $\hat{\rho}^{dj}$ WUs (assuming the grids are co-extensive, unlike those in Sec. 7).

Relaxation sweeps make up most of the multi-grid computational work. The only other process that consumes any significant amount of computations is the $I^{k-1}_k$ and $I^k_{k-1}$ interpolations. It is difficult to measure them precisely in WUs, but their total work is always considerably smaller than the total relaxation work. In the example in Appendix B, the interpolation work is about 20% of the relaxation work. Usually the percentage is even lower, since relaxing Poisson problems is particularly inexpensive. To unify our estimates and measurements we will therefore define the multi-grid convergence rate $\bar{\mu}$ as the factor by which the errors are reduced per one WU of relaxation, ignoring any other computational work (which is never more than 30% of the total work).

The multi-grid rate of convergence may be estimated by a full local mode analysis. The following is a simplified analysis, which gives a good approximation. We assume that the relaxation sweep over any grid $G^k$ affects error components $e^{i\theta\cdot x}$ only in the range $\frac{\pi}{h_{k-1}} \le |\theta| \le \frac{\pi}{h_k}$, where

$$(6.1) \qquad \theta = (\theta_1, \theta_2, \ldots, \theta_d), \quad \theta\cdot x = \sum_{j=1}^{d} \theta_j x_j, \quad |\theta| = \max_{1\le j\le d} |\theta_j| .$$

(The $\theta/h_k$ of Sec. 3 and Appendix A is denoted here $\theta$, to unify the discussion of all levels.) In fact, if proper interpolation scheme is used (see Sec. A.2) only components in the range $|\theta| \leq (1+\varepsilon)\pi/h_{k-1}$, say, are affected by interactions with coarser grids. But if proper residual-weighting is also used (to make $\bar{\sigma} = 1$; cf. Sec. A.4) then the combined action of the coarse-grid correction cycles and the $G^k$ relaxation sweeps yields convergence rates which are slowest at $|\theta| = \pi/h_{k-1}$ (cf. App. A). For such $\theta$ the coarse-grid cycles have neutral effect, since $\bar{\sigma} = 1$, hence the convergence rate is indeed as affected only by the $G^k$ relaxation sweeps.

One relaxation sweep over $G^k$ reduces the error components in the range $\frac{\pi}{h_{k-1}} \leq |\theta| \leq \frac{\pi}{h_k}$ by the smoothing factor $\bar{\mu}$. (See Sec. 3. If the smoothing rate near a boundary is slower than $\bar{\mu}$, which is not the usual case, smoothing may be accelerated there by partial relaxation sweeps - cf. Sec. A.9.) Thus a multi-grid cycle with s relaxation sweeps on each level reduces all error components by the factor $\bar{\mu}^s$. The amount of work units expended in these sweeps is

$$s + s\rho^{\wedge d} + s\rho^{\wedge 2d} + \ldots + s\rho^{\wedge (M-1)d} < \frac{s}{1-\rho^{\wedge d}} .$$

Hence, the multi-grid convergence rate is

$$\overset{\circ}{\mu} = \bar{\mu}^{(1-\rho^{\wedge d})} , \tag{6.2}$$

which is not much bigger than $\bar{\mu}$. In case $\bar{\sigma}>1$, the effective smoothing rate $\tilde{\mu}$ (see (A.8)) should replace $\bar{\mu}$ in this estimate.

Estimate (6.2) is not rigorous, but is simple to compute and very realistic. In fact, numerical experiments (Sec. 6.4-5) usually show slightly faster (smaller) rates $\overset{\circ}{\mu}$, presumably because the worst combination of Fourier components is not always present.

The theoretical multi-grid convergence rates, for various representative cases, are summarized in Table 1.

Explanations to Table 1. The first column specifies the difference operator and the dimension d. $\Delta_h$ denotes the central second-order ((2d+1)-point) approximation, and $\Delta_h^{(4)}$ the fourth-order ((4d+1)-point "star") approximation, to the Laplace operator. $\Delta_h^2$ is the central 13-point approximation to the biharmonic operator. The operators $\partial_x$, $\partial_y$, $\partial_{xx}$ and $\partial_{yy}$ are the usual central second-order approximations to the corresponding partial-differential operators. $\partial_x^-$ is the backward approximation. Upstream differencing is assumed for the inertial terms of the Navier stokes equations; central differencing for the viscosity terms, forward differencing for the pressure terms, and backward differencing for the continuity equation. Rh is the Reynolds number times the mesh-size.

The second column specifies the relaxation scheme and the relaxation parameter $\omega$. SOR is Successive Over Relaxation, which for $\omega$=1 is the Gauss-Seidel relaxation. xLSOR (yLSOR) is Line SOR, with lines in the

TABLE 1. Theoretical smoothing and MG-convergence rates.

| $L_h$ | d | Relax. Scheme | ω | $\hat{\rho}$ | $\bar{\mu}$ | $\overset{o}{\mu}$ | $\lvert \ln \overset{o}{\mu} \rvert^{-1}$ | add | mult | $W_M$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\Delta_h$ | 1 | SOR | 1 | 1:3 | .557 | .693 | 2.73 | 2 | 1 | 9.0 |
| | | | | 1:2 | .477 | .668 | 2.49 | 3 | 2 | 6.9 |
| | | | | 2:3 | .378 | .723 | 3.08 | 3 | 2 | 7.5 |
| | 2 | SOR | 1 | 1:3 | .667 | .697 | 2.77 | 4 | 1 | 6.8 |
| | | | .8 | 1:2 | .552 | .640 | 2.24 | 5 | 2 | 4.1 |
| | | | 1 | | .500 | .595 | 1.92 | 4 | 1 | 3.5 |
| | | | 1.2 | | .552 | .640 | 2.24 | 5 | 2 | 4.1 |
| | | | 1 | 2:3 | .400 | .601 | 1.96 | 4 | 1 | 2.9 |
| | | LSOR | 1 | 1:2 | .447 | .547 | 1.66 | 8 | 4 | 3.1 |
| | | ADLR | 1 | | .386 | .490 | 1.40 | 8 | 4 | 2.6 |
| | | | .8 | | .456 | .555 | 1.70 | 8 | 4 | 3.1 |
| | | SD | .8 | 1:2 | .600 | .682 | 2.61 | 5 | 2 | 4.8 |
| | | WSD | 1.17, .195 | | .220 | .321 | 0.88 | 9 | 3 | 1.6 |
| | | | 1.40, .203 | | .506 | .600 | 1.96 | 9 | 3 | 3.6 |
| | 3 | SOR | 1 | 1:3 | .738 | .746 | 3.42 | 6 | 1 | 7.8 |
| | | | | 1:2 | .567 | .608 | 2.01 | 6 | 1 | 3.7 |
| | | | | 2:3 | .441 | .562 | 1.73 | 6 | 1 | 2.0 |
| $\Delta_h^{(4)}$ | 2 | SOR | .8 | 1:2 | .581 | .665 | 2.46 | 9 | 3 | 9.1 |
| | | | 1 | | .534 | .625 | 2.13 | 8 | 2 | 7.9 |
| | | | 1.2 | | .582 | .666 | 2.46 | 9 | 3 | 9.1 |
| | | LSOR | 1 | | .484 | .580 | 1.84 | 14 | 7 | 6.8 |
| | 3 | SOR | 1 | | .596 | .636 | 2.21 | 12 | 2 | 7.0 |
| $\partial_{xx} + 2\partial_x\partial_y + \partial_{yy}$ | 2 | SOR | 1 | 1:2 | . 62 | .699 | 2.79 | 8 | 2 | 5.2 |
| | | LSOR,ADLR | | | .447 | .547 | 1.66 | 12 | 5 | 3.1 |
| $\Delta_h^2$ | 2 | SOR | 1 | 1:2 | .802 | .847 | 6.04 | 12 | 3 | 11.1 |
| | | | 1 | 2:3 | .666 | .798 | 4.43 | 12 | 3 | 6.5 |
| | | WSD | 1.552, .353 | 1:2 | .549 | .638 | 2.22 | 17 | 5 | 4.1 |
| | | | 1.4 , .353 | | 1.03 | div. | div. | 17 | 5 | div. |
| | | WSDΔ | 1.552, .353 | | .549 | .638 | 2.22 | 14 | 4 | 4.1 |
| NAVIER - STOKES | | CSOR | | | | | | | | |
| Rh = 0 | 2 | downstr. | 1, .5 | 1:2 | .800 | .846 | 5.98 | 18 | 6 | 11.0 |
| any | | | 1, .5 | | .800 | .846 | 5.98 | 33 | 16 | 11.0 |
| 100 | | | 1.1, .5 | | 1.73 | div. | div. | 33 | 16 | div. |
| 100 | | | .8, .5 | | .93 | .947 | 18.4 | 33 | 16 | 34.0 |
| 10 | | upstream | 1, .5 | | .884 | .912 | 10.8 | 33 | 16 | 20.0 |
| 100 | | | 1, .5 | | .994 | .995 | 220. | 33 | 16 | 400. |
| 100 | | | .8, .5 | | .984 | .988 | 83. | 33 | 16 | 150. |
| 0 | 3 | downstr. | 1, .5 | | .845 | .863 | 6.79 | 33 | 8 | 10.7 |
| any | | | 1, .5 | | .845 | .863 | 6.79 | 60 | 25 | 10.7 |
| 10 | | upstream | 1, .5 | | .874 | .889 | 8.49 | 60 | 25 | 13.4 |
| 100 | | | 1, .5 | | .989 | .990 | 100. | 60 | 25 | 160. |

TABLE 1. (Cont'd. Here d=2, $\hat{\rho}$=1:2)

| $L_h$ | Relax. Scheme | $\omega$ | $\bar{\mu}$ |
|---|---|---|---|
| $\partial_{xx} + \varepsilon\partial_{yy}$, $\varepsilon<<1$ | SOR, xLSOR | any | $1- 0(\varepsilon)$ |
| $a\partial_{xx} + c\partial_{yy}$ $(q = \min(\frac{a}{c}, \frac{c}{a}))$ | yLSOR | 1 | $\max(5^{-1/2}, \frac{a}{a+2c})$ |
| | ADLR | | $5^{-1/4}\ (1+2q)^{-1/2}$ |
| | SD, yLSD, ADLSD | 1 | 1 |
| | SD | $(2q+2)/(3q+2)$ | $(q+2)/(3q+2)$ |
| | yLSD | $(2a+2c)/(2a+3c)$ | $(2a+c)/(2a+3c)$ |
| | ADLSD | 2/3, 2/3 | $\leq 3^{-1/2} = .577$ |
| $\Delta_h - \frac{\eta}{h}\partial_x$ | yLSOR | 1 | $\max\left(\frac{1-\eta}{3-\eta}, \left[\frac{1-\eta+\eta^2/4}{5+\eta+\eta^2/4}\right]^{1/2}\right)$ |
| $\Delta_h - \frac{\eta}{h}\partial_x^-$ $(\eta>o)$ | yLSOR+ | 1 | $\max\left(\frac{1}{3}, [5+6\eta+2\eta^2]^{-1/2}\right)$ |
| | yLSOR- | | $\max\left(\frac{1}{3}, \left\lvert\frac{1+\eta}{2+\eta+i}\right\rvert\right)$ |
| | yLSORs | | $\leq 3^{-1/2} = .577$ |
| Navier - Stokes with large Rh in 2 or 3 dimensions | SOR (pressure corrected by the continuity equation), downstream or up-stream, with any relaxation parameters. | | $\geq 1 - \frac{2}{Rh}$ |

x (y) direction. yLSOR+, yLSOR- and yLSORs indicate, respectively, relaxation marching forward, backward and symmetrically (alternately forward and backward). CSOR means Collective SOR (see Sec. 3 in [3]) and the attached $\omega$'s are $\omega_1$ for the velocity components and $\omega_2$ for the pressure. ADLR denotes Alternating Direction Line Relaxation (a sweep of xLSOR followed by a sweep of yLSOR). SD is Simultaneous Displacement (Jacobi) relaxation, WSD is Weighted Simultaneous Displacement with the optimal weights as specified in Sec. 3.3 above (and with other weights, to show the sensitivity). WSD$\Delta$ (for $\Delta_h^2$) is like WSD, except that residuals are computed in less operations by making first a special pass that computes $\Delta_h u$. yLSD is y-lines relaxation with simultaneous displacement, ADLSD is the corresponding alternating-direction (yLSD alternating with xLSD) scheme.

The next columns list $\hat{\rho} = h_k : h_{k+1}$ (see discussion below), the smoothing rate $\bar{\mu}$ as defined by (3.8), and the multi-grid convergence rate $\overset{\circ}{\mu}$, calculated by (6.2). We also list $|\log \overset{\circ}{\mu}|^{-1}$, which is the theoretical number of relaxation Work Units required to reduce the error by the factor e, and $W_M$, the overall multi-grid computational work (see Sec. 6.3). To make comparisons of different schemes possible, we also list, for each case, the number of operations per grid point per sweep. This number times n (the number of points on $G^M$) give the number of operations in a Work Unit. We list only the basic number of additions and multiplications (counting shifts as multiplications), thus ignoring the operations of transferring information, indexing, etc., which may add up to a significant amount of operations, but which are too computer- and program-dependent to be specified. Also, we assumed that the right-hand sides $f^k$, including $f^M$, are stored in the most efficient form (e.g., $h^2 f^M$ is actually stored). Note that the SOR operation count is smaller for $\omega$=1 (Gauss-Seidel) than for any other $\omega$.

Numbers in this table were calculated by Allan S. Goodman, at IBM Thomas J. Watson Research Center. A more extensive list is in preparation.

Mesh-size ratio optimization. Examining Table 1, and many other unlisted examples, it is evident that the mesh-size ratio $\hat{\rho}$ = 1:2 is close to optimal, yielding almost minimal $|\log \overset{\circ}{\mu}|^{-1}$ and minimal $W_M$. This ratio is more convenient and more economic in the interpolation processes (which are ignored in the above calculations) than any other efficient ratio. In practice, therefore, the ratio $\hat{\rho}$ = 1:2 should always be used, giving also a very desirable standardization.

6.3. Over-All Multi-Grid Computational Work. Denote by $W_M$ the computational work (in the above Work Units) required to solve the $G^M$ problem ((2.2), k=M) to the level of its truncation errors $\tau^M$ (cf. Sec. A.8). If the problem is first solved on $G^{M-1}$ to the level $\tau^{M-1}$, and if the correct order of interpolation is used to interpolate the solution to $G^M$ (so that unnecessary high-frequencies are not excited; cf. Sec. A.2, and in particular (A.7) for i=1) then the residuals of this first $G^M$ approximation are $O(\tau^{M-1})$. The computational work required to reduce them to $O(\tau^M)$ is $\log O(\tau^M/\tau^{M-1}) \,/\, \log \overset{\circ}{\mu}$. Hence,

$$(6.3) \qquad W_M = W_{M-1} + \log \frac{\tau^M}{\tau^{M-1}} \Big/ \log \overset{\circ}{\mu} .$$

Similarly, we can solve the $G^{M-j}$ problem expending work

$$(6.4) \qquad W_{M-j} = W_{M-j-1} + \hat{\rho}^{jd} \log \frac{\tau^{M-j}}{\tau^{M-j-1}} \Big/ \log \overset{\circ}{\mu}$$

(since a $G^{M-j}$ work unit is $\hat{\rho}^{jd}$ times the $G^M$ unit). If we use p-order approximation, then

$$(6.5) \qquad \frac{\tau^k}{\tau^{k-1}} \leq \frac{O(h_k^p)}{O(h_{k-1}^p)} = O(\hat{\rho}^p) .$$

Hence, using (6.4) for j=0,1,2,...M-1 and neglecting $W_o$,

$$W_M \leq (1+\hat{\rho}^d+\hat{\rho}^{2d}+\ldots)\ p \log \hat{\rho} \Big/ \log \overset{\circ}{\mu} .$$

Or, by (6.2),

$$W_M \leq \frac{p \log \hat{\rho}}{(1 - \hat{\rho}^d)^2 \log \bar{\mu}} .$$

(The same $\hat{\rho}$ was assumed in computing the first approximation and in the improvement cycles. This of course is not necessary.)

Typical values of this theoretical $W_M$ are shown in Table 1 above. In actual computations a couple of extra Work Units are always expended in solving a problem, because we cannot make non-integral number of relaxation sweeps or MG cycles, and also because we usually solve to accuracy below the level of the truncation errors.

For 5-points Poisson problems, for example, the following procedure gives a $G^M$ solution with residuals smaller than $\tau^M$. (i) Obtain $u^{M-1}$ on $G^{M-1}$, with residuals smaller than $\tau^{M-1}$. (ii) Starting with the cubic interpolation $u^M \leftarrow I_{M-1}^M u^M$ (preferably by using the difference operator itself; cf. [7]), make a MG correction cycle such as Cycle C with $\eta$=o (i.e., switching to $G^{k-1}$ after two sweeps on $G^k$), with $I_k^{k-1}$ transfer by injection (cf. Sec. A.4) and $I_{k-1}^k$ by linear interpolation, and with "convergence" on $G^k$ defined as obtained after the first sweep following a return from $G^{k-1}$. A precise count shows Step (ii) to require $30n + O(n^{1/2})$ operations, where n is the number of points in $G^M$. Thus, the total number of operations is

$$(1 + \frac{1}{4} + \frac{1}{16} + \ldots)30n + O(n^{1/2}) \leq 40n + O(n^{1/2}).$$

Incidentally, none of these operations is a full multiplication: only additions and shifts (multiplications or divisions by 2 or 4) are used.

The theoretical $W_M$ for this problem (sixth line in Table 1) amounts to only 17.5n operations, since it ignores interpolation work (10.3n operations in the above procedure) and allows non-integral numbers of sweeps and cycles. In fact, numerical tests showed the above algorithm to yield residuals considerably below the truncation errors. (The only cases in which the residuals approached 50% of the truncation errors were cases with high smoothness, in which the correct MLAT discretization would be different; namely, of higher order. (Cf. Sec. 8 and the remark following formula (A.7).)

6.4. Numerical Experiments: Elliptic Problems. A typical numerical experiment is shown in Appendix B, including the FORTRAN program and the computer output. The output shows a multi-grid convergence rate

$$\overset{o}{\mu} = \left(\frac{.009051}{28.1}\right)^{\frac{1}{12.92}} = .537$$

which is close to, and slightly faster than, the theoretical value $\overset{o}{\mu} = .595$ shown in Table 1.

Many numerical experiments with various elliptic difference equations in various domains were carried out at the Weizmann Institute in 1970-1972, with the collaboration of Y. Shiftan and N. Diner. Some representative results were reported in [2], and many others in [11]. These experiments were made with other variants of the multi-grid algorithm (variants A and B), but their convergence rates agree with the same theoretical rates $\overset{o}{\mu}$. The experiments with equations of the form $aU_{xx} + cU_{yy}$, with a>>c, showed poor convergence rates, since the relaxation scheme used was Gauss-Seidel, and not the appropriate line relaxation (cf. Sec. 3.1). Some of these rates were better than predicted by the mode analysis, because the grids were not big enough to show the worst behavior. The convergence rates found in the experiments with the biharmonic equation were also rather poor (although nicely bounded away from 1, independently of the grid size), again because we used Gauss-Seidel relaxations and injections instead of the appropriate schemes (cf. Sec. 3.3 and A.4). All these points were later clarified by mode analyses, which fully explain all the experimental results. In solving the stationary Navier-Stokes equations, as reported in [2], SOR instead of CSOR was employed (cf. table 1 above), and an additional over-simplification was done by using, in each multi-grid cycle, values of the nonlinear terms from previous cycle, instead of using the FAS scheme (Sec. 5).

Nevertheless, these experiments did clearly demonstrate important features of the multi-grid method: The rate of convergence was essentially insensitive to several factors, including the shape of the domain $\Omega$, the right-hand side F (which has some influence only at the first couple of cycles; cf. Sec. A.2) and the finest mesh-size $h_M$ (except for mild variations when $h_M$ is large). The experiments indicated that the order I of the interpolations $I_{k-1}^{k}$ should be the order of the elliptic equation, as shown in Sec. A.2 below. (Note that in [2] the order was defined as the degree $\ell$ of the polynominal used in the interpolation, whereas here $I = \ell+1$.)

More numerical experiments are now being conducted at the Weizmann Institute in Israel and at IBM Research Center in New York, and will

be reported elsewhere. We will briefly report here only an extreme case of the multi-grid tests - the solution of transonic flow problems.

6.5. Numerical Experiments: Transonic Flow Problems. These experiments were conducted in 1974 at the Weizmann Institute with J.L. Fuchs, and recently at the NASA Langley Research Center in collaboration with Dr. Jerry South while the present author was visiting the Institute for Computer Applications in Science and Engineering (ICASE). They are preliminarily reported in [12], and will be further reported elsewhere. One purpose of this work was to examine the performance of the multi-grid method in a problem that is not only nonlinear, but more significantly, is also of mixed (elliptic-hyperbolic) type and contains discontinuities (shocks).

We considered the transonic small-disturbance equation in conservation form

$$(6.7) \qquad [(K-\bar{K}\phi_x)\ \phi_x]_x + c\ \phi_{yy} = 0,$$

for the velocity disturbance potential $\phi(x,y)$ outside an airfoil. Here $K = (1-M_\infty^2) / \tau^{2/3}$, $\bar{K} = \frac{1}{2}(\gamma+1)\ M_\infty^2$, $M_\infty$ is the free-stream Mach number, and $\gamma=1.4$ is the ratio of specific heats. $\tau$ is the airfoil thickness ratio, assumed to be small. $c=1$, unless the y coordinate is stretched. The airfoil, in suitably scaled coordinates, is located at $\{y=0,\ |x| \le \frac{1}{2}\}$, and we consider nonlifting flows, so that the problem domain can, by symmetry, be reduced to the half-plane $\{y \ge 0\}$, with boundary conditions

$$(6.8) \qquad \phi(x,y) \to 0 \qquad \text{as } x^2+y^2 \to \infty$$

$$(6.9) \qquad \phi_y(x,0) = \begin{cases} 0, & \text{for } |x| > \frac{1}{2}, \\ F'(x), & \text{for } |x| < \frac{1}{2}, \end{cases}$$

where $\tau F(x)$ is the airfoil thickness function which we took to be parabolic. Equation (6.7) is of hyperbolic or elliptic type depending on whether $K - 2\bar{K}\phi_x$ is negative or positive (supersonic or subsonic).

The difference equations we used were essentially the Murman's conservative scheme ([9]; for a recent account of solution methods, see [8]), where the main idea is to adaptively use upstream differencing in the hyperbolic region and central differencing in the elliptic region, keeping the system conservative. For relaxation we used vertical (y) line relaxation, marching in the stream direction. The multi-grid solution was programmed both in the CS (Sec. 4) and the FAS (Sec. 5) modes, with practically the same results. We used cubic interpolation for $I_k^{k+1}$ and injection for $I_k^{k-1}$.

Local mode analysis of the linearized-freezed difference equations and vertical-forward line relaxation gives the smoothing rate

$$(6.10)\qquad \bar{\mu} = \max\left\{\left|\frac{b_+}{b_+ + b_- + ib_-}\right|, \frac{b_+}{2c+b_+}\right\}, \quad b_{\pm}(x) = K - 2\bar{K}\,\phi_x\left(x \pm \frac{h}{2}\right),$$

at elliptic (subsonic) points, and $\bar{\mu} = 0$ at supersonic points. We were interested in cases where $K<1$ and $\phi_x \geq 0$, and hence, in smooth elliptic regions ($b_+ \sim b_-$) without coordinate stretching we get $\bar{\mu} \sim 1/|2+i| = 0.45$ and $\overset{o}{\mu} = \bar{\mu}^{3/4} = 0.55$.

The actual convergence rates, observed in our experiments with moderately supercritical flows ($M_\infty = 0.7$ and $M_\infty = 0.85$, $\tau = 0.1$) on a 64x32 grid, were $\overset{o}{\mu} = 0.52$ to $0.53$, just slightly faster than the theoretical value. (See detailed output in [12]. The work count in [12] is slightly different, counting also the work in the $I^k_{k+1}$ transition).

