REPRODUCED AT GOVERNMENT EXPENSE

(13)

AD-A154 047 ARO Report 85-1

# TRANSACTIONS OF THE SECOND ARMY

# CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

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 FILE COPY

DTIC ELECTE MAR 21 1985 S D

Sponsored by

The Army Mathematics Steering Committee

on behalf of

THE CHIEF OF RESEARCH, DEVELOPMENT

AND ACQUISITION

AR 03 18 077

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>ARO Report 85-1 | 2. GOVT ACCESSION NO.<br>AD-A154047 | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>Transactions of the Second Army Conference on Applied Mathematics and Computing | | 5. TYPE OF REPORT & PERIOD COVERED<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>February, 1985<br>13. NUMBER OF PAGES<br>952 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)<br>U.S. Army Research Office<br>P. O. 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 construed 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 First Army Conference on Applied Mathematics and Computing. It contains most of the papers in 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)

| | |
|---|---|
| computer-aided design | deformation theory |
| finite element methods | Green's function |
| flow problems | gun tube analysis |
| binary relations | periodic waves |
| code iterations | T-matrice |
| MACSYMA | front tracking |
| dynamical problems | explosions |
| Pascal | algorithms |
| perturbation analysis | screw calculus |
| chaos | splines |
| Cauchy problems | spectral analysis |
| variational principle | robotics |
| flame problems | eigenfunctions |
| elastic-plastic problems | confidence bounds |

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

U. S. ARMY RESEARCH OFFICE

Report No. 85-1

February 1985

# TRANSACTIONS OF THE SECOND ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

Hosts

Benet Weapons Laboratory
and
Rensselaer Polytechnic Institute

Held at

The Communications Center of the

Rensselaer Polytechnic Insitute

Troy, New York

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.

U.S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

Accession For
NTIS GRA&I
DTIC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

COPY INSPECTED

# FOREWORD

The Second Army Conference on Applied Mathematics and Computing was held ~~on 22-25 May 1984~~ in the Communications Center at Rensselaer Polytechnic Institute. It featured several current research areas, including robotics, continuum mechanics, and innovative computational methods. Also a special session was held on constitutive equations for high strain rate problems and The conference was initially planned for three days, but in order to accommadate the unusually large number of contributed papers, namely seventy-three, an additional day was scheduled.

The Army Mathematics Steering Committee (AMSC) is the sponsor of this conference, and it had as its hosts the Benet Weapons Laboratory and Rennselaer Polytechnic Institute. Drs. San Li Pu and John Vasilakis served as representatives for the Benet Weapons Laboratory, while Professors Donald A. Drew and Joseph E. Flaherty were the local chairpersons for the other host. Members of the AMSC take this opportunity to thank these gentlemen for all their time and work preparing for and conducting this well run scientific meeting.

The Series of annual meetings entitled Army Conferences on Applied Mathematics and Computing combines two former symposia, namely the Conferences of Army Mathematicians and the Numerical Analysis and Computers Conferences. Picking topics to be emphasized at the present conference, the organizing committee selected areas of research that span the fields of the earlier conferences. This point is well brought out by the following list of invited speakers together with the titles of their addresses.

| SPEAKERS AND AFFILIATIONS | TITLE OF ADDRESSES |
|---|---|
| Professor Ferdinand Freudenstein<br>Columbia University | Computer-Aided Mechanisms Analysis and Design, |
| Professor George C. Sih<br>Lehigh University | Scaling of Size and Time Associated with Damage Prediction, |
| Professor John W. Hutchinson<br>Harvard University | Methods for Analyzing the Mechanical Properties of Nonlinear Two Phase Composite Materials, → cont next page. |

| | |
|---|---|
| Professor H. T. Kung<br>Carnegie-Mellon University | Parallel Computations, Computational Complexity and Very Large Scale Integration, |
| Professor John Hopcroft<br>Cornell University | Mathematical Foundations for Robotics, |
| Professor D. P. Bertsekas<br>Massachusetts Institute of Technology | Distributed Asynchronous Algorithms, and |
| Professor A. Jameson*<br>Princeton University | Computational Methods for for Transonic Flows. |

Members of the AMSC were very pleased with the number and the quality of the papers presented at this conference. They were also pleased to have so many of said papers submitted for publication in the Transactions of this meeting. These interesting and informative articles can reach, in this printed form, persons who were unable to attend the conference; and for those in attendance this technical manual offers an opportunity to study in depth their contents.

*We are sorry that Professor Jameson was unable to present his paper at this meeting.

TABLE OF CONTENTS


| Title | Page |
|---|---|
| Foreward | iii |
| Table of Contents | v |
| Program | ix |
| Computer-Aided Mechanisms Analysis and Design<br>Ferdinand Freudenstein | 1 |
| Validation of a Multilayered Viscoelastic Seismic Propagation Model for Short to Medium Ranges<br>Ben L. Carnes | 13 |
| Buckled Elastica in Contact - Finite Element Solutions<br>Arthur R. Johnson and Claudia J. Quigley | 37 |
| Optimal Control Techniques for Computing Stationary Flows of Viscoelastic Fluids with Memory<br>Patrick Le Tallec | 49 |
| Programming with Binary Relations and an Associated Algebra of Programs<br>P ul Broome | 61 |
| Code Iteration for Noisy Channels<br>A. Brinton Cooper, III | 75 |
| Application of MACSYMA to Kinematics and Mechanical Systems<br>M. A. Hussain and B. Noble | 85 |
| Dynamic Instability of the Flexible Coupler of a Four-Bar Mechanism<br>Iradj G. Tadjbakhsh | 105 |
| An Introduction to the Scientific Computing Language Pascal-SC<br>L. B. Rall | 117 |
| Optimal Corrections of a Damped Linear Oscillator Under Random Perturbation<br>P. L. Chow and J. L. Menaldi | 149 |
| Non-Periodic Conditions for Chaos and Snap-Back Repellers<br>Nam P. Bhatia and Walter O. Egerland | 159 |


*This Table of Contents lists only the papers that are published in this Technical Manual. For a list of all the papers presented at the First Army Conference on Applied Mathematics and Computing, see the Agenda

| Title | Page |
|---|---|
| On the Numerical Solution of Singularly Perturbed Linear Two-Point Boundary-Value Problems<br>B. S. Ng and W. H. Reid | 165 |
| Self Similar Solutions for a Degenerate Cauchy Problem<br>Klaus Hollig and John A. Nohel | 177 |
| Variational Principle for Penetrator Dynamics Using Bilinear Functional and Adjoint Formulation<br>C. H. Shen | 185 |
| Representation of Two-Phase Flows by Averaging<br>Aivars Celmins | 199 |
| Numerical Investigation of the Stability of Diffusion Flames Near Extinction and Ignition<br>Y.S. Choi, C. Laine-Schmidt and G.S.S. Ludford | 225 |
| Fluid Mechanics of Quenching<br>Donald A. Drew, Ronald Brent, Susan Melly, William Schroeder and Stephen Wells | 231 |
| Beyond Stefan Problems: The Structure of Sheared Solidification Fronts<br>F. S. Hall and G.S.S. Ludford | 253 |
| Shock-Induced Thermal Runaway<br>T. L. Jackson and A. K. Kapila | 261 |
| The Stefan Problem of Detonation Theory<br>A.A. Oyediran and G.S.S. Ludford | 273 |
| Finite Increment Formulation of the Prandtl-Reuss Constitutive Equations<br>Russell L. Mallett | 285 |
| Constitutive Features of Solids at Shock-Wave Loading Rates<br>Dennis E. Grady | 303 |
| Examples and Significance of Change of type in Viscolasticity<br>Daniel D. Joseph, Michael Renardy, and Jean-Clausde Saut | 313 |
| A More Accurate Solution to the Elastic-Plastic Problem of Pressurized Thick-Walled Cylinders<br>Peter C. T. Chen | 319 |
| A Refined Shear Deformation Theory for Laminated Anisotropic Plates<br>J. N. Ready | 331 |
| Anisoparametric Finite Element Methodology for Penalty Constraint Formulations in Solid Mechanics<br>Alexander Fessler | 343 |
| Importance of Crack Tip Shape in Elastic-Plastic Fracture Analysis<br>Dennis M. Tracey and Colin E. Freese | 359 |

| Title | Page |
|---|---|
| Kinematic Hardening Applied to Non-Proportional Loading<br>Charles S. White | 373 |
| The Inverse Gaussian Pulse in the Experimental Determination of Linear System Green's Functions<br>Alfred S. Carasso and Nelson N. Hsu | 389 |
| Simulations of Special Interior Ballistic Phenomena with and Without Heat Transfer to the Gun Tube Wall<br>Budi Heiser and James A. Schmitt | 405 |
| An Analytical Model of Periodic Waves in Shallow Water--Summary<br>Harvey Segur and Allan Finkel | 459 |
| On an Improved T-Matrix Approach to Study the Scalar Scattering Response of Doubly Periodic Surfaces<br>A. Lakhtakia, V.K. Varadan and V.V. Varadan | 471 |
| Solitary, Periodic and Chaotic Waves in Thin Films<br>S. P. Lin and O. Suryadevara | 483 |
| Discontinuous Dependence of Solution on Boundary Conditions for Large Amplitude Shock Waves<br>T. C. T. Ting | 501 |
| Front Tracking and Two Dimensional Riemann Problems: A Conference Report<br>James Glimm, Christian Klingenberg, Oliver McBryan, Bradley Plohr, David Sharp and Sara Yaniv | 513 |
| Migration of the Gas Globe from Underwater Explosions: The Effects of Drag and Radiative Energy Loss<br>K. C. Heaton | 535 |
| A Local Refinement Element Method for Time Dependent Partial Differential Equations<br>Joseph E. Flaherty and Peter K. Moore | 585 |
| Post-Buckling Analysis of an Elastica with One and Many Critical Loads<br>Iradj G. Tadjabkhsh | 597 |
| A Mesh Moving Technique for Time Dependent Partial Differential Equations in Two Space Dimensions<br>David C. Arney and Joseph E. Flaherty | 611 |
| Numerical Simulation of Fluid Ejection from a Spinning Cylinder<br>Paul D. Fedele | 635 |
| A Numerical Algorithm for the Multidimensional, Multiphase, Viscour Equations of Interior Ballistics<br>James A. Schmitt | 649 |

| Title | Page |
|---|---|
| Note on Evaporation in Porous Media<br>R. E. Meyer | 693 |
| Jet-Contaminant Interaction in Confined Geometries<br>Lang-Mann Chang | 713 |
| A Methodology for the Development of Fire Control Equations for Guns and Rockets Fired From Aircraft<br>Harold J. Breaux | 729 |
| Singular Value Decomposition for Solution of Differential-Algebraic Equations of Mechanical System Dynamics<br>Neel K. Mani and Edward J. Haug | 737 |
| Application of Screw Calculus to the Evaluation of Manipulator Workspace<br>L. M. Hsia and Ting W. Lee | 765 |
| Recursive Gradient Estimation Using Splined for Navigation of Autonomous Vehicles<br>C. W. Shen | 787 |
| On Phase Transitions with Interfacial Energy<br>Morton E. Gurtin | 809 |
| A Computational Method for Field Detection of Unknown Substances<br>Edward W. Ross | 815 |
| Mathematical Foundations for Robotics<br>John E. Hopcroft | 835 |
| Dynamic Response in an Elastic-Plastic Projectile Due to Normal Impact<br>P.C.T. Chen, J.E. Flaherty and J.D. Vasilakis | 839 |
| On the Asympotic Analysis of Travelling Shocks and Phase Boundaries in Elastic Bars<br>Thomas J. Pence | 859 |
| Eigenfunctions at a Singular Point in Transversely Isotropic Materials under Axisymmetric Deformations<br>T.C.T. Ting, Yijing Jin, and S. C. Chou | 875 |
| Highly Viscous Fluid Flow in a Spinning and Nutating Cylinder<br>Thorwald Herbert | 883 |
| Computing Sets<br>Charles R. Leake | 895 |

| Title | Page |
|---|---|
| Calculation of Lower Confidence Bounds or System Reliability<br>Joseph V. Michalowicz | 903 |
| B-Splines of Nonuniform Triangulations<br>Chalres K. Chui | 939 |
| A Model for Asynchronous Distributed Computation<br>Dimitri P. Bertsekas | 943 |
| List of Participants | 949 |

SECOND ARMY CONFERENCE

on

APPLIED MATHEMATICS AND COMPUTING

MAY 22-25, 1984

RENSSELAER POLYTECHNIC INSTITUTE
TROY, NEW YORK 12181

**********AGENDA**********

Tuesday, May 22, 1984

0815-0845 REGISTRATION - Communications Center (CC) Grand Hall

0845-0900 OPENING REMARKS - CC318

0900-1000 GENERAL SESSION I - CC318

CHAIRPERSON: Edward W. Ross, Jr., US Army Natick Research and Development Center, Natick, Massachusetts

- COMPUTER-AIDED MECHANISMS ANALYSIS AND DESIGN

  Ferdinand Freudenstein, Columbia University, New York, New York

1000-1020 BREAK - CC Grand Hall

1020-1200 TECHNICAL SESSION I - Continuum Mechanics - CC318

CHAIRPERSON: Roshdy Barsoum, Army Mechanics and Materials Research Center, Watertown, Massachusetts

- VALIDATION OF A MULTI-LAYERED VISCOELASTIC SEISMIC PROPAGATION MODEL FOR SHORT TO MEDIUM RANGES

  Ben L. Carnes, US Army Waterways Experiment Station, Vicksburgh, Mississippi

- BUCKLED ELASTICA IN CONTACT-FINITE ELEMENT SOLUTIONS

  A. R. Johnson and C. J. Quigley, US Army Materials and Mechanics Research Center, Watertown, Massachusetts

- VISCOPLASTICITY BASED ON OVERSTRESS WITH A DIFFERENTIAL GROWTH LAW FOR THE EQUILIBRIUM STRESS

  M. Sutcu and E. Krempl, Rensselaer Polytechnic Institute, Troy, New York



PREVIOUS PAGE IS BLANK

Tuesday, May 22, 1984

1020-1200 TECHNICAL SESSION I - Continuum Mechanics (Cont'd)

- STRESS INTENSITY FACTORS FOR A CIRCULAR RING WITH UNIFORM ARRAY OF RADIAL CRACKS OF UNEQUAL DEPTH

  S. L. Pu, Benet Weapons Laboratory, Watervliet, New York

- OPTIMAL CONTROL TECHNIQUES FOR COMPUTING STATIONARY FLOWS OF VISCOELASTIC FLUIDS AT HIGH WEISSENBERG NUMBER

  P. Le Tallec, Mathematics Research Center, Madison, Wisconsin

Tuesday, May 22, 1984

1020-1200 TECHNICAL SESSION II - Computer Science and Related Topics - CC324

CHAIRPERSON: Idelle Peterson, Aviation Research and Development Command, St. Louis, Missouri

- PROGRAMMING WITH BINARY RELATIONS AND AN ASSOCIATED ALGEBRA OF PROGRAMS

  Paul H. Broome, Aberdeen Proving Ground, Maryland

- CODE ITERATION FOR NOISY CHANNELS

  A. Brinton Cooper, III, Aberdeen Proving Ground, Maryland

- APPLICATIONS OF COMPUTER SYMBOL MANIPULATION IN KINETICS, MECHANISMS AND COMPUTER GEOMETRY

  M. A. Hussain, General Electric Co., Schenectady, New York, and Ben Noble, University of Wisconsin, Madison

- AN INTRODUCTION TO THE SCIENTIFIC COMPUTING LANGUAGE PASCAL-SC

  L. B. Rall, University of Wisconsin-Madison

- DYNAMIC STABILITY OF FLEXIBLE MECHANISMS

  I. G. Tadjbakhsh, Rensselaer Polytechnic Institute

1200-1330 LUNCH

1330-1430 GENERAL SESSION II - CC318

CHAIRPERSON: Dr. Tony S. C. Chou, Army Mechanics and Materials Research Center, Watertown, Massachusetts

- SCALING OF SIZE AND TIME ASSOCIATED WITH MATERIAL DAMAGE PREDICTION

  George C. Sih, Lehigh University, Bethlehem, Pennsylvania

1430-1450 BREAK - CC Grand Hall

Tuesday, May 22, 1984

1450-1650 TECHNICAL SESSION III - Differential Equations - CC318

CHAIRPERSON: Y. Nakano, Cold Region Research and Engineering Lab, Hanover, New Hampshire

- OPTIMAL CORRECTIONS FOR A DAMPED LINEAR OSCILLATOR UNDER RANDOM PERTURBATIONS

  P. L. Chow and J. L. Menaldi, Wayne State University, Detroit, Michigan

- ASYMPTOTIC METHODS IN SOLID MECHANICS

  Julian Davis, US Army Armament Research and Development Center, Dover, New Jersey

- NONPERIODIC CONDITIONS FOR CHAOS AND SNAP-BACK REPELLERS

  Walter O. Egerland, Aberdeen Proving Ground, Maryland

- ON THE NUMERICAL SOLUTION OF LINEAR TWO-POINT BOUNDARY-VALUE PROBLEMS

  B. S. Ng, Indiana University - Purdue University, Indianapolis, Indiana and W. H. Reid, The University of Chicago, Chicago, Illinois

- SELF-SIMILAR SOLUTIONS FOR DEGENERATE PARABOLIC PROBLEMS

  John A. Nohel, University of Wisconsin-Madison

- VARIATIONAL PRINCIPLE FOR PENETRATOR DYNAMICS USING BILINEAR FUNCTIONAL AND ADJOINT FORMULATION

  C. N. Shen, Benet Weapons Laboratory, Watervliet, New York

Tuesday, May 22, 1984

1450-1650 TECHNICAL SESSION IV - Combustion and Multiphase Flow - CC324

CHAIRPERSON: Barry Fishburn, Large Caliber Weapons Systems Laboratory, ARDC, Dover, New Jersey

- REPRESENTATION OF TWO-PHASE FLOWS BY AVERAGING

  Aviars Celmins, Aberdeen Proving Ground, Maryland

- NUMERICAL INVESTIGATION OF THE STABILITY OF DIFFUSION FLAMES NEAR EXTINCTION AND IGNITION

  Y. S. Choi, C. Laine, and G. S. S. Ludford, Cornell University, Ithaca, New York

## Tuesday, May 22, 1984

1450-1650 TECHNICAL SESSION IV - Combustion and Multiphase Flow (Cont'd)

- FLUID MECHANICS OF QUENCHING

  Donald A. Drew, Ronald Brent, Susan Melly, William Schroeder, and Stephen Wells, Rensselaer Polytechnic Institute, Troy, New York

- BEYOND STEFAN PROBLEMS: STRUCTURE OF THE SOLIDIFICATION FRONT

  F. S. Hall AND G. S. S. Ludford, Cornell University, Ithaca, New York

- SHOCK-INDUCED THERMAL RUNAWAY

  T. Jackson and A. Kapila, Rensselaer Polytechnic Institute, Troy, New York

- THE STEFAN PROBLEM OF DETONATION THEORY

  A. A. Oyediran and G. S. S. Ludford, Cornell University, Ithaca, New York

## Wednesday, May 23, 1984

0830-1000 SPECIAL SESSION - Constitutive Equations - CC318

CHAIRPERSON: Dennis Tracey, Army Mechanics and Materials Research Center, Watertown, Massachusetts