For highly supercritical flows ($M_\infty = 0.95$, $\tau = 0.1$) the MG convergence rate deteriorated, although it was still 3 times faster than solution by line relaxation alone. The worse convergence pattern is explainable in terms of the mode analysis for the elliptic region immediately behind the shock, where $b_+ >> b_-$, yielding $\bar{\mu}$ closer to 1. Also, the fast changes in $\phi_x$ in that region gives $\sigma > 1$ (see Sec. A.1), i.e., the coarse grid cycles actually magnify the Fourier component with $\theta = (\frac{\pi}{2}, 0)$, the same component for which $\bar{\mu}$ is closer to 1. This worse behavior in this restricted region further affected our computations because we did not use separate stopping tests for this region as we should (see Sec. A.6). A correct multi-grid algorithm for this problem should, therefore, include symmetric selective line relaxation (see Sec. 3.2), or partial relaxation sweeps (see Sec. A.9), or both, in addition to residual weighting (Sec. A.4).

Coordinate stretching, which transforms the bounded computational domain to the full half plane, gave difference equations that again exhibited slow multi-grid convergence rate. This, too, is explainable by the mode analysis. For example, in the regions where the $y$ coordinate is highly stretched, $c$ in (6.7) becomes very small and hence $\bar{\mu}$ in (6.10) is close to 1. The theoretical remedies: alternating-direction line relaxations and partial relaxation sweeps. The latter was tried in one simple situation (stretching only the x coordinate), and indeed restored the convergence rate of the corresponding unstretched case.

## 7. NON-UNIFORM GRIDS.

Many problems require very different resolution in different parts of their domains. Special refinement of the grid is required near singular points, in boundary layers, near shocks, and so on. Coarse grids (with higher approximation order) should be used where the solution is smooth, or in subdomains far from the region where the solution is accurately

needed, etc. A general method for locally choosing mesh-sizes and approximation orders is described in Sec. 8. An important feature of the method is adaptivity: the grid may change during the solution process, adapting itself to the evolving solution. In this section, we propose a method of organizing non-uniform grids so that the local refinement is highly flexible. The main idea is to let the sequence of uniform grids $G^0, G^1, \ldots, G^M$ (cf. Sec. 2) be open-ended and non-coextensive (i.e., finer levels may be introduced on increasingly narrower subdomains to produce higher local refinement, and coarser levels may be introduced on increasingly wider domains to cover unbounded domains), and, furthermore, to let each of the finer levels be defined in terms of suitable local coordinates. The multi-grid FAS process remains practically as before (Sec. 5), with similar efficiency. Also discussed is a method which employs this grid organization for "segmental refinement", a multi-grid solution process with substantially reduced storage requirement.

7.1. Organizing Non-Uniform Grids. How are general non-uniform grids organized for actual computations? There are two popular approaches: One, usually used with the finite element method, is to keep the entire system very flexible, allowing each grid point to be practically anywhere. This requires a great deal of bookkeeping: grid-points' locations and pointers to neighbors need to be stored; sweeping over the grid is complicated; obtaining the coefficients of the difference equations (or the local "stiffness") may require lengthy calculations, especially where the grid is irregular; and these calculations should be repeated each relaxation sweep, or else additional memory areas should be allocated to store the coefficients. Also, it is more difficult to organize a multi-grid solution on a completely general grid (see, however, Secs. 7.3 and A.5), and complete generality is not necessary for obtaining any desired refinement pattern.

Another approach for organizing a non-uniform grid is by a coordinate transformation, with a uniform grid being used over the transformed domain. On such grids, topologically still rectangular, the multi-grid method can be implemented in the usual way, the lines of $G^{k-1}$ being every other line of $G^k$. Decisions (stopping criteria, residual weighting, relaxation mode and relaxation directions) should be based on the transformed difference equations. Very often, however, coordinate transformation does not offer enough flexibility. A local refinement is not easy to produce, unless it is a one-dimensional refinement, or a tensor product of one-dimensional refinements. The difficulties are enlarged in adaptive procedures, where it should be inexpensive to change local mesh-sizes several times in the solution process. Moreover, the transformation usually makes the difference equation much more complicated (requiring additional storage for keeping coefficients, or additional work in recomputing them every sweep), especially when the transformation does become sophisticated (i.e., adaptive, and not merely a product of one-dimensional transformations), and in particular if higher-order approximations should be used in some or all subdomains.

Thus, be it in the original or in some transformed domain, one would like to have a convenient system for local refinements, with minimal bookkeeping and efficient methods for formulating and solving difference equations. The following system is proposed (and then generalized, in Secs. 7.3, 7.4):

A non-uniform grid is a union of uniform sub-grids, $G^0$, $G^1$,...,$G^M$, with corresponding mesh-sizes $h_0$, $h_1$,...,$h_M$. Usually $h_k:h_{k+1} = 2:1$ and every other grid line of $G^{k+1}$ is a grid line of $G^k$. Unlike the description in Sec. 2, however, the sub-grids are not necessarily extended over the same domain. The domain of $G^{k+1}$ may be only part of the domain of $G^k$ (but not vice versa). Thus we may have different levels of refinement at different subdomains.

For problems on a bounded domain $\Omega$, several of the first (the coarsest) sub-grids may extend to the entire domain $\Omega$. That is, they do not serve to produce different levels of refinement, but they are kept in the system for serving in the multi-grid process of solving the difference equations. $G^0$ should be coarse enough to have its system of difference equations relatively inexpensive to solve (i.e., requiring less than $O(\Sigma n_k)$ operations, where $n_k$ is the number of grid points in $G^k$. But cf. Sec. 4.1). The finer sub-grids typically extend only over certain subdomains of $\Omega$, not necessarily connected. Generally, $G^k$ is stretched over those subdomains where the desired mesh-size is $h_k$ or less. Thus, very fine levels (e.g., with M=20, so that $h_M=2^{-20}h_0$) may be introduced, provided they are limited to suitably small subdomains.

Such a system is very flexible, since grid refinement (or coarsening) is done by extending (or contracting) uniform sub-grids. There are several possible ways of storing functions on a (possibly disconnected) uniform grid, allowing for easy grid changes. For example, each string (i.e., connected row or column) of function values can be stored separately, at an arbitrary place in one big storing area, with a certain system of pointers leading from one string to the next. The extra storage area needed for these pointers is small compared with the area needed for storing the function values themselves. One such system, with subroutines for creating, changing and interpolating between the grids, is now under construction, and will be reported elsewhere.

If the (original or transformed) problem's domain is unbounded, we usually put suitable boundary conditions on some finite, "far enough" artificial boundary. In the present system, we do not have to decide in advance where to place the artificial boundary: We can extend (or contract) the coarsest sub-grid(s) as the solution evolves. Moreover, we can add increasingly coarser levels ($G^{-1}$, $G^{-2}$,...) to cover increasingly wider domains, if required by the evolving solution. In this way, we may reach computational domains of large diameter R, by adding only O(log R) grid points (assuming the desired mesh-size, out at distance r, is proportional to r, or larger. This should usually be the case, especially if appropriate higher-order approximations are used at large distances).

There appears to be a certain waste in the proposed system, as one function value may be stored several times, when its grid point belongs to several levels $G^k$. This is not the case. First, because the amount of such extra storage is small (less than $2^{-d}$ of the total storage; see (4.6)). Moreover, the stored values are exactly those needed for the multi-grid process of solution: In fact, in that process, the values stored for different levels at the same grid-point are not identical; they only converge to the same value as the process proceeds.

7.2. The Multi-Grid Algorithm on Non-Uniform Grids. The following is a description of the modification in the FAS multi-grid algorithm (Sec. 5) in case of a non-uniform grid with the above structure. The algorithm remains almost the same, except that the difference equations (5.2-4) are changed to take account of the fact that the levels $G^k$ do not necessarily cover the same domain. Denoting by $G^k_m$ the set of points of $G^k$ which are inner points of a finer level $G^m$ (i.e., points where the $G^m$ difference equations are defined [3]; see Figure 4), the modified form of the difference equations on $G^k$ is

$$L^k u^k = \bar{F}^k \ , \qquad \Lambda^k u^k = \bar{\Phi}^k \ , \tag{7.1}$$

where

$$\bar{F}^k = F^k \quad \text{and} \quad \bar{\Phi}^k = \Phi^k \quad \text{in } G^k - G^k_{k+1} \text{ and for } k=M, \tag{7.2}$$

$$\bar{F}^k = F^k_{k+1} \quad \text{and} \quad \bar{\Phi}^k = \Phi^k_{k+1} \quad \text{in } G^k_{k+1} \ , \tag{7.3}$$

$$F^k_m = I^k_m (\bar{F}^m - L^m u^m) + L^k (I^k_m u^m) \ , \tag{7.4}$$

$$\Phi^k_m = I^k_m (\bar{\Phi}^m - \Lambda^m u^m) + \Lambda^k (I^k_m u^m) \ . \tag{7.5}$$

$F^k$ and $\Phi^k$, as in Sec. 2, are the $G^k$ approximation to the original right-hand sides F and $\Phi$, respectively.

Observe that, by (7.2-3), each intermediate level $G^k$ plays a double role: on the subdomain where the finer sub-grid $G^{k+1}$ is not defined, $G^k$ plays the role of the finest grid and the difference equation there is an approximation to the original differential equation. At the same time, on the subdomain where finer sub-grids are present, $G^k$ serves for calculating the coarse-grid correction. These two roles are not confused owing to the FAS mode, in which the correction $v^k$ is only implicitly computed, its equation being actually written in terms of the full approximation $u^k$. In other words, $F^k_m$ may be regarded as the usual $G^k$ right-hand side $(F^k)$, corrected to achieve $G^m$ accuracy in the $G^k$ solution. Indeed

$$F^k_m - I^k_m \bar{F}^m = L^k (I^k_m u^m) - I^k_m (L^m u^m) \ , \tag{7.6}$$

which is the $G^m$ approximation to the $G^k$ truncation error.

---

[3] We use the term "inner", and not "interior", because these points may well be boundary points. Indeed, at boundary points difference equations are defined, although they are of a special type, called boundary conditions. The only $G^m$ points where $G^m$ difference equations are not defined are points on or near the internal boundary of $G^m$; i.e., the boundary beyond which the level $G^m$ is not defined, but some coarser levels are. If the grid-lines of $G^k$ do not coincide with grid lines of $G^m$, $G^k_m$ is defined as the set of points of $G^k$ to which proper interpolation from inner points of $G^m$ is well-defined. For m>M, $G^k_m$ is empty.

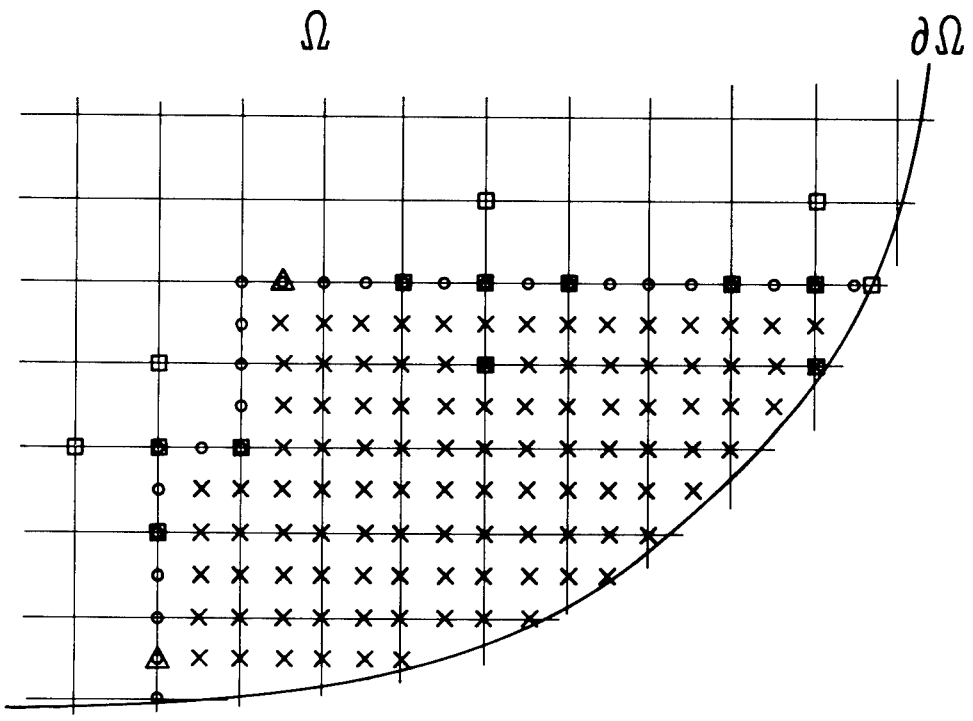



Figure 4. Example of Non-Uniform Grid.

A section of the domain $\Omega$ and its boundary $\partial\Omega$ is shown, covered with a coarser grid $G^k$ (line intersections) and a finer grid $G^{k+1}$ (crosses and circles). For the case of a 5-points (or 9-points "box") difference equations, $G^{k+1}$ inner points are marked with crosses, its outer points with circles. (For convenient interpolation, outer points should lie on $G^k$ lines). At outer points belonging to $G^k$, the converged solution satisfies the $G^k$ difference equations, such as the 5-point relations indicated by squares. At other outer points, such as those shown with triangles, the solution is always an interpolation from values at adjacent $G^k$ points. (Note that starting values at outer points should be such that these interpolation relations are satisfied. The FAS interpolation steps will then automatically preserve these relations.)

The only other modification required in applying Cycle C to non-uniform grids is in the convergence switching criteria. See Sec. A.10.

When converged, the solution so obtained satisfies equations (2.2) in the inner part of $G^k - G_{k+1}$, (k=0,1,...,M). On outer (i.e., non-inner) points the solution automatically satisfies either a coarser-grid difference equation (if the point belongs to a coarser grid) or a coarser grid interpolation relation. (See Figure 4). Note that, in this procedure, difference equations should be defined on uniform grids only. This is an important advantage. Difference equations on equi-distant points are much simpler, more accurate. The basic weights for each term (e.g., the weights (1,2,1) for the secord order approximation to $\partial^2/\partial x^2$) can be read from small standard tables; whereas on a general grid those weights should be recomputed (or stored) separately for each point, and they are very complicated for high-order approximations.

Another advantage is that the relaxation sweeps, too, are on uniform grids only. This simplifies the sweeping, and is particularly important where symmetric and alternating-direction sweeps are required (cf. Sec. 3).

Numerical experiments indicate that the typical multi-grid convergence rates, measured by the overall error reduction per work unit and predicted by local mode analysis (cf. Sec. 6), are retained in multi-grid solutions on non-uniform grids. The work unit, though, is somewhat different: It is the computational work of one sweep on all levels, not only on $G^M$, since here $G^M$ may make up only a small part of the points of the final non-uniform grid.

7.3. Finite-Elements Generalization. The structure and solution process outlined above can be generalized in various ways. An important generalization is to employ piece-wise uniform, rather than strictly uniform, levels.

Quite often, especially in problems that use finite-elements discretizations, the "basic" partition $G^0$ (e.g., the coarsest triangulation) of the domain is a non-uniform one, but one which is particularly suitable for the geometry of the problem. Finer levels $G^1, G^2, \ldots$, are defined as uniform refinements of that basic level; e.g., $h_k = 2^{-k} h_0$; so that $h_k$ is constant within each basic element.

Having defined the levels $G^k$ in this manner, the rest may in principle be as before: The actual composite grid may use only certain, arbitrary portions of each level; i.e., the actual subgrids $G^k$ need not be co-extensive, allowing for adaptive refinements. Coarser levels ($G^{-1}$, $G^{-2}$ ...,) may be added if the basic level $G^0$ is not coarse enough for full-speed multi-grid solution. (Although there is no general algorithm for coarsening a non-uniform $G^0$, and usually $G^0$ is coarse enough).

Data structures, similar to the uniform case may be used, but should be constructed separately for each basic element (or each set of identical basic elements).

The multi-grid algorithm is the same as in Sec. 7.2. The discrete equations are thus defined separately for each level. The reproduction of these equations during relaxation is not as convenient as in the strictly uniform case, but still, in the interior of any basic element the equations can readily be read from fixed tables, one table for each set of identical basic elements.

7.4. Local Transformations. Another important generalization of the above structure is to subgrids which are defined each in terms of another set of variables. For example, near a boundary or an interface, the most effective local discretizations are made in terms of local coordinates in which the boundary (or interface) is a coordinate line. In particular, with such coordinates it is easy to formulate high-order approximations near the boundary; or to introduce mesh sizes that are different across and along the interface (or the boundary layer); etc. Usually it is easy to define suitable local coordinates, and uniformly discretize them, but it is more difficult to patch together all these local discretizations.

A multi-grid method for patching together a collection of local grids $G_1, G_2, \ldots, G_m$ (each being uniform in its own local coordinates) is to relate them all to a basic grid $G_0$, which is uniform in the global coordinates and stretches over the entire domain. The relation is essentially as above (Sec. 7.2); namely, finite-difference equations are separately defined in the inner points of each grid, and the FAS multi-grid process automatically combines them together through its usual interpolation periods.

A remark: To a given collection of local grids we may have to add intermediate grids to obtain fast multi-grid convergence. That is, if a given local grid $G_k$ is much finer than the basic grid $G_0$, we have to add increasingly coarser grids, all of them uniform grids in the same local coordinates, such that the coarsest of them has a mesh size which is (in the global coordinates) nowhere much smaller than the basic mesh size $h_0$. Similarly, if the basic global grid $G_0$ is not coarse enough, the usual multi-grid sequence of global grids $G^0, G^1, \ldots, G^M = G_0$ should be introduced. Thus, in each set of coordinates we will generally have several grids.

Such a system offers much flexibility. Precise treatment of boundaries and interfaces by the global coordinates is not required. The local coordinates may be changed in course of computations, e.g., to fit a moving interface. New sets of local coordinates may be introduced (or deleted) as the need arises.

The data structure required for creating, changing and employing such grids is basically again just any data structure suitable for changeable uniform grids. This, however, should be supplemented by tables for the local transformations, such that one can efficiently (i) reproduce the local difference equation, and (ii) interpolate from local to global grid points, and vice-versa.

7.5. Segmental Refinement. The multi-grid algorithm for non-uniform grids (Sec. 7.2) can be useful even in the case of uniform grids, if the computer memory is not sufficiently large to store the finer levels.

"Segmental refinement" is the refinement of one subdomain at a time. To see why and how this is possible, observe that with the FAS mode (Sec. 5) the full solution $u^M$ is obtained on all grids. But on a coarser grid $G^k$, the $u^M$ solution satisfies "corrected" difference equation, with $\bar{F}^k = F_M^k$ replacing $F^k$. It is therefore not necessary to keep the fine grid, once $F_M^k$ has been computed.

The corrected forcing function $F_M^k$ can be computed by segmental refinement. Refining only one subdomain, one can use the algorithm above (Sec. 7.2) to obtain a multi-grid solution, including the values of $F_M^k$ in the refined subdomain. Keeping this $F_M^k$ (instead of $F^k$), one can then discard this refinement, and refine a second subdomain. And so on, through a sequence of subdomains covering the entire domain.

Since subsequent subdomain refinements change the solution everywhere, some further changes are also due in the values of $F_M^k$ on former subdomains. However, at points inner to (and few meshes away from the boundary of) such a former subdomain, these further changes are much smaller than the first correction $F_M^k - F^k$, since they represent changes in the $G^k$ truncation error due to small smooth changes in the solution, while the first correction represents the full $G^k$ truncation error. Thus, if the refinement segments are chosen so that neighboring segments overlap (several mesh intervals into each other), then the further corrections may be ignored. If extra accuracy is desired, another cycle of segmental refinements may be performed. Another way of viewing this technique is to observe that the roll of the finer levels, relative to the coarser ones, is only to liquidate high-frequency error components which cannot be "seen" on the coarser grids. These components have a short (just few mesh sizes) coupling range, and can therefore be computed at any point by refining only few neighboring meshes.

With this technique one can operate the multi-grid algorithm almost in its full efficiency, using a storage area which is much smaller than that of the finest grid. This has been confirmed by preliminary (one-dimensional) numerical tests.

In principle, the required storage area can be reduced to only a constant cube, of $J^d$ locations, on each level (where even J=20 probably offer enough overlap without substantial reduction in efficiency). Thus, the overall storage requirement can in principle be reduced to

$$J^d \{1 + \log \frac{R}{h} / \log J\}$$

locations, where h is the finest mesh-size and R is the diameter of the domain. No external memory is needed.

## 8. ADAPTIVE DISCRETIZATION TECHNIQUES.

The previous section described a flexible data structure and solution process which facilitate implementation of variable mesh-sizes h. The difference equations in that process are always defined at inner points of uniform subgrids, which make it easy to employ high and variable approximation orders p. How, then, mesh-sizes and approximation-orders are to be chosen? Should boundary layers, for examples, be resolved by the grid? What is their proper resolution? Should we use high-order of approximation at such layers? How to detect such layers automatically? In this section we propose a general framework for automatic selection of h and p in a (nearly) optimal way. In the next (Sec. 9) we will study some special cases, and show how this proposed system automatically resolve or avoid from resolving thin layer, depending on the alleged goal of the computations.

8.1 Basic Principles. We will treat the problem of selecting the discretization parameters h and p (and possibly other parameters, see Sec. 8.4) as an optimization problem: We will seek to minimize a certain error estimator E, subject to a given amount of computational work W. (Or, equivalently, minimize the work W to obtain a given level E of the error estimator. We will see that the actual control quantity is neither E nor W, but their rate of exchange.) It is important, however, to promptly emphasize that we should not take this optimization too pedantically it is enough, for instance, to obtain E which is one or two orders of magnitudes larger than the minimum (or, equivalently, to invest work W which is by some fraction more than theoretically needed. Note below that $\log(1/E_{min})$ is usually proportional to W). Full optimization is not our purpose, is enormously harder and, in fact, is self-defeating, since it requires too much computational work to be invested in controling h and p. We will aim at having the control work much smaller than the actual numerical work W, using the optimization problem only as a loose directive for sensible discretization.

The Error Estimator E is a functional that estimates the overall error in solving the differential boundary-value problem, in terms of any given numerical approximation. In principle, such a functional should be furnished whenever a problem is submitted for numerical solution; in practice, it is seldom provided. To have such an estimator depends on having a clear and well-defined idea about the goal of the computations, i.e., an idea about what error norm we intend to minimize. Given the goal, even roughly, we can usually formulate E quite easily. We assume that the numerical approximation $U^h$ is in some suitable neighborhood of the true solution (this is a necessary and justifiable assumption; see Sec. 8.2), so that E can be written as a linear functional

$$E = \int_{\Omega} G(x)\ \tau(x)\ dx\ . \tag{8.1}$$

$\tau(x)$ is a local estimate of the truncation error i.e., the error by which the numerical solution $U^h$ fails to satisfy the differential equation LU=F; or more conveniently, the error by which the differential solution U fails to satisfy the discrete equation $L^h U^h = F$. That is,

$$\tau(x) = \left| LU(x) - L^h U(x) \right| . \tag{8.2}$$

G(x) is the non-negative "error-weighting function" (or distribution), through which the computations goal should be expressed.

The choice of G can be crude. In fact, multiplying G by a constant does not change our optimization problem. Also, we can make large errors, up to one or two orders of magnitudes, in the relative values of G at two different points, since we are content in having E only to that accuracy. What matter are only large changes in G, e.g., near boundaries. For example, if we have a uniformly elliptic problem of order m, and if we are interested in computing good approximations to U and its derivatives up to order $\ell$ and upto the boundary, then a suitable choice is

$$G(x) = d_x^{\,m/2-\ell} \tag{8.3}$$

where $d_x$ is the distance of x from the boundary. (The formula should be suitably modified near a boundary corner). This and similar choices of G are easily found by local one-dimensional crude analysis of the relation between a perturbation in the equations and the resulting perturbation in the quantity we wish to approximate. Even though crude, such choice of G would specify our goal much closer than people usually bother to. Moreover, we can change G if we learn that it fails to properly weigh a certain region of the computation; it can serve as a convenient control, conveying our intentions to the numerical discretization and solution.