- FINITE ELEMENT ANALYSIS OF DUCTILE RUPTURE

  Alan Needleman, Brown University, Providence, Rhode Island

- NUMERICAL SIMULATIONS OF LARGE DEFORMATIONS AT HIGH STRAIN RATES

  John Mescall, Army Materials and Mechanics Research Center, Watertown, Massachusetts

- INCREMENTAL FORMULATIONS OF ELASTIC-PLASTIC CONSTITUTIVE LAWS

  Russell L. Mallet, Rensselaer Polytechnic Institute, Troy, New York

1000-1020 BREAK - CC Grand Hall

1020-1150 SPECIAL SESSION - Constitutive Equations (Cont'd) - CC318

- CONSTITUTIVE FEATURES OF SOLIDS AT SHOCK-WAVE LOADING RATES

  Dennis E. Grady, Sandia National Laboratories, Albuquerque, New Mexico

Wednesday, May 23, 1984

1020-1150 SPECIAL SESSION - Constitutive Equations (Cont'd) - CC318

- CONSTITUTIVE EQUATIONS FOR HOT-WORKING

  Lallit Anand, Massachusetts Institute of Technology, Cambridge, Massachusetts

- EXAMPLES AND SIGNIFICANCE OF CHANGE OF TYPE IN VISCOELASTICITY

  D. D. Joseph, M. Renardy and J. C. Saut, University of Minnesota, Minneapolis, Minnesota

1150-1220 A SHORT MOVIE ON FLOW OF TWO FLUIDS - CC318

D. D. Joseph, University of Minnesota, Minneapolis, Minnesota

1220-1330 LUNCH

Wednesday, May 23, 1984

1330-1530 TECHNICAL SESSION V - Continuum Mechanics - CC318

CHAIRPERSON: Thomas E. Simkins, Benet Weapons Laboratory, Watervliet, New York

- A MORE ACCURATE SOLUTION TO THE ELASTIC-PLASTIC PROBLEM OF PRESSURIZED THICK-WALLED CYLINDERS

  P. C. T. Chen, Benet Weapons Laboratory, Watervliet, New York

- BURSTING PRESSURE OF A THICK-WALLED CYLINDER SUBJECTED TO TORSIONAL LOADS

  Shih C. Chu, US Army Armament Research and Development Center, Dover, New Jersey

- A REFINED THEORY FOR LAMINATED ANISOTROPIC PLATES

  J. N. Reddy, Virginia Polytechnic Institute and State University, Blacksburg, Virginia

- ANISOPARAMETRIC FINITE ELEMENT METHODOLOGY FOR PENALTY CONSTRAINT FORMULATIONS IN SOLID MECHANICS

  Alexander Tessler, Army Materials and Mechanics Research Center, Watertown, Massachusetts

- IMPORTANCE OF CRACK TIP SHAPE IN ELASTIC-PLASTIC FRACTURE ANALYSIS

  Dennis M. Tracey and Colin E. Freese, Army Materials and Mechanics Research Center, Watertown, Massachusetts

Wednesday, May 23, 1984

1330-1530 TECHNICAL SESSION V - Continuum Mechanics (Cont'd) - CC318

- KINEMATIC HARDENING APPLIED TO NON-PROPORTIONAL LOADING

  Charles S. White, Army Materials and Mechanics Research Center, Watertown, Massachusetts

Wednesday, May 23, 1984

1330-1530 TECHNICAL SESSION VI - Heat Transfer and Inverse Problems - CC324

CHAIRPERSON: Kevin O'Neil, Cold Region Research and Engineering Lab, Hanover, New Hampshire

- PROBE WAVEFORMS AND DECONVOLUTION IN THE EXPERIMENTAL DETERMINATION OF LINEAR SYSTEM GREEN'S FUNCTIONS

  Alfred S. Carasso and Nelson N. Hsu, National Bureau of Standards, Washington, DC

- SPECTRAL ANALYSIS OF THE SCATTERING OF ELASTIC WAVES FROM A FLUID-FILLED CYLINDER

  P. P. Delsanto, J. D. Alemar, and E. Rosario, University of Puerto Rico, Mayaguez, Puerto Rico

- CONTROL VOLUME BASED FINITE ELEMENT METHODS FOR FLUID FLOW AND HEAD TRANSFER -- AN OVERVIEW

  C. Prakash, Rensselaer Polytechnic Institute, Troy, New York

- APPLICATION OF THE LOCALLY ANALYTIC DIFFERENCING SCHEME TO SOME TEST PROBLEMS FOR THE CONVECTION-DIFFUSION EQUATION

  C. Prakash, Rensselaer Polytechnic Institute, Troy, New York

- SIMULATIONS OF SPECIAL INTERIOR BALLISTIC PHENOMENA WITH AND WITHOUT HEAT TRANSFER TO THE GUN TUBE WALL

  Rudi Heiser, Fraunhofer-Institut fuer Kurzzeitdynamik, Weil am Rhein, West Germany and James A. Schmitt, Aberdeen Proving Ground, Maryland

- COMMENTS ON FINITE ELEMENT METHOD AND BANDWIDTH WITH REFERENCE TO TRANSIENT HEAD CONDUCTION

  Rao Yalamanchili, US Army Armament Research and Development Center, Dover, New Jersey

1530-1550 BREAK - CC Grand Hall

## Wednesday, May 23, 1984

1550-1650 GENERAL SESSION III - CC318

CHAIRPERSON: Joseph E. Flaherty, Benet Weapons Lab and Rensselaer Polytechnic Institute, Troy, New York

- METHODS FOR ANALYZING THE MECHANICAL PROPERTIES OF NONLINEAR TWO-PHASE COMPOSITE MATERIALS

  John W. Hutchinson, Harvard University, Cambridge, Massachusetts

## Thursday, May 24, 1984

0830-1030 TECHNICAL SESSION VII - Wave Propagation - CC318

CHAIRPERSON: Charles Bowden, Redstone Arsenal, Huntsville, Alabama

- NONLINEAR HYPERBOLIC WAVE PROBLEMS WITH INPUT SETS II

  E. Adams, Institut fur Angewandte Mathematik, Universit at Karlsruhe, Federal Republic of Germany, and W. F. Ames, Georgia Institute of Technology, Atlanta, Georgia

- AN ANALYTICAL MODEL OF TWO-DIMENSIONAL, DOUBLY-PERIODIC WAVES IN SHALLOW WATER

  Harvey Segur, Aeronautical Research Associates of Princeton, Inc., Princeton, New Jersey, and Allan Finkel, Thomas Watson Research Center, Yorktown Heights, New York

- ON THE RAYLEIGH HYPOTHESIS AND WAVE SCATTERING BY PERIODICALLY CORRUGATED SURFACES

  Akhlesh Lakhtakia, Vijay V. Varadan and Vasundara V. Varadan, The Pennsylvania State University, University Park, Pennsylvania

- PERIODIC AND LOCALLY CHAOTIC WAVES

  S. P. Lin and O. Suryadevara, Clarkson College of Technology, Potsdam, New York

- ON NONLINEAR WAVE PACKETS NEAR DIRECT RESONANCE

  Yan-Chow Ma, TRW Space and Technology Group, One Space Park, Redondo Beach, California

- DISCONTINUOUS DEPENDANCE OF SOLUTION ON BOUNDARY CONDITIONS FOR LARGE AMPLITUDE SHOCK WAVES

  T. C. T. Ting, University of Illinois at Chicago, Chicago, Illinois

## Thursday, May 24, 1984

0830-1030 TECHNICAL SESSION VIII - Numerical Solutions to P.D.E.'s - CC324

CHAIRPERSON: Gary Carofano, Benet Weapons Laboratory, Watervliet, New York

- FRONT TRACKING: VALIDATION AND NEW DEVELOPMENTS

  J. Glimm, O. McBryan, B. Plohr, S. Yaniv, Courant Institute of Mathematical Sciences, New York University, New York, New York

- MIGRATION OF THE GAS GLOBE FROM UNDERWATER EXPLOSIONS

  K. C. Heaton, Defence Research Establishment Valcartier, Courcelette, Quebec

- HIGH RESOLUTION SCHEMES FOR COMPRESSIBLE EULER AND NAVIER-STOKES EQUATIONS

  Stanley Osher, University of California, Los Angeles, California

- AN UNCONDITIONALLY STABLE SECOND ORDER FINITE DIFFERENCE METHOD FOR THE TIME DEPENDENT SIMULATION OF P-N-JUNCTION

  Christian A. Ringhofer, University of Wisconsin-Madison, Wisconsin

- AN ADAPTIVE LOCAL REFINEMENT FINITE ELEMENT METHOD FOR PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS

  Joseph E. Flaherty and Peter K. Moore, Rensselaer Polytechnic Institute, Troy, New York

- A NONLINEAR EIGENVALUE PROBLEM IN ELASTIC INSTABILITY

  I. G. Tadjbakhsh, Rensselaer Polytechnic Institute

1030-1100 BREAK - CC Grand Hall

## Thursday, May 24, 1984

1100-1200 GENERAL SESSION IV - CC318

CHAIRPERSON: Harry Reid, Aberdeen Proving Ground, Maryland

- PARALLEL COMPUTATIONS, COMPUTATIONAL COMPLEXITY AND VERY LARGE SCALE INTEGRATION

  H. T. Kung, Carnegie-Mellon University, Pittsburgh, Pennsylvania

1200-1330 LUNCH

## Thursday, May 24, 1984

1330-1550 TECHNICAL SESSION IX - Stability, Transport and Diffusion in Fluids - CC318

CHAIRPERSON: Orazio Sindoni, Chemical Systems Laboratory, Edgewood, Maryland

- NUMERICAL SIMULATION OF FLUID EJECTION FROM A SPINNING CYLINDER

  Paul D. Fedele, Aberdeen Proving Ground, Maryland

- ELLIPTICALLY DESINGULARIZED VORTEX REPRESENTATION AND MERGER FOR THE TWO-DIMENSIONAL EULER EQUATIONS

  M. V. Melander, A. S. Styczek and N. J. Zabusky, University of Pittsburgh, Pittsburgh, Pennsylvania

- EVAPORATION IN POROUS MEDIA

  R. E. Meyer, University of Wisconsin-Madison, Wisconsin

- A NUMERICAL ALGORITHM FOR THE MULTIDIMENSIONAL, MULTIPHASE EQUATIONS OF INTERIOR BALLISTICS

  James A. Schmitt, Aberdeen Proving Ground, Maryland

- A STUDY OF THE CRITICAL LAYER IN A ROTATING PAYLOAD

  Raymond Sedney, Aberdeen Proving Ground, Maryland

- JET-CONTAMINANT INTERACTION IN CONFINED GEOMETRIES

  Lang-Mann Chang, Aberdeen Proving Ground, Maryland

- HIGHLY VISCOUS FLUID FLOW IN A SPINNING AND NUTATING CYLINDER

  Thorwald Herbert, Virginial Polytechnic Institute and State University, Blacksburg, Virginia

## Thursday, May 24, 1984

1330-1550 TECHNICAL SESSION X - Control, Mechanisms and Robotics - CC324

CHAIRPERSON: Julian J. Wu, Army Research Office, Research Triangle Park, North Carolina

- A METHODOLOGY FOR THE DEVELOPMENT OF FIRE CONTROL EQUATIONS FOR GUNS AND ROCKETS FIRED FROM AIRCRAFT

  Harold J. Breaux, Aberdeen Proving Ground, Maryland

- SINGULAR VALUE DECOMPOSITION FOR SOLUTION OF DIFFERENTIAL-ALGEBRAIC EQUATIONS OF MECHANICAL SYSTEM DYNAMICS

  Neel K. Mani and Edward J. Haug, University of Iowa, Iowa City, Iowa

1330-1550 TECHNICAL SESSION X - Control, Mechanisms and Robotics (Cont'd) - CC324

- APPLICATION OF SCREW CALCULUS TO THE EVALUATION OF MANIPULATOR WORKSPACE

  L. M. Hsia, California State University, Los Angeles, California, and Ting W. Lee, University of Notre Dame, Notre Dame, Indiana

- RECURSIVE GRADIENT ESTIMATION USING SPLINES FOR NAVIGATION OF AUTONOMOUS VEHICLES

  C. N. Shen, Benet Weapons Laboratory, Watervliet, New York

- DYNAMICS AND CONTROLS IN DESIGN AND DEVELOPMENT

  Ronald R. Beck and Gerald W. Jackson, US Army Tank-Automotive Command, Warren, Michigan

- APPLICATION OF QUATERNION ALGEBRA TO VEHICLE DYNAMICS

  Roger A. Wehage, US Army Tank-Automotive Command, Warren, Michigan

- ON PHASE TRANSITIONS WITH INTERFACIAL ENERGY

  Morton Gurtin, Carnegie Mellon, Pittsburgh, Pennsylvania and MRC, Madison, Wisconsin

1330-1550 POSTER SESSIONS - CC232

- A COMPUTATIONAL METHOD FOR FIELD-DETECTION OF UNKNOWN SUBSTANCES

  Edward W. Ross, Jr., US Army Natick R & D Center, Natick, Massachusetts

- IMPLICIT DIFFERENCE APPROXIMATION OF QUASILINEAR EVOLUTION EQUATIONS

  Michael G. Crandall, University of Wisconsin-Madison, Wisconsin

- AUTOMATED NUMERICAL ANALYSIS OF HYDROGEOLOGIC DATA

  James D. Crabtree, USAE Waterways Experiment Station, Vicksburg, Mississippi

1550-1610 BREAK

1610-1710 GENERAL SESSION V - CC318

CHAIRPERSON: Ronald L. Racicot, Benet Weapons Laboratory, Watervliet, New York

- MATHEMATICAL FOUNDATIONS FOR ROBOTICS

  John Hopcroft, Cornell University, Ithaca, New York

## Friday, May 25, 1984

0800-1010 TECHNICAL SESSION XI - Penetration Mechanics - CC318

CHAIRPERSON: John Vasilakis, Benet Weapons Laboratory, Watervliet, New York

- A TWO-STEP VERSION OF STRANG-GOTTLIEB TECHNIQUES FOR DEFORMABLE LAGRANGIAN MESHES

  S. Hanagud and H. P. Chen, Georgia Institute of Technology, Atlanta, Georgia

- DYNAMIC RESPONSE IN A ELASTIC-PLASTIC PROJECTILE DUE TO NORMAL IMPACT

  P. C. T. Chen, J. E. Flaherty and J. D. Vasilakis, Benet Weapons Laboratory, Watervliet, New York

- ASYMPTOTIC ANALYSIS OF TRAVELLING SHOCKS AND PHASE BOUNDARIES IN ELASTIC BARS

  Thomas J. Pence, University of Wisconsin, Madison, Wisconsin

- EIGENFUNCTIONS AT A SINGULAR POINT IN TRANSVERSELY ISOTROPIC MATERIALS UNDER AXI-SYMMETRIC DEFORMATIONS

  T. C. T. Ting, Yinjing Jin and S. C. Chou, University of Illinois at Chicago, Chicago, Illinois

- PENETRATION PROBLEMS - VIS-A-VIS ISOPARAMETRIC FINITE ELEMENTS

  R. Natarajan, School of Engineering and Architecture, Tuskegee Institute, Tuskegee, Alabama

## Friday, May 25, 1984

0800-1010 TECHNICAL SESSION XII - Approximation and Data Analysis - CC324

CHAIRPERSON: Jack Pollin, US Military Academy, West Point, New York

- MULTIVARIATE INTERPOLATION OF SCATTERED DATA BY FUNCTIONAL MINIMIZATION

  Peter Alfeld, University of Utah, Salt Lake City, Utah

- MATHEMATICAL PROPERTIES OF COMPUTATIONAL SETS

  Charles R. Leake, US Army Concepts Analysis Agency, Bethesda, Maryland

- NORMAL SOLUTIONS OF LARGE SCALE LINEAR PROGRAMS AND INEQUALITIES

  O. L. Mangasarian, University of Wisconsin-Madison, Wisconsin

## Friday, May 25, 1984

0800-1010 TECHNICAL SESSION XII - Approximation and Data Analysis (Cont'd) - CC324

- CALCULATION OF LOWER CONFIDENCE BOUNDS ON SYSTEM RELIABILITY

  Joseph V. Michalowicz, Harry Diamond Laboratories, Adelphi, Maryland

- B-SPLINES ON TYPE -1 AND TYPE -2 TRIANGULATIONS

  Charles K. Chui, Texas A and M University, College Station, Texas

1010-1030 BREAK - CC Grand Hall

## Friday, May 25, 1984

1030-1230 GENERAL SESSION VI - CC318

CHAIRPERSON: Jagdish Chandra, US Army Research Office, Research Triangle Park, North Carolina

- DISTRIBUTED ASYNCHRONOUS ALGORITHMS

  D. P. Bertsekas, Massachusetts Institute of Technology, Cambridge, Massachusetts

- COMPUTATIONAL METHODS FOR TRANSONIC FLOWS

  A. Jameson, Princeton University, Princeton, New Jersey

1230 ADJOURN

# COMPUTER-AIDED MECHANISMS ANALYSIS AND DESIGN

Ferdinand Freudenstein
Department of Mechanical Engineering
Columbia University
New York, New York, 10027

ABSTRACT. The modern development of the subject of mechanisms has been influenced heavily by high-speed computation. This is due primarily to the nonlinearity, complexity and variety of the mechanical elements involved. Recent developments and research in this area will be reviewed with emphasis on conceptual design, kinematic analysis and synthesis and dynamic analysis.

I. INTRODUCTION. The subject of mechanisms is in some respects very old, while in others it is remarkably new. Mechanical invention is already in evidence in the earliest civilizations. The evolution of mechanical design from a purely intuitive "mechanic art" into an engineering discipline constitutes one of the significant current engineering developments. We begin this review with a survey of the creative phase of mechanisms design.

II. THE CONCEPTUAL DESIGN OF MECHANISMS. The creative phase of mechanisms design is probably the most elusive and challenging. Attempts to systematize this process traditionally involve the functional classification of mechanisms, such as atlases of mechanisms. Beginning with C. Babbage in 1826 (2) isolated attempts have appeared aimed at the development of a systematic, abstract and useful representation and classification of mechanisms (e.g. 1,8,14). Beginning in 1964-65 graph theory was introduced for the representation of kinematic structure (3-6,9-13,15). In this approach links are represented by vertices, joints by edges and the edge connection of vertices corresponds to the joint connection of links, edges being labeled according to joint type and the fixed link being identified as well. For example the familiar swinging-block mechanism and its graph are shown in Figs. 1a and 1b, respectively. In the latter figure the number of each vertex corresponds to the link identification of Fig. 1a; the symbols R,P denote pin joints and sliding joints, respectively; and the small circle around vertex 2 represents the fixed link identification.

Graphs can be defined analytically and hence stored in the memory of a computer. The enumeration of mechanisms can thus be reduced to a combinatorial problem in graph enumeration with different mechanisms corresponding to non-isomorphic graphs. In this way it is possible to create mechanisms in a systematic manner with only minimal assumptions regarding their nature (such as the degree of freedom of the motion, the number of moving elements and the admissible joint types). In this approach to creative design, therefore, we first concentrate on kinematic structure, enumerate potentially useful mechanisms and then evaluate these in the light of functional considerations.

We illustrate the procedure in a specific case: the development of a variable-stroke engine mechanism (11). A variable-stroke engine would have potentially improved fuel economy due to the fact that load control is achieved by varying piston stroke, thus eliminating inlet-throttling and reducing pumping losses during short-stroke operation. Hence, in this case a mechanism is desired for converting rotary to variable-stroke reciprocating motion while maintaining a constant or nearly constant compression ratio.

In this problem we consider single-degree-of-freedom mechanisms with pin joints and/or sliding joints and we limit the search to plane mechanisms. The desired mechanism must be adjustable so as to provide stroke control while maintaining the desired compression ratio, favorable force transmission and acceptable dynamic characteristics under all operating conditions. From the standard degree-of-freedom equation for mechanisms it follows that we need to search for mechanisms with (4+2n) links and (4+3n) joints where n is a positive integer. To avoid undue complexity we restrict the search to n=1 or 2. Even so the number of mechanisms created in the search is in excess of one hundred. Restrictions as to the number of admissible sliding pairs, floating links etc. limit the search to about 40 mechanisms. Fig. 2 shows some of the graphs associated with six-link configurations (n=1) and sketches of the corresponding variable-stoke mechanisms, while Fig. 3 shows the same for the eight-link structures (n=2). The mechanisms were then evaluated on the basis of functional requirements. Known configurations, such as prior-art patents, were recreated, thus providing a valuable check both on the search as well as on the prior art. Three mechanisms were found which were potentially acceptable. Of these the most favorable was judged to be the mechanism shown in Fig. 4, which was then investigated in greater detail. The final design of this configuration, shown in Fig. 5, was awarded a U.S. patent (#4,270,495, 1981).

This approach has been used with success in recent times in several applications including the development of novel casement window mechanisms (7) and wheel dampers (16). Current research efforts are concerned with the partial automation of this process. This involves automation of the determination (enumeration) of the kinematic structures or graphs satisfying the search specification and the dimensioning and mechanical analysis of the mechanism created in the search. The latter would include interactive-graphics capabilities and potentially also artificial-intelligence techniques (expert systems) which would permit the evaluation of individually tailored search specifications and functional evaluation criteria.

This approach to conceptual design also has important implications in the evaluation of patents and in recognizing similarities and differences in mechanisms.

III. KINEMATIC AND DYNAMIC ANALYSIS OF MECHANISMS. The computer-aided kinematic and dynamic analysis of mechanisms has been greatly aided by the development of large-scale design codes. These solve the kinematic loop-closure equations of mechanisms to determine displacements, velocities

and accelerations. The dynamics is determined from the Lagrangian equations of motion, usually in variational form. In this way complicated mechanisms can be analyzed economically and design alternatives evaluated by parameter variation. This has made possible the economical analysis of both mechanisms and mechanical systems of extraordinary complexity in the design stage.

Table 1 summarizes the basic characteristics of some major design codes.

Table 1; Codes for the kinematic and dynamic analysis and synthesis of mechanisms (all with interactive graphics capability).

| CODE | DEVELOPER | BASIC CAPABILITY | SOME TYPICAL APPLICATIONS |
|---|---|---|---|
| ADAMS | M.A. Chace and associates (U. of Michigan; Mechanical Dynamics Inc.) | Kinematic and dynamic analysis of three-dimensional mechanisms and mechanical systems. | Vehicle dynamics, agricultural equipment, general machinery, mechanisms, robotics. |
| DADS | E.J. Haug and associates (U. of Iowa) | Kinematic and dynamic analysis of plane and three-dimensional mechanisms and mechanical systems. | Complex vehicles,mechanisms, agricultural machinery, intermittent mechanisms, large-scale mobile equipment. |
| DRAM | M.A. Chace and Associates (U. of Michigan; Mechanical Dynamics Inc.) | Kinematic and dynamic analysis of plane mechanisms and mechanical systems. | Accident reconstruction, general machinery, nuclear core simulation, aircraft mechanisms, vehicles. |
| IMP | J.J. Uicker and associates (U. of Winconsin) | Kinematic and dynamic analysis of plane and three-dimensional mechanisms and mechanical systems. | Automotive suspensions, sewing machine drives, vehicle dynamics, general machinery, robotics. |
| KYNSYN | R.E. Kaufman (U. of Washington) | Kinematic synthesis of plane linkage mechanisms (determination of mechanism proportions for a particular motion sequence via a microcomputer setup). | Sizing of plane linkage mechanisms so as to obtain a prescribed motion sequence, e.g. in production processes general machinery etc. |
| LINCAGES | A.G. Erdman (U. of Minnesota). | Plane kinematic analysis and synthesis of linkage mechanisms. | Variable-speed drives, window actuators, biomechanical devices, linkage mechanisms. |

Mechanisms which have been analyzed with the aid of computer-aided techniques in recent years also include artificial limbs, complex planetary gear trains, shaft couplings and many others.

In addition to the computer-aided design of mechanisms and mechanical systems, a substantial area of research and development is concerned with basic mechanical components. As we shall see in the following remarks, this subject is still filled with challenges.

IV. MECHANICAL COMPONENTS. Despite the sophistication of the analysis capability in mechanisms and mechanical systems, there are important areas in which our understanding of mechanisms remains lacking. For these the development of effective computer-aided design methods hinges on a better understanding of their fundamental mechanics.

Amongst the mechanisms in this category there are many basic machine components, such as cam-follower systems, universal joints, constant-velocity shaft couplings, variable-speed transmissions and others. Many of these were established long ago, as were the simpler aspects of their motions. These components are skillfully proportioned on the basis of many years of experience, as well as life and wear testing. The sizing of these components in the design stage, however, remains a challenging objective. As technology continues to advance at a rapid pace, mechanical devices are expected to operate reliably at ever increasing speeds and loads and the rational and predictive design of these mechanisms has become increasingly necessary. Amongst other things this includes the determination of the internal force and torque reactions of these components under both static and dynamic loading and the ability to predict limiting speeds before the onset of unacceptable wear rates and/or destructive impacts and overloads.

In belt drives, for example, the exponential belt-tension equation, which was derived by Euler, is still representative of the state of the art in many aspects of this field. In universal joints, the internal force and torque reactions are not as yet properly understood. In many constant-velocity shaft couplings there are a substantial number of ball or roller elements with complex motions involving line or point contact. Their analysis is only just beginning. The development of the fundamental mechanics of these components presents exciting challenges in rigid-body mechanics, elasticity theory and stress analysis. Eventually such efforts can be expected to lead to more efficient and powerful computer-aided design procedures, which in turn will accelerate utilization of these components under more severe operating conditions.

Another area which is still largely unexplored is the logical design of mechanisms. Many mechanisms perform sohisticated logical functions e.g. in production machinery involving decisions concerning sequences and timing in cutting, loading, transfer operations etc. This subject has received some attention in recent years (12), but its potential remains to be recognized by the engineering profession.

And finally let us consider the area of three-dimensional mechanisms. Although much progress has been made, especially with computational approaches reducing the complexity and tediousness of their design analysis, much remains to be done. For example, the skew four-bar linkage (Fig. 6) is a basic spatial linkage used to connect non-parallel non-intersecting shafts. It is known from experience that depending on proportions the input and output links function either as cranks (i.e. can make complete rotations) or rockers (i.e. can only oscillate). An analysis of the dimensional restrictions which govern the rotatability of the links is both basic and extremely difficult.

In the area of robotics the analysis of workspace represents another very active area. Workspace is a complicated function of three-dimensional linkage geometry. The determination of robotic workspace for given linkage proportions can be handled by the previously mentioned three-dimensional codes. But the synthesis of robot dimensions for optimum workspace characteristics is as yet in the research stage. As the fundamental mechanics of these areas continues to develop we can expect corresponding advances in their computer-aided design.

V. CONCLUSION. This has been a brief survey of trends in the computer-aided design of mechanisms. Many of the subjects which have been discussed (e.g. the conceptual design of mechanisms) involve areas of research in which the author and his students have been involved. A significant part of this research has been sponsored by the Army Research Office. The author would like to take this opportunity to express his appreciation for this support.

REFERENCES


1. Assur, L.V. "Investigation of plane hinged mechanisms with lower pairs from the point of view of their structure and classification", Izv. S. Peterburgskogo Politekh. Inst. Imp. Petra Velikago. 1913, 20, Pt. I, pp. 329-385; 20, Pt. II, pp. 681-635; 1914, 21,pp. 187-283; 22, pp. 177-257; 1915,23, pp. 1-169.
2. Babbage, C. "On a method of expressing by signs the action of machinery", Phil. Trans. Roy. Soc. 2, 1826, p.250.
3. Buchsbaum, F. and Freudenstein, F. "Synthesis of Kinematic structure of geared kinematic chains and other mechanisms", J. of Mechanisms 5, 1970, pp. 357-392.
4. Crossley, F.R.E. "The permutations of kinematic chains of eight members or less from the graph-theoretic viewpoint", Dev. in Theor. and Appl. Mechanics, 2, W.A. Shaw, Editor, Pergamon Press, 1965, pp. 467-486.
5. Davies, T.H. "An extension of Manolescu's classification of planar kinematic chains and mechanisms of mobility $M \geq 1$, using graph theory", J. Mechanisms 2, 1968, pp. 87-100.
6. Dobrjanskyj, L. and Freudenstein, F. "Some applications of graph theory to the structural analysis of mechanisms", Trans. ASME 89B, J. Eng. Ind., 1967 pp. 153-158.
7. Erdman, A.G., Nelson, E., Peterson, J. and Bowen, J. "Type and dimensional synthesis of casement window mechanisms", ASME Paper 80-DET-78, 1980; see also Mechanical Engineering, Dec. 1981, pp. 46-55.
8. Franke, R. "Vom Aufbau der Getriebe", Vol. I, VDI, Düsseldorf, 1951.
9. Freudenstein, F. "An application of Boolean algebra to the motion of epicyclic drives", Trans. ASME 93B, J. Eng. Ind., 1971, pp. 176-182.
10. Freudenstein, F. and Dobrjanskyj, L. "On a theory for the type synthesis of mechanisms, Proc. 11th. Int'l. Cong. Applied Mechanics, Springer, 1965, pp. 420-428.
11. Freudenstein, F. and Maki, E.R. "Development of an optimum variable-stroke internal-combustion engine mechanism from the viewpoint of kinematic structure", Trans. ASME 105, J. Mechanisms, Transmissions and Automation in Design, 1983, pp. 259-266.
12. Freudenstein, F. and Söylemez. E. "The multiport lever; a mechanical logical element", Trans. ASME 89B, J. Eng. Ind., 1977, pp. 353-359.
13. Freudenstein, F. and Woo, L.S. "Kinematic structure of mechanisms", in 'Basic Questions of Design Theory', North-Holland Publ. Co., Amsterdam, 1974, pp. 241-264.
14. Reuleaux, F. "Kinematics of Machinery", translated by A.B.W. Kennedy, Macmillan, London, 1876.
15. Woo. L.S. "Type synthesis of plane linkages", Trans. ASME 89B, J. Eng. Ind., 1967, pp. 158-172.
16. Yan, H.S. and Chen, J.J. "Creative design of a wheel damping mechanism", Proc. 8th. O.S.U. Applied Mechanisms Conference, St. Louis, Mo., Sept., 1983, Paper # 74.

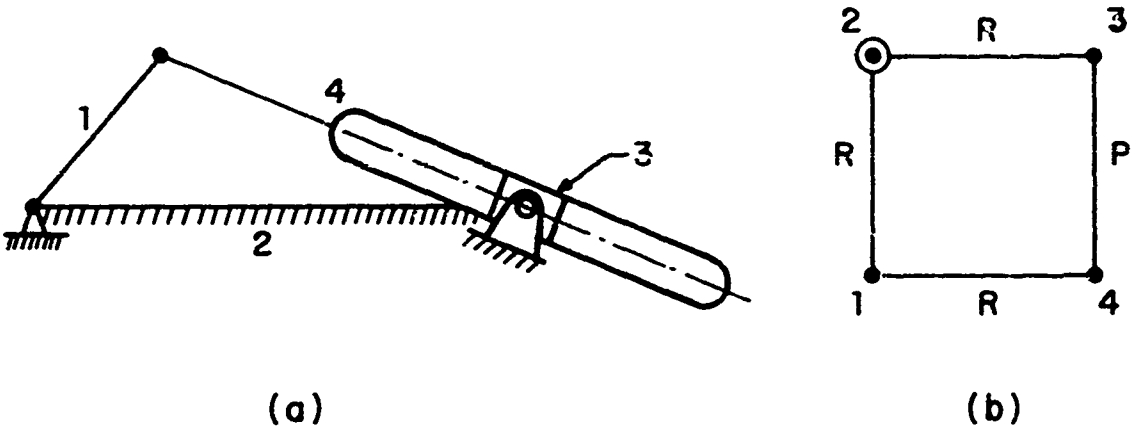



FIG. 1

A-ME-3688 W.L.

FIG. 1 SWINGING-BLOCK LINKAGE ( Fig. 1a) AND GRAPH (Fig. 1b).

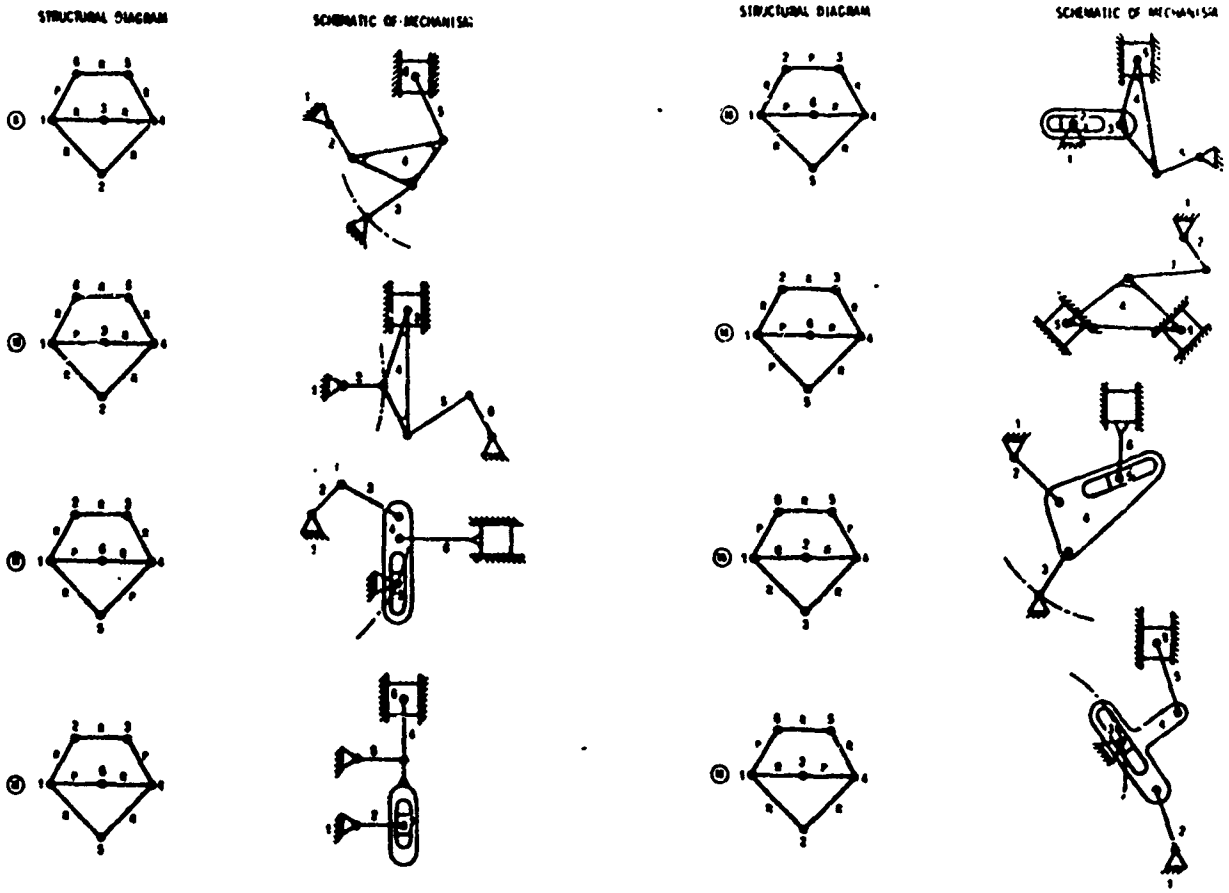


FIG.2

STRUCTURAL AND SCHEMATIC DRAWINGS OF SOME SIX-LINK VARIABLE-STROKE MECHANISMS

Reprinted with the permission of the American Society of Mechanical Engineers from "Development of an optimum variable-stroke internal-combustion engine mechanism from the viewpoint of kinematic structure" by F. Freudenstein and E.R. Maki. Trans. ASME; J. of Mechanisms, Transmissions and Automation in Design 105, 1983, pp. 259-266.

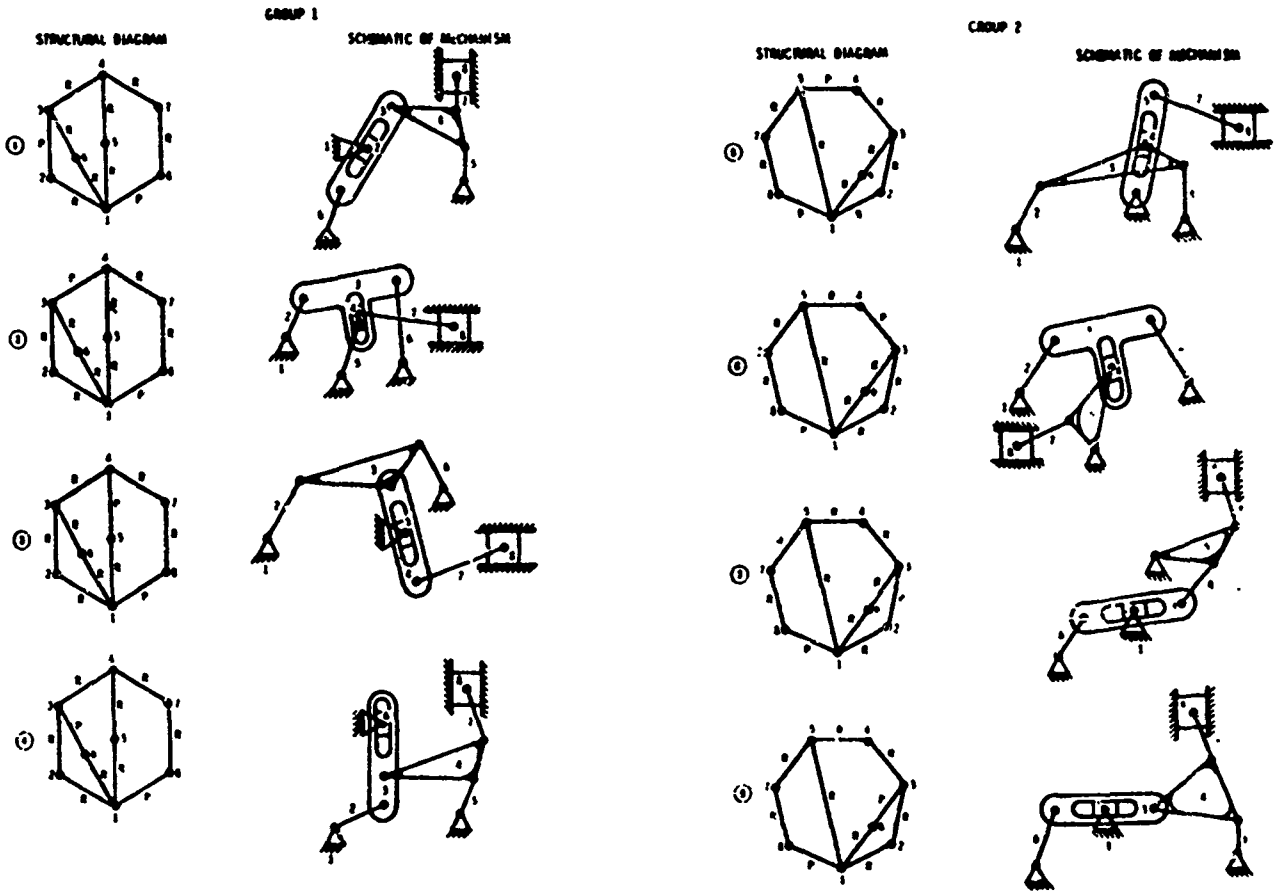



FIG.3.