The Work Functional W. In solving the discrete equations by the multigrid method, the main overall computational work is the number of Work Units invested in relaxations, times the amount of computations in each Work Unit (see Sec. 6). If the discretization and relaxation schemes are suitable, the number of Work Units is almost independent of the relaxation parameters h and p. (See e.g., the rate $\bar{\mu}$ for $\Delta_h^{(4)}$ vs. $\Delta_h$ in Table 1 above). Since for our optimization problem we need W only up to a multiplicative constant, we can take into account only the amount of computations in a single Work Unit, i.e., the work in one relaxation sweep over the domain. The local number of grid points per unit volume is $h(x)^{-d}$, and the amount of computation at each grid point is a function w(p(x)), where p(x) is the local order of approximation. Hence, we can regard the work functional as being

$$W = \int_\Omega \frac{w(p(x))}{h(x)^d}\, dx. \tag{8.4}$$

Global Optimization Equations. Treating the discretization parameters as spatial variables, h(x) and p(x), the Euler equations of minimizing E for fixed W are

$$\frac{\partial E}{\partial h(x)} + \lambda \frac{\partial W}{\partial h(x)} = 0 \tag{8.5a}$$

$$\frac{\partial E}{\partial p(x)} + \lambda \frac{\partial W}{\partial p(x)} = 0 , \tag{8.5b}$$

where $\lambda$ is a constant (the Lagrange multiplier). It is easily seen that $\lambda$ is actually the marginal rate of exchange between work and optimal accuracy, i.e.,

$$(8.6) \qquad \lambda = -\frac{dE_{min}}{dW} = E\,\frac{d\,\log\frac{1}{E}}{dW},$$

and the meaning of (8.5) is that we cannot lower E by trading work (e.g., by taking smaller h at one point and larger at another, keeping W constant, or trading a change in h with a change in p).

Equations (8.5) make some essential simplifications in the optimization problem: They regard h and p as defined at all points $x\in\Omega$; Also, h and p are assumed to be continuous variables, whereas in practice they are discrete. (p should be a positive integer, in some schemes a positive even integer. Values of h are restricted by some grid-organization considerations.) These simplifications are crucial for our approach, and they are altogether justified by the fact that we are content in having only an approximate optimum. The practical aspect, of choosing permissible h and p close to the solution of (8.5), is discussed in Sec. (8.3). One restriction we should, however, take into account in the basic equations, namely, the restriction

$$(8.7) \qquad p_o \le p(x) \le p_1(x).$$

Without such a restriction, the optimization equations may give values of p which cannot be approximated by permissible values. $p_o$ is usually 1 or (in symmetric schemes) 2. The upper bound $p_1$ may express the highest feasible order due to round-off errors; or the highest order for which we actually have appropriate (stable) discretization formulae, with special such restriction near boundaries (hence the possible dependence of $p_1$ on the position x). With this restriction, Euler equation (8.5b) should be rewritten as

$$(8.8) \qquad \frac{\partial E}{\partial p(x)} + \lambda\,\frac{\partial w}{\partial p(x)} \begin{cases} \ge 0, & \text{if } p(x) = p_o \\ = 0, & \text{if } p_o < p(x) < p_1(x) \\ \le 0, & \text{if } p(x) = p_1(x). \end{cases}$$

Local Optimization Equations. Substituting (8.1) and (8.4) into (8.5a) and (8.8), we get the following equations at each point $x \in \Omega$:

$$(8.9a) \qquad G\,\frac{\partial \tau}{\partial h} - \frac{\lambda d\, w(p)}{h^{d+1}} = 0$$

$$(8.9b) \qquad G\,\frac{\partial \tau}{\partial p} + \frac{\lambda w'(p)}{h^{d}} \gtreqless 0,$$

where the equality-inequality sign, in (8.9b) and hereinafter, corresponds to the three cases introduced in (8.8). In principle, the pair of equations (8.9) determines, for each $x \in \Omega$, the local values of the pair $(h,p)$, once $\lambda$ is given.

Thus $\lambda$ is our global control parameter. Choosing larger $\lambda$, we will get an optimized grid with less work and poorer accuracy; lowering $\lambda$, we invest more work and get higher accuracy. For each $\lambda$, however, we get (approximately) the highest accuracy for the work invested. In principle $\lambda$ should be given by whoever submits the problem for numerical solution; i.e., he should tell at what rate of exchange he is willing to invest computational work for additional accuracy (see (8.5)). In practice this is not done, and $\lambda$ usually serves as a convenient control parameter (see Secs. 8.2 and 8.3).

To compute h and p from (8.9) we should know the behavior of $\tau$ as a function of h and p. Generally,

$$\tau(x,h,p) \overset{\sim}{\sim} t(x,p)\, h^p, \tag{8.10}$$

where $t(x,p)$ depends on the equations and on the solution. Since it is assumed that all our numerical approximations are in some neighborhood of the solution (see Sec. 8.2), we may assume that the truncation-error estimates, automatically calculated by the multi-grid processing (see (5.7), for example), give us local estimates for $t(x,p)$. In practice, we never actually solve (8.9), but use these relations to decide upon changes in h and p (see Sec. 8.3), so that we need to estimate $\tau(x,h,p)$ only for h and p close to the current $h(x)$ and $p(x)$.

8.2. Continuation Methods. Continuation methods are generally used in numerical solutions of nonlinear boundary value problems. A certain problem-parameter, $\gamma$ say, is introduced, so that instead of a single isolated problem we consider a continuum of problems, one problem $P(\gamma)$ for each value of $\gamma$ in an interval $\gamma_o \le \gamma \le \gamma_*$, where $P(\gamma_o)$ is easily solvable (e.g., it is linear), and $P(\gamma_*)$ is the target (the given) problem. The continuation method of solution is to advance $\gamma$ from $\gamma_o$ to $\gamma_*$ in steps $\delta\gamma$. At each step we use the final solution of the previous step (or extrapolation from several previous steps) as a first approximation in an iterative process for solving $P(\gamma)$. The main purpose of such continuation procedures is to ensure that the approximations we use in the iterative process are always "close enough" to the solution (of the current $P(\gamma)$), so that some desirable properties are maintained. Usually $\gamma$ is some natural physical parameter (the Reynolds number, the Mach number, etc.) in terms of which either the differential equations or the boundary conditions, or both, are expressed.

The continuation process is not a waste, for several reasons. In many cases, the intermediate problems $P(\gamma)$ are interesting by themselves, since they correspond to a sequence of cases of the same physical problem. More importantly, in solving non-linear discretized problems the continuation process is not only a method of computing the solution, but also, in

effect, the only way to define the solution, i.e., the way to select one out of the many solutions of the non-linear algebraic system. The desired solution is defined as the one which is obtained by continuous mapping from $[\gamma_o, \gamma_*]$ to the solution space with a given solution at $\gamma_o$ (e.g., the single solution, if $P(\gamma_o)$ is linear). By the continuation process, we keep every intermediate numerical solution in the vicinity of a physical solution (to an intermediate problem), hence the target numerical solution is, hopefully, near the target physical solution, and is not some spurious solution of the algebraic system. Thus, although sometimes we may get away without a continuation process (simply because a starting solution is "close enough", so that the continuation may be done in just one step), in principle a continuation process must be present in any numerical solution of non-linear problems. Moreover, such a process is usually inexpensive, since it can be done with crude accuracy, so that its intermediate steps usually total less computational work than the final step of computing an accurate solution to $P(\gamma_*)$.

A continuation process is necessary, in principle, not only for non-linear problems, but also for linear problems with grid adaptation. In fact, when h or p are themselves unknown, the discrete problem is nonlinear, even if the differential problem is linear.

In our system, a continuation process with crude accuracy and little work is automatically obtained by selecting a large value for the control parameter $\lambda$ (cf. Sec. 8.1). Then, in the final step ($\gamma=\gamma_*$), $\lambda$ is decreased to refine the solution. Thus, the overall process may be viewed as a multi-grid process of solution, controlled by the two parameters $\gamma$ and $\lambda$.

The most efficient way of changing $\gamma$ is probably to change it as soon as possible (e.g., when the multi-grid processing exhibits convergence to a crude tolerance) and to control the step-size $\delta\gamma$ by some automatic procedure, so that $\delta\gamma$ is sharply decreased when divergence is sensed (in the multi-grid processing), and slowly increased otherwise.

In changing $\gamma$ it is advisable to keep the residuals as smooth as possible, since higher frequency components are more expensive to liquidate (lower components being liquidated on coarser grids). Thus, for example, if a boundary condition should be changed while changing $\gamma$, it is advisable to introduce this change into the system at a stage when the algorithm is to start working on the coarsest grid.

$\gamma$-Extrapolation. In some cases the given problem ($\gamma=\gamma_*$) is much too difficult to solve, e.g., because the differential solution fluctuates on a scale too fine to be resolved. In such cases one is normally not interested in the details of the solution but rather in a certain functional of the solution. It is sometimes possible in such cases to solve the problem for certain values of $\gamma$ far from $\gamma_*$, and to extrapolate the corresponding functional values to $\gamma=\gamma_*$.

8.3. Practice of Discretization Control. The main practical restrictions imposed on the theoretical discretization equations (8.9) are the following: The approximation order p should be a positive integer. In many problems p is also restricted to be even, since odd orders are less efficient. The mesh-size function h(x) should be such that a reasonable grid can be constructed with it. Thus, in the grid structure outlined in Sec. 7.1, h is restricted to be of the form $h=2^k h_0$, where k is an integer. Also, in the multi-grid discretization method outlined in Sec. 7.2, any uniform subgrid truly influences the global solution only if it is large enough, i.e., if at least some of its inner points belong also to coarser grids. These discretization restrictions will actually help us in meeting another practical requirement, namely, the need to keep the control-work (computer work invested in testing for and affecting discretization reformulations) small compared with the numerical work (relaxation sweeps and interpolations).

The practical adaptive procedure is proposed to be generally along the following lines:

A. Testing. In the multi-grid solution process (possibly incorporating a continuation process), at some natural point we get an estimate of the decrease in the error estimator E introduced by the present discretization parameters. For example, in FAS Cycle C (see its flowchart in Fig. 2), at the point where new $\bar{F}^k$ is computed, the quantity

$$-\Delta E = G \left| \bar{F}^k - I^k_{k+1} \bar{F}^{k+1} \right| \tag{8.11}$$

at each point may serve as a local estimate for the decrease in E per unit volume (cf. (8.1) and (5.7)), owing to the refinement from $h_k$ to $h_{k+1}$. Each such decrease in E is related to some additional work $\Delta W$ (per unit volume). For example, the refinement from $h_k$ to $h_{k+1}$ requires the additional work

$$\Delta W = \frac{w(p)}{h_{k+1}^d} - \frac{w(p)}{h_k^d} \qquad \text{(per unit volume)}. \tag{8.12}$$

Hence we compute the ratio of exchanging accuracy per work $Q = -\Delta E / \Delta W$. At regions where this ratio is much bigger than $\lambda$ (the control rate of exchange; cf. Sec. 8.1) we say that the present parameter ($h_{k+1}$ in the example) is highly profitable and it is worth trying to further refine the discretization (e.g., introduce there the subgrid $G^{k+2}$ with $h_{k+2} = h_{k+1}/2$). At regions where $Q$ is much smaller than $\lambda$ we may coarsen the discretization (abolish the $G^{k+1}$ subgrid).

Extrapolated tests. More sophisticated tests may be based on assuming the truncation error to have some form of dependence on h and p, such as (8.10) above. Instead of using $\Delta E$ and $\Delta W$ at the previous change (from $h_k$ to $h_{k+1}$, in the above example) we can then anticipate the corresponding values $\overline{\Delta E}$ and $\overline{\Delta W}$ at the next change (from $h_{k+1}$ to $h_{k+2}$), which are the more appropriate values in testing whether to make that next change.

Thus, in the above example, assuming (8.10) and $h_{k+2} = h_{k+1}/2 = h_k/4$, we get $\overline{\Delta E} = 2^{-p}\,\Delta E$, $\overline{\Delta W} = 2^{d}\,\Delta W$, and hence

$$(8.13)\qquad \bar{Q} = \frac{-\overline{\Delta E}}{\overline{\Delta W}} = 2^{-p-d}\,Q = \frac{h_{k+1}^{d}\, G\, |\bar{F}^{k} - I_{k+1}^{k}\,\bar{F}^{k+1}|}{w(p)\,(2^{d}-1)\,2^{p}}\ .$$

The extrapolated ratio $\bar{Q}$ is used in testing for grid changes. This may seem risky, since it depends on assuming (8.10). But in fact there is no such risk, because we can see from (8.13) that testing with $\bar{Q}$ is not that much different from testing with $Q$. (In fact, if p is constant, testing with $\bar{Q}$ is equivalent to testing with $Q$ against another constant $\lambda$.) And the test with $Q$ does not presume (8.10); it only assumes that the finer ($G^{k+1}$) approximation is considerably better than the coarser one, so that their difference roughly corresponds to an added accuracy due to the refinement. Note also that the multi-grid stopping criteria ((A.17) or (A.20) in App. A) are precisely such that $Q$ can be reliably computed from the final approximation.

B. Changing the discretization. The desired grid changes are first just recorded (e.g., incidentally to the stage of computing $\bar{F}^{k}$) and only then they are simultaneously introduced, taking into account some organizational and stabilizational considerations: A change (e.g., refinement) is introduced only if there is a point where the change is "overdue" (e.g., a point where $\bar{Q} > 10\lambda$). Together with such a point the change is then also introduced at all neighbor (and neighbor of neighbor, etc.) points where the change is "due" (e.g., where $\bar{Q} > 3\lambda$). The changed subgrid ($G^{k+2}$ in the above example) is then augmented as follows: (i) Around each new grid point we add extra points, if necessary, so that the grid point (corresponding to a $G^{k+1}$ point where a refinement was due) becomes an inner point (cf. Sec. 7.2) in the new subgrid ($G^{k+2}$). (ii) Holes are filled; that is, if, on any grid line, a couple of points are missing in between grid points, these missing points are added to the grid.

The control work in this system is negligible compared with, say, the work of relaxing over $G^{k+1}$, because: (i) The tests are made in the transition from $G^{k+1}$ to $G^{k}$, which takes place only once per several $G^{k+1}$ relaxation sweeps. (ii) $Q$ is computed and tested only at points of the coarser grid $G^{k}$, and at each such point the work is smaller than the relaxation work per point. (iii) Changing the discretization is itself inexpensive since it is done by extending or contracting uniform grids (cf. Sec. 7.1), the main work being in interpolating the approximate solution to the new piece of uniform subgrid.

8.4 Generalizations. In some problems it is not enough to adapt h and p. Sometimes different increments $h^{(1)}, h^{(2)}, \ldots, h^{(d)}$ should be used at the d different directions, and each $h^{(j)}$ should be separately adapted. Basically the same procedures as above can be used to test and execute, for example, a change from $h^{(j)}$ to $h^{(j)}/2$. More generally, one would like to adapt the local coordinates (cf. Sec. 7.4), e.g., near discontinunities. Automatic procedures for such adaptation have not been so far developed, but are conceivable.

Other discretization parameters, such as the centering of each term in the difference operator, may be treated adaptively. (In fact, such adaptive discretization is already in use in mixed-type problems, where it was introduced by Murman to obtain stability. See, e.g., [9]). In problems with unbounded domains, the discrete domain may be determined adaptively (with increasingly coarser levels; cf. Sec. 7.1), using a procedure that decides to extend the domain if the previous extension was highly profitable in terms of $-\Delta E/\Delta W$. In many problems, some terms in the difference operator can altogether be discarded on most levels $G^k$. In particular, in singularly perturbed problems, the highest order terms may be kept only on the finest-narrowest levels. Decision can again be made in terms of $-\Delta E/\Delta W$, in an obvious way.

## 9. ADAPTIVE DISCRETIZATION: CASE STUDIES.

To get a transparent view of the discretization patterns and the accuracy-work relations typical to the adaptive procedures proposed above, we consider now several test cases which are simple enough to be analyzed in closed forms. That is, we consider problems with known solutions and simple behavior of the local truncation errors, and we calculate the discretization functions $h(x)$ and $p(x)$ that would be selected by the local optimization equations (8.9), and the resulting relation between the error estimator E and the computational work W.

9.1 Uniform-Scale Problems. A problem is said to have the uniform scale $\eta(x)$ if the local truncation error (8.2) has the behavior

$$\tau(x,h,p) \overset{\sim}{\sim} t(x) \left[\frac{h}{\eta(x)}\right]^p, \qquad (p_o \le p \le p_1) \ . \tag{9.1}$$

Such a behavior occurs, for example, when the solution is a trigonometric or exponential function $\exp(\theta \cdot x)$, where $\theta$ is either a constant or a slowly varying function (see example in Sec. 9.2). We will also assume for simplicity that (see (8.4))

$$w(p) = w_o p^{\ell} \tag{9.2}$$

Usually $\ell=1$, since the number of terms in the difference equations, and hence also the amount of computer operations at each grid point, are proportional to p. $\ell=2$ is appropriate if we assume that we have to increase the precision of our arithmetic when we increase p. Rescaling W, we can assume that $w_o = 1$.

Using (9.1-2) in equations (8.9) we get

$$G\tau = \lambda d\, p^{\ell-1} h^{-d}, \tag{9.3a}$$

$$G\tau \log \frac{h}{\eta} + \lambda \ell\, p^{\ell-1} h^{-d} \gtreqless o \ . \tag{9.3b}$$

Hence, denoting by $\tilde{p}$ the value of p that satisfies

(9.4) $$p^{\ell-1}\, e^{\ell p/d} = \lambda^{-1}\, Gt\, \eta^{d}\, e^{-\ell}\, d^{-1} ,$$

we have

(9.5a) $$h = \eta\, e^{-\ell/d} , \qquad p = \tilde{p}, \quad \text{if} \quad p_o \le \tilde{p} \le p_1 ,$$

(9.5b) $$h = (\lambda d\, p_o^{\ell-1} \eta^{p_o} t^{-1} G^{-1})^{1/(p_o+d)} , \quad p = p_o, \quad \text{if} \quad \tilde{p} \le p_o ,$$

(9.5c) $$h = (\lambda d\, p_1^{\ell-1} \eta^{p_1} t^{-1} G^{-1})^{1/(p_1+d)} , \quad p = p_1, \quad \text{if} \quad p_1 \le \tilde{p} .$$

Notice that at any point either p or h, but never both, is "adaptive", i.e., dependent of $\lambda$. Where p is adaptive ($p_0 \le p = \tilde{p} \le p_1$), h is fixed and each "scale cube" $\eta^d$ is divided into $e^{\ell}$ mesh cells.

Assume now further that the computer precision is unlimited (which is never really the case, but may provide insight), so that $\ell=1$ and $p_1=\infty$. If sufficiently high accuracy is desired, then $\lambda$ is sufficiently small to have $\tilde{p}>p_o$, so that (9.5a) applies. By (8.1) and (9.3a) this implies

(9.6) $$E = \lambda de \int \eta^{-d} dx ,$$

and hence, by (8.6),

(9.7) $$E = C_o e^{-W/(ed\int \eta^{-d} dx)} = C_o e^{-c\, \beta^{d} W} ,$$

where $\beta$ is some average value of the scale $\eta(x)$. In this (idealized) case, E decreases exponentially with W. For realistic W this convergence rate becomes poor when $\beta$ is very small, as in singularly perturbed problems. In such problems, however, for realistic W (9.5a) no longer applies, and another rate of convergence, independent of $\beta$, takes over (see Sec. 9.3).

9.2. One-Dimensional Case. Consider a 2-point boundary-value problem

(9.8) $$\frac{\eta}{2} \frac{d^2U}{dx^2} + \frac{dU}{dx} = 0 , \qquad \text{in } 0<x<1,$$

with constant $\eta>0$ and with boundary conditions U(0) and U(1) such that the solution is $U = e^{-2x/\eta}$. An elliptic (stable) difference approximation to such an equation can be central for $\eta \ge h$ but should be properly directed for $\eta<h$. (The first order term being the main term, the second order term should be differenced backward relative to it with approximation order $p' = p - [\log\eta/\log h]$. See [4] and Sec. 3.2 in [3]). In either case, the truncation error is approximately

(9.9) $$\tau(x,h,p) = t(x)\left(\frac{h}{\eta}\right)^p, \quad \text{where } t(x) = \frac{1}{2\eta} e^{-2x/\eta} .$$

We now choose the error weighting function to be

(9.10) $$G(x) \equiv 1,$$

which would be the choice (see (8.3)) when one is interested in accurate computation of boundary first-order derivatives (corresponding, e.g., to boundary pressure or drag, in some physical models). We again assume no precision limitations, so that $\ell=1$ and $p_1=\infty$. We take $p_o=2$ since second-order is no more expensive than first-order approximations. Inserting these into (9.5) we get

(9.11a) $$h = \frac{\eta}{e}, \quad p = \log\frac{1}{2\lambda} - 1 - \frac{2x}{\eta}, \quad \text{for} \quad 0 < x \le x_o ,$$

(9.11b) $$h = \frac{\eta}{e} e^{2(x-x_o)/(3\eta)}, \quad p = 2, \quad \text{for} \quad x_o \le x < 1 ,$$

where

(9.11c) $$x_o = \frac{\eta}{2}\left(\log\frac{1}{2\lambda} - 3\right) .$$

If $x_o \ge 1$, then (9.11a) applies throughout, and hence

(9.12) $$W = \int_o^1 \frac{p}{h}\,dx = \frac{e}{\eta}\left(\log\frac{1}{2\lambda} - 1 - \frac{1}{\eta}\right)$$

(9.13) $$E = \int_0^1 \tau dx = \frac{\lambda e}{\eta} = \frac{1}{2\eta} e^{-\frac{\eta}{e}W - \frac{1}{\eta}}$$

and the condition $x_o \ge 1$ itself becomes, by (9.11c, 12),

(9.14) $$W \ge \left(2+\frac{1}{\eta}\right)\frac{e}{\eta} .$$

Thus, if W satisfies (9.14), E converges like (9.13).

9.3. Singular Perturbation: Boundary Layer Resolution. When $\eta$ is very small, problem (9.8) is singularly perturbed, and its solution has a boundary layer near $x=0$. The above mesh-size $h=\eta/e$ is too small to be practical. Indeed, in the optimal discretization (9.11), for small $\eta$ we get small $x_o$, and an "external region" $x_o \le x < 1$ is formed where the mesh size grows exponentially from $\eta/e$. The small mesh size is used only to resolve the boundary layer. In this simplified problem the solution away from the boundary layer (i.e., for $x \gg \eta$) is practically constant, so that indefinitely large h is suitable. Usually h will grow exponentially, as in (9.11b), from $h = \frac{\eta}{e}$ to some definite value suitable for the external region. In the transition region we have $p=2$, i.e., the minimal

order of differencing is used in the region where h changes. This may be useful in practical implementations.

From (9.11) and (9.9) we get for small $\eta$

$$W = \int_0^1 \frac{p}{h}\,dx \approx \frac{e}{4}\left(\log\frac{1}{2\lambda}\right)^2 \tag{9.15}$$

$$E = \int_0^1 \tau\,dx = \frac{\lambda e}{2}\log\frac{1}{2\lambda} \approx \left(\frac{e}{4}W\right)^{1/2} e^{-\left(\frac{2}{e}W\right)^{1/2}} \tag{9.16}$$

where the integrals are separately calculated in $(o,X_o)$ and $(X_o,1)$. Thus, E converges exponentially as a function of $W^{1/2}$ instead of W, but this rate is independent of $\eta$ and does not deterioriate as $\eta \to 0$.