STRUCTURAL AND SCHEMATIC DRAWINGS OF SOME EIGHT-LINK VARIABLE-STROKE MECHANISMS.

Reprinted with the permission of the American Society of Mechanical Engineers from "Development of an optimum variable-stroke internal-combustion engine mechanism from the viewpoint of kinematic structure" by F. Freudenstein and E.R. Maki; Trans. ASME, J. of Mechanisms, Transmissions and Automation in Design 105, 1983, pp. 259-266.

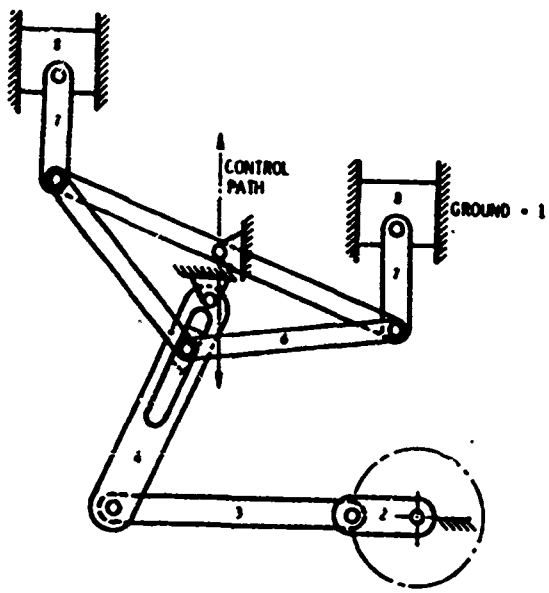



FIG. 4.

OPTIMUM VARIABLE-STROKE MECHANISM

Reprinted with the permission of the American Society of Mechanical Engineers from "Development of an optimum variable-stroke internal-combustion engine mechanism from the viewpoint of kinematic structure" by F. Freudenstein and E.R. Maki, Trans.ASME, J. of Mechanisms, Transmissions and Automation in Design 105, 1983, pp. 259-266.

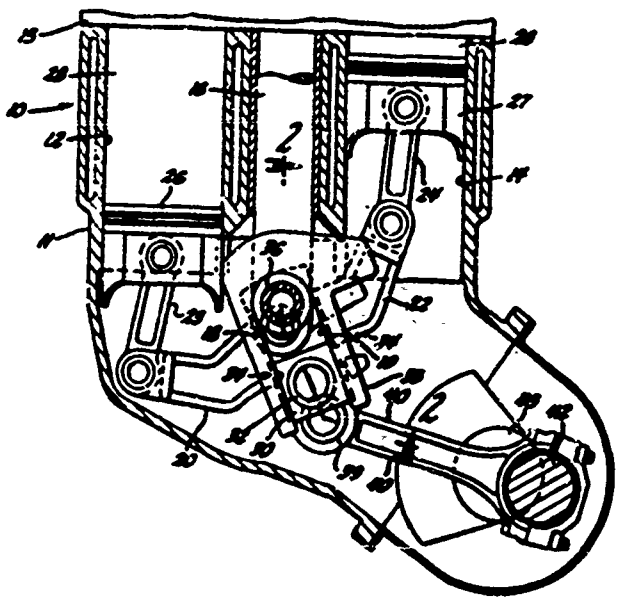


FIG. 5

FINAL DESIGN OF OPTIMUM VARIABLE-STROKE MECHANISM OF FIG.4 (U.S. PAT. #4,270,495; 1981)

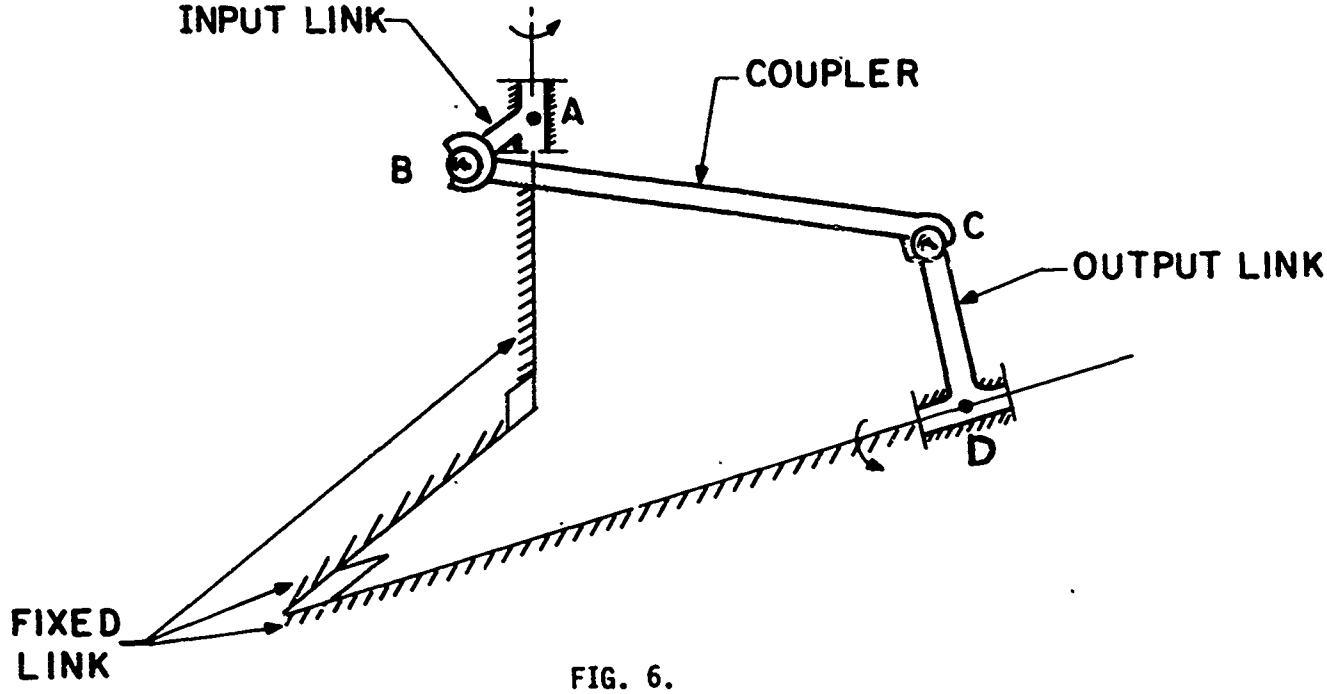



FIG. 6.

SKEW FOUR-BAR LINKAGE

A-ME-3687 JD

# VALIDATION OF A MULTILAYERED VISCOELASTIC SEISMIC PROPAGATION MODEL FOR SHORT TO MEDIUM RANGES

Ben L. Carnes, P.E.
Environmental Systems Division
Environmental Laboratory
U.S. Army Engineer Waterways Experiment Station
Vicksburg, MS 39180

Abstract. Field tests were conducted to provide a complete data base for validation of the WES Seismic Propagation Model. The model was developed in 1973 to predict seismic signals resulting from a variety of seismic sources, but had not been validated for short to medium ranges (up to 10 km). Model predictions compared favorably with tests results and after a refinement of site physical properties and internal damping factor (or g), calculations were even more compatible. The model will now be available for making predictions at other sites with a greater degree of reliability. The validation procedure included preliminary prediction, testing and verification of data, and model refinement and validation.

Introduction. The Rayleigh wave components of seismic waves are generally considered as the primary energy conveyance that activates seismic sensors. The amount of energy generated by a target of military interest, such as a vehicle, military equipment, or a specific military operation or activity, is site dependent, and the effect of terrain on Rayleigh wave propagation is known to be substantial. A mathematical computer model was developed at the U. S. Army Engineer Waterways Experiment Station (WES) to predict microseismic signals in various terrain materials for application to sensor systems (Lundien and Nikodem 1973). The model uses site information including compression and shear wave velocities and material density and thickness for multiple layers, along with seismic source data, to simulate the generation, coupling, propagation, and decoupling of seismic signals (surface waves) from the source to the sensor. This concept is illustrated in Figure 1, which shows the seismic wave propagating from the source to the sensor.

Model Description. The model develops mathematics within a framework of theoretical mechanics to simulate the Rayleigh component of a seismic wave in a layered medium. Rayleigh waves, or elastic surface waves, are formed when a bounding surface exists on an elastic half space. The particle motion of Rayleigh waves is elliptically retrograde in a plane perpendicular to the bounding surface and parallel to the direction of propagation. The Rayleigh wave propagation velocity is smaller than that of compression or shear waves and in most soils is generally about 90 percent of the shear wave velocity, which in turn is about 40 to 50 percent of the compression wave velocity. A Rayleigh wave's energy per unit area (usually expressed as a form of particle motion) decreases with increasing range (as it spreads over an increasing area). This is termed geometric attenuation, which causes a decrease in particle motion amplitude proportional to $1/\sqrt{r}$ where $r$ is the range from the seismic source. Of the total seismic energy input by the source, approximately two-thirds is transmitted away by a Rayleigh wave.

In addition to geometric attenuation, there is also a loss of energy from internal friction in which some of the mechanical energy is converted into heat. This internal damping increases with frequency. Rayleigh waves with short wavelengths (high frequencies) are attenuated more rapidly with depth than those with long wavelengths (low frequencies). When a Rayleigh wave travels in a horizontally layered medium in which its velocity is constant within a layer but different from layer to layer, the wave will be dispersed as it propagates. Because of dispersion, the Rayleigh wave can have a low velocity (and high frequency) in the surface layer and a high velocity (and low frequency) in the basement layer, depending on layer properties.

Another effect of a horizontally layered medium is that elastic waves originating at the surface may be reflected or refracted by the layers. These changes cause energy losses that also affect attenuation. The waves can also be polarized or even converted to other wave forms. The effects of macrogeometric features can be simulated by frequency-dependent amplitude transmission coefficients.

Closed-form equations for computing the seismic signal are derived from classical equations of motion to define the stress-strain relationship in the propagating medium. The development of an economical and practical mathematical simulation of the phenomena shown in Figure 1 requires several assumptions for simplification. The assumptions describing the geometry and composition of the system are: (a) the medium is composed of homogeneous and isotropic horizontal layers, (b) all layer boundaries are smooth and abrupt, and (c) major discontinuities at the surface are large compared with the signal wavelength. The assumptions describing wave motion are: (a) all predictions are made for a point source, which means that source contact dimensions are small compared to the signal wavelength, (b) all motions are at a level low enough to remain within the elastic realm of the material, and (c) the calculations are made using axial symmetry (all inputs are vertical or horizontal).

The equations of motion used herein represent the propagation of a disturbance that involves both equivoluminal and irrotational motion. Separate wave equations can then be obtained for each type of motion. These equations are translated to cylindrical coordinates so that axial symmetry can be applied. They are then solved by applying boundary values. The end result is the general expression shown in Figure 2. It should be noted that the more general form of this expression makes use of Fourier transforms to widen the scope of problems that can be evaluated.

The conceptual steps for prediction of seismic signals are shown in the block diagram in Figure 3. This diagram is a generalization of the actual computational sequence used in the computer program. The essential steps shown in Figure 3 are used to explain the computational procedures that are basically composed of two parts: (a) computation of the site coefficients, and (b) transformation of the input forcing functions to a predicted signal.

The site coefficients that must be determined to predict a signal include (a) wave numbers, (b) source coupling coefficients, (c) Hankel function (i.e., the transmission coefficients), and (d) surface macrogeometry coefficients.

After the source coupling, transmission, and surface macrogeometry coefficients are computed, they are combined for each wave number of interest. This is accomplished by taking products of the three coefficients for each frequency and mode and summing the model results, one sum for each frequency.

The site coefficients described above are used with a stress-time function and particle motion coefficients to arrive at a prediction of particle velocity, particle displacement, or particle acceleration (see Figure 2). The stress-time function and the site coefficients are combined to arrive at a prediction of a signal of particle motion versus time at some range r . To arrive at a predicted signal in time domain, the inverse Fourier transform must be taken of the product of combined site coefficients, the stress-time Fourier coefficients, and the particle motion coefficients. This time domain signal is repetitive as a result of the periodic input source.

Test Program. Since its development, the WES seismic model has been used for analysis of data for numerous seismic sensor systems that deal only with short ranges (up to 1 km). Present military systems require knowledge of activities producing seismic signals out to 10 km and beyond. The Army Seismic Attenuation Test (SAT) program was conducted in the spring of 1983 at the U. S. Army White Sands Missile Range to provide a database of long-range (up to 10-km) data for model refinement and to document other phenomena. The WES seismic model was used to make predictions for the tests, using preliminary seismic data for the alluvial valley test site. These predictions are shown in Figures 6-8 as phase velocity, transmission coefficients, and transfer functions. Tests were conducted using an electrohydraulic vibrator, a vacuum impulse loader, an M-35 vehicle out to 1 km, and explosions from 1 to 10 km. Data were recorded at a stationary array of geophones and on a portable group of geophones emplaced near the source. Approximately 2000 channels of data were recorded during this extensive test program (See Figure 8 for site layout).

Analysis. The analog data were digitized and then processed using several signal processing techniques, such as band pass filtering, Fourier transforms, and cross correlation to analyze attenuation, transfer function, source/array spectral densities, time of arrival (mode analysis), and time/frequency analysis. Figures 8-11 show some of the processed data, while Figures 12-14 show attenuation curves for vibration and explosive tests and calculated transfer functions. Model verification was begun upon completion of the data analysis.

To compare actual data with the model predictions, processed data were used to calculate attenuation ratios for various tests, as can be seen in Figures 15 and 16. The model was then used to match attenuation ratios for the same frequencies by varying the internal damping factor (IDF). The results can be seen in Figures 18 and 19, which show IDF versus frequency. A rule-of-thumb estimate of 0.03 for IDF was used to make all predictions for the SAT program. The data reveal that the 0.03 value is not a bad estimate, but that the calculations could be refined somewhat by having an IDF of 0.02 or a value dependent on range.

The effect on the predicted transfer functions as a result of changing the IDF can be seen in Figure 19. The short-range (0.1-, 0.5-, and 1.0-km) site transfer functions for an IDF of 0.02 compare very favorably with both the actual data and the functions for an IDF of 0.03. The long-range (2-, 4-, and 10-km) site transfer functions for an IDF of 0.02 compare more favorably with the actual data than the long-range functions for an IDF of 0.03.

Conclusions. This analysis is not complete. Further model refinements now being attempted include varying IDF with frequency and range and using overall site information determined by the field tests. Further data analysis also being accomplished with the field test data includes averaging of data for a number of geophones to reduce the problem of background noise on data that are totally masked by wind noise; verifying Rayleigh wave motion using vertical, radial, and transverse geophone data to plot particle motion; and calculating the point at which transfer functions are affected by background noise.

The conclusions reached in this study are: seismic waves are dispersed such that very low-frequency (2- to 3-Hz) signals which travel in deep, high-velocity layers are the only component remaining beyond 4 km; the resolution of seismic signals measured out to 10 km is below 0.5 Hz; and wind noise is a problem in making seismic measurements over long ranges for wind speeds above 8 m/sec.

This study has been conducted as a "generic" study in order that the results can be applied to many programs, especially those for which the WES seismic model can be used to extrapolate and predict information for short- and medium-range seismic signals. In addition, the field test program has enabled the use of the WES seismic model for prediction of medium-range seismic information, and upon completion of the current analysis will provide not only model verification but also insight into capabilities of military equipment for seismic detection.

This study has been documented in a WES Technical Report entitled "An Analysis of Short- to Medium-Range Seismic Attenuation Tests Using a Multilayered Viscoelastic Seismic Propagation Model" (Carnes and Lundien 1984).

## References

Carnes, B. L., and Lundien, J. R. 1984. "An Analysis of Short- to Medium-Range Seismic Attenuation Tests Using a Multilayered Viscoelastic Seismic Propagation Model" in preparation, U. S. Army Engineer Waterways Experiment Station, CE, Vicksburg, Miss.

Lundien, J. R., and Nikodem, H. 1973. "A Mathematical Model for Predicting Microseismic Signals in Terrain Materials," Technical Paper M-73-4, U. S. Army Engineer Waterways Experiment Station, CE, Vicksburg, Miss.