9.4. Singular Perturbation without Boundary-Layer Resolution. To see the effect of choosing different error weighting functions, consider again the above problem (Secs. 9.2, 9.3), but with the choice $G(x) = x$. This choice is typical to cases where one is not interested in calculating boundary derivatives of the solution (see (8.3)). We then get

$$\tilde{p} = \log\frac{x}{2\lambda} - 1 - \frac{2x}{\eta} \leq \log\frac{\eta}{4\lambda} - 2\ . \tag{9.17}$$

Therefore, for small $\eta$ and reasonable $\lambda$, $\tilde{p} < o$ and $p=2$ for all x. Hence, no resolution of the boundary layer is formed. Indeed, by (9.5b), for very small $\eta$ (singular-perturbation case)

$$\left(\frac{h}{\eta}\right)^3 = \frac{2\lambda}{x}e^{2x/\eta} \geq \frac{4\lambda e}{\eta} >> 1 \tag{9.18}$$

so that $h>>\eta$ . In practical situation where the solution in the external region is not constant, the actual mesh-size will be determined by the external regime.

9.5. Boundary Corners. Consider the two-dimensional Poisson equation $\Delta U=F$ with smooth F and homogeneous boundary conditions, near a boundary corner with angle $\pi/\alpha$, $\frac{1}{2} \leq \alpha < 1$. Denoting by r the distance from the corner, at small r the solution U is $O(r^\alpha)$, and so is also the error weighting function G (if accuracy is sought in the solution, but not in its derivatives near the boundary). Hence, $\tau = O(h^p r^{\alpha-p-2})$ and $\partial\tau/\partial h = O(h^{p-1} r^{\alpha-p-2})$. If we fix the order of approximation p, then the optimal mesh-spacing derived from (8.9a) is

$$h = O(\lambda^{1/(p+2)}\, r^\beta), \quad \beta = \frac{p+2-2\alpha}{p+2}\ . \tag{9.19}$$

Hence, by (8.4) and (8.1) the total work and total error contribution from a region of radius r around the corner are, respectively,

$$W = \int \frac{p}{h^2}\, dxdy \quad = \quad O(\lambda^{-2/(p+2)}\ r^{2-2\beta}) \ ,$$

$$E = \int G\ \tau\ dxdy \quad = \quad O(\lambda^{p/(p+2)}\ r^{2-2\beta}) \ .$$

Hence the relation $E \overset{\sim}{\sim} W^{-p/d}$ (the usual relation in d-dimensional smooth problem with p-th order approximation) still holds uniformly. The corner does not "contaminate" the global convergence.

In the practical grid organization (Sec. 8.3) finer levels $G^k$ with increasingly smaller mesh-sizes $h_k = 2^{-k} h_o$ will be introduced near the corner. By (9.19), the level $G^k$ will extend from the corner to a distance $r_k = C\ \lambda^{2\alpha-p-2}\ h_k^{1/\beta}$. Since $\beta<1$, for small $h_k$ we get $h_k > r_k$. This gives us in practice a natural stopping value for the refinement process: The finest mesh-size near the corner is such that $h_k \overset{\sim}{\sim} 4r_k$, so that level $G^k$ still has an inner point belonging to $G^{k-1}$.

9.6. Singularities. Like boundary corners, all kinds of other problem singularities, when treated adaptively, cause no degradation of the convergence rate (of E as function of W).

Consider for example the differential equation LU=F where F is smooth except for a jump discontinuity at x=0. Whatever the approximation order p, the system will find $-\Delta E$ (see (8.11)) to be O(1) at all points whose difference equation include values on both sides of the discontinuity. At these points further refinements will, therefore, be introduced as long as $-\Delta E/\Delta W > O(\lambda)$. Thus, around x=0, some fixed number (depending only on p) of mesh points will be introduced at each level $G^k$, until a mesh size $\bar{h} = O(\lambda^{1/d})$ is reached. The total amount of added work is therefore proportional to the number of levels introduced, which is $O(\log \bar{h})$. The error contribution of the discontinuity is $O(\bar{h}^{-d})$, which is exponentially small in terms of the added work.

This and similar analyses show that the adaptive scheme retains its high-order convergence even when the problem is only piecewise smooth, or has algebraic singularities, etc.

## 10. HISTORICAL NOTES AND ACKNOWLEDGEMENTS.

Coarse-grid acceleration techniques were recommended and used by several authors, including Southwell [24,13,14], Stiefel [15], Fedorenko [5], Ahamed [19], Wachspress [17], de la Valée Poissin [16] and Settari and Aziz [24].

Southwell called his technique "block" and more generally "group relaxation", described it as "almost essential to practical success", and gave heuristic explanation as well as practical implementation methods based on variational considerations ("the aim being to reduce the total energy by as great an amount as possible"). He also depicted procedures of "advance to a finer net" [14]. Techniques of multiplicative coarse-grid corrections (special-cases of which appeared in [14], [19]) were developed by Wachspress ([17], Chapter 9), who called them "variational techniques". This work motivated several studies, by Froelich, Wagner, Nakamura and Reed (see a brief survey in [18]) and was applied in nuclear reactor design computations.

All these were two-level methods. The multi-grid idea was introduced by Fedorenko [6], mainly for theoretical purposes. Namely, he rigorously proved that $W(n,\varepsilon)$, the number of operations required to reduce the residuals, of a Poisson problem on a rectangular grid with n points, by a factor $\varepsilon$, is $O(n|\log\varepsilon|)$. Bakhvalov [1] generalized this result to any second-order elliptic operator with continuous coefficients. For large n, this is the best possible result - except for the actual value of the coefficient. The Fedorenko estimate can be written as

$$W(n,.01) \leq 210000n + W(10^6,.01),$$

and the Bakhvalov constants are still much larger. For admissible values of n these estimates are therefore far worse than estimates obtained in other methods, and they did not encourage any development of the method. Fedorenko experimented with a two-level algorithm only, and seemed to imply that for practical grid sizes ADI may be more efficient. He did not realize the true practical potential, in both efficiency and programming simplification, of a full, systematic multi-grid approach. (It can be proved that $W(n,.01) \leq 106n$, and in practice $W(n,.01) \approx 50n$ is obtainable. See App. C).

The first full multi-grid algorithms and numerical tests were described in [2]. Our original approach was to regard the finer levels as "correcting" the coarser level (cf. Secs. 1, 7.2 and 7.5 above). For uniform non-adaptive grids this approach turns out to be equivalent to the one implied by [6], but fundamentally it is different and more powerful, since the process is not confined to a fixed discrete system.

A systematic multi-grid approach for a restricted class of problems, with somewhat different procedures of relaxation and transfer to coarser grids, is described in [21]. The multi-grid method is also portrayed in [23].

Adaptive discretization procedures were introduced by several authors. See for example [10], [20], [21] and references in [21]. The present approach is different, not only in its multi-level setting, but also in its basic criteria and procedures.

It is my pleasure to acknowledge the help I received from my students and colleagues throughout the work reported here. Yosef Shiftan, Nathan Diner, Yehoshua Fuchs and Dan Ophir in the Weizmann Institute; Jerry South in NASA Langley Research Center; and Will Miranker, Don Quarles, Fred Gustavson and Allan Goodman at IBM Thomas J. Watson Research Center - thank you all. I am also grateful for valuable discussions I had with Olof Widlund, Eugene Wachspress, Antony Jameson, Perry Newman, Jim Ortega, Ivo Babuska and Werner Rheinboldt.

REFERENCES:


[1] N.S. Bakhvalov, On the convergence of a relaxation method with natural constraints on the elliptic operator, Zh. vychisl. Mat. mat. Fiz. 6 (1966), 861-885.

[2] A. Brandt, Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems, Proc. 3rd. Int. Conf. Numerical Methods in Fluid Mechanics (Paris, 1972), Lecture Notes in Physics 18, Springer Verlag, Berlin, pp. 82-89.

[3] A. Brandt, Multi-Level Adaptive Techniques, IBM Research Report RC6026 (1976).

[4] A. Brandt, Elliptic difference operators and smoothing rates, in preparation.

[5] R.P. Fedorenko, A relaxation method for solving elliptic difference equations. Zh vychisl. Mat. mat. Fiz. 1 (1961), 922-927.

[6] R.P. Fedorenko, The speed of convergence of one iterative process, Zh. vychisl. Mat. mat. Fiz. 4 (1964), 559-564.

[7] J.M. Hyman, Mesh Refinement and Local Inversion of Elliptic Partial Differential Equations, LA-UR-75-1737, Los Alamos Scientific Laboratory (1975).

[8] A. Jameson, Numerical solution of nonlinear partial differential equations of mixed type, Third Symposium on the Numerical Solution of Partial Differential Equations (SYNSPADE), Univ. of Maryland (1975).

[9] E.M. Murman, Analysis of embedded shock waves calculated by relaxation methods. Proc. AIAA Computational Fluid Dynamics Conf., Palm Springs, California (1973), pp. 27-40.

[10] C.E. Pearson, On non-linear ordinary differential equations of boundary layer type, J. Math. Phys. 47 (1968), 351-358.

[11] Y. Shiftan, Multi-Grid Method for Solving Elliptic Difference Equations, MSc. Thesis (in Hebrew), Weizmann Institute of Science, Rehovot, Israel (1972).

[12] J.C. South, Jr. and A. Brandt, Application of a Multi-Level Grid Method to Transonic Flow Calculations, ICASE Report 76-8 (March 1976), ICASE, NASA Langley Research Center, Hampton, Virginia.

[13] R.V. Southwell, Relaxation Methods in Engineering Science, Oxford University Press, 1940.

[14] R.V. Southwell, Relaxation Methods in Theoretical Physics, Clarendon Press, Oxford (1946).

[15] E. Stiefel, Über einige Methoden der Relaxationsrechnung, Z. Angew. Math. Phys. 3 (1952), 1-33.

[16] F. de La Vallée Poussin, An accelerated relaxation algorithm for iterative solution of elliptic equations, SIAM J. Numer. Anal. 5 (1968), 340-351.

[17] E.L. Wachspress, Iterative Solution of Elliptic Systems, Prentice-Hall, New Jersey (1966).

[18] E.L. Wachspress, Variational acceleration of linear iteration, Proc. Army Workshop, Watervliet Arsenal, 1974.

[19] S.V. Ahamed, Accelerated Convergence of numerical solution of linear and nonlinear vector field problems, Computer J. 8 (1965), 73-76.

[20] I. Babuska, W. Rheinboldt and C. Mesztenyi, Self-Adaptive Refinements In the Finite Element Method, University of Maryland, Computer Science Technical Report TR-375 (1975).

[21] P.O. Frederickson, Fast Approximate Inversion of Large Sparse Linear Systems, Math Report 7-75, Lakehead University (1975).

[22] M. Lentini and V. Pereyra, An adaptive finite difference solver for nonlinear two point boundary problems with mild boundary layers, Stanford University Computer Science Dept. STAN-CS-75-530 (1975).

[23] R.A. Nicolaides, On multiple grid and related techniques for solving discrete elliptic systems, J. Comp. Phys. 19 (1975), 418-431.

[24] A. Settari and K. Aziz, A generalization of the additive correction methods for the iterative solution of matrix equations, SIAM J. Numer. Anal. 10 (1973), 506-521.

[25] R.V. Southwell, Stress Calculation in frameworks by the method of Systematic relaxation of contstraints, Parts I and II, Proc. Roy. Soc. (A) 151 (1935), 56-95.

## APPENDIX A. INTERPOLATIONS AND STOPPING CRITERIA: ANALYSIS AND RULES.

The multi-grid algorithms described above (Secs. 4 and 5) need to be supplemented with some rules of interpolations and stopping criteria. More specifically, for the interpolation $I_k^{k-1}$, transferring weighted residuals from a fine grid $G^k$ to the next coarser grid $G^{k-1}$, we should prescribe the weights, while for $I_{k-1}^k$, interpolating corrections from $G^{k-1}$ back to $G^k$, the method and order of interpolation should be prescribed. Stopping criteria should define convergence at the various levels and detect slow convergence rates. Numerical tests show that the parameters to be used are very robust: Full efficiency of the multi-grid algorithm is obtained for stopping parameters that do not depend on the geometry and the mesh size, and which may change over a wide range (see, e.g., Appendix B), provided the correct forms of the stopping criteria are used, and some basic rules of interpolation are observed. To find the correct forms and rules, and to determine the stopping parameters, we have to analyze the Coarse-Grid Correction (CGC) cycle, which consists of interpolating ($I_k^{k-1}$) the residuals to the coarser grid $G^{k-1}$, where the residual problem is solved, and then interpolating ($I_{k-1}^k$) that solution back as a correction to the $G^k$ approximation.

We can use a local mode analysis (for the linearized, coefficient-freezed difference equations), similar to the example in Sec. 3. Such an analysis may be inaccurate for the lowest frequency modes, for which the interaction with the boundary is significant. But these lowest modes are of little significance in our considerations, since they are efficiently approximated on the coarsest grids with little computational work, and since care will be taken (i) to choose interpolation schemes that do not convert small low-frequency errors into large high-frequency errors; and (ii) to stop relaxation sweeps before low-frequency error components become so large that they significantly feed the high frequencies (e.g., by boundary and non-linear interactions). In fact, we will see that the dominant components (i.e., the components that are slowest to converge in the combined process of relaxation and coarse-grid corrections) are the Fourier components $e^{i\theta\cdot x/h}$ for which $|\theta|$ is close to $\hat{\rho}\pi$, where (in a general d-dimensional problem)

$$\text{(A.0)} \qquad \theta = (\theta_1,\theta_2,\ \ldots,\theta_d), \qquad \theta\cdot X = \sum_{j=1}^{d} \theta_j x_j, \qquad |\theta| = \max_{1\le j\le d} |\theta_j| ,$$

$$h = h_k = \hat{\rho} h_{k-1} .$$

These components feed on each other in the interpolation processes between $G^k$ and $G^{k-1}$, they are slower to converge by relaxation, and in the CGC cycles they may even diverge.

To simplify the discussion we will assume that the mesh-size ratio has its usual value $\hat{\rho}=\frac{1}{2}$, which is the only one to be used in practice (cf. Sec. 6.2).

A.1. Coarse Grid Amplification Factors. For any given set of difference operators $L^k$ and a multi-grid scheme, a local mode analysis of the complete MG cycle can be made (cf. App. C), and the various parameters can be optimized. The essential information can, however, be obtained from a much simpler analysis that treat separately the two main processes, relaxation sweeps and CGC cycles. The smoothing rate $\bar{\mu}$ (see Sec. 3) is the main quantity describing the relaxation sweeps. The CGC local mode analysis is summarized below (for algebraic details see Sec. 4.5 of [3]).

In the CGC analysis, together with each basic Fourier component $e^{i\theta\cdot x/h}$ $(0 < |\theta| \le \frac{\pi}{2})$ we should treat all the $G^k$ components that coincide with it on $G^{k-1}$, i.e., all components $e^{i\theta'\cdot x/h}$ $(0 < |\theta'| \le \pi)$ such that $\theta'_j \equiv \theta_j \pmod{\pi}$ for $j=1,2,\ldots,d$. We call such component $\theta'$ a harmonic of $\theta$. We are especially interested in those harmonics that are not separated from $\theta$ by the relaxation sweeps, e.g., the set

$$T_\theta = \left\{ \theta' \equiv \theta \pmod{\pi} : \ \mu(\theta') \ge \mu(\theta)^2 \right\}.$$

Denote by $|T_\theta|$ the number of members in this set. (Usually $|T_\theta| = 2^\alpha$, where $\alpha$ is the number of coordinates j for which $|\theta j| \overset{\sim}{\sim} \frac{\pi}{2}$). In terms of the $\theta$ Fourier components and its harmonics, the CGC cycle has two effects:

(i) Assuming the components not in $T_\theta$ to be comparatively small when the CGC cycle is entered, the set of components in $T_\theta$ is transformed in the cycle by a certain matrix, whose spectral radius turns out to be

$$\text{(A.1)} \qquad \sigma(\theta) = \begin{cases} \sigma_o(\theta), & \text{if } |T_\theta| = 1, \\ \max(1, \sigma_o(\theta)), & \text{if } |T_\theta| > 1, \end{cases}$$

where

$$\text{(A.2)} \qquad \sigma_o(\theta) = \left| 1 - \sum_{\theta' \in T_\theta} R(\theta,\theta')\, B_k(\theta')\, B_{k-1}(2\theta)^{-1}\, \rho(\theta') \right| .$$

The functions $\rho(\theta')$, $R(\theta,\theta')$ and $B_\ell(\theta)$ are the "symbols" of $I_k^{k-1}$, $I_{k-1}^{k}$ and $L^\ell$, respectively, i.e.,

$$I_k^{k-1}\, e^{i\theta'\cdot x/h} = \rho(\theta')\, e^{i\theta\cdot x}, \qquad \text{(cf. (A.10) below)},$$

$$\text{(A.3)} \qquad I_{k-1}^{k}\, e^{i\theta\cdot x/h} = \sum_{\theta' \equiv \theta \,(\mathrm{mod}\,\pi)} R(\theta,\theta')\, e^{i\theta'\cdot x/h},$$

$$L^\ell\, e^{i\theta\cdot x/h_\ell} = B_\ell(\theta)\, e^{i\theta\cdot x/h_\ell}, \qquad (\ell = k, k-1).$$

(If L is a system of equations, and the right-hand side of (A.2) is therefore a matrix, then $\sigma_o(\theta)$ is meant to be the spectral radius of that matrix). For small $|\theta|$ we have $|T_\theta| = 1$ and hence

$$\text{(A.4)} \qquad \sigma(\theta) = \sigma_o(\theta) = 1 - \rho(o) + O\left(|\theta|^p + |\theta|^I\right),$$

where p is the approximation order of $L_k$ and $L_{k-1}$ (or the minimum of the two) and I is the order of the $I_{k-1}^{k}$ interpolation (I=2 for linear interpolation, etc.). The principal CGC amplification factor is

$$\text{(A.5)} \qquad \bar{\sigma} = \max_{o \le \theta \le \frac{\pi}{2}} \sigma(\theta)$$

$$= \max(1, \bar{\sigma}_o), \qquad \text{where} \quad \bar{\sigma}_o = \max_{o \le \theta \le \frac{\pi}{2}} \sigma_o(\theta).$$

(ii) The CGC cycles also generate new secondary harmonics $\theta''$ of $T_\theta$. The rate of generating these, i.e., the ratio of the new $\theta''$ amplitude to the old amplitude of the combined harmonics, turns out to be

$$\text{(A.6)} \qquad \sigma_1(\theta'') = \left| R(\theta,\theta'')\, B_k(\theta'')\, R_{k-1}(2\theta)^{-1}\, \rho(\theta'') \right|$$

$$= O(|\theta|^{I-m}) ,$$

where m is the order of the differential equations.

It follows from (A.4,6) that if $\rho(o) = 1$, as it is always chosen to be (cf. Sec. A.4), and if $I \geq m$, then components with small $|\theta|$ are very efficiently reduced in the multi-grid process.

A.2. The Coarse-to-Fine Interpolation $I^k_{k-1}$. On the other hand, it follows from (A.6) that if $I<m$ then even a small and smooth residual function may produce large high-frequency residuals, and significant amount of computational work will be required to smooth them out. This effect was clearly shown in numerical experiments ([2], [11]). Hence we have

The Basic Rule: The order of interpolation should be no less than the order of the differential equations. ($I \geq M$.) In particular, polynominal interpolation should be made with polynomials of degree $\geq m-1$.

Higher interpolation orders ($I>M$) are desired in the initial stages of solving a problem, when the residuals are (locally) smooth. For instance, in regions where the given problem has smoothness of order q (i.e., $F(x) = \Sigma A_\theta e^{i\theta\cdot x/h}$, $A_\theta = O(|\theta|^{-q} h^q)$), in order to ensure that the high-frequency residuals remain $O(h^q)$, at the i-th interpolation from $u^{M-1}$ to $u^M$ the order should be

$$\text{(A.7)} \qquad I \geq m + \max[q-(i-1)p,\ o].$$

(In fact, as long as $q > ip$, this interpolation need not be followed by $G^k$ relaxation sweeps, since the low-frequency amplitudes are still dominant. Relaxation would only feed from these low components to high frequency ones, causing additional work later. Still better, however, instead of this multi-grid mode without intermediate $G^k$ relaxation, is to make a higher-order correction on $G^{k-1}$).

Eventually, however, the smoothness of F (which is the original residual function) is completely lost in subsequent residuals and the convergence of components in the dominant range ($|\theta| \gtrsim \frac{\pi}{2}$) becomes our main concern. For these components, higher interpolation orders ($I>m$) is no more effective than the minimal order ($I=m$). This again was exhibited in numerical experiments ([2], [10]), which confirmed that the multi-grid efficiency is not improved (except in the $\lceil q/p \rceil$ first cycles) by using $I > m$.

An efficient method to implement high-order interpolations in case of equations of the form $\Delta^m U = F$ is to base the interpolation on suitably-rotated difference approximations. See [14, p. 53] and [7].

A.3. The Effective Smoothing Rate. The smoothing rate $\bar{\mu}$ was defined in (3.8) as the slowest convergence rate for all components not represented at the coarser level. More relevant, however, is the slowest rate among all components for which the coarse-grid correction is not effective, namely,

$$\text{(A.8)} \qquad \tilde{\mu} = \max \{\mu(\theta) : \frac{\pi}{2} \le |\theta| \le \pi \quad \text{or} \quad \sigma_o(\theta) \ge 1\} ,$$

which we call the "effective smoothing rate". It is clear, on one hand, that no rate faster than $\tilde{\mu}$ can be generally obtained as rate of convergence per $G^k$ relaxation sweep, no matter how well and how often the $G^{k-1}$ problem is solved. On the other hand, the rate $\tilde{\mu}$ can actually be attained (or approached) by correctly balancing the number of relaxation sweeps in between CGC cycles (see Sec. A.6). In most cases (all cases examined by us) one can make $\sigma_o(\theta) \le 1$ for all $|\theta| \le \frac{\pi}{2}$ by proper choice of $I_k^{k-1}$ (see Sec. A.4), and it is therefore justifiable to use $\bar{\mu}$ as the effective rate when relaxation schemes are studied by themselves.

A.4. The Fine-to-Coarse Weighting of Residuals ($I_k^{k-1}$), and the Coarse-Grid Operator $L^{k-1}$. The transfer of the $G^k$ residuals $r^k = f^k - L^k u^k$ to the coarser grid $G^{k-1}$, to serve there as the right-hand side $f^{k-1}$ (see Sec. 4, Step e) can be made in many ways. Generally $f^{k-1}$ is defined as some weighted average of the residuals in neighboring $G^k$ points:

$$\text{(A.9)} \qquad f^{k-1}(x) = I_k^{k-1} r^k(x) = \sum \rho_\nu r^k(x+\nu h) ,$$

where $\nu=(\nu_1,\nu_2,\ldots,\nu_d)$ , $\nu_j$ integers, and the summation is over a small set. In terms of these weights, $\rho(\theta)$ in (A.2) is given by

$$\text{(A.10)} \qquad \rho(\theta) = \sum \rho_\nu e^{i\theta\cdot\nu} .$$

The coarse grid operator $L^{k-1}$ can also be chosen in many ways, e.g., as some weighted average of the operator $L^k$ in neighboring points.

How are these choices to be made? The main purpose should be to minimize $\bar{\sigma}$, but without investing two much computational work in the weighting. Usually, it is preferable to adjust $\rho_\nu$ and not $L^{k-1}$, because this provides enough control on $\bar{\sigma}$ (cf. (A.2)) and because complicating $L^{k-1}$ adds many more computations and gets increasingly complicated as one advances to still coarser levels. For the programmer, using the same operators at all levels is an important simplification (cf. App. B), especially for non-linear problems.

It is clear from (A.4) that we should take $\rho(o) = \Sigma \rho_\nu = 1$. There is no apriori restriction, however, on the signs of the weights $\rho_\nu$. The trivial weighting

(A.11) $\qquad \rho_o=1, \quad \rho_\nu=0 \text{ for } \nu\neq 0; \qquad \rho(\theta) \equiv 1,$

called injection, has an important advantage in saving computations, not only because the weighting itself is saved, but mainly because it requires the computation of $r^k$ only at the $G^{k-1}$ points, while other weighting schemes compute $r^k$ at all $G^k$ points, an additional work comparable to one $G^k$ relaxation sweep.

Examples. For symmetric second-order equations, injection should usually be used. For the 5-point Laplace operator, for example, if we take $I^{k-1}_k$ to be injection, $I^k_{k-1}$ linear interpolation and $L^{k-1}$ also a 5-point Laplace operator, we get $\bar\sigma = \bar\sigma_o = 1$, the minimal possible value. Any weighting is a pure waste, including the "optimal" weighting

(A.12) $\qquad \rho_{00} = \frac{1}{2}, \quad \rho_{01} = \rho_{0-1} = \rho_{10} = \rho_{-10} = \frac{1}{8}, \quad \rho_{\alpha\beta} = o \text{ for } |\alpha|+|\beta| > 1,$

which minimized $\bar\sigma_o$, giving $\bar\sigma_o = \frac{1}{3}$, but does not lower $\bar\sigma$. Numerical tests (modifying the program of Appendix B) indeed showed no improvement by weighting. If, however, the equation has strong variation, making $B_{k-1}$ quite different from $B_k$, we may get for injection $\bar\sigma = \bar\sigma_o > 1$, while weighting (A.12) will keep $\bar\sigma_o$ safely below 1, giving $\bar\sigma = 1$.

For higher-order equations, non-trivial weighting offers an important advantage. If, for example, $L^k$ and $L^{k-1}$ are 13-points biharmonic operators and $I^k_{k-1}$ is cubic interpolation, then $\bar\sigma = 3$ for injection, while $\bar\sigma = 1$ for the weighting

$$\rho_{01} = \rho_{0-1} = \rho_{10} = \rho_{-10} = \frac{1}{4}, \qquad \rho_{\alpha\beta} = 0 \quad \text{for} \quad |\alpha|+|\beta| \neq 1.$$

A.5. Finite Elements Procedures. The main difference between finite-element and finite-difference multi-grid procedures is in the interpolation schemes. In the finite-element case, interpolation procedures follow automatically from the variational formulation and the definition of the approximation spaces $S^k$ (corresponding to the levels $G^k$). Usually, $S^k$ is a subspace of $S^{k-1}$. The coarse-to-fine interpolation is, therefore, simply the identity operation. Also, if the variational problem in $S^k$ is to minimize $A_k(v^k)$, then, for any given approximation $v^k$, the correction problem in the coarser space $S^{k-1}$ is, simply, to minimize

(8.13) $\qquad A_{k-1}(v^{k-1}) = A_k(v^k + v^{k-1}) .$

Example. Consider the standard example, where $S^k$ is the space of piecewize linear functions on the triangulation $G^k$ and $A_k$ is a Dirichlet integral whose minimization is equivalent to the difference equation $\Delta^k v^k = F^k$, $\Delta^k$ being the 5-point Laplacian. Computing $A_{k-1}$ by (A,B), it turns out to be equivalent to the equation $\Delta^{k-1} v^{k-1} = I_k^{k-1} (F^k - \Delta^k v^k)$, where $I_k^{k-1}$ has the weights (cf. Sec. A.4).

$$\rho_{00} = \frac{1}{4}, \qquad \rho_{01} = \rho_{11} = \rho_{10} = \rho_{0-1} = \rho_{-1\,-1} = \rho_{-10} = \frac{1}{8}.$$

These weights give the same multi-grid convergence rate as injection (and are, therefore, redundant).

A.6. Criteria for Slow Convergence Rates. (A) Relaxation sweeps, say on $G^k$, should be discontinued, and a switch should be made to a coarse-grid correction, when the rate of convergence becomes slow; e.g., when

$$\text{(A.14)} \qquad \frac{\text{residual norm}}{\text{residual norm a sweep earlier}} \geq \eta \equiv \frac{\bar{\sigma}^{-3} + 3\tilde{\mu}}{\bar{\sigma}^{-3} + 3}$$

The norm here is a suitable (e.g., $L_2$, $L_\infty$ or (A.18)) discrete measure, usually of the "dynamic" residuals, that is, residuals computed incidentally to the relaxation process. $\tilde{\mu}$ and $\bar{\sigma}$ are defined in (A.8) and (A.5), respectively. Usually, one can choose the $I_k^{k-1}$ weighting so that $\bar{\sigma}=1$, in which case $\tilde{\mu}=\mu$. In any case, (A.14) is designed to ensure that, on one hand, the CGC cycle is delayed enough to make its $\bar{\sigma}$ magnification small compared with the intermediate reduction by relaxation sweeps. On the other hand, for $\theta$ with $\mu(\theta)$ considerably slower than $\tilde{\mu}$, the CGC cycles are still sufficiently frequent to compensate for the slower $\mu$, since their reduction rate $\sigma(\theta)$ decreases rapidly ((A.4) with $\rho(0)=1$). the stopping rule (A.14) also prevents low error frequencies from dominating relaxation, thus avoiding significant feeding from low to high frequencies (through boundary and nonlinear interactions).

If the "stopping rate" $\eta$ varies over the domain of computations (as a result of variations in L, in case of nonlinear or non-constant-coefficients problems), the largest $\eta$ should be chosen for the stopping criterion (A.14). If log $\eta$ changes too much over the domain (which should not happen when a proper relaxation scheme is used), then (A.14) must be checked separately in subdomains, and partial sweeping (see Sec. A.9) might be used.

An appropriate value of $\eta$ may also easily be found by direct trial and error. Such value is typical to the (locally linearized, coefficient-freezed) problem, is independent of either h, $\Omega$ or F, and may therefore be found, once for all, on a moderately coarse grid. In some nonlinear problems the value may need some adjustment as the computations proceed. Whenever the coarse-grid corrections seem to be ineffective, $\eta$ should be increased, e.g., to $(1+3\eta)/4$. Generally, the overall multi-grid convergence rate is not much sensitive to increasing $\eta$: At worst, the rate may become $\eta$ instead of the theoretically best rate $\max_\Omega \tilde{\mu}^{3/4}$ (cf. Sec. 6.2).

For the Poisson equation with Gauss-Seidel relaxation, for example, we have $\bar{\sigma}=1$, $\tilde{\mu}=\bar{\mu}=.5$, hence $\eta=.625$. The example in Appendix B shows that the optimal MG convergence rate $\tilde{\mu}^{3/4} \approx .595$ is indeed attained. Experimenting with this program gave similar results for any smaller $\eta$ (the reason being that the minimal number of two sweeps at each level is good enough in this problem), while for any $\eta \leq .95$ the total amount of computational work was no more than twice the work at $\eta=.62$.

(B) Another way to decide upon discontinuation of relaxation is to directly measure the smoothness of the residuals. The switch to coarser grids can be made, for instance, when differences between residuals at neighboring points are small compared with the residuals themselves.

A.7. Convergence Criteria on Coarser Grids. In the CGC mode analysis above it was assumed that the problem on the coarser grid $G^{k-1}$ was fully solved and then interpolated as a correction to the $G^k$ approximation. In the actual multi-grid algorithm (Sec. 4) we solve the $G^{k-1}$ problem iteratively, stopping the iterations when some convergence criterion is met. This criterion should roughly detect the situation at which more improvement (per unit work) is obtained by relaxing on the $G^k$ grid (after interpolating) then by further iterating the $G^{k-1}$ problem (before interpolating). A crude mode analysis (similar to Sec. 4.6.2 in [3]) shows that such a criterion is

$$(A.15) \qquad ||r^{k-1}|| \leq \delta \, ||r^k||, \qquad \delta = \frac{\bar{\sigma}\,(1 - \mu_k^{\,2^{-d}})}{\bar{S}\,(\mu_k^{\,2^{-d}} - \mu_{k-1})},$$

where d is the dimension, $\bar{\sigma}$ is given by (A.5),

$$\bar{S} = \max_{|\theta| \leq \frac{\pi}{2}} \left| \sum_{\theta' \in T_\theta} R(\theta,\theta')\, B_k(\theta')\, B_{k-1}(2\theta)^{-1}\, \rho(\theta') \right|$$

and $\mu_\ell = \tilde{\mu}^{(1-2^{-d})}$ on the $G^\ell$ grid (cf. (A.8)). $||r^{k-1}||$ is any norm of the current residuals in the $G^{k-1}$ problem, while $||r^k||$ is the corresponding norm in the $G^k$ problem. It is important that these norms are comparable: They should be discrete approximations to the same continuum norms. Also, if $r^{k-1}$ are the "dynamic" residuals (i.e., computed incidentally to the last $G^{k-1}$ relaxation sweep, using latest available values of the relaxed solution) then $r^k$ should be the $G^k$ dynamic residuals, unlike the residuals transferred to $G^{k-1}$ (to define $f^{k-1}$; cf. Sec. A.4) which must be "static" residuals (i.e., computed over the grid without changing the solution at the same time). If, however, $r^k$ and $r^{k-1}$ are static and dynamic, respectively, the parameter $\delta$ in (A.15) should be multiplied by a certain factor $\bar{\beta}$ (see Sec. 4.6.2 in [3]).

The stopping criterion (A.15) is based on the assumption that error components with $|\theta| = \frac{\pi}{2}$ dominates the process. In the first $\lceil q/p \rceil$ CGC cycles, however, lower components are dominant, and the main consideration is to converge them. Hence, at that initial state, the $G^{k-1}$ convergence criteria should be

$$(A.16) \qquad ||r^{k-1}|| \leq ||\tau^{k-1}||$$

where $\tau^{k-1}$ are the $G^{k-1}$ truncation errors (cf. Sec. A.8).

The key factor $\delta$ can also be found by trial and error. Like $\eta$ above, it is essentially independent of h, $\Omega$ and F, and may, therefore, be found once for all by tests on moderately coarse grids. Numerical experiments show that the overall multi-grid efficiency is not much sensitive to very large variations in $\delta$ and, in particular, $\delta$ may be lowered by orders of magnitudes without large changes in the efficiency. For example:

For the 5-points Poisson equation with Gauss-Seidel relaxation, injection and linear interpolations, (A.15) yields $\delta = .219$. Numerical experiment (e.g., with the program in Appendix B) show that with any $.001 \leq \rho \leq .5$ the computational work is no more than 25% above the work with $\rho = .22$, and no more than 100% extra work for any $.0001 \leq \rho \leq .7$.

A.8. Convergence on the Finest Grid. On the finest grid $G^M$ the solution is usually considered converged when the (static) residuals are of the order of the truncation error, in some appropriate norm. One way to estimate the truncation error is to measure them on coarser grids by (5.7), and extrapolate (taking into account that they are $O(h^p)$). Another, related but more straightforward criterion is to detect when the $G^M$ solution has contributed most of its correction to the $G^{M-1}$ solution. In the FAS algorithm the natural place to check is when a new $\bar{F}^{M-1}$ is computed, the convergence test being

$$(A.17) \qquad ||\bar{F}^{M-1} - \bar{F}^{M-1}_{\text{previous}}|| \ll ||\bar{F}^{M-1} - I^{M-1}_{M} F^{M}|| .$$

The norm here may be any ($L_2$, $L_\infty$, etc.), but the most relevant one is the discrete version of the norm (cf. Sec. 8.1)

$$(A.18) \qquad ||f|| = \int G(x) \, |f(x)| dx$$

A.9. Partial Relaxation Sweeps. A partial relaxation sweep over $G^k$ is a relaxation sweep that may skip some subdomains of $G^k$. (Unlike "selective" relaxation sweeps, which in principle pass through all the grid points, although corrections may not be introduced in some of them. Cf. Sec. 3.2. A partial sweep may be selective, too.)

Partial sweeps are not used much in standard relaxation calculations. Usually, a slow-to-converge subdomain is coupled to other subdomains and therefore cannot be relaxed separately. In the multi-grid process, however, only high-frequency error components are to be reduced by relax-

ation, and this can be done separately in subdomains: With regard to high-frequencies, subdomains are practically decoupled. Hence, in the multi-grip process, partial sweeps are potentially very important. In fact, high-frequency amplitudes may vary greatly over the domain, especially if $\bar{\mu}$ and $\bar{\sigma}$ vary much, or if high-frequency error components are introduced at boundaries, making partial sweeping there very desirable.

Partial sweeping may be performed by applying a criterion for slow convergence (Sec. A.6) separately in subdomains. (If the connected region of partial relaxation is small, $\eta$ in (A.14) should be changed to $(\bar{\sigma}^3\hat{\mu}+3\hat{\mu})/(\bar{\sigma}^3+3)$, where $\hat{\mu}$ is the largest amplification factor for Fourier components on the relaxed region.) A subdomain may be excluded from subsequent relaxation sweeps if slow convergence is shown simultaneously on that subdomain and on all neighboring subdomains. Under relaxation may be used to phase-out the relaxed region (cf. [3], Sec. 4.6.4). The subdomains may be chosen quite arbitrarily, but each of them should be large enough (at least 4x4) to allow for separate smoothing.

## A.10. Convergence Criteria on Non-uniform Grids

When $G^k$ and $G^{k-1}$ are not coextensive (i.e., the domain covered by $G^k$ is only part of the $G^{k-1}$ domain; cf. Sec. 7.2), the convergence criteria (Secs. A.7-8) should be slightly modified. First, in (A.15), $||r^k||$ is not a comparable norm, since it may be measured on a much narrower subdomain. Instead, one can use the test

$$||r^{k-1}|| < \delta||r_1^{k-1}||/\eta, \tag{A.19}$$

where $||r_1^{k-1}||$ is the residual norm computed on $G^{k-1}$ at the first relaxation sweep after switching from $G^k$. The division by $\eta$ in (A.19) is designed to compensate for the fact that $||r_1^{k-1}||$ is computed a sweep later than $||r^k||$.

The other modification is in (A.17), where it was assumed that $G^M$ is the finest level everywhere. Generally, the convergence test can be, for example,

$$||\bar{F}^k - \bar{F}^k_{\text{previous}}|| << ||\bar{F}^k - I^k_{k+1}\bar{F}^{k+1}||, \qquad \text{for all } k = (0,1,\ldots,M-1), \tag{A.20}$$

where the norms are taken over $G^k_{k+1}$ (or, more precisely, over $G^k_{k+1} - G^{k+1}_{k+2}$).

## APPENDIX B: SAMPLE MULTI-GRID PROGRAM AND OUTPUT.

This simple program of Cycle C (written in 1974 by the author at the Weizmann Institute) illustrates multi-grid programming techniques and exhibits the typical behavior of the solution process. For a full description of Cycle C, see Sec. 4 or the flowchart in Fig. 1.

The program solves a Dirichlet problem for Poisson equation on a rectangle. The same 5-point operator is used on all grids. The $I^{k-1}_{k}$ residuals transfer is the trivial one (injection), the $I^{k}_{k-1}$ interpolation is linear. The higher interpolation (A.7) and the special stopping criterion (A.16), recommended for the first [q/p] cycles, are not implemented here.

For each grid $G^k$ we store both $v^k$ and $f^k$ (k=1,2,...,M). For handling these arrays $f^k$ is also called $v^{k+M}$. The coarsest grid has NX0 x NY0 intervals of length H0 each. Subsequent grids are defined as straight refinements, with mesh sizes H(k) = H0/2**(k-1). The function F(x,y) is the right-hand side of the Poisson equation. The function G(x,y) serves both as the Dirichlet boundary condition ($\Phi^M$) and as the first approximation ($u^M_o$). The program cycles until the $L_2$ norm of the residuals on $G^M$ is reduced below TOL, unless WORK exceeds WMAX. After each relaxation sweep on any grid $G^k$, a line is printed out showing the level k, the $L_2$ norm of the ("dynamic") residuals computed in course of this relaxation, and WORK, which is the accumulated relaxation work (where a sweep on the finest grid is taken as the work unit).

Note the key role of the GRDFN and KEY subroutines. The first is used to define a grid ($v^k$), i.e., to allocate for it space in the general vector Q (where IQ points to the next available location), and to store its parameters. To use grid $v^k$, CALL KEY(k,IST,M,N,H) retrieves the grid parameters (dimension MxN and mesh-size H) and sets the array IST(i) so that $v^k_{ij}$ = Q(IST(i)+j). This makes it easy to write one routine for all grids $v^k$; see for example, Subroutine PUTZ(k). Or to write the same routines (RELAX, INTADD, RESCAL) for all levels.

To solve on the same domain problems other than Poisson, the only subroutines to be changed are the relaxation routine RELAX and the residual injection routine RESCAL, the latter being just a slight variation of the first.

For different domains, more general GRDFN and KEY subroutines should be written. A general GRDFN subroutine, in which the domain characteristic function is one of the parameters, has been developed, together with the corresponding KEY routine. This essentially reduces the programming of any multi-grid solution to programming a usual relaxation routine.

# CYCLE C

```
      PROGRAM CYCLE C
      EXTERNAL G,F
      CALL MULTIG(3,2,1.,6,.01,30.,G,F)
      STOP
      END

      FUNCTION F(X,Y)
      F=SIN(3.*(X+Y))
      RETURN
      END

      FUNCTION G(X,Y)
      G=COS(2.*(X+Y))
      RETURN
      END

      SUBROUTINE MULTIG(NX0,NY0,H0,M,TOL,WMAX,U1,F)
      EXTERNAL U1,F
      DIMENSION EPS(10)
      DO 1 K=1,M
      K2=2**(K-1)
      CALL GRDFN(K,NX0*K2+1,NY0*K2+1,H0/K2)
    1 CALL GRDFN(K+M,NX0*K2+1,NY0*K2+1,H0/K2)
      EPS(M)=TOL
      K=M
      WU=0
      CALL PUTF(M,U1,0)
      CALL PUTF(2*M,F,2)
    5 ERR=1.E30
    3 ERRP=ERR
      CALL RELAX(K,K+M,ERR)
      WU=WU+4.**(K-M)
      WRITE(6,4)K,ERR,WU
    4 FORMAT(' LEVEL',I2,'   RESIDUAL NORM=',1PE10.3,'    WORK=',0PF7.3)
      IF(ERR.LT.EPS(K))GOTO 2
      IF (WU.GE.WMAX)RETURN
      IF(K.EQ.1.OR. ERR/ERRP.LT. .6)GOTO 3
      CALL RESCAL(K,K+M,K+M-1)
      EPS(K-1)=.3*ERR
      K=K-1
      CALL PUTZ(K)
      GOTO 5
    2 IF (K.EQ.M)RETURN
      CALL INTADD(K,K+1)
      K=K+1
      GOTO 5
      END

      SUBROUTINE GRDFN(N,IMAX,JMAX,HH)
      COMMON/GRD/NST(20),IMX(20),JMX(20),H(20)
      DATA IQ/1/
      NST(N)=IQ
      IMX(N)=IMAX
      JMX(N)=JMAX
      H(N)=HH
      IQ=IQ+IMAX*JMAX
      RETURN
      END

      SUBROUTINE KEY(K,IST,IMAX,JMAX,HH)
```

Right-hand side of the equation

Boundary values and first approximation

Multi-grid algorithm (see Fig. 1)

$\eta$=.6

$\delta$=.3

Define an IMAX × JMAX array $v^N$ .

```
      COMMON/GRD/NST(20),IMX(20),JMX(20),H(20)
      DIMENSION IST(1)
      IMAX=IMX(K)
      JMAX=JMX(K)
      IS=NST(K)-JMAX-1
      DO 1 I=1,IMAX
      IS=IS + JMAX
    1 IST(I)=IS
      HH=H(K)
      RETURN
      END
```

Set IST such that

$$v^k(I,J) = Q(IST(I) + J),$$

and set IMAX = IMX(K)

JMAX = JMX(K)

HH = H(K)

```
      SUBROUTINE PUTF(K,F,NH)
      COMMON Q(18000),IST(600)
      CALL KEY (K,IST,II,JJ,H)
      H2=H**NH
      DO 1 I=1,II
      DO 1 J=1,JJ
      X=(I-1)*H
      Y=(J-1)*H
    1 Q(IST(I)+J)=F(X,Y)*H2
      RETURN
      END
```

$$v^K \leftarrow H(K)^{NH} \cdot F^K$$

```
      SUBROUTINE PUTZ(K)
      COMMON Q(18000),IST(200)
      CALL KEY(K,IST,II,JJ,H)
      DO 1 I=1,II
      DO 1 J=1,JJ
    1 Q(IST(I)+J)=0.
      RETURN
      END
```

$$v^K \leftarrow 0$$

```
      SUBROUTINE RELAX(K,KRHS,ERR)
      COMMON Q(18000),IST(200),IRHS(200)
      CALL KEY(K,IST,II,JJ,H)
      CALL KEY(KRHS,IRHS,II,JJ,H)
      I1=II-1
      J1=JJ-1
      ERR=0.
      DO 1 I=2,I1
      IR=IRHS(I)
      IO=IST(I)
      IM=IST(I-1)
      IP=IST(I+1)
      DO 1 J=2,J1
      A=Q(IR+J)-Q(IO+J+1)-Q(IO+J-1)-Q(IM+J)-Q(IP+J)
      ERR=ERR+(A+4.*Q(IO+J))**2
    1 Q(IO+J)=-.25*A
      ERR=SQRT(ERR)/H
      RETURN
      END
```

A Gauss-Seidel Relaxation sweep on the equation

$$\Delta_h v^K = v^{KRHS}$$

giving

$$ERR = ||\text{residuals}||_{L_2}$$

```
      SUBROUTINE INTADD(KC,KF)
      COMMON Q(18000),ISTC(200),ISTF(200)
      CALL KEY(KC,ISTC,IIC,JJC,HC)
      CALL KEY(KF,ISTF,IIF,JJF,HF)
      DO 1 IC=2,IIC
      IF=2*IC-1
      JF=1
```

Linear interpolation and addition

$$v^{KF} \leftarrow v^{KF} + I^{KF}_{KC} v^{KC}$$

```
      IFO=ISTF(IF)
      IFM=ISTF(IF-1)
      ICO=ISTC(IC)
      ICM=ISTC(IC-1)
      DO 1 JC=2,JJC
      JF=JF+2
      A=.5*(Q(ICO+JC)+Q(ICO+JC-1))
      AM=.5*(Q(ICM+JC)+Q(ICM+JC-1))
      Q(IFO+JF)   = Q(IFO+JF)+Q(ICO+JC)
      Q(IFM+JF)   = Q(IFM+JF)+.5*(Q(ICO+JC)+Q(ICM+JC))
      Q(IFO+JF-1)=Q(IFO+JF-1)+A
    1 Q(IFM+JF-1)  = Q(IFM+JF-1)+.5*(A+AM)
      RETURN
      END

      SUBROUTINE RESCAL(KF,KRF,KRC)
      COMMON Q(18000),IUF(200),IRF(200),IRC(200)
      CALL KEY(KF,IUF,IIF,JJF,HF)
      CALL KEY(KRF,IRF,IIF,JJF,HF)
      CALL KEY(KRC,IRC,IIC,JJC,HC)
      IIC1=IIC-1
      JJC1=JJC-1
      DO 1 IC=2,IIC1
      ICR=IRC(IC)
      IF=2*IC-1
      JF=1
      IFR=IRF(IF)
      IFO=IUF(IF)
      IFM=IUF(IF-1)
      IFP=IUF(IF+1)
      DO 1 JC=2,JJC1
      JF=JF+2
      S=Q(IFO+JF+1)+Q(IFO+JF-1)+Q(IFM+JF)+Q(IFP+JF)
    1 Q(ICR+JC)=4.*(Q(IFR+JF)-S+4.*Q(IFO+JF))
      RETURN
      END
```