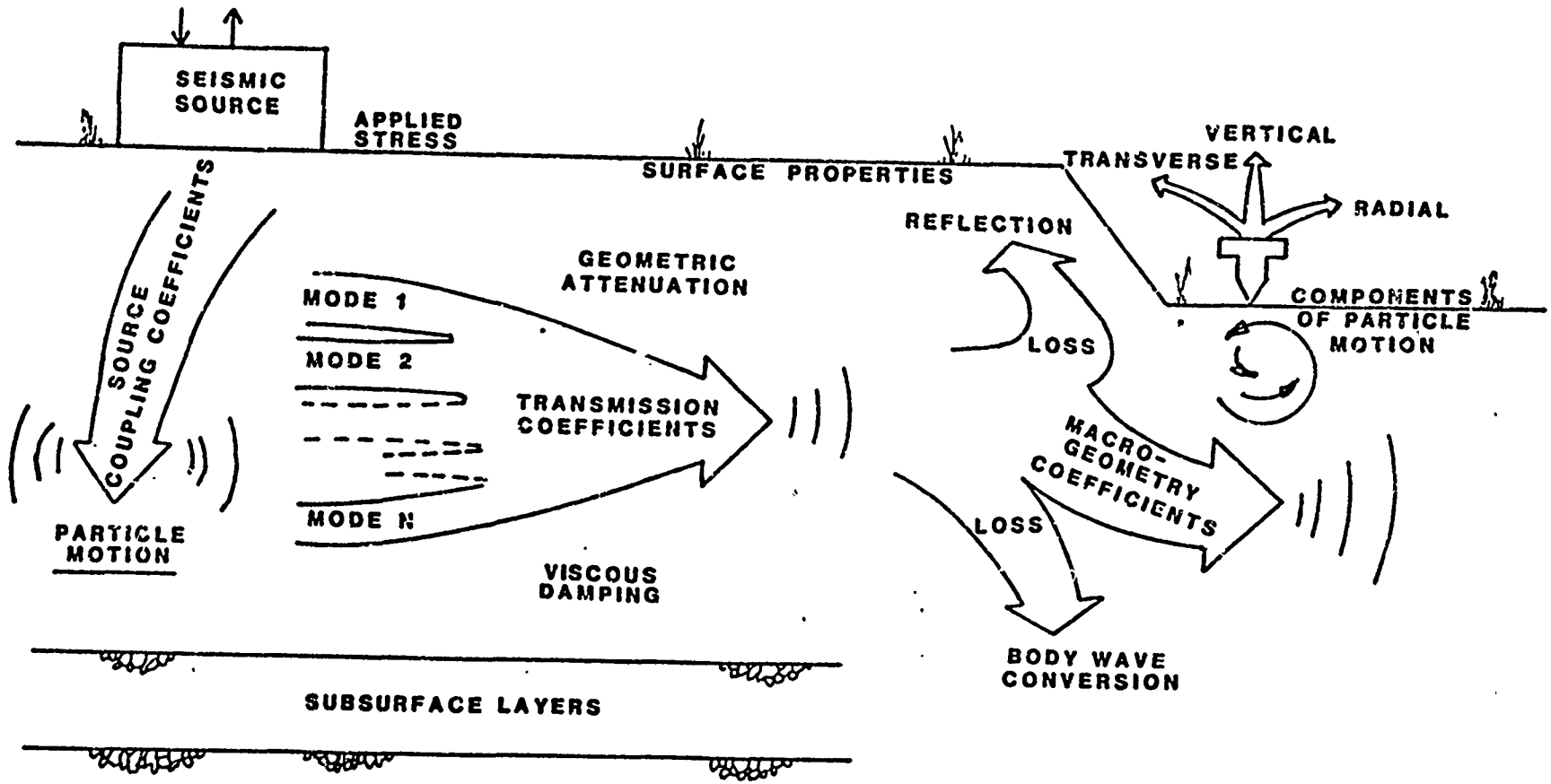



Figure 1. Concept of seismic wave propagation.

$$A_{\ell p}(r,t) = \sum_{n=-a}^{n=a} \left[\omega_n e^{i(\pi/2)}\right]^{p-1} D_n e^{i\omega_n t} \sum_{m=1}^{m=b} S_{m,n} B_{m,n,\ell} H^{(1)}_{(\ell-1)}(\kappa_{m,n} \cdot r)$$

where

$A_{\ell,p}(r,t)$ = particle motion, in cm, as a function of range r and time t for $\ell^{th}$ component and $p^{th}$ derivative

$\ell$ = 1, 2; 1 for vertical component and 2 for radial component

p = 0, 1, 2; 0 for particle displacement, 1 for particle velocity, and 2 for particle acceleration

r = range from source, m

t = time, sec

n = -a, -a+1, ... -1, 0, +1, +2, ... a-1, a; frequency numbers

$\omega_n$ = circular frequency, rad/sec

$i = \sqrt{-1}$

$D_n$ = Fourier coefficient for source signal:

$$D_n = \frac{1}{T}\int_0^T f(t)e^{i\omega t}\, dt$$

T = period of source signal, sec

f(t) = source signal, dynes

m = 1, 2, ... b; mode numbers

$S_{m,n}$ = surface macrogeometry coefficients

$B_{m,n,\ell}$ = source coupling coefficients = $\omega\pi\kappa^2_{m,n}\ \mathrm{Res}_{\kappa=\kappa_{m,n}}[Q(\kappa)]$ for $\ell = 2$

$= -\omega\pi_{m,n}\ \mathrm{Res}_{\kappa=\kappa_{m,n}}[W(\kappa)]$ for $\ell = 1$

$H^{(1)}_{(\ell-1)}(\kappa_{m,n} \cdot r)$ = Hankel function (Bessel function of third kind) of first type and order $(\ell-1)$ for argument $(\kappa_{m,n} \cdot r)$

$\kappa_{m,n}$ = wave number for $m^{th}$ mode and $n^{th}$ frequency

Figure 2. General solution

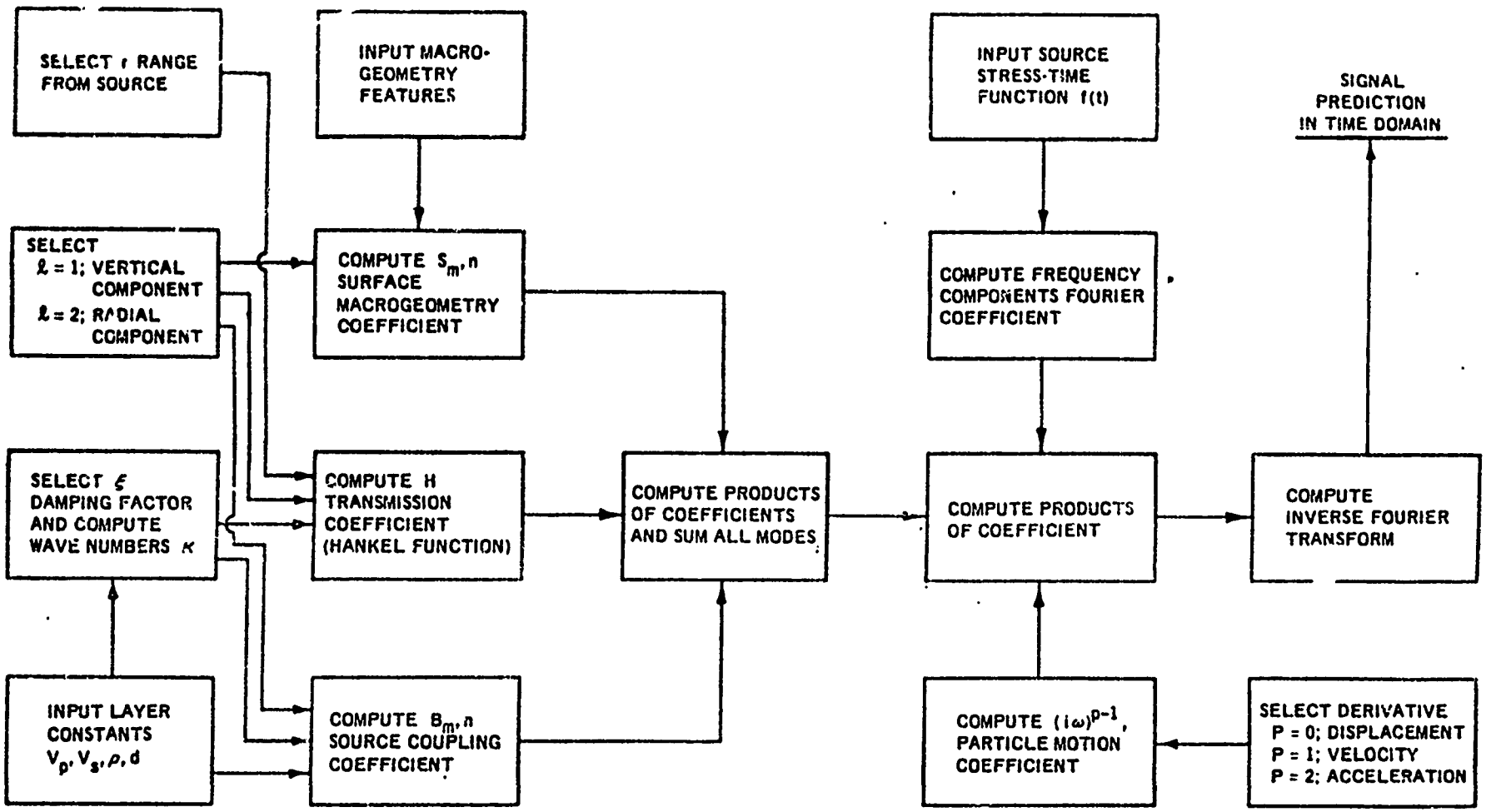
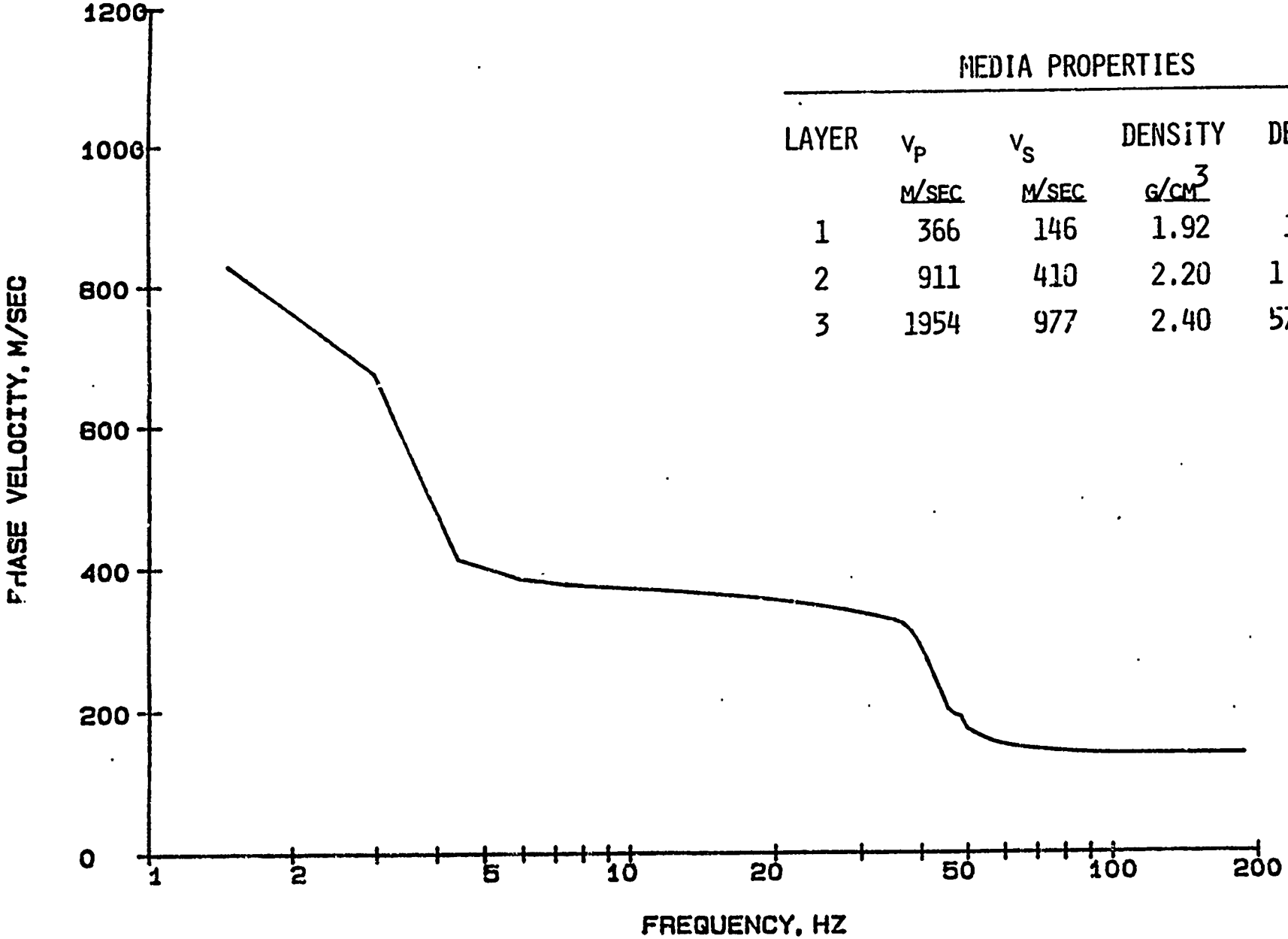



Figure 3. Prediction of Seismic Signals




MEDIA PROPERTIES

| LAYER | $V_P$ M/SEC | $V_S$ M/SEC | DENSITY G/CM$^3$ | DEPTH M |
|---|---|---|---|---|
| 1 | 366 | 146 | 1.92 | 1.52 |
| 2 | 911 | 410 | 2.20 | 1.52-57.0 |
| 3 | 1954 | 977 | 2.40 | 57.0+ |

Figure 4. Rayleigh wave phase velocity calculated for SAT site by WES seismic model

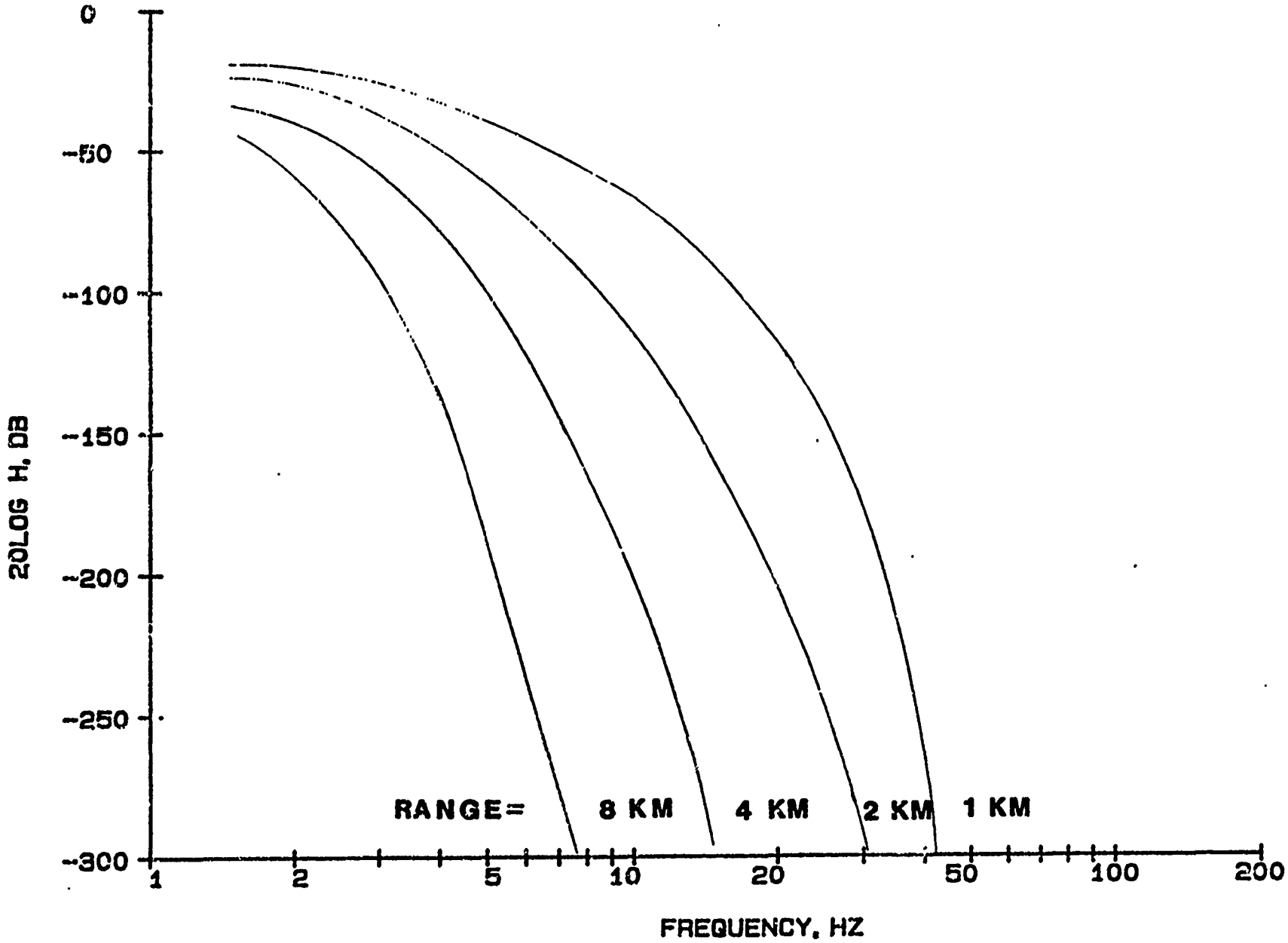



Figure 5. Transmission coefficient H vs. frequency predicted by model

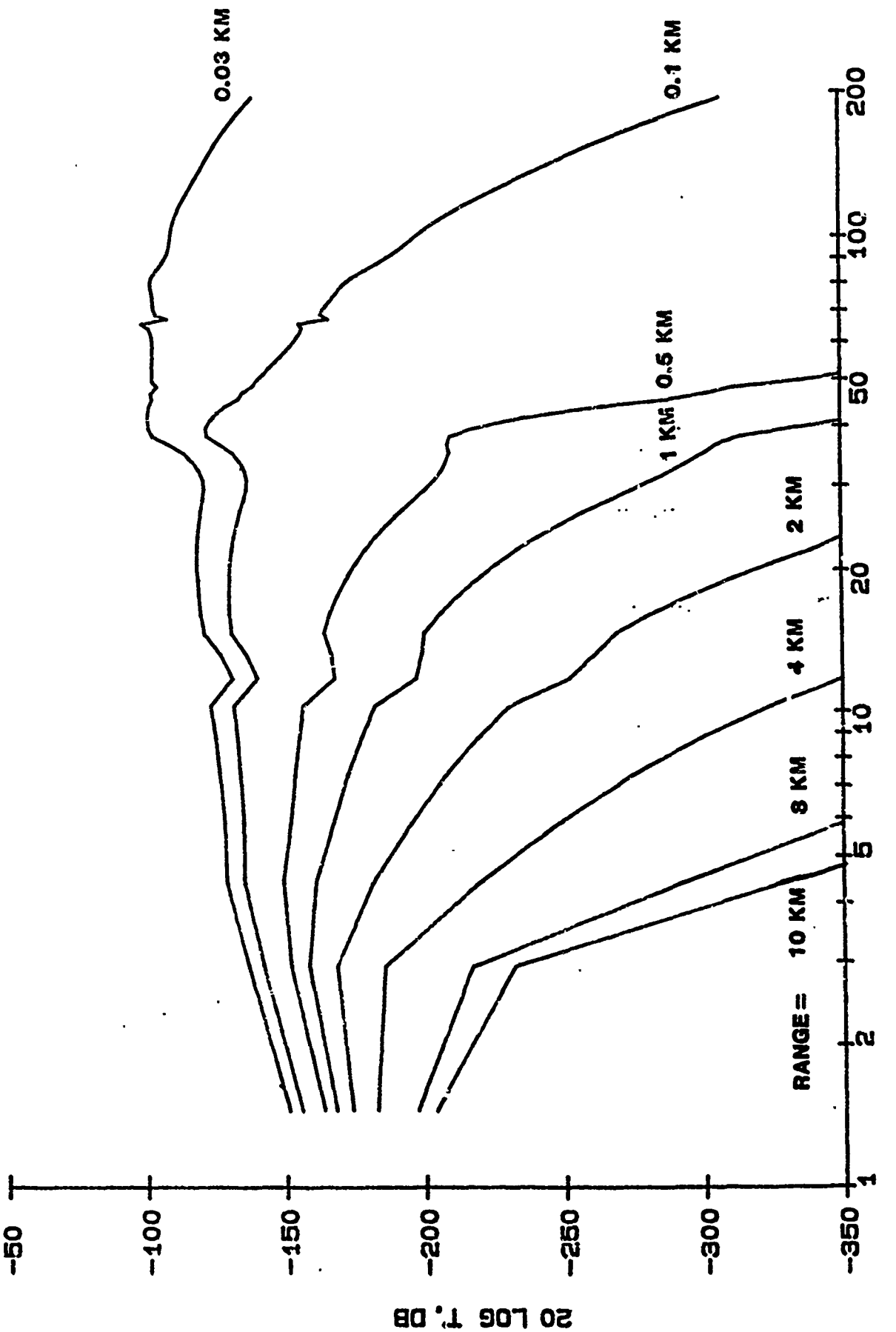



Figure 6. Predicted transfer functions for SAT site.

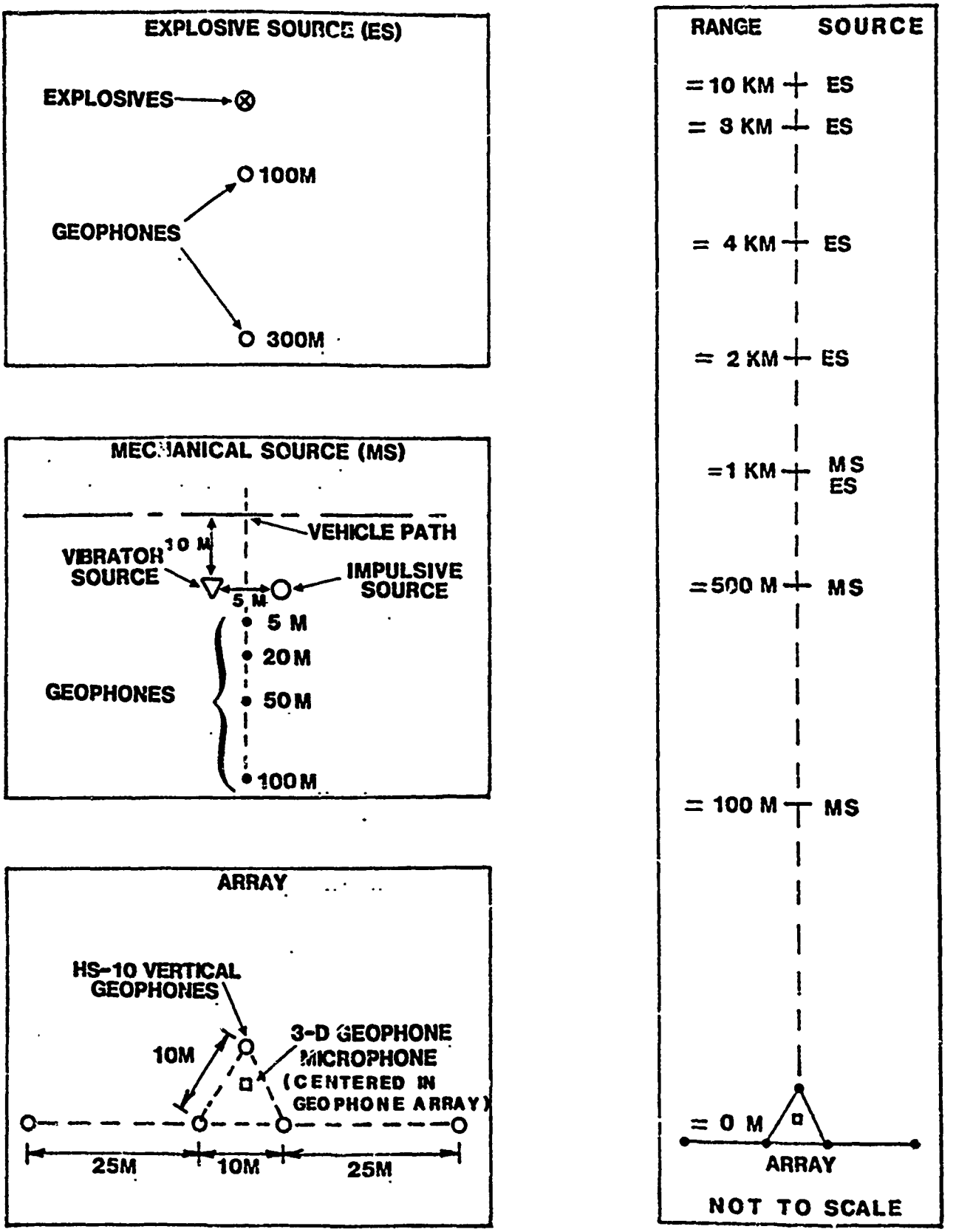



Figure 7. SAT site layout.