**Residuals injection**

$$v^{KRC} \leftarrow I_{fine}^{coarse}\,(v^{KRF} - \Delta_h v^{KF})$$

```
LEVEL 6   RESIDUAL NORM= 2.814E+01   WORK=   1.000
LEVEL 6   RESIDUAL NORM= 2.764E+01   WORK=   2.000
LEVEL 5   RESIDUAL NORM= 2.659E+01   WORK=   2.250
LEVEL 5   RESIDUAL NORM= 2.555E+01   WORK=   2.500
LEVEL 4   RESIDUAL NORM= 2.317E+01   WORK=   2.563
LEVEL 4   RESIDUAL NORM= 2.095E+01   WORK=   2.625
LEVEL 3   RESIDUAL NORM= 1.649E+01   WORK=   2.641
LEVEL 3   RESIDUAL NORM= 1.285E+01   WORK=   2.656
LEVEL 2   RESIDUAL NORM= 7.626E+00   WORK=   2.660
LEVEL 2   RESIDUAL NORM= 3.840E+00   WORK=   2.664
LEVEL 3   RESIDUAL NORM= 5.058E+00   WORK=   2.680
LEVEL 4   RESIDUAL NORM= 8.006E+00   WORK=   2.742
LEVEL 4   RESIDUAL NORM= 2.545E+00   WORK=   2.805
LEVEL 5   RESIDUAL NORM= 9.736E+00   WORK=   3.055
LEVEL 5   RESIDUAL NORM= 2.464E+00   WORK=   3.305
LEVEL 6   RESIDUAL NORM= 1.064E+01   WORK=   4.305
LEVEL 6   RESIDUAL NORM= 2.442E+00   WORK=   5.305
LEVEL 6   RESIDUAL NORM= 2.399E+00   WORK=   6.305
LEVEL 5   RESIDUAL NORM= 2.351E+00   WORK=   6.555
LEVEL 5   RESIDUAL NORM= 2.303E+00   WORK=   6.805
LEVEL 4   RESIDUAL NORM= 2.173E+00   WORK=   6.867
LEVEL 4   RESIDUAL NORM= 2.043E+00   WORK=   6.930
LEVEL 3   RESIDUAL NORM= 1.739E+00   WORK=   6.945
LEVEL 3   RESIDUAL NORM= 1.453E+00   WORK=   6.961
LEVEL 2   RESIDUAL NORM= 9.889E-01   WORK=   6.965
LEVEL 2   RESIDUAL NORM= 6.183E-01   WORK=   6.969
LEVEL 1   RESIDUAL NORM= 2.760E-01   WORK=   6.970
LEVEL 1   RESIDUAL NORM= 5.170E-02   WORK=   6.971
LEVEL 2   RESIDUAL NORM= 2.292E-01   WORK=   6.975
LEVEL 3   RESIDUAL NORM= 5.465E-01   WORK=   6.990
LEVEL 4   RESIDUAL NORM= 7.710E-01   WORK=   7.053
LEVEL 4   RESIDUAL NORM= 1.163E-01   WORK=   7.115
LEVEL 5   RESIDUAL NORM= 8.657E-01   WORK=   7.365
LEVEL 5   RESIDUAL NORM= 1.058E-01   WORK=   7.615
LEVEL 6   RESIDUAL NORM= 9.059E-01   WORK=   8.615
LEVEL 6   RESIDUAL NORM= 1.052E-01   WORK=   9.615
LEVEL 6   RESIDUAL NORM= 1.012E-01   WORK=  10.615
LEVEL 5   RESIDUAL NORM= 9.759E-02   WORK=  10.865
LEVEL 5   RESIDUAL NORM= 9.452E-02   WORK=  11.115
LEVEL 4   RESIDUAL NORM= 8.710E-02   WORK=  11.178
LEVEL 4   RESIDUAL NORM= 7.960E-02   WORK=  11.240
LEVEL 3   RESIDUAL NORM= 6.389E-02   WORK=  11.256
LEVEL 3   RESIDUAL NORM= 4.931E-02   WORK=  11.271
LEVEL 2   RESIDUAL NORM= 2.916E-02   WORK=  11.275
LEVEL 2   RESIDUAL NORM= 1.622E-02   WORK=  11.279
LEVEL 2   RESIDUAL NORM= 1.017E-02   WORK=  11.283
LEVEL 3   RESIDUAL NORM= 1.949E-02   WORK=  11.299
LEVEL 4   RESIDUAL NORM= 3.128E-02   WORK=  11.361
LEVEL 4   RESIDUAL NORM= 8.843E-03   WORK=  11.424
LEVEL 5   RESIDUAL NORM= 3.710E-02   WORK=  11.674
LEVEL 5   RESIDUAL NORM= 8.486E-03   WORK=  11.924
LEVEL 6   RESIDUAL NORM= 4.007E-02   WORK=  12.924
LEVEL 6   RESIDUAL NORM= 9.051E-03   WORK=  13.924
```

# OUTPUT

Error reduction by a factor greater than 10 per cycle.

Each cycle costs 4.3 WU

**Insensitivity:** Results would be practically the same for any $.005 \leq \delta \leq .5$ or any $0 \leq \eta \leq .65$

## APPENDIX C. RIGOROUS BOUND TO MODEL-PROBLEM CONVERGENCE RATE.

We consider the model problem: 5-points Poisson equation $\Delta_h U_h = F$ on a $(n_1+1) \times (n_2+1)$ rectangular grid $G^M$ with Dirichlet boundary conditions. Let $n_j = 2^M N_j$ and let $G^k$ be the $(2^k N_1+1) \times (w^k N_2+1)$ uniform grid on the same domain, with mesh size $h_k = 2^{-k} h_o$, $(k=o, 1, \ldots, M)$. We will estimate the convergence rate and work in one multi-grid cycle $C^M$.

The cycle $C^M$ is defined inductively as follows: (i) Make r relaxation sweeps on the $G^M$ approximate solution $u^M$. To facilitate the rigorous Fourier analysis we choose as our relaxation the Weighted Simultaneous Displacement (WSD, or "weighted Jacobi") method with the optimal weights $\omega_{00}=48/41$, $\omega_{01}=\omega_{0-1}=\omega_{10}=\omega_{-10}=8/41$ (see Sec. 3.3). (ii) Inject (cf. Sec. A.4) the residual problem to $G^{M-1}$. (iii) Get an approximate solution $v^{M-1}$ to this $G^{M-1}$ problem by <u>two</u> $C^{M-1}$ cycles, starting from the zero approximation. (iv) Correct $u^M \leftarrow u^M + I^M_{M-1} v^{M-1}$, where $I^M_{M-1}$ is linear interpolation.

It is easily calculated that one WSD sweep amplifies the Fourier component $\exp(i\theta \cdot x/h_M)$ of the residual by the factor

$$\mu(\theta) = 1 - (2 - \cos\theta_1 - \cos\theta_2)\ (24 + 8\cos\theta_1 + 8\cos\theta_2)\ /\ 41.$$

Denote by $A(\theta)$ the amplitude, before the $C^M$ cycle, of the $\theta=(\theta_1,\theta_2)$ component of the residual. Acutally present on the grid $G^M$ are only components of the form $\theta = (\alpha_1\pi/n_1, \alpha_2\pi/n_2)$, $(\alpha_j = \pm 1, \pm 2, \ldots, \pm(n_j-1))$, and their amplitudes $A(\theta_1,\theta_2) = -A(\theta_1,-\theta_2) = -A(-\theta_1,\theta_2)$ are real (assuming two of the boundary lines to lie on the axes). Since $\mu(\theta_1,\theta_2) = \mu(\pm\theta_1,\pm\theta_2)$ is real, the r relaxation sweeps operate separately on each residual mode, transforming its amplitude $A(\theta)$ to $A'(\theta) = \mu(\theta)^r A(\theta)$.

For any component $\theta=(\theta_1,\theta_2)$ such that $|\theta| = \max(|\theta_1|,|\theta_2|) \le \pi/2$, denote $\theta^1 = (\theta_1,\theta_2)$, $\theta^2 = (\theta_1 \pm \pi, \theta_2)$, $\theta^3 = (\theta_1, \theta_2 \pm \pi)$, $\theta^4 = (\theta_1 \pm \pi, \theta_2 \pm \pi)$, where each $\pm$ sign is chosen so that $|\theta^\ell| < \pi, (\ell=1,2,3,4)$. Of these four "harmonics", only the $\theta^1$ mode appears on $G^{M-1}$, its amplitude there (in the right-hand side of the $G^{M-1}$ residual problem formed in Step (ii)) being

$$A_\theta = A'(\theta^1) + A'(\theta^2) + A'(\theta^3) + A'(\theta^4). \tag{C.1}$$

Let $\varepsilon_k$ denote an upper bound to the factors by which any $C^k$ cycle reduces the $L_2$ norm of the residuals on $G^k$. In particular, the two $C^{M-1}$ cycles (Step (iii)) are equivalent to solving a $G^{M-1}$ problem with amplitudes $a_\theta$ instead of $A_\theta$, where

(C.2) $$\sum_{|\theta| \le \frac{\pi}{2}} |a_\theta - A_\theta|^2 \le \varepsilon_{M-1}^4 \sum_{|\theta| \le \frac{\pi}{2}} A_\theta^2 .$$

Hence, interpolating the computed correction from $G^{M-1}$ to $G^M$ (Step (iv)), the new residual amplitudes are easily calculated to be

$$\bar{A}(\theta^\ell) = A'(\theta^\ell) - S(\theta^\ell)\, a_\theta$$

$$= \mu(\theta^\ell)^r A(\theta^\ell) - S(\theta^\ell) A_\theta + S(\theta^\ell)\,(A_\theta - a_\theta), \qquad (\ell=1,2,3,4,),$$

where

$$S(\theta) = \frac{(1+\cos\theta_1)(1+\cos\theta_2)(4-2\cos\theta_1-2\cos\theta_2)}{4-2\cos 2\theta_1 - 2\cos 2\theta_2} .$$

Hence

(C.3) $$\sum_\ell \bar{A}(\theta^\ell)^2 \le 2q^2 \sum_\ell A(\theta^\ell)^2 + 2 \sum_\ell S(\theta^\ell)^2 (A_\theta - a_\theta)^2 ,$$

where $q$ is any upper bound to the spectral radii of the $4\times$ matrices $Q(\theta)$, defined by

$$Q_{\ell m}(\theta) = (\delta_{\ell m} - S(\theta^\ell))\, \mu(\theta^m)^r , \qquad (1 \le \ell, m \le 4).$$

Denoting $\beta_j = 1 - \cos^2\theta_j$, it is easy to check that

(C.4) $$\sum_\ell S(\theta^\ell)^2 = \frac{1}{4}\left(1 + \frac{\beta_1^2 + \beta_2^2}{(\beta_1+\beta_2)^2} - \frac{\beta_1\beta_2}{\beta_1+\beta_2}\right) \le \frac{1}{2} .$$

Hence, summing (C.3) over the relevant range of $\theta$, using (C.2) and (C.4) and then (C.1), we obtain

$$\sum_{|\theta| \le \pi} \bar{A}(\theta)^2 \le 2q^2 \sum_{|\theta| \le \pi} A(\theta)^2 + \varepsilon_{M-1}^4 \sum_{|\theta| \le \frac{\pi}{2}} A_\theta^2$$

$$\le (2q^2 + \gamma\varepsilon_{M-1}^4) \sum_{|\theta| \le \pi} A(\theta)^2 ,$$

where $\gamma$ is any upper bound to all $\sum_\ell \mu(\theta^\ell)^{2r}$, $(0 < |\theta| \le \pi/2)$.

Thus, we have obtained the bound

(C.5) $$\varepsilon_M^2 = 2q^2 + \gamma\varepsilon_{M-1}^4 .$$

A simple computer program confirms the bounds $q^2 = (7/41)^r$ and $\gamma = 1 + 3(7/41)^{2r}$. Choose r=3. From (C.5) it now follows, by induction on M, that $\varepsilon_M \le .101$ .

The number of operations in the $C^M$ cycle is $W_M \leq (12r+3)n_1n_2 + 2W_{M-1}$. Hence, by induction on M, $W_M \leq (24r+6)n_1n_2$. We thus have in summary

Theorem. The above $C^M$ cycle reduces the $L_2$ error by a factor $\leq .101$ and costs 78 operations (additions and multiplications) per ($G^M$) grid point.

The theorem can be improved (to .1 reduction in only 53 operations-per-point) by defining the $C^M$ cycle to consist of r+M relaxation sweeps and only one $C^{M-1}$ cycle, and choosing large r. (Employing arbitrarily large r pays only with simultaneous-displacement schemes on rectangular domains, where there is no feed from low to high frequencies).

In practice, .1 reduction is obtained in about 26 operations. (See App. B. The Gauss-Seidel sweep employed there can be done in 5 operations-per-point. But for every 3 sweeps on $G^k$ the interpolations $I^k_{k-1}$ and $I^{k-1}_k$ are also performed, each costing an average of 6/4 operations per point. Hence, a work unit in App. B should be considered as representing $(3\times5+3)/3 = 6$ operations). These operations involve only additions and shifts.

# SINGULAR VALUE DECOMPOSITION: APPLICATIONS AND COMPUTATIONS*

Gene H. Golub
and
Franklin T. Luk
Stanford University, Stanford, California

ABSTRACT. The Singular Value Decomposition (SVD) of a rectangular matrix is described. Several problems arising in data analysis are given and their solution is given in terms of the SVD. Numerical methods are discussed for computing the decomposition for dense and sparse matrices.

1. INTRODUCTION. This paper is concerned with the singular value decomposition of a given matrix. The decomposition is very useful although it may not be as familiar as some of the other matrix decompositions. We shall describe the decomposition, give some specific examples of its applications, and suggest some methods to compute the decomposition.

There are many matrix decompositions that are useful in mathematical applications. A very familiar one is the QR decomposition of a square matrix $A$:

$$A = QR ,$$

where $Q$ is an orthogonal matrix and $R$ is an upper triangular matrix. There are several numerical schemes to compute this decomposition. We could use the Gram-Schmidt method; the columns of $Q$ are the orthogonal columns generated by the process. Another way to generate $Q$ and $R$ is through the use of Householder transformations.

Another familiar decomposition is the reduction of a square matrix to its Jordan canonical form:

$$A = XJX^{-1} ,$$

where $X$ is nonsingular and $J$ is a block diagonal matrix in which

*This work was in part supported by U.S. Army Research Grant DAHC04-75-G-0185.

each diagonal matrix is an elementary Jordan block $J_r(\lambda_i)$, viz.

$$J_r(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & & & \\ & \lambda_i & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda_i & 1 \\ & & & & \lambda_i \end{pmatrix} \quad r \times r$$

This decomposition has been used extensively in the study of stability of differential equations. Unfortunately, there does not appear to be any good numerical algorithm to compute the decomposition (Golub and Wilkinson [14]).

Finally, we shall discuss the singular value decomposition of an $m \times n$ matrix $A$:

$$A = U\Sigma V^t ,$$

where $U$ is an $m \times m$ orthogonal matrix, $V$ is an $n \times n$ orthogonal matrix, and $\Sigma$ is an $m \times n$ matrix with non-negative elements down the main diagonal and zeros everywhere else. For our discussion, we shall assume that $A$ has at least as many rows as columns so that $m \geq n$, although this is not always the case. There are many proofs of this decomposition, for instance, in the book by Forsythe and Moler [6]. A very clear and useful discussion is given in the book by Lanczos: "Linear Differential Operators" [17].

It is not very difficult to see that $U$ consists of the eigenvectors of $AA^t$, $V$ consists of the eigenvectors of $A^tA$ and the diagonal elements $\sigma_i$, $1 \leq i \leq n$, of $\Sigma$ are the non-negative square roots of the eigenvalues of $A^tA$. We assume the $\sigma_i$'s are arranged in such a way that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 ,$$

where $r$ is the rank of the matrix $A$.

The singular values and the eigenvalues of a given matrix can frequently differ. Consider an $m \times m$ matrix

$$A = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & & 0 \end{pmatrix} .$$

The matrix $A$ is of rank m-1 but all its eigenvalues equal 0. However, (m-1) singular values of $A$ equal 1 and only one singular value is zero. Hence the number of non-zero eigenvalues of a matrix gives a lower bound on its rank, whereas the number of non-zero singular values of a matrix is its rank.

2. APPLICATIONS. In this section we shall discuss some applications of the singular value decomposition (cf. Golub [8]).

A. Let $\mathcal{U}_m$ be the set of all $m \times m$ orthogonal matrices. We wish to replace a given $m \times m$ matrix $A$ by an $m \times m$ orthogonal matrix $Q$ that is near $A$. In order to study the nearness of one matrix with respect to another matrix, we introduce a norm; we use the Frobenius norm of a matrix, viz.,

$$\|A\| = \Big( \sum_{i,j} |a_{ij}|^2 \Big)^{1/2} .$$

We shall use this matrix norm throughout this discussion. Our problem then consists of the following: let $A$ be an arbitrary $m \times m$ matrix; determine $Q \in \mathcal{U}_m$ such that

$$\|A - Q\| \leq \|A - X\| \qquad \text{for any } X \in \mathcal{U}_m .$$

This problem is important in factor analysis and has also found applications in aeronautics (cf. Bar-Itzhack [1]).

The solution to the problem is fairly simple. It is as follows: if

$$A = U\Sigma V^t,$$

then we replace all the singular values by 1 and write

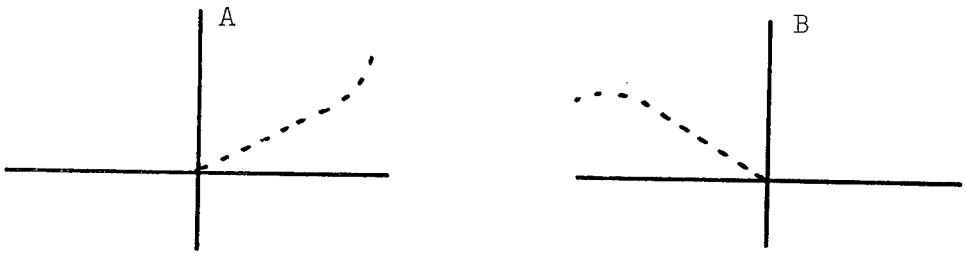
$$Q = UIV^t .$$

It is well-known that the singular values of an orthogonal matrix all equal 1. Now,

$$\begin{aligned} \|A - Q\| &= \|U\Sigma V^t - UIV^t\| \\ &= \|\Sigma - I\| \quad \text{since the Frobenius norm is unitarily invariant}^{1)} \\ &= [(\sigma_1-1)^2 + (\sigma_2-1)^2 + \cdots + (\sigma_n-1)^2]^{1/2} ; \end{aligned}$$

this value then is a measure of the departure from orthogonality of a given matrix. The result is true for all unitarily invariant norms (Fan and Hoffman [5]).

B. We consider the following important generalization of problem A. Let $A$ be an $m \times n$ matrix associated with a set of data and let $B$ be obtained from $A$ through a rotation of the data. The following figure may represent a typical situation:




1) A norm is said to be unitarily invariant if $\|AU\| = \|VA\| = \|A\|$ where $U^*U = I$ and $V^*V = I$.

Our idea is to replace $A$ by $BQ$, that is, we wish to replace $A$ by a rotation of $B$. We want to determine $Q \in \mathcal{U}_n$ such that

$$\|A - BQ\| = \min .$$

The solution is again given in terms of the singular value decomposition. Green [15] and Schönemann [21] showed that if

$$B^t A = U\Sigma V^t$$

and

$$Q = UV^t ,$$

then

$$\|A - BQ\| \leq \|A - BX\| \quad \text{for all} \quad X \in \mathcal{U}_n .$$

C. Let $\mathcal{M}_{m,n}^{(k)}$ be the set of all $m \times n$ matrices of rank $k$. Assume $A \in \mathcal{M}_{m,n}^{(r)}$. We want to determine $B \in \mathcal{M}_{m,n}^{(k)}$ $(k \leq r)$ such that

$$\|A - B\| \leq \|A - X\| \quad \text{for all} \quad X \in \mathcal{M}_{m,n}^{(k)} .$$

In other words, we want to approximate the matrix $A$ with a matrix of lower rank and we want the best approximation for the fixed rank. The solution is given in terms of the singular value decomposition.

Let $A = U\Sigma V^t$, then $B = U\Omega_k V^t$, where

$$\Omega_k = \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_k & \\ & & & & \end{pmatrix} \quad m \times n$$

Now

$$\|A - B\| = \|U\Sigma V^t - U\Omega_k V^t\|$$

$$= \|\Sigma - \Omega_k\|$$

$$= (\sigma_{k+1}^2 + \cdots + \sigma_n^2)^{1/2} .$$

Mirksy [18] showed that the above result is true for all unitarily invariant norms.

Consider the following example. Let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-10} \end{pmatrix} .$$

Mathematically, the matrix is of rank 2. But the following rank 1 matrix

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

differs from $A$ by only $10^{-10}$ and is the closest matrix of rank 1 to A.

D. The singular value decomposition also enters in the computation of the pseudo-inverse of a matrix. An $n \times m$ matrix $X$ is a pseudo-inverse of an $m \times n$ matrix $A$ if it satisfies the following four relations:

(i) $AXA = A$,

(ii) $XAX = X$,

(iii) $(AX)^t = AX$,

(iv) $(XA)^t = XA$ .

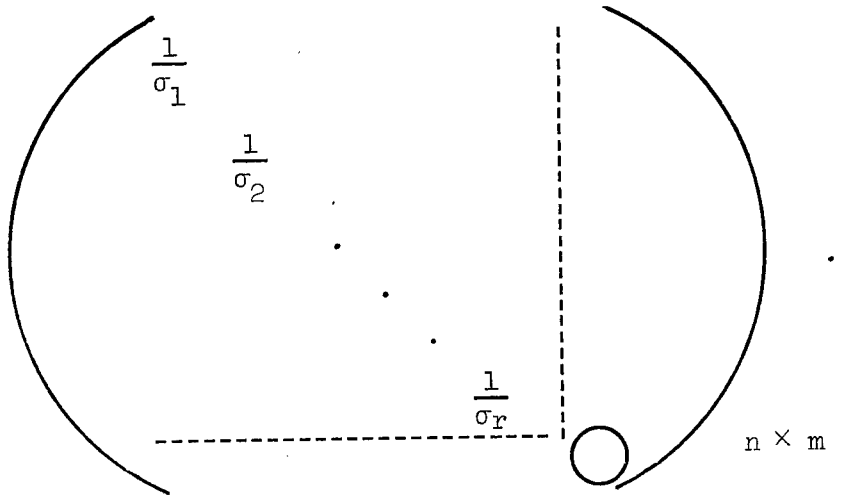
The pseudo-inverse $X$ is unique and we denote it by $A^+$. We can easily verify that given

$$A = U\Sigma V^t ,$$

we always have

$$A^+ = V\Lambda U^t ,$$

where

$\Lambda =$ 


If $A$ is square and of full rank, then $A^+ = A^{-1}$.

Consider the following problem. Suppose we have an m-vector $\underset{\sim}{b}$ and an $m \times n$ matrix $A$. We would like to determine an n-vector $\underset{\sim}{x}$ such that

$$\|A\underset{\sim}{x} - \underset{\sim}{b}\|_2 = \min.^{2)}$$

If $A$ is not a matrix of full rank, we do not have a unique solution to the problem. Let

$$\chi = \{\underset{\sim}{x} \mid \|A\underset{\sim}{x} - \underset{\sim}{b}\|_2 = \min\} .$$

We would like to determine $\hat{\underset{\sim}{x}} \in \chi$ such that $\|\hat{\underset{\sim}{x}}\|_2$ is a minimum.

---

2) $\|\underset{\sim}{y}\|_2 = (\sum_{i=1}^{n} y_i^2)^{1/2}$ for $\underset{\sim}{y} = (y_1, y_2, \ldots, y_n)^t$ .

The solution is given by $\hat{\underset{\sim}{x}} = A^{+}\underset{\sim}{b}$. Hence if we had $A^{+}$, it would be fairly simple to compute a sequence of solutions $\{\hat{\underset{\sim}{x}}_j\}$ given the sequence of data $\{\underset{\sim}{b}_j\}$.

Unfortunately, the pseudo-inverse of a matrix is not a continuous function of the elements of $A$. If we let

$$A(\epsilon) = \begin{pmatrix} 1 & 0 \\ 0 & \epsilon \end{pmatrix},$$

where $\epsilon > 0$, then

$$A^{+}(\epsilon) = \begin{pmatrix} 1 & 0 \\ 0 & \epsilon^{-1} \end{pmatrix}$$

But

$$A^{+}(0) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence for a small positive $\epsilon$, we see that $A^{+}(\epsilon)$ is quite different from $A^{+}(0)$. Thus the computation of the pseudo-inverse is quite an ill-conditioned problem.

If we want to compute the pseudo-inverse in a stable way, we must impose some additional conditions. We shall give one possibility which seems quite satisfactory.

Suppose we are given a matrix $A$ but we also know that the matrix is really some matrix $B$ plus some perturbation $\Delta$, viz.,

$$A = B + \Delta .$$

We do not know $B$ but we know some bound on the error:

$$\|\Delta\| \leq \eta ;$$

for example, this would happen if the elements of $A$ were empirical data with known uncertainties. We wish to determine $\hat{B}$ such that

$$\|A - \hat{B}\| \leq \eta ,$$

and

$$\text{rank}(\hat{B}) = \min .$$

The solution is given by the singular value decomposition. If we write

$$B_k = U\Omega_k V^t ,$$

then

$$\hat{B} = B_p$$

if

$$\sigma_{p+1}^2 + \cdots + \sigma_r^2 \leq \eta^2 ,$$

and

$$\sigma_p^2 + \sigma_{p+1}^2 + \cdots + \sigma_r^2 > \eta^2 .$$

Note that although

$$\|A - \hat{B}\| \leq \eta ,$$

yet

$$\|A^+ - \hat{B}^+\| = \left( \frac{1}{\sigma_{p+1}^2} + \cdots + \frac{1}{\sigma_r^2} \right)^{1/2} .$$

E. We may use the singular value decomposition to solve homogeneous equations. Suppose $A$ is an $m \times n$ matrix of rank $r$. Let

$$AV = U\Sigma .$$

We partition $V$ into an $n \times r$ matrix $V_1$ and an $n \times (n-r)$ matrix $V_2$, i.e.

$$V = (V_1, V_2) ,$$

and

$$A(V_1, V_2) = (U\Sigma_1, 0) ,$$

where

$$\Sigma_1 = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \\ \hline & \bigcirc & \end{pmatrix}_{m \times r} .$$

Then

$$AV_2 = \bigcirc ,$$

and we have found an orthogonal basis for the null space of $A$.

Given a set of eigenvalues of a square matrix, we need to solve a set of homogeneous equations in order to find the eigenvectors. Golub and Wilkinson [14] used the idea to compute the Jordan canonical form of a matrix.

Often, we wish to know which columns of a given matrix $A$ are linearly independent. If $A$ is a set of measurements and if some columns are dependent, we may want to determine which are the dependent columns, eliminate them and obtain a linearly independent set of measurements. The singular value decomposition can be very effective for this purpose.

Let $A \in \mathcal{R}_{m,n}^{(n-1)}$ and let the last column of $A$ consist of all zeros. We find

$$V_2 = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} ,$$

from which we see we should eliminate the last column of $A$.

In general, we want to take $V_2$ and perform Gaussian elimination with complete pivoting on $V_2^t$ such that

$$PV_2^t\Pi = (\nabla \square) ,$$

where

$\nabla$ is an $(n-r) \times (n-r)$ upper triangular matrix

$\square$ is an $(n-r) \times r$ matrix, and

$\Pi$ is an $n \times n$ permutation matrix.

Then if

$$A\Pi = (A_1, A_2) ,$$

we can decide that the columns of $A_2$ form a linearly independent basis for the columns of A. This and other problems of dependence are discussed extensively in a paper by Golub, Klema and Stewart [10].

F. Another problem is the following. Consider

$$\max_{p,q \neq 0} \frac{p^t A q}{\|p\|_2 \, \|q\|_2} .$$

It is not difficult to see that the maximal value of the normalized bilinear form is $\sigma_1$, which is attained when $p = u_1$ and $q = v_1$, where $\sigma_1$ is the largest singular value of A, and $u_1$, $v_1$ are the corresponding left and right singular vectors, respectively.

Let X be an $m \times s$ matrix and Y be an $m \times t$ matrix. Consider

$$\xi = Xu \quad \text{and} \quad \eta = Yv .$$

The angle $\theta$ between $\xi$ and $\eta$ is given by

$$\cos \theta = \frac{\xi^t \eta}{\|\xi\|_2 \, \|\eta\|_2} .$$

We can choose $\xi$ and $\eta$ to maximize the normalized inner product. We call the maximal value the canonical correlation and the corresponding angle (say $\hat{\theta}$) the angle between the two subspaces U and V.

We can determine $\hat{\theta}$ very easily using the singular value deomposition. We compute the QR decomposition of X and Y, viz.

$$X = QR \qquad \text{and} \qquad Y = PS \ .$$

Then

$$\hat{\theta} = \cos^{-1}(\sigma_{max}(Q^tP)) \ .$$

The computation can be carried out even when $X$ and $Y$ have less than full rank (Björck and Golub [2]).

G. One further application of the singular value decomposition is in computing the parameter $\lambda$ in ridge regression using the cross validation technique (Golub, Wahba and Heath [13]).

Given an $m \times n$ matrix $K$ of rank $r$ and an $m$-vector $g$. We wish to minimize

$$\varphi(\underset{\sim}{f}) = \|g - K\underset{\sim}{f}\|_2^2 + \lambda\|\underset{\sim}{f}\|_2^2 \ .$$

Using the variational technique, we see $\varphi(\underset{\sim}{f})$ attains its minimum at $\underset{\sim}{f} = \hat{\underset{\sim}{f}}$ where $\hat{\underset{\sim}{f}}$ satisfies

$$(K^tK + \lambda I)\hat{\underset{\sim}{f}} = K^tg \ .$$

Hence we have a ridge regression problem. The question is how to choose $\lambda$. One possibility is to try to estimate $\lambda$ from the data; we shall describe one method based on cross validation. We shall see how the singular value decomposition of $K$ aids us in both choosing $\lambda$ and solving for $\hat{\underset{\sim}{f}}$ for the chosen value of $\lambda$.

Let $K^{(j)}$ denote the $(m-1) \times n$ matrix obtained by leaving out the $j$-th row of $K$, and let $g^{(j)}$ denote an $(m-1)$-vector obtained by leaving out the $j$-th component of $g$, viz.

$$K^{(j)} = \begin{pmatrix} \underset{\sim}{k}_1^t \\ \vdots \\ \underset{\sim}{k}_{j-1}^t \\ \underset{\sim}{k}_{j+1}^t \\ \vdots \\ \underset{\sim}{k}_m^t \end{pmatrix}$$

and

$$\underset{\sim}{g}^{(j)} = \begin{pmatrix} g_1 \\ \vdots \\ g_{j-1} \\ g_{j+1} \\ \vdots \\ g_m \end{pmatrix} ,$$

where $\underset{\sim}{k}_j^t$ is the j-th row of K.

Let $\hat{\underset{\sim}{f}}^{(j)}(\lambda)$ denote the solution to

$$(K^{(j)t}K^{(j)} + \lambda I)\, \hat{\underset{\sim}{f}}^{(j)}(\lambda) = K^{(j)t}\underset{\sim}{g}^{(j)} .$$

The <u>cross-validation weighted square error</u> $CV(\lambda)$ is defined by

$$CV(\lambda) = \sum_{j=1}^{m} w_j [g_j - \underset{\sim}{k}_j^t \hat{\underset{\sim}{f}}^{(j)}(\lambda)]^2 ,$$

where

$$w_j \geq 0.$$

We wish to choose $\lambda$ such that $CV(\lambda)$ is a minimum. We see

$$CV(\lambda) = \sum_{j=1}^{m} w_j [g_j - \underset{\sim}{k}_j^t (K^{(j)t}K^{(j)} + \lambda I)^{-1} K^{(j)t} \underset{\sim}{g}^{(j)}]^2$$

$$= \sum_{j=1}^{m} w_j [g_j - \underset{\sim}{k}_j^t (K^tK + \lambda I - \underset{\sim}{k}_j\underset{\sim}{k}_j^t)^{-1} (K^t - \underset{\sim}{k}_j\underset{\sim}{e}_j^t)\underset{\sim}{g}]^2 ,$$

where $\underset{\sim}{e}_j = (\underbrace{0,\ldots,0,1,0,\ldots,0}_{j})^t$.

We apply the Sherman-Morrison formula to obtain

$$(K^tK + \lambda I - \underset{\sim}{k}_j\underset{\sim}{k}_j^t)^{-1} = (K^tK + \lambda I)^{-1} + \alpha_j^{-1}(K^tK + \lambda I)^{-1}\underset{\sim}{k}_j\underset{\sim}{k}_j^t(K^tK + \lambda I)^{-1},$$

where

$$\alpha_j = 1 - \underset{\sim}{k}_j^t (K^t K + \lambda I)^{-1} \underset{\sim}{k}_j$$

and

$$\alpha_j \neq 0 \quad \text{by assumption .}$$

After some additional computations, we get

$$CV(\lambda) = \|B(\lambda)[I - K(K^t K + \lambda I)^{-1} K^t] g\|^2 ,$$

where $B(\lambda)$ is an $m \times m$ matrix given by

$$\{B(\lambda)\}_{ii} = w_i^{1/2} [1 - \underset{\sim}{k}_i^t (K^t K + \lambda I)^{-1} \underset{\sim}{k}_i]^{-1} ,$$

and $\{B(\lambda)\}_{ij} = 0 \qquad \text{for } i \neq j .$

We factorize $K$ as

$$K = U \Sigma V^t ,$$

i.e., the singular value decomposition of $K$. Then

$$CV(\lambda) = \|B(\lambda)[g - U\Sigma(\Sigma^t \Sigma + \lambda I)^{-1} \Sigma^t \hat{g}]\|^2 ,$$

where $\hat{g} = U^t g$. Now,

$$\{B(\lambda)\}_{ii} = w_i^{1/2} [1 - \underset{\sim}{k}_i^t V(\Sigma^t \Sigma + \lambda I)^{-1} V^t \underset{\sim}{k}_i]^{-1} .$$

But since

$$KV = U\Sigma ,$$

we obtain

$$\{B(\lambda)\}_{ii} = w_i^{1/2} (1 - \sum_{j=1}^{r} u_{ij}^2 \varphi_j(\lambda))^{-1} ,$$

where

$$\varphi_j(\lambda) = \frac{\sigma_j^2}{\sigma_j^2 + \lambda},$$

and $\sigma_j$'s are the singular values of $K$. Finally,

$$CV(\lambda) = \sum_{i=1}^{m} w_i \left[ \frac{g_i - \sum_{j=1}^{r} u_{ij}\varphi_j(\lambda)\hat{g}_j}{1 - \sum_{j=1}^{r} u_{ij}^2 \varphi_j(\lambda)} \right]^2,$$

which is very easy to evaluate.

For a chosen value of $\lambda$, we may solve the ridge regression problem easily using the singular value decomposition of $K$. We have

$$(K^tK + \lambda I)\hat{\underset{\sim}{f}} = K^t\underset{\sim}{g},$$

which reduces to

$$V(\Sigma^t\Sigma + \lambda I)V^t\hat{\underset{\sim}{f}} = V\Sigma^t\hat{\underset{\sim}{g}}.$$

Hence

$$\hat{\underset{\sim}{f}} = V(\Sigma^t\Sigma + \lambda I)^{-1}\Sigma^t\hat{\underset{\sim}{g}}$$

$$= \sum_{j=1}^{r} \frac{\sigma_j \hat{g}_j}{\sigma_j^2 + \lambda} \underset{\sim}{v}_j,$$

where

$$V = (\underset{\sim}{v}_1, \underset{\sim}{v}_2, \ldots, \underset{\sim}{v}_n).$$

Many numerical experiments have been carried out in [13].

## 3. COMPUTING THE SINGULAR VALUE DECOMPOSITION OF A DENSE MATRIX

Our basic tool is the Householder transformation. Consider a matrix $P^{(1)}$ of the form

$$P^{(1)} = I - 2\underset{\sim}{u}^{(1)}\underset{\sim}{u}^{(1)t},$$

where

$$\|\underset{\sim}{u}^{(1)}\|_2 = 1.$$

Note that the matrix $P^{(1)}$ is symmetric and orthogonal. Let $A^{(1)}$ denote the original matrix. We construct $P^{(1)}$ to annihilate all elements below the diagonal in the first column of $A^{(1)}$:

$$P^{(1)}A^{(1)} = \begin{pmatrix} \alpha_1 & a'_{12} & \cdots & a'_{1n} \\ 0 & a'_{22} & \cdots & a'_{2n} \\ 0 & a'_{32} & \cdots & a'_{3n} \\ \vdots & \vdots & & \vdots \\ 0 & a'_{m2} & & a'_{mn} \end{pmatrix} \equiv A^{(3/2)} .$$

We next apply a Householder transformation $Q^{(1)}$ on the right of $A^{(3/2)}$, and our idea is to eliminate all elements to the right of the (1,2) position in the first row of $A^{(3/2)}$ without disturbing the zero elements in the first column:

$$A^{(3/2)}Q^{(1)} = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ 0 & a''_{22} & a''_{23} & \cdots & a''_{2n} \\ 0 & a''_{32} & a''_{33} & \cdots & a''_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a''_{m2} & a''_{m3} & & a''_{mn} \end{pmatrix} \equiv A^{(2)} .$$

Our process continues with

$$A^{(k+1/2)} = P^{(k)} A^{(k)} ,$$

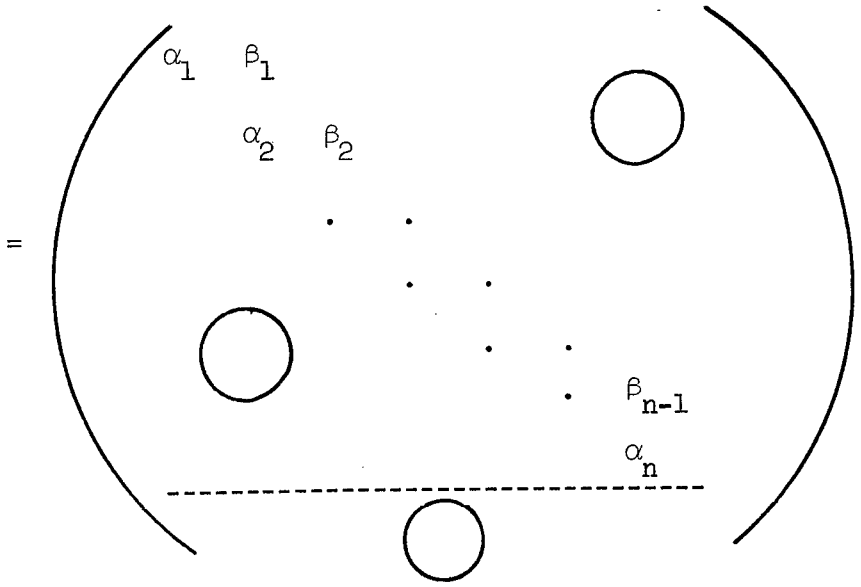
where the effect of $P^{(k)}$ is to eliminate all elements below the diagonal in the k-th column of $A^{(k)}$, and with

$$A^{(k+1)} = A^{(k+1/2)} Q^{(k)} ,$$

where the effect of $Q^{(k)}$ is to eliminate all elements to the right of the $(k,k+1)$ position in the k-th row of $A^{(k+1/2)}$.

The end result is that we have $n$ transformations on the left ($(n-1)$ transformations if $m = n$), and $(n-2)$ transformations on the right of $A$:

$$J = P^{(n)} \cdots P^{(1)} A Q^{(1)} \cdots Q^{(n-2)}$$




We now apply the QR method due to Francis [7] and Kublanovskaya [16] (Golub and Kahan [9]) so that

$$J = X \Sigma Y^t ,$$

i.e., the singular value decomposition of $J$. If we write

$$P = P^{(1)} \cdots P^{(n)} \qquad \text{and} \qquad Q = Q^{(1)} \cdots Q^{(n-2)} ,$$

then

$$A = PJQ^t$$
$$= (PX)\ \Sigma(QY)^t$$
$$= U\Sigma V^t ,$$

where

$$U = PX, \qquad V = QY .$$

The first program to do the above computations was given by Golub and Reinsch [12]. A version for complex matrices was given by Businger and Golub [3] A program for real matrices is available in Release 2 of EISPACK [4].

4. COMPUTING THE SINGULAR VALUE DECOMPOSITION OF LARGE SPARSE MATRICES. We have several possibilities for computing the singular value decomposition of a large and sparse matrix. In most problems, we want only the few greatest singular values of a large matrix; for instance, in image reconstruction, the order of the matrix frequently exceeds 10,000 but only very few, generally less than 100, of the greatest singular values are of physical significance.

A. Standard Lanczos algorithm. The best available algorithm for computing a few of the greatest singular values of a large sparse matrix, say A, is the Lanczos algorithm. The algorithm uses the matrix A only in the computation of the matrix-vector product $A\underset{\sim}{x}$ or $A^t\underset{\sim}{x}$ given a vector $\underset{\sim}{x}$. Hence we can use the sparsity of A to compute the products very efficiently. Unlike other methods that transform the matrix, the Lanczos algorithm preserves the matrix's sparse structure and works well even if the matrix is so large that it has to be stored on some auxiliary device (e.g. magnetic disk or tape).

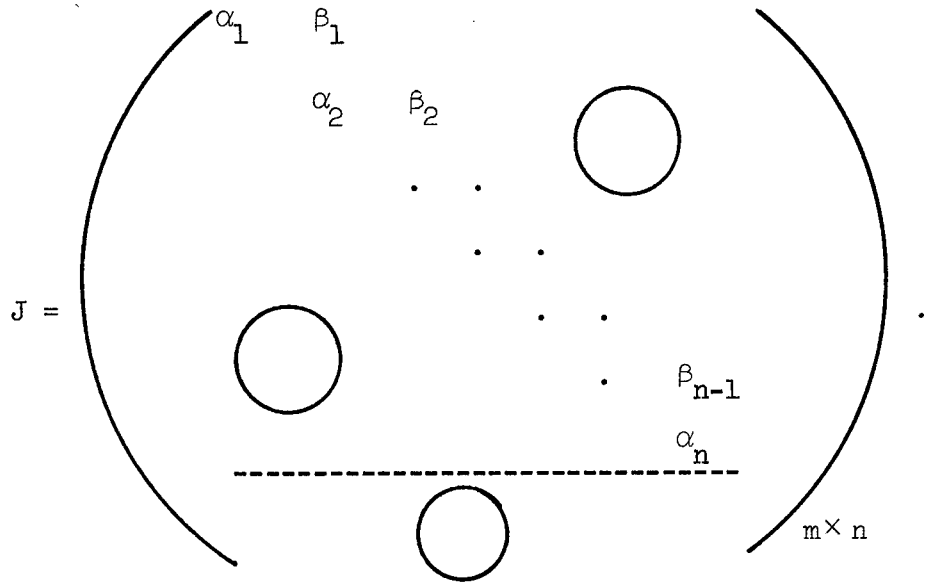
We use the Lanczos algorithm to bidiagonalize a given $m \times n$ matrix A:

$$A = PJQ^t ,$$

where

$$P^tP = I_m , \qquad Q^tQ = I_n ,$$

and




We can expand the two resultant equations:

$$AQ = PJ \qquad \text{and} \qquad P^tA = JQ^t ,$$

in terms of the columns $\underset{\sim}{p}_i$ of P and $\underset{\sim}{q}_i$ of Q to yield

$$\left.\begin{aligned} A\underset{\sim}{q}_1 &= \alpha_1\underset{\sim}{p}_1 , \\ A\underset{\sim}{q}_{i+1} &= \beta_i\underset{\sim}{p}_i + \alpha_{i+1}\underset{\sim}{p}_{i+1} , \\ \underset{\sim}{p}_i^tA &= \alpha_i\underset{\sim}{q}_i^t + \beta_i\underset{\sim}{q}_{i+1}^t , \\ \underset{\sim}{p}_n^tA &= \alpha_n\underset{\sim}{q}_n^t . \end{aligned}\right\} \quad i = 1,2,\ldots,n-1 ,$$

So our algorithm is

(1) Choose $\underset{\sim}{q}_1$ such that $\|\underset{\sim}{q}_1\|_2 = 1$.

Set

$$\underset{\sim}{w}_1 = A\underset{\sim}{q}_1 ,$$

$$\alpha_1 = \|\underset{\sim}{w}_1\|_2 ,$$

$$\underset{\sim}{p}_1 = \alpha_1^{-1}\underset{\sim}{w}_1 .$$

(2) For $i = 1,2,\ldots,s-1$ $(2 \leq s \leq n)$, compute

$$z_i = A^t p_i - \alpha_i q_i ,$$

$$\beta_i = \|z_i\|_2 ,$$

$$q_{i+1} = \beta_i^{-1} z_i ,$$

$$w_{i+1} = A q_{i+1} - \beta_i p_i ,$$

$$\alpha_{i+1} = \|w_{i+1}\|_2 ,$$

$$p_{i+1} = \alpha_{i+1}^{-1} w_{i+1} .$$

For some $s \leq n$, we denote

$$J^{(s)} = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ & \alpha_2 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \alpha_{s-1} & \beta_{s-1} \\ & & & & \alpha_s \end{pmatrix} ,$$

$$P^{(s)} = (p_1, p_2, \cdots, p_s) ,$$

and

$$Q^{(s)} = (q_1, q_2, \cdots, q_s) .$$

We now apply the QR method on $J^{(s)}$ so that

$$J^{(s)} = X^{(s)} \Sigma^{(s)} Y^{(s)t}$$

i.e., the singular value decomposition of $J^{(s)}$. Let

$$\Sigma^{(s)} = \begin{pmatrix} \sigma_1^{(s)} & & & \\ & \sigma_2^{(s)} & & \\ & & \ddots & \\ & & & \sigma_s^{(s)} \end{pmatrix}$$

where $\sigma_1^{(s)} \geq \sigma_2^{(s)} \geq \cdots \geq \sigma_s^{(s)} \geq 0$,

$$X^{(s)} = (\underset{\sim}{x}_1^{(s)}, \underset{\sim}{x}_2^{(s)}, \ldots, \underset{\sim}{x}_s^{(s)}),$$

and

$$Y^{(s)} = (\underset{\sim}{y}_1^{(s)}, \underset{\sim}{y}_2^{(s)}, \ldots, \underset{\sim}{y}_s^{(s)}) .$$

The $\sigma_1^{(s)}$, $P^{(s)}\underset{\sim}{x}_1^{(s)}$, and $Q^{(s)}\underset{\sim}{y}_1^{(s)}$ are usually accurate approximations to the largest singular value, and the corresponding left and right singular vectors, respectively, of A. We may apply the Kaniel-Paige theory [19] to show that if $\theta > 0$ is the angle between $\underset{\sim}{q}_1$ and $\underset{\sim}{y}_1$, then

$$\sigma_1 - \epsilon_1^2 \leq \sigma_1^{(s)} \leq \sigma_1 ,$$

where

$$\epsilon_1^2 = \frac{2\sigma_1 \tan^2 \theta}{T_{s-1}^2(\frac{1+\gamma}{1-\gamma})} ,$$

$T_{s-1}$ is the (s-1)-st Chebyshev polynomial of the first kind, and

$$\gamma = \frac{\sigma_1 - \sigma_2}{2\sigma_1} .$$

We construct an example to show how $\sigma_1^{(s)}$ generally approximates $\sigma_1$ well even for a small s. Let $\sigma_1 = 1.0$, $\sigma_2 = 0.9$, $s = 20$ and $\theta = \cos^{-1} 0.1$. Then

$$\tan^2 \theta = \frac{1 - 0.1^2}{0.1^2} = 99 ,$$

$$\gamma = \frac{1.0 - 0.9}{2(1.0)} = 0.05 ,$$

$$\frac{1 + \gamma}{1 - \gamma} \doteq 1.105$$

and

$$T_{19}(1.105) \doteq 2.8 \times 10^3 .$$

Hence

$$\epsilon_1^2 \doteq \frac{2 \cdot 1 \cdot 99}{(2.8 \times 10^3)^2} \doteq 2.5 \times 10^{-5}$$

and

$$\sigma_1 - 0.000025 \leq \sigma_1^{(20)} \leq \sigma_1 .$$

Since $n$ is usually very large, we often choose some $s \ll n$ subject to storage availability. If our convergence criterion for the singular value is not satisfied, we may use $Q^{(s)} y_1^{(s)}$ as the new initial vector and restart the Lanczos algorithm. Since the accuracy of our approximation is bounded by $\tan \theta$, where $\theta$ is the angle between our initial vector and $y_1$, we expect to obtain better approximations if we iterate the Lanczos algorithm. If $z_i$ (or $w_i$) $= 0$ for some $i \leq s$, we could continue the algorithm by choosing some $z_i$ (or $w_i$) orthogonal to all the previous $z_j$'s (or $w_j$'s), $j < i$. We could also choose to terminate the algorithm because $z_i$ (or $w_i$) $= 0$ usually means some singular values have converged.