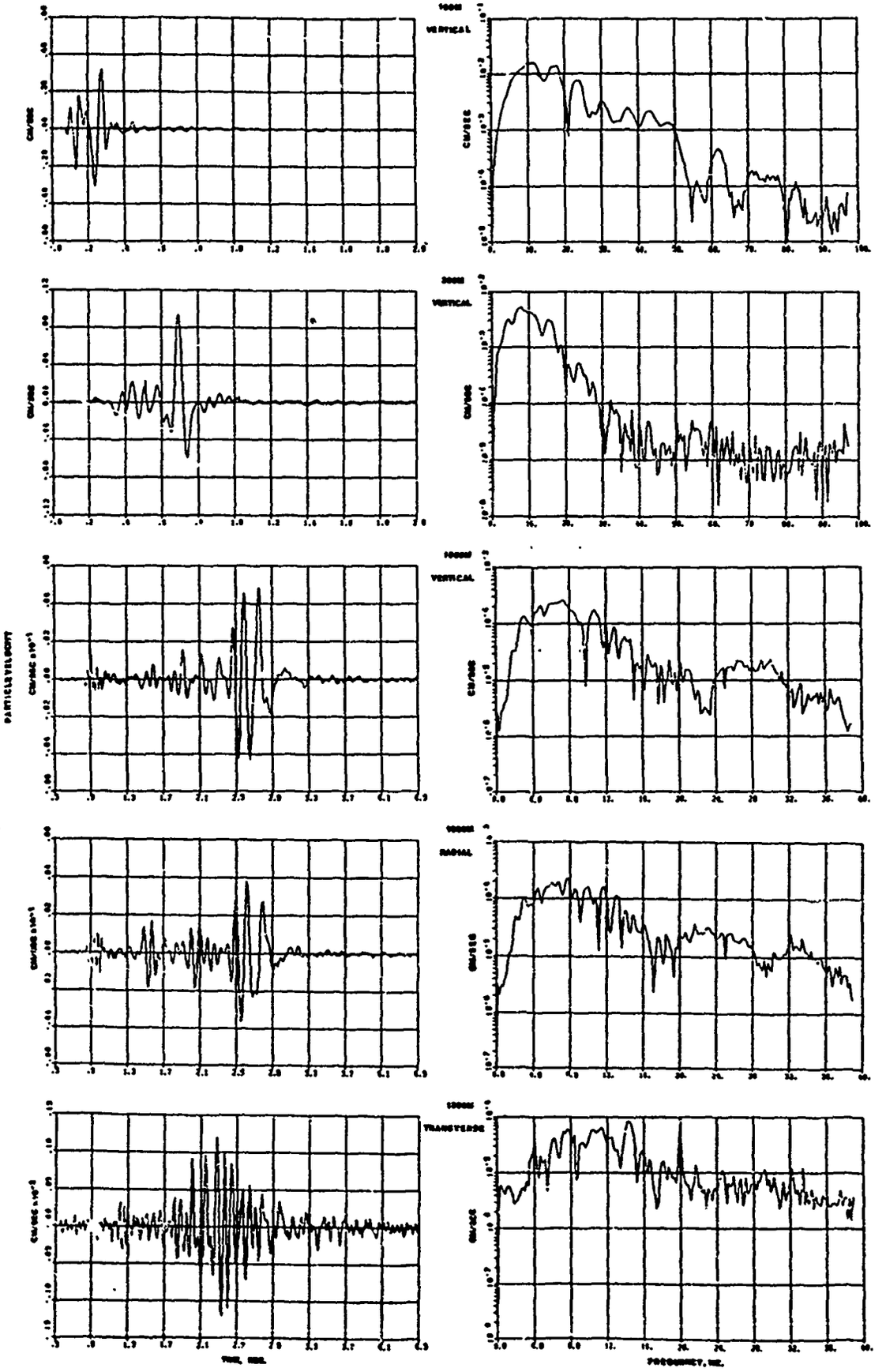



Figure 8. Test 133, Explosive Source at 1km

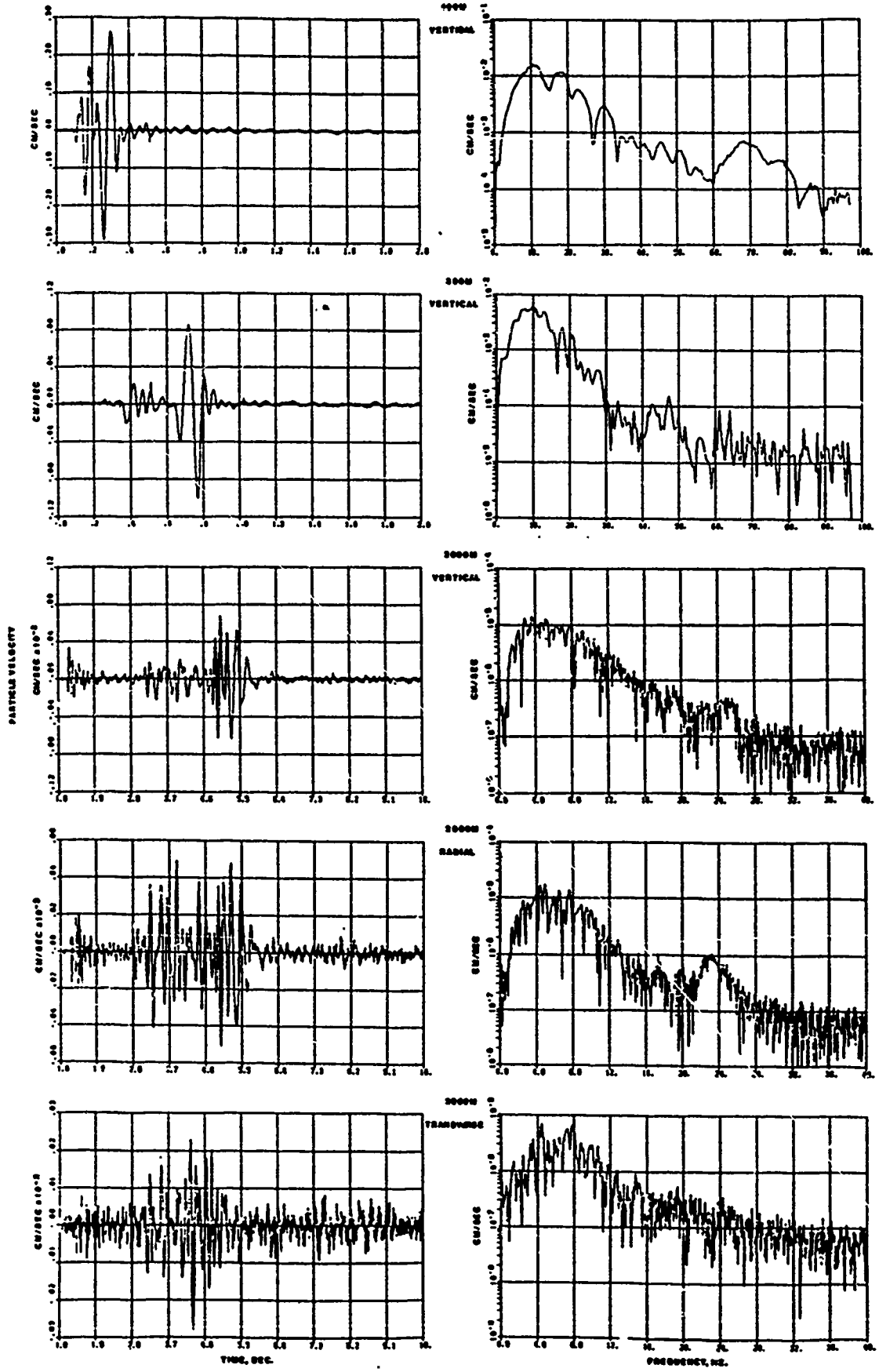



Figure 9. Test 132, Explosive Source at 2 km

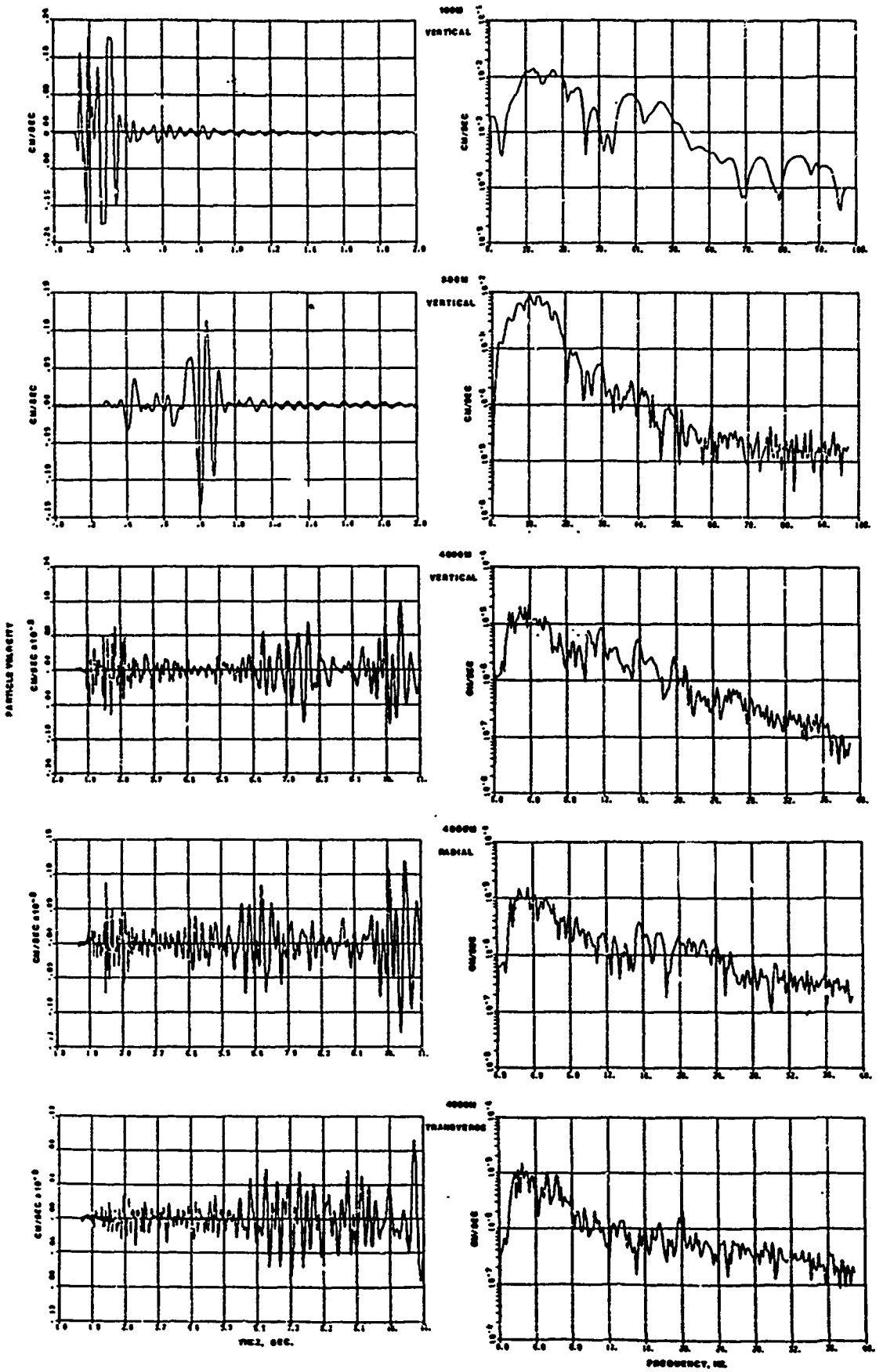



Figure10.. Test 131, Explosive Source at 4 km

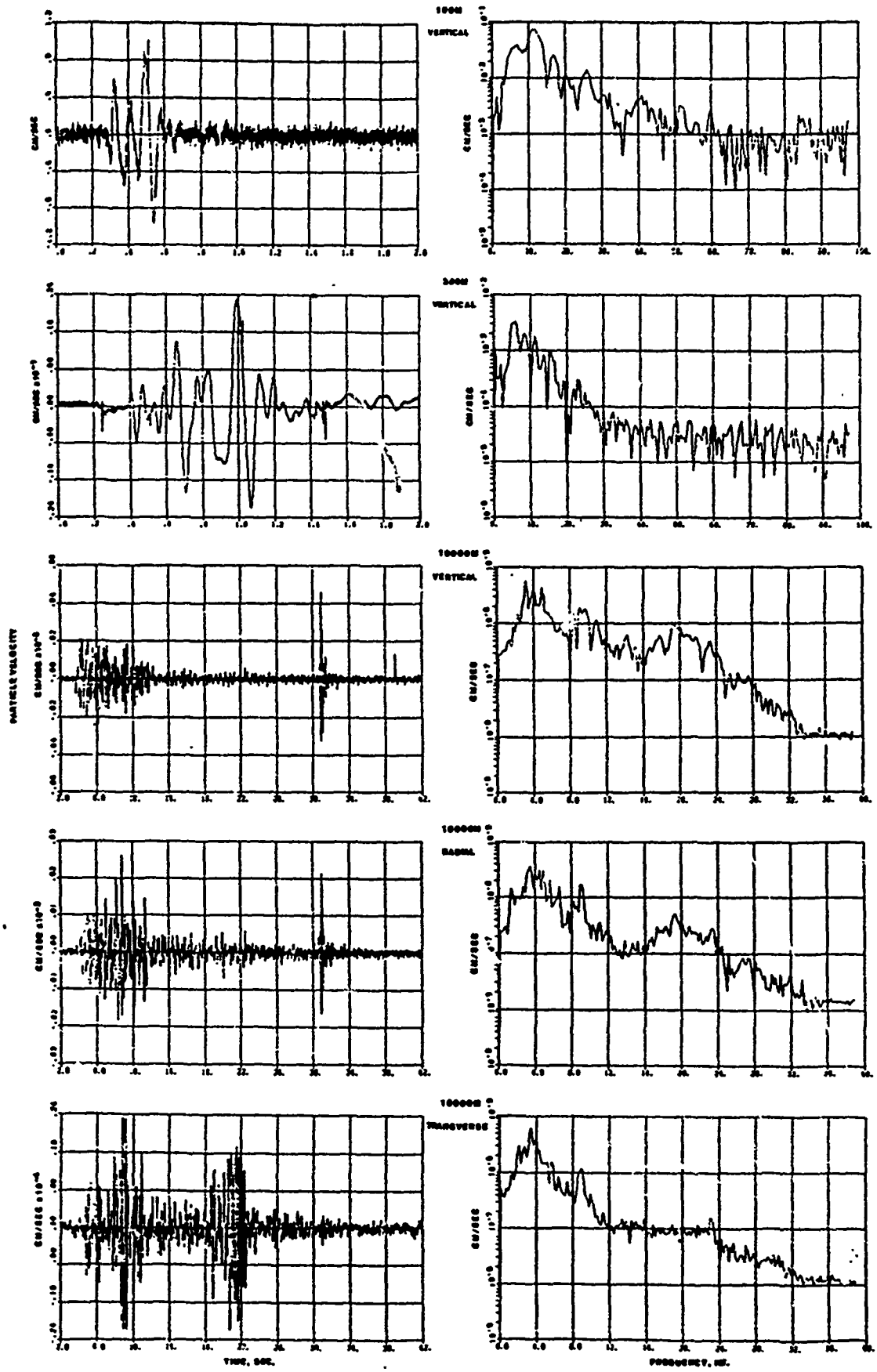


Figure 11. Test 136, Explosive Source at 10 km

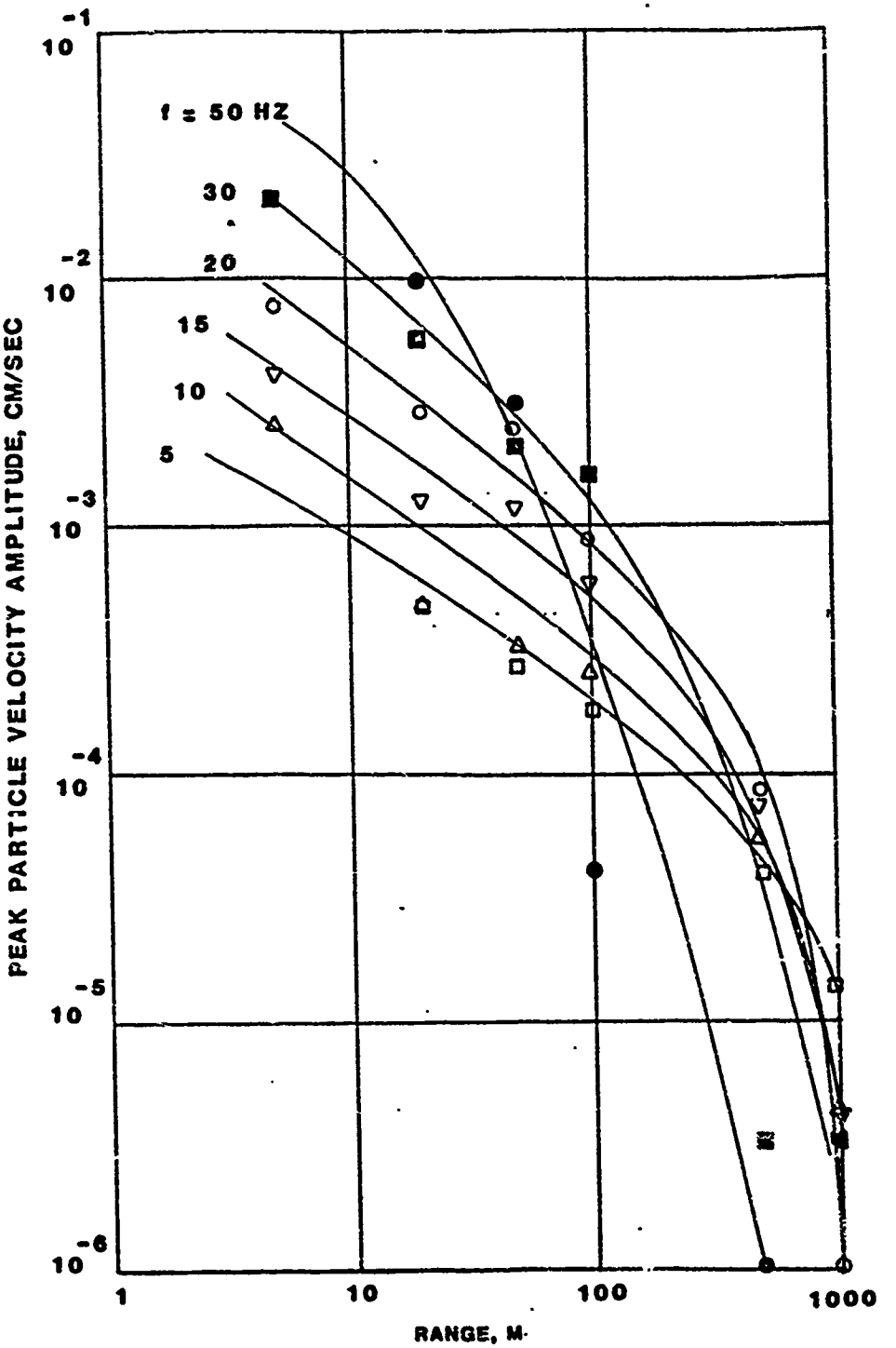



Figure 12. Attenuation curves for discrete frequency vibration tests

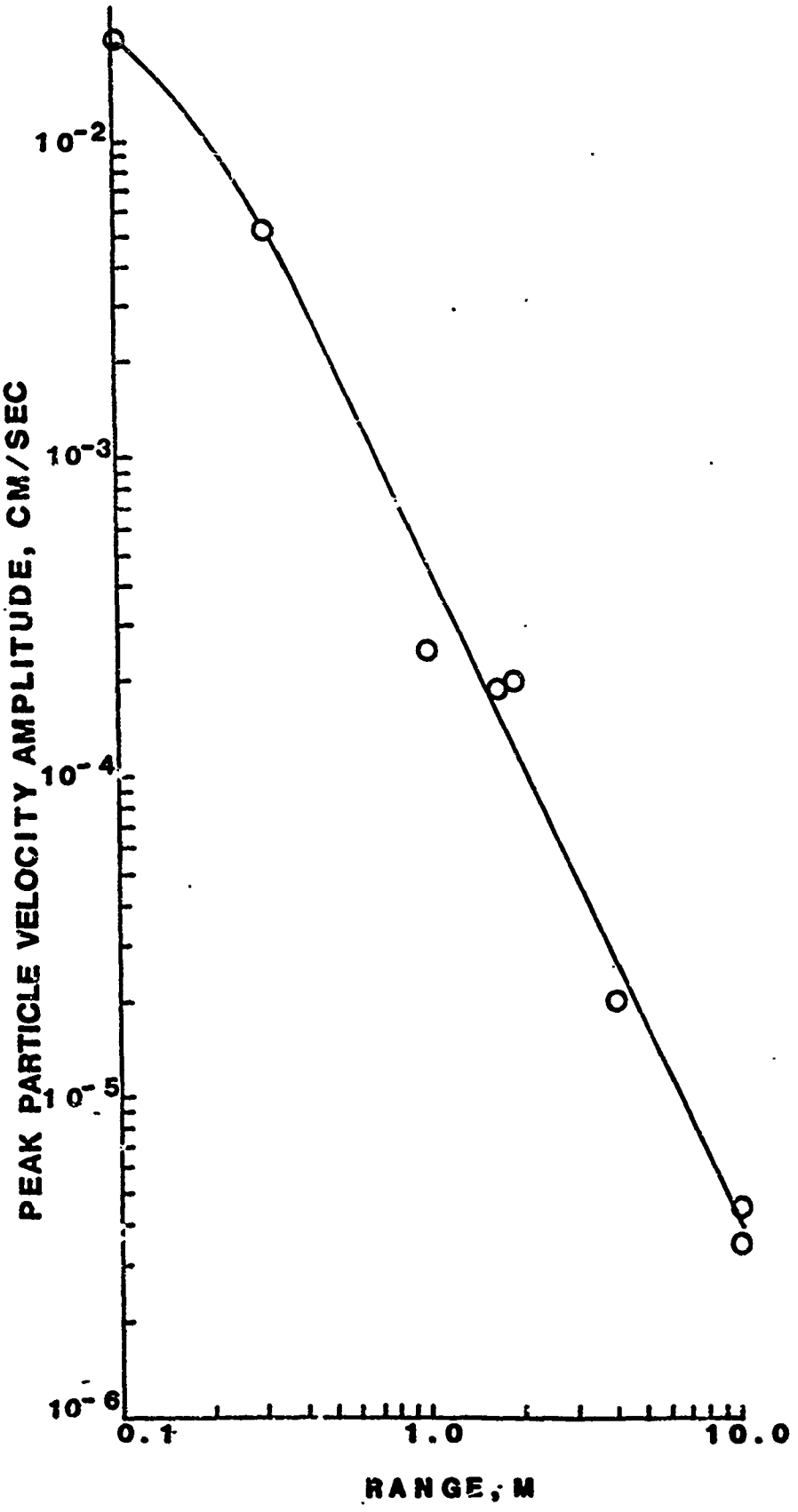



Figure 13. Attenuation curve generated from explosive-source test data

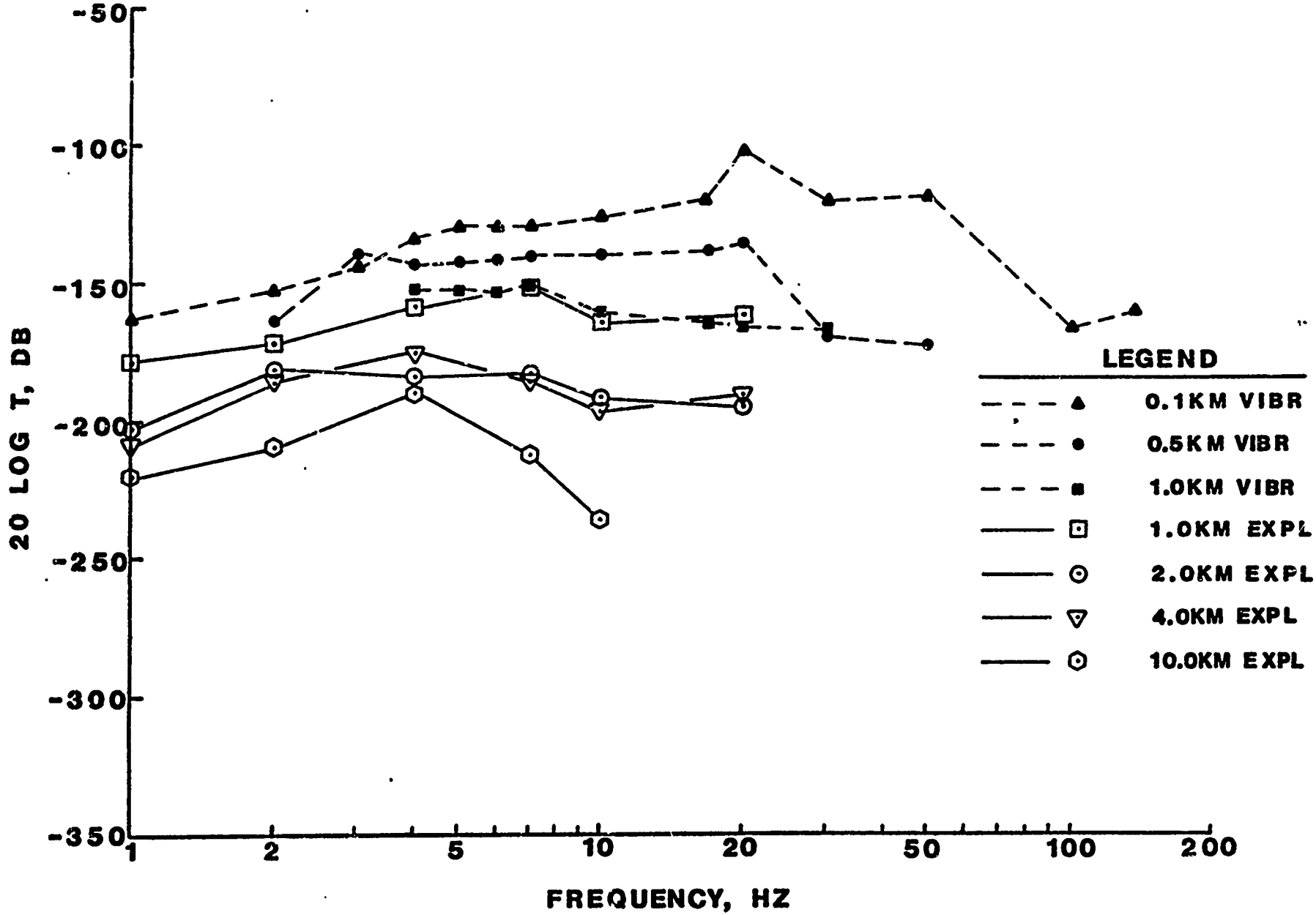



Figure 14. Transfer function T vs. frequency calculated for discrete frequency vibration tests (top three curves) and for explosive - source tests.

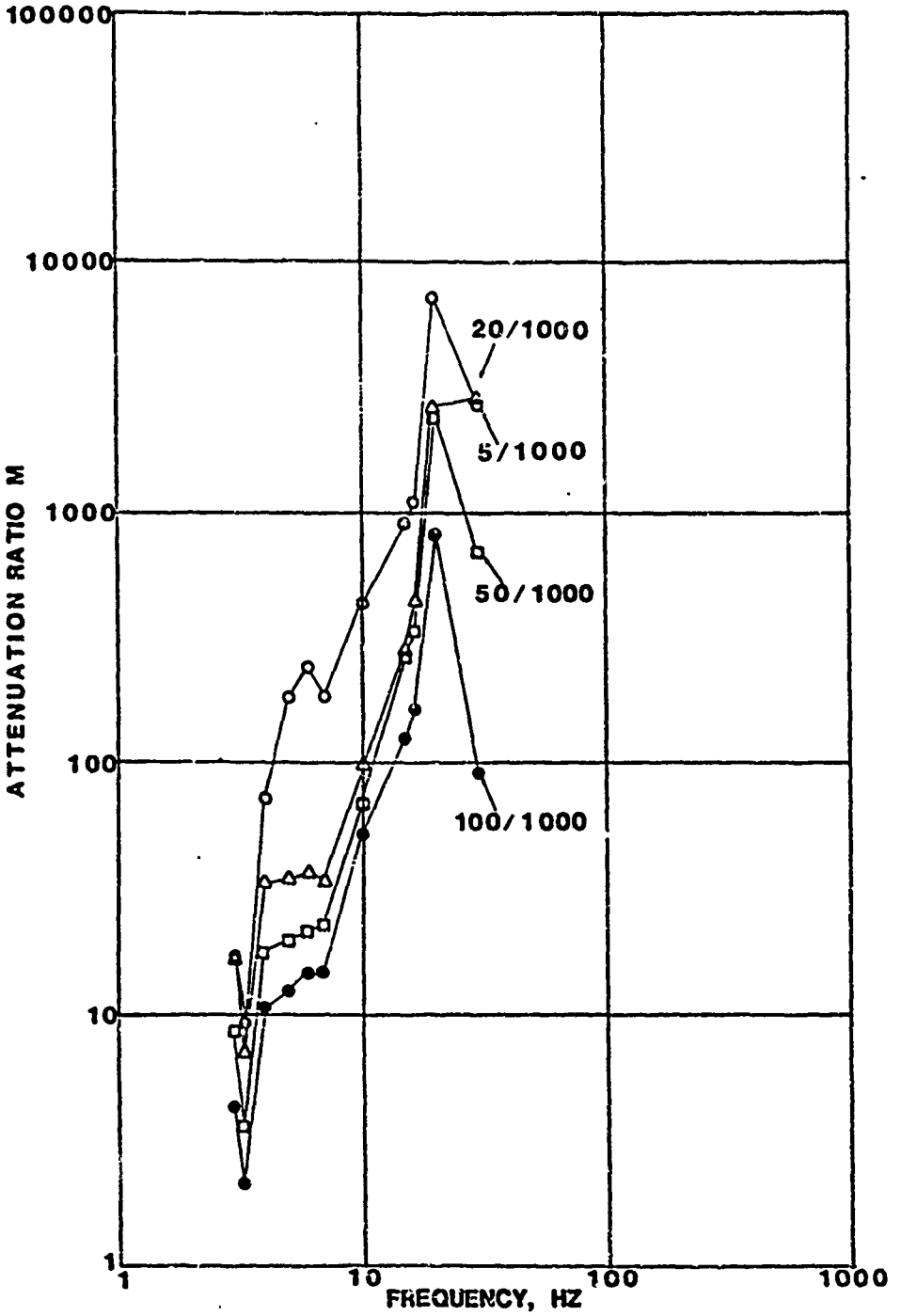



Figure 15. Attenuation ratio M vs. frequency for 1-km discrete frequency vibration tests. Numbers indicate locations of compared data.

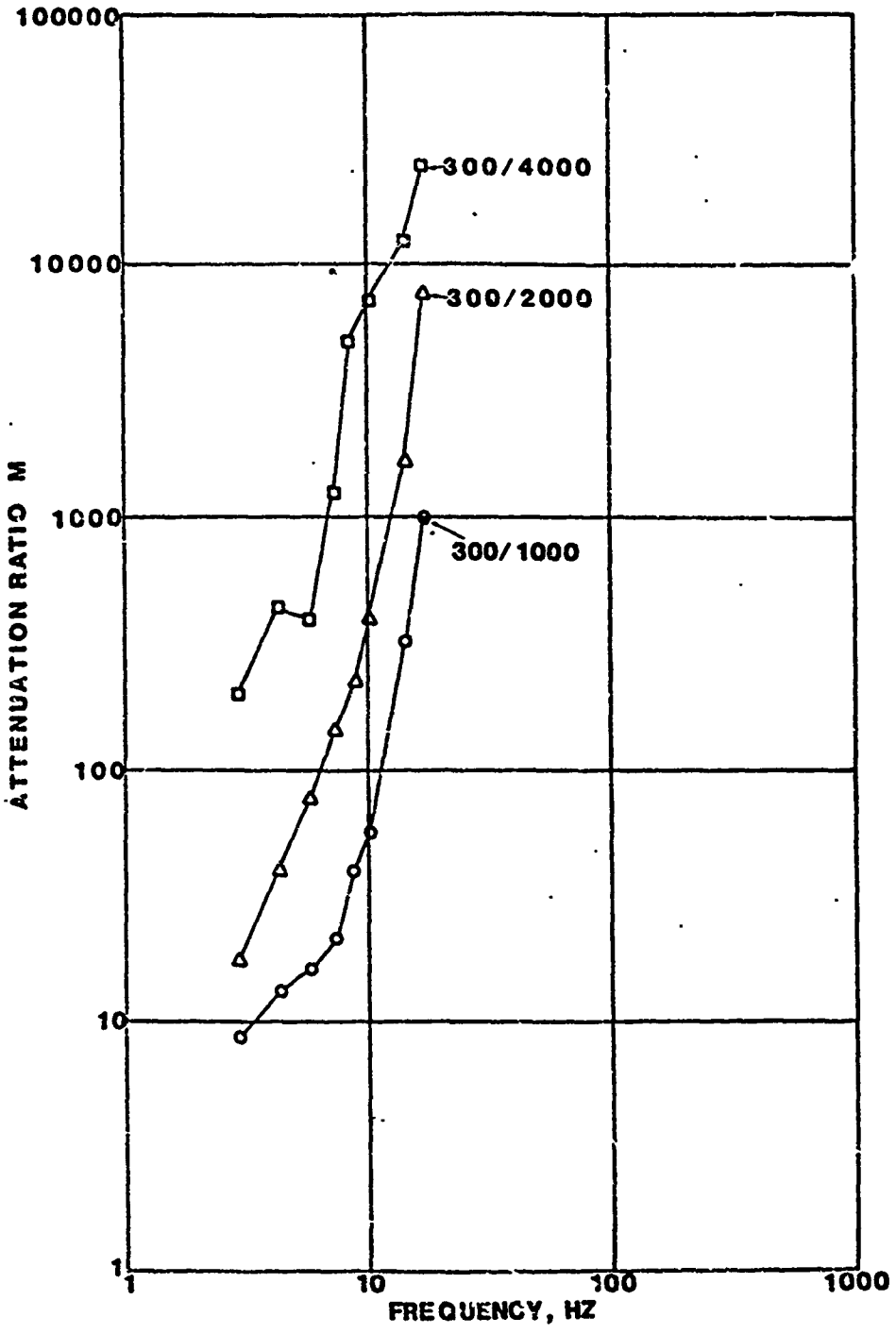



Figure 16. Attenuation ratio M vs. frequency for 1-km, 2-km, and 4-km explosive-source test data. Numbers indicate locations of compared data.