The sequences of vectors $\{p_i\}$ and $\{q_i\}$ form orthogonal sets in exact arithmetic. Hence theoretically, we need only to keep the most recent pairs of $p_i$'s and $q_i$'s in memory, providing great savings in storage. Unfortunately, the sequences $\{p_i\}$ and $\{q_i\}$ generally lose orthogonality very quickly due to cancellation errors in the computations of the $z_i$'s and $w_i$'s. A remedy is to reorthogonalize the most recently computed $p_i$ (or $q_i$) with respect to all

the previous $\underset{\sim}{p}_j$'s (or $\underset{\sim}{q}_j$'s), $j < i$. But this task is expensive in both execution time and storage, because we must now store all the computed $\{\underset{\sim}{p}_i\}$ and $\{\underset{\sim}{q}_i\}$ in memory. Paige [19] argues against the necessity of reorthogonalization, but the matter is still a subject of controversy.

B. Block Lanczos algorithm. In many cases we may save work if we iterate with a block of vectors instead of a single vector. The saving could be considerable if we were computing a multiple singular value. In general, if we had some a priori knowledge of the singular value spectrum, we could choose an appropriate block size with good gains. Computer experiments (Golub, Luk and Overton [11]) show that if we want several of the largest singular values, we often gain by choosing a block size $p > 1$. Also, if the matrix is stored on an auxiliary device, we may make some gains in efficiency if we multiply the matrix into several vectors simultaneously.

In a similar way to the standard Lanczos algorithm, our block version reduces the matrix $A$ to a block bidiagonal form. We start with an arbitrary $n \times p$ matrix $Q_1$, and perform a QR factorization of the product $AQ_1$:

$$P_1 A_1 = AQ_1 ,$$

where $P_1$ is an $m \times p$ matrix such that $P_1^t P_1 = I$ ,

and $A_1$ is a $p \times p$ upper triangular matrix.

Our algorithm continues with

$$\left.\begin{aligned} Q_i B_{i-1} &= A^t P_{i-1} - Q_{i-1} A_{i-1}^t , \\ \text{and} \qquad P_i A_i &= AQ_i - P_{i-1} B_{i-1}^t , \end{aligned}\right\} \quad i = 2,3,\ldots,s ,$$

where $Q_i B_{i-1}$ and $P_i A_i$ are the QR factorizations of the respective right-hand sides, and

$Q_i$ is an $n \times p$ matrix such that $Q_i^t Q_i = I$,

$P_i$ is an $m \times p$ matrix such that $P_i^t P_i = I$,

and both $B_{i-1}$ and $A_i$ are $p \times p$ upper triangular matrices.

We have tacitly assumed $p \times s \leq n$. We consider the $ps \times ps$ block tridiagonal matrix $J^{(s)}$:

$$J^{(s)} = \begin{pmatrix} A_1 & B_1^t & & & \\ & A_2 & B_2^t & & \\ & & \ddots & \ddots & \\ & & & A_{s-1} & B_{s-1}^t \\ & & & & A_s \end{pmatrix},$$

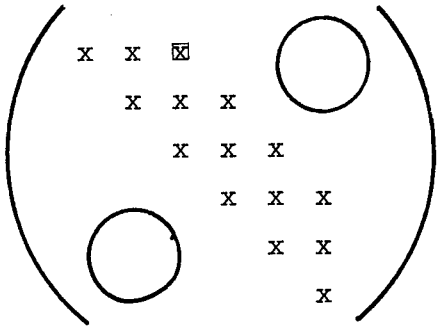
which is also banded upper triangular with bandwidth = p+1.

We can reduce $J^{(s)}$ to bidiagonal form using the Householder transformations. We can also use plane rotations to reduce $J^{(s)}$ to bidiagonal form to take advantage of the sparse banded structure of $J^{(s)}$. A plane rotation in the (i,j)-plane is an orthogonal matrix $P_{ij}$ of the form

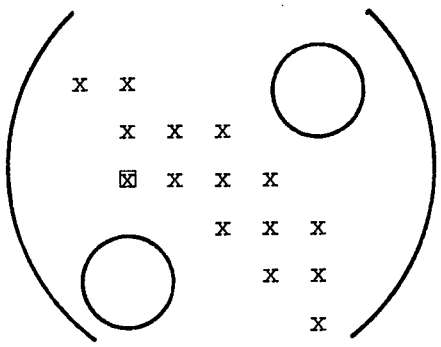
$$P_{ij} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & \gamma & \cdots & \sigma & & & & \\ & & \vdots & 1 & \vdots & & & & \\ & & -\sigma & \cdots & \gamma & & & & \\ & & & & & 1 & & & \\ & & & & & & \ddots & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ i \\ \\ j \\ \\ \\ \\ \end{matrix},$$

where $\gamma^2 + \sigma^2 = 1$. It is easy to verify that given a vector $\underset{\sim}{x}$ we can choose $\gamma$ and $\sigma$ such that $P_{ij}$ annihilates the j-th component of $\underset{\sim}{x}$. We give a simple example to demonstrate how we can reduce $J^{(s)}$ using plane rotations.

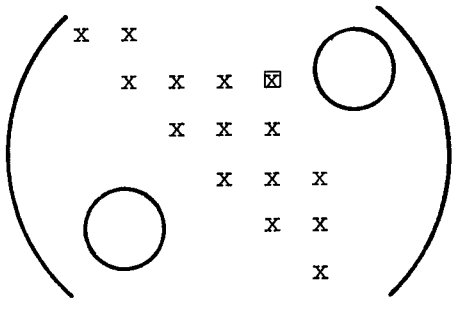
Suppose we have the following $6 \times 6$ matrix

$\hat{A} =$ 

We construct a plane rotation $Q_{23}$, postmultiplying $\hat{A}$ to annihilate the (1,3) element. The rotation creates a non-zero element in the (3,2) position, i.e.,

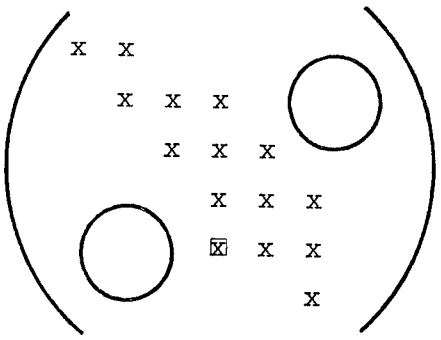
$\hat{A}Q_{23} =$ 

Now we apply a plane rotation $P_{23}$, premultiplying $\hat{A}Q_{23}$ to eliminate the (3,2) element. A new non-zero element appears in the (2,5) position:

$P_{23}\hat{A}Q_{23} =$ 

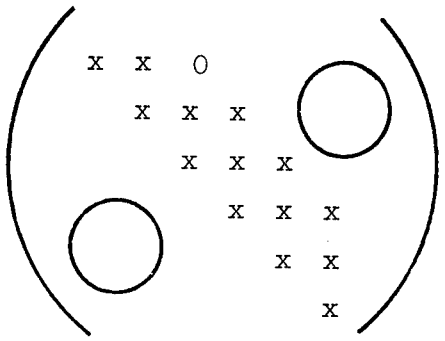
We construct $Q_{45}$ to annihilate the new nonzero (2,5) element from the right:

$$P_{23}\hat{A}Q_{23}Q_{45} = \quad$$



An appropriate plane rotation $P_{45}$ from the left will annihilate the newly created (5,4) element without creating new nonzero elements:

$$P_{45}P_{23}\hat{A}Q_{23}Q_{45} = \quad$$



We say we have "chased" away the (1,3) element of $\hat{A}$ (cf. Rutishauser [20]).

We may determine the singular value decomposition of the resultant bidiagonal matrix using the QR method. Using a theorem due to Underwood [22], we can show that the $p$ largest singular values of $J^{(s)}$ are usually accurate approximations to the $p$ largest singular values of $A$. In fact, if $\sigma_{min} > 0$ is the smallest singular value of $Q_1^t V_1$, where $V_1$ consists of the first $p$ columns of $V$, then for $k = 1,2,\ldots,p$,

$$\sigma_k - \epsilon_k^2 \leq \sigma_k^{(s)} \leq \sigma_k ,$$

where

$$\epsilon_k^2 = (\sigma_1 + \sigma_k) \frac{\tan^2 \theta}{T_{s-1}^2\left(\frac{1+\gamma_k}{1-\gamma_k}\right)} ,$$

$$\theta = \cos^{-1} \sigma_{min} ,$$

$$\gamma_k = \frac{\sigma_k - \sigma_{p+1}}{\sigma_k + \sigma_1} ,$$

and $T_{s-1}$ is the (s-1)-st Chebyshev polynomial of the first kind.

We consider an example similar to the one we have given in the previous section. Let $\sigma_1 = 1.0$, $\sigma_2 = 0.9$, $\sigma_3 = 0.5$, $p = 2$, $s = 10$ and $\theta = \cos^{-1} 0.1$. Then

$$\tan^2 \theta = 99,$$

$$\gamma_1 = \frac{\sigma_1 - \sigma_3}{\sigma_1 + \sigma_1} = \frac{0.5}{2.0} = 0.25,$$

$$\gamma_2 = \frac{\sigma_2 - \sigma_3}{\sigma_2 + \sigma_1} = \frac{0.4}{1.9} \doteq 0.21 ,$$

$$\frac{1 + \gamma_1}{1 - \gamma_1} = \frac{1.25}{0.75} \doteq 1.67 ,$$

$$\frac{1 + \gamma_2}{1 - \gamma_2} \doteq \frac{1.21}{0.79} \doteq 1.53 ,$$

$$T_9(1.67) \doteq 10^4 ,$$

$$T_9(1.53) \doteq 3.7 \times 10^3 .$$

Hence

$$\epsilon_1^2 \doteq \frac{2 \times 99}{10^8} \doteq 2.0 \times 10^{-6} ,$$

and

$$\epsilon_2^2 \doteq \frac{1.9 \times 99}{(3.7 \times 10^3)^2} \doteq 1.4 \times 10^{-5} .$$

Comparing the two examples, we can see how a proper choice of the block size would save us work with the same limitation on

storage space. In general, a good choice of $p$ depends on the singular value spectrum, the number of singular values desired, and the availability of memory. If there is a cluster of $\hat{p}$ largest singular values, it usually pays to choose $p = \hat{p}$. Often, the knowledge is not available and a satisfactory rule appears to be choosing $p$ equal to the number of singular values we want to compute. Our tests [11] show that the reorthogonalization of each recently computed $P_i$ $(Q_i)$ with respect to all the previous $P_j$'s $(Q_j$'s), $j < i$, is necessary for accurate results. We therefore must keep all the $P_i$'s and $Q_i$'s in memory, effectively bounding the value $p \times s$.

An algorithm will soon be published [11].

REFERENCES


[1] I. Y. Bar-Itzhack, "Iterative optimal orthogonalization of the strapdown matrix," IEEE Trans. Aerospace and Electronic Systems 11 (1975), 30-37.

[2] A. Björck and G. Golub, "Numerical methods for computing angles between linear subspaces," Math. Comp. 27 (1973), 579-594.

[3] P. Businger and G. Golub, Algorithm 358, Singular value decomposition of a complex matrix, ACM 12 (1969), 564-565.

[4] EISPACK Release 2, Argonne National Laboratory, Argonne, Illinois, 1976.

[5] K. Fan and A. Hoffman, "Some metric inequalities in the space of matrices," Proc. Amer. Math. Soc. 6 (1955), 111-116.

[6] G. Forsythe and C.B. Moler, "Computer Solution of Linear Algebraic Systems," Prentice-Hall, Englewood Cliffs, N.J., 1967.

[7] J.G.F. Francis, "The QR transformation: a unitary analogue to the LR transformation. I, II," Computer J. 4 (1961/1962), 265-271, 332-345.

[8] G. Golub, "Least squares, singular values and matrix approximations," Aplikace Mathematiky 13 (1968), 44-51.

[9] G. Golub and W. Kahan, "Calculating the singular values and pseudo-inverse of a matrix," J. SIAM Numer. Anal. 2 (1965), 205-224.

[10] G. Golub, V. Klema and G.W. Stewart, "Rank degeneracy and least squares problems," University of Maryland Rep. TR 456, University of Maryland, College Park, Md., 1976.

[11] G. Golub, F. Luk and M. Overton, "A block Lanczos algorithm to determine the largest singular values of a large sparse matrix," work in progress.

[12] G. Golub and C. Reinsch, (1970, HACLA/I/10), "Singular value decomposition and least squares solutions," Numer. Math. 14 (1970), 403-420.

[13] G. Golub, G. Wahba, and M. Heath, "Generalized cross-validation as a method for choosing a good ridge parameter," to be published as a Stanford CS report.

[14] G. Golub and J. Wilkinson, "Ill-conditioned eigensystems and the computation of the Jordan canonical form," Stanford University Rept. CS 478, Stanford University, Stanford, Ca., 1975, to appear in SIAM Review.

[15] B. Green, "The orthgonal approximation of an oblique structure in factor analysis," Psychometrika 17 (1952), 429-440.

[16] V. N. Kublanovskaja, "Some algorithms for the solution of the complete problem of eigenvalues," V. Vycisl. Mat. i. Mat. Fiz. 1 (1961), 555-570.

[17] C. Lanczos, "Linear differential operators," Van Nostrand, London, 1961.

[18] L. Mirsky, "Symmetric gauge functions and unitarily invariant norms," Quart. J. Math. Oxford (2), 11 (1960), 50-59.

[19] C. Paige, "The computation of eigenvalues and eigenvectors of very large sparse matrices," Ph.D. Dissertation, The University of London, 1971.

[20] H. Rutishauser, "On Jacobi rotation patterns," Proc. Sym. Applied Math., 15 (1963), 219-239.

[21] P. Schönemann, "A generalized solution of the Orthogonal Procrustes problem," Psychometrika 31 (1966), 1-10.

[22] R. Underwood, "An iterative block Lanczos method for the solution of large sparse symmetric eigenproblems," Ph.D. thesis, Stanford University, Rept. CS 496, Stanford University, Stanford, Ca., 1975.

# LIST OF ATTENDEES

22nd Conference of Army Mathematicians
12-14 May 1976

Dr. John C. Amazigo
Rensselaer Polytechnic Institute
Troy, NY 12181

Dr. Merle M. Andrew
AF Office of Scientific Research
Building 410
Bolling AFB, DC 20332

Oscar L. Bowie
Army Materials & Mechanics Research Center
Watertown, MA 02172

Professor E. J. Brunelle
Dept. of Mechanics
Rensselaer Polytechnic Institute
Troy, NY 12181

Professor J. Buckmaster
University of Illinois
Urbana, IL 61801

Dr. Hans F. Bueckner
General Electric Company
Building 285-102
Schenectady, NY 12309

Dr. Aivars Celmins
USA Blllistic Research Laboratories
Aberdeen Proving Ground, MD 21005

Dr. Jagdish Chandra
US Army Research Office
PO Box 12211
ATTN: DRXRO-MA
Aberdeen Proving Ground, MD 27709

Y. M. Chen
Dept. of Applied Math & Statistics
State University of New York at Stony Brook
Stony Brook, NY 11794

Dr. Bart Childs
Texas A&M University
PO Box 6206
Texarkana, TX 75501

Dr. Shih-Chi Chu
GEN Thomas J. Rodman Laboratory
Rock Island Arsenal
Rock Island, IL 61201

Robert H. Coberly
GEN Thomas J. Rodman Laboratory
Rock Island Arsenal
Rock Island, IL 61201

Dr. Paul Davis
Dept. of Mathematics
Worcester Polytechnic Institute
Worcester, MA 01609

Professor J. B. Diaz
Rensselaer Polytechnic Institute
Troy, NY 12181

Germano DiLeonardo
General Electric Company--KAPL
River Road
Schenectady, NY 12309

Professor Richard C. DiPrima
Dept. of Mathematical Sciences
Rennselaer Polytechnic Institute
Troy, NY 12181

Dr. Donald Drew
Rensselaer Polytechnic Institute
Troy, NY 12181

Dr. J. Edwards
General Electric Company
Schenectady, NY 12309

Professor A. C. Eringen
Princeton University
Princeton, NJ 68540

Professor B. Fleishman
Dept. of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12181

K. R. Gandhi
Army Materials and Mechanics Research Center
Watertown, MA 02172

Ross Gingrich
Dept. of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12181

Marvin J. Goldstein
Naval Underwater Systems Center
New London Laboratory
New London, CT 06320

C. Maxson Greenland
ATTN: SAREA-PL-S, Plans Office
Edgewood Arsenal
Aberdeen Proving Ground, MD 21010

Dr. T. N. E. Greville
Mathematics Research Center
610 Walnut Street
Madison, WI 53706

Craig Hunter
ATTN: DRCPM-PBM-T-PA
Picatinny Arsenal
Dover, NJ 07801

Professor Thomas Kailath
Stanford University
Stanford, CA 94305

A. K. Kapila
Cornell University
Thurston Hall
Ithaca, NY 14853

Professor R. B. Kelman
Dept. of Computer Science
Colorad State University
Ft. Collins, CO 80521

Dr. Leon Kotin
Communication/Automatic Data Processing Lab
USA Electronics Command
Ft. Monmouth, NJ 07703

Dr. Frank Sheafen Kuo
Construction Engineering Research Lab
Champaign, IL

Dr. Badrig Kurkjian
DARCOM
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Siegfried H. Lehnigk
USA Missile Command
Redstone, AL 35809

Eric L. Leese
Dept. of National Defence
Ottawa, Canada

Dr. T. J. Mahar
Courant Institute
251 Mercer Street
New York, NY

Professor Bernard Matkowsky
Rennsselaer Polytechnic Institute
Troy, NY 12181

John F. Mescall
USA Materials and Mechanics Research Center
Watertown, MA 02172

Colonel Lothrop Mittenthal
US Army Research Office
Research Triangle Park, NC 27709

Dr. Donald M. Neal
Army Materials & Mechanics Research Center
Watertown, MA 02172

Professor Ben Noble, Director
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53706

Dr. Walter Pressman
USA Electronics Command
Ft. Monmouth, NJ 07703

Professor Louis B. Rall
Mathematics Research Center
University of Wisconsin
610 Walnut Street
Madison, WI 53706

Dr. Bart Rice
Dept. of Defense
9800 Swage Road
Ft. Meade, MD 20755

Professor James R. Rice
Brown University
Division of Engineering-Box D
Providence, RI 02912

Dr. T. P. Rich
USA Materials & Mechanics Research Center
Watertown, MA 02172

Professor Stephen M. Robinson
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53706

Professor J. Barkley Rosser
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53706

Professor Richard S. Sacher
Dept. of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12181

Dr. Edward Saibel
US Army Research Office
PO Box 12211
Research Triangle Park, NC 27709

W. D. Scharf
ATTN: AMXDO-RCC
Harry Diamond Laboratories
2800 Powder Mill Road
Adelphi, MD 20783

Randy Jay Schuetz
DARCOM
Intern Training Center
Red River Army Depot
Texarkana, TX 75501

Professor M. Slemrod
Dept. of Mathematics
Rensselaer Polytechnic Institute
Troy, NY 12181

Dr. Ram P. Srivastav
Dept. of Applied Math and Statistics
State University of New York at Stony Brook
Stony Brook, NY 11794

Roy Streit
Naval Underwater Systems Center
New London, CT 06320

Dr. Shunsuko Takagi
USA Cole Regions R&E Laboratory
Hanover, NH

Dr. James L. Thompson
USA Tank-Automotive Development Center
ATTN: DRDTA-RHMM
Warren, MI 48090

Dennis Tracey
USA Materials & Mechanics Research Center
Watertown, MA 02172

Dr. E. Wachspress
General Electric Company
Schenectady, NY

Dr. James N. Walbert
Materiel Testing Directorate
ATTN: STEAP-MT-G
Aberdeen Proving Ground, MD 21005

John H. Walker
Dept. of Maintenance Effectiveness
DARCOM Intern Training Center
Red River Army Depot
Texarkana, TX 75501

Dr. Richard A. Weiss
Waterways Experiment Station
Vicksburg, MS 39180

Larry Whatley
DARCOM Intern Training Center
Red River Army Depot
Texarkana, TX 75501

James M. Wilkes
USA White Sands Missile Range
ATTN: STEWS-TE-QC
White Sands Missile Range, NM 88002

Roger F. Willis
TRASANA
White Sands Missile Range, NM 88002

Dr. R. A. Willoughby
IBM Research
PO Box 218
Yorktown Heights, NY 10598

Mrs. Emma Wineholt
USA Ballistic Research Laboratory
ATTN: DRXBR-IB
Aberdeen Proving Ground, MD 21005

Dr. Rao Yalamanchili
GEN Thomas J. Rodman Laboratory
Rock Island Arsenal
Rock Island, IL 61201

Dr. C. C. Yang
Naval Research Lab
Washington, DC 20375

Dr. Ting N. Lee
George Washington University
Washington, DC

CPT Wayne Parrish
US Military Academy
West Point, NY

MAJ Anthony Quattromani
US Military Academy
West Point, NY

# WATERVLIET ARSENAL ATTENDEES

Peter C. T. Chen
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

Eugene E. Coppola
SARWV-RD-RM
Watervliet Arsenal
Watervliet, NY 12189

T. E. Davidson
SARWV-RR-ME
Watervliet Arsenal
Watervliet, NY 12189

Garry C. Carofano
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

Alma M. Gray
SARWV-RR-PS
Watervliet Arsenal
Watervliet, NY 12189

D. M. Gray
SARWV-RR-PS
Watervliet Arsenal
Watervliet, NY 12189

M. A. Hussain
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

David P. Kendall
SARWV-RR-ME
Watervliet Arsenal
Watervliet, NY 12189

William E. Lorensen
SARWV-RR-C
Watervliet Arsenal
Watervliet, NY 12189

Dr. L. Meisel
SARWV-RR-PS
Watervliet Arsenal
Watervliet, NY 12189

G. Peter O'Hara
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

G. A. Pflegl
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

San Li Pu
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

Ronald L. Racicot
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

Charles R. Thomas
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

Thomas E. Simkins
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

Royce W. Soanes
SARWV-RR-C
Watervliet Arsenal
Watervliet, NY 12189

Michael Strack
SARWV-RD-PD
Watervliet Arsenal
Watervliet, NY 12189

Joseph F. Throop
SARWV-RR-ME
Watervliet Arsenal
Watervliet, NY 12189

John D. Vasilakis
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

Julian J. Wu
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189

John Zweig
SARWV-RR-AMM
Watervliet Arsenal
Watervliet, NY 12189



*U.S. GOVERNMENT PRINTING OFFICE: 1977-741-223/4262 REGION NO. 4

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>ARO Report Number 77-1 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE *(and Subtitle)*<br>TRANSACTIONS OF THE TWENTY-SECOND CONFERENCE OF ARMY MATHEMATICIANS | | 5. TYPE OF REPORT & PERIOD COVERED<br>Interim Technical Report<br>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<br>Army Mathematics Steering Committee on Behalf of the Chief of Research Development and Acquisition | | 12. REPORT DATE<br>January 1977<br>13. NUMBER OF PAGES<br>612 |
| 14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*<br>US Army Research Office DRXRO<br>PO Box 12211<br>Research Triangle Park, NC 27709 | | 15. SECURITY CLASS. *(of this report)*<br>Unclassified<br>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-second 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)*

fracture mechanics
crack surface
finite element method
finite difference solutions
free-boundary problems
numerical integration
probability model
electron microscope
secure voice program
control problems
scattering problems
combustion
shock
input-output methods
elastic-plastic deformation
beam theory
fast transforms
Bessel functions
continued fractions
orthogonal matrices
response of payments
penetrators of armor
combat models
best fit methods
regression of n-th order differential equations
multi-level adaptive techniques
least squares and robust regression

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE 613 UNCLASSIFIED