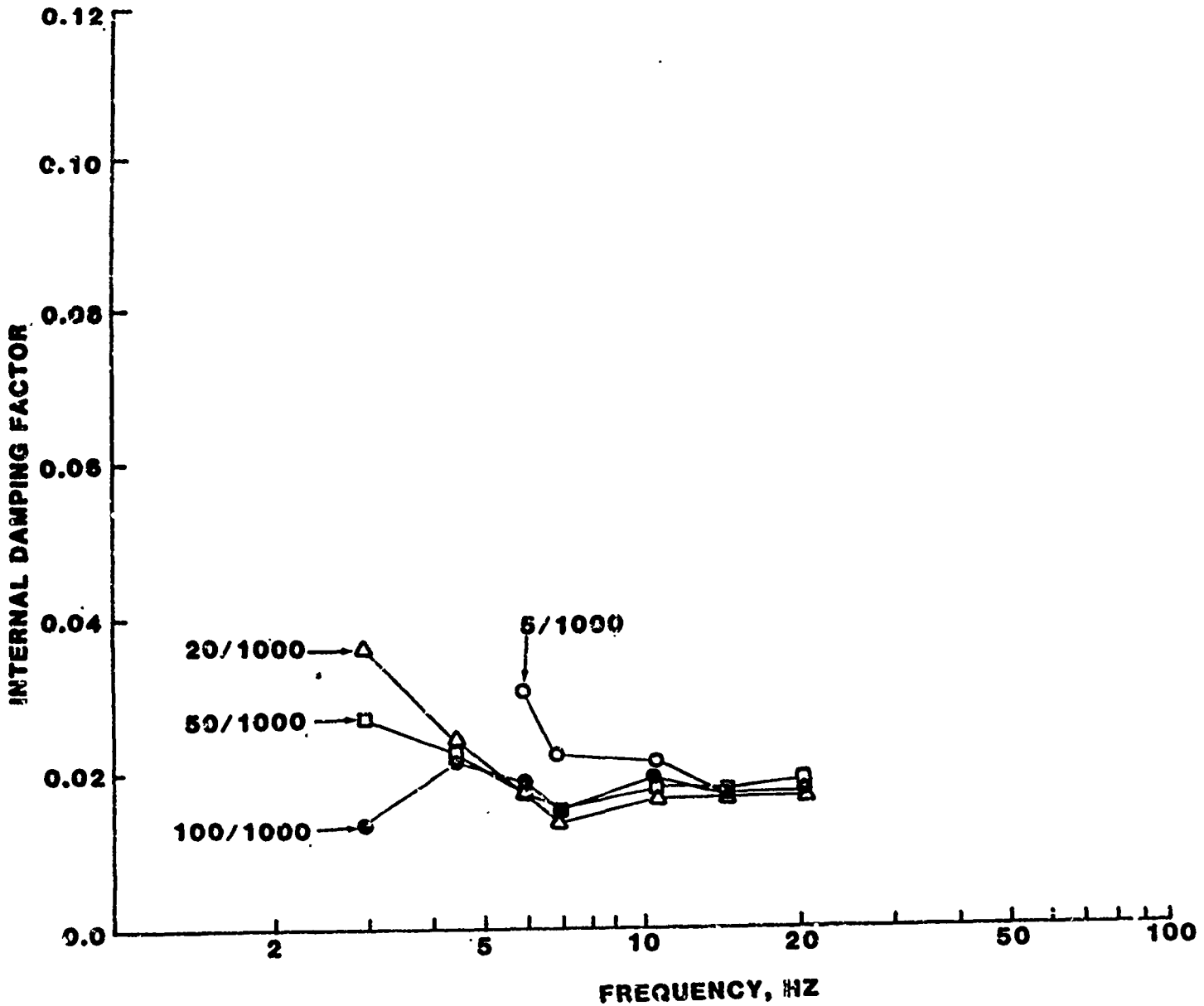



Figure 17. Internal damping factor vs. frequency calculated from attenuation ratio for 1-km discrete frequency test data. Numbers indicate locations of compared data.

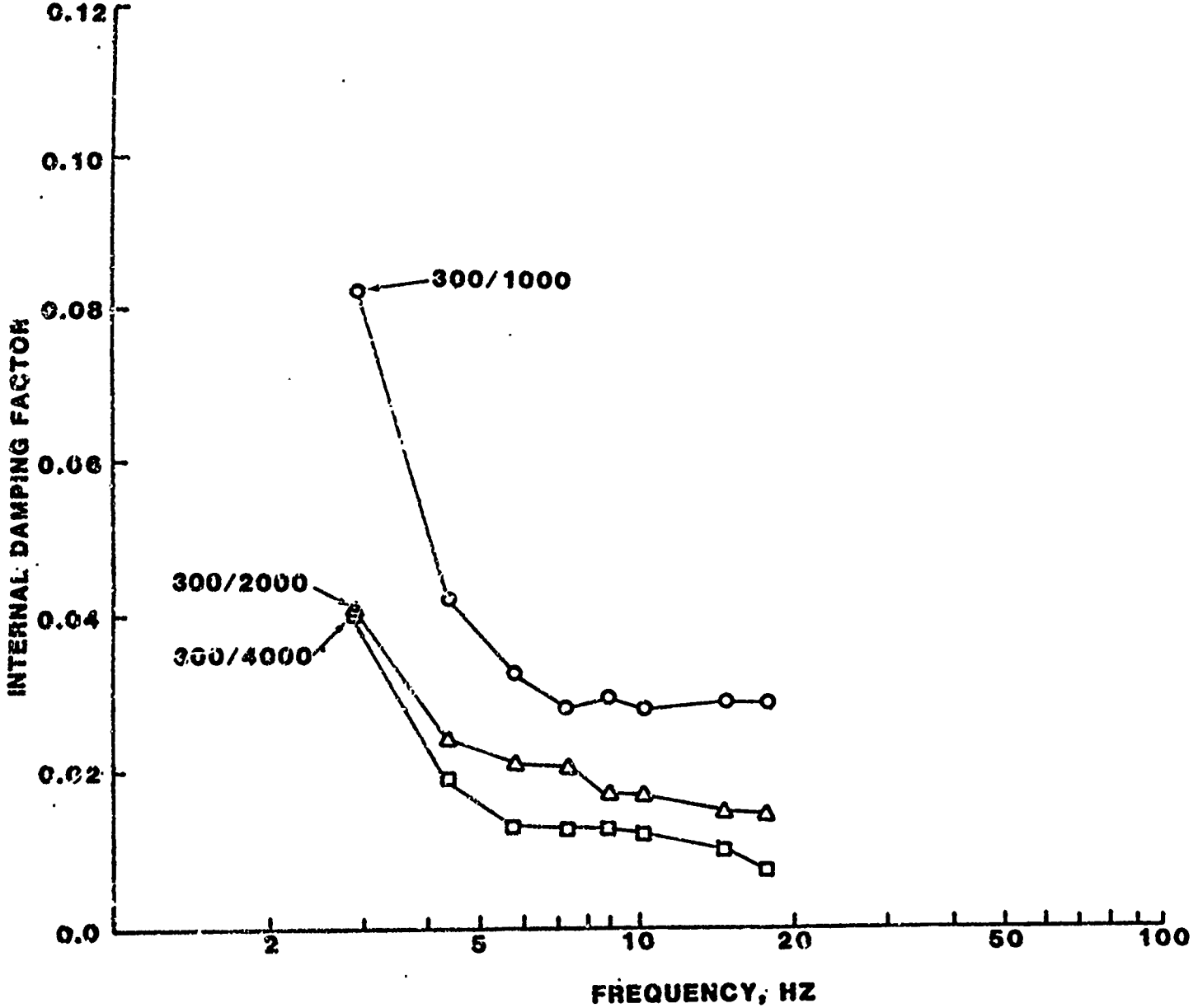



Figure 18. Internal damping factor vs. frequency calculated from attenuation ratio for explosive-source testsdata. Numbers indicate locations of compared data.

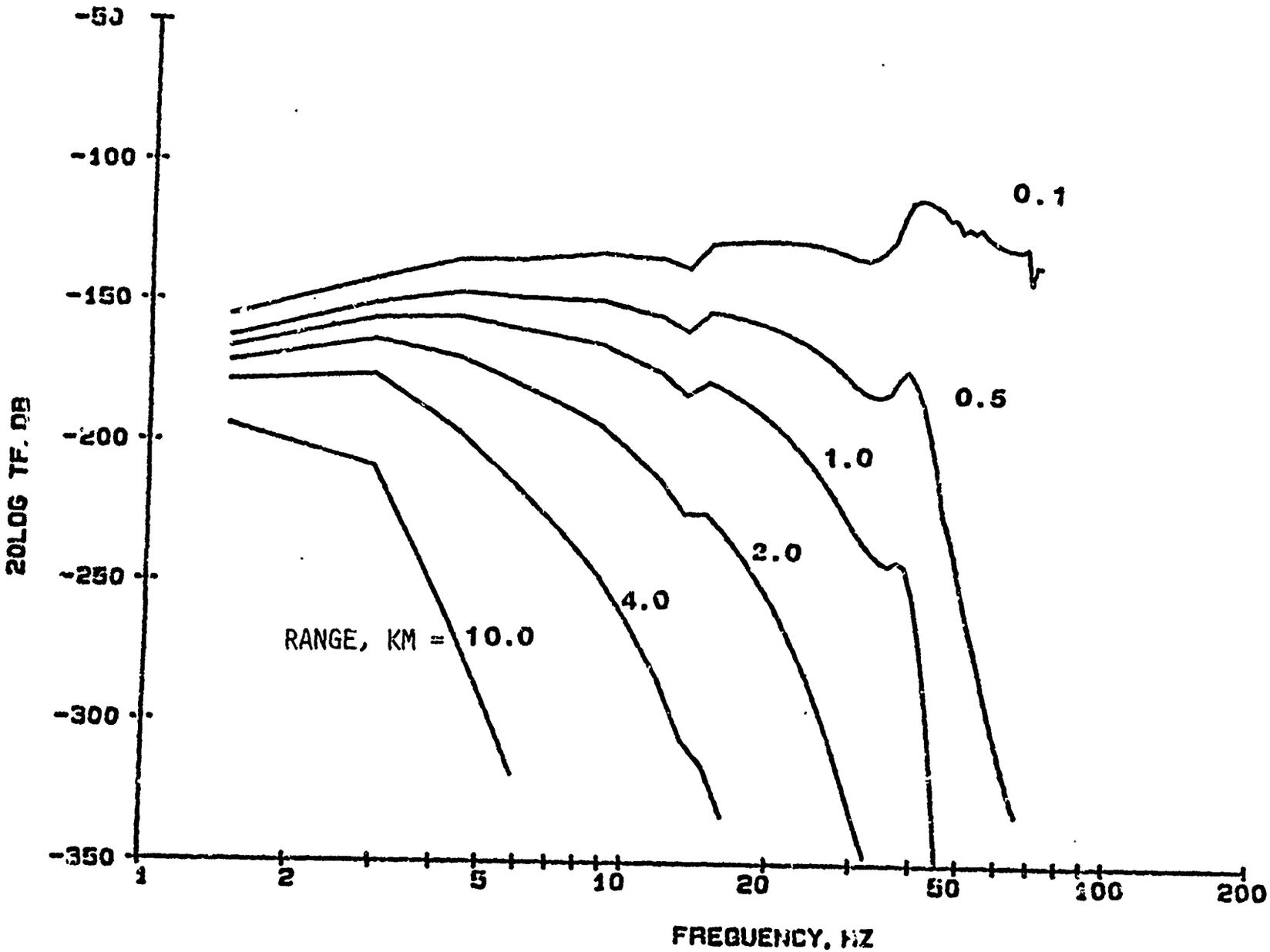



Figure 19. Transfer functions with internal damping factor = 0.02

# BUCKLED ELASTICA IN CONTACT - FINITE ELEMENT SOLUTIONS

Arthur R. Johnson
and
Claudia J. Quigley

US Army Materials and Mechanics Research Center
DRXMR-SMM
Watertown, Massachusetts 02172

ABSTRACT The analysis of manufacturing processes for metal sheets, stresses in elastomer bumpers and other geometrically nonlinear problems require numerical techniques which can deal with continuously changing regions of contact. The most common method of dealing with the contact problem is to introduce gap elements in the contact region. These elements provide either zero or very large stiffness between two degrees of freedom. The methods for determining gap element stiffness include computations which determine if contact has been made. In this effort, a revised simplex method for quadratic programming is investigated for treating contact problems. The method allows the contact problem to be treated directly as a minimization problem subject to an inequality constraint without the need to define gap elements. To test the effectiveness of the algorithm the problem of the buckled elastica in contact with itself was solved. An existing finite element formulation for the extensional elastica based on the minumim potential energy principle is used to model the buckled elastica. Solutions to the contact problem are also obtained by a penalty method using gap elements. Results from both numerical methods are compared to theoretical results.

INTRODUCTION The problem of a beam deforming near a rigid barrier was recently solved by Westbrook [1]. In his analysis, the potential energy for a beam in its quadratic finite element form and a linear constraint equation (the rigid boundary) are joined to form a quadratic programming problem which is solved by Rusin's revised simplex method for quadratic programming [2]. A total Lagrangian finite element formulation for the extensional elastica (large deformations) was recently developed by Fried [3]. In this effort, the potential energy for the total Lagrangian finite element formulation is expanded so that the contact problem can be approximated by a quadratic programming problem. Successive solutions to this quadratic programming problem converge to the contact solution. The problem of a buckled elastica in contact with itself was solved theoretically by Flaherty and Keller [4]. This problem presents an interesting test for numerical algorithms intended for contact problems in the presence of large deformations. Contact occurs at a point, not a region, and the location of the contact point moves in a nonuniform manner as the load increases. Numerical solutions are obtained using both the simplex method and the more commonly used penalty method. Both methods compare well with the asymptotic solutions presented by Flaherty and Keller (Ref. 4).

The finite element formulation for the elastica is presented and is used to compute known values of the initial buckling load and the initial contact load. The effect of axial stiffness on the buckling and contact loads is investigated and an axial stiffness is selected which allows the inextensional elastica to be approximated. The formulation for the analysis of



PREVIOUS PAGE IS BLANK

geometrically nonlinear problems in contact by the quadratic programming method is presented. The penalty method used is also presented and the effect of the penalty parameter is numerically investigated. Both methods are used to solve the post contact problem for the buckled elastica. Comments are given regarding the implementation of each method.

FINITE ELEMENT MODEL: The finite element model for the extensional elastica developed by Fried (Ref. 3) was applied to the initially flat elastica and used to compute the initial buckling and contact loads for the end loaded flat elastica. A short description of the element is given here for completeness. The coordinate systems used to describe the elastica and the element are shown in Figure 1. The element is mapped to the interval $0 \leq \xi_1 \leq 1$ by $S = S_1 + h\xi_1$. The element nodal variables are

$$U_e^T = ( X_1, \dot{X}_1, Y_1, \dot{Y}_1, X_2, \dot{X}_2, Y_2, \dot{Y}_2 ) \tag{1}$$

The values of X and Y are interpolated by

$$X(\xi_1) = U_e^T \phi(\xi_1) \tag{2}$$

and $$Y(\xi_1) = U_e^T \psi(\xi_1)$$

where

$$\phi(\xi_1) = ( \phi_1 \; \phi_2 \; 0 \; 0 \; \phi_3 \; \phi_4 \; 0 \; 0)$$

$$\psi(\xi_1) = (0 \; 0 \; \phi_1 \; \phi_2 \; 0 \; 0 \; \phi_3 \; \phi_4)$$

$$\phi_1 = 1 - 3\xi_1 + 2\xi_1^3 \qquad \phi_2 = \xi_1 - 2\xi_1^2 + \xi_1^3$$

$$\phi_2 = 3\xi_1^2 - 2\xi_1^3 \qquad \phi_4 = -\xi_1^2 + \xi_1^3$$

Then, the potential energy for an element with unit properties is approximated as

$$\Pi_e = h \sum_{j=1}^{3} \omega_j \left( \frac{1}{2} K_j^2 + \frac{1}{2} c \epsilon_j^2 \right) - W_e \tag{3}$$

where $\kappa = \alpha \beta^{-3/2}$ = the curvature

$\epsilon = h^{-1} \beta^{1/2} - 1$ = the axial strain

$\alpha = \dot{X}\ddot{Y} - \dot{Y}\ddot{X}$ $\qquad ( \dot{} ) = \frac{d}{d\xi_1}$

$\beta = \dot{X}^2 + \dot{Y}^2$

$c$ = axial stiffness/bending stiffness

$W_e$ = work done by external forces on an element
$\omega_i$ = integration weight.

The above form of the potential energy can be differentiated with respect to the nodal variables to obtain an element gradient of the potential energy. That is,

$$g_e = \frac{\partial \Pi_e}{\partial u_e} \tag{4}$$

Similarly, the element tangent matrix can be obtained by differentiating the element gradient vector. That is,

$$k_e = \frac{\partial g_e}{\partial u_e} = \frac{\partial^2 \Pi_e}{\partial u_e^2} \tag{5}$$

These element gradient and tangent matrices can be assembled to form global gradient and tangent matrices. The global potential energy is minimum when the gradient is null. The locations of the minimums are found by using the Newton-Raphson method. That is, given a displacement vector $u_1$ we find a displacement $u_2$ whose gradient is closer to zero by computing

$$u_2 = u_1 - K_1^{-1} g_1 \tag{6}$$

The above method was used to obtain solutions for the initial buckling, large deformation and initial contact of the elastica. The deformed shape of the elastica for end loads of 40, 60 and 72 are shown in Figure 2. Note, the initial theoretical buckling load is $4\pi^2$ and the initial contact load is 72.187. To demonstrate that this extensible elastica formulation can be used to accurately represent the inextensible elastica the effect of the magnitude of c on the buckling and contact load was determined, see Figure 3. With $c = 10^5$ the finite element values for the buckling and contact loads have converged to three significant digits to the theoretical values using only a coarse mesh of 14 elements.

NONLINEAR CONTACT USING QUADRATIC PROGRAMMING: Quadratic programming has been used to solve contact problems in linear elasticity where potential energy (P.E.) expression and contraint equation (C.E.) can be expressed as follows.

$$\Pi(x) = \frac{1}{2} x^T K x - p^T x \qquad \text{P.E.} \tag{7}$$

$$x \geq 0 \qquad \text{C.E.} \tag{8}$$

The paper by Rusin describes a particular form of quadratic programming which can be adapted to finite element contact problems. In the linear problem, K is a positive definite symmetric matrix and p is the load vector. The algorithm finds the minimum of $\Pi(x)$ such that equation (8) is also satisfied.

The geometrically nonlinear contact problem is not in the form given by equations (7) and (8). However, if we expand the nonquadratic form of the potential energy, in a Taylor expansion near the contact solution we can solve the contact problem by solving a sequence of linear problems. Figure 4 shows a Newton-Raphson solution to the nonlinear buckled elastica problem obtained without consideration of contact and the near-by contact solution. Note, in this Lagrangian formulation x is a configuration vector, not a displacement vector. With $x_o$ = the vector obtained by minimizing the potential energy and x = the vector near $x_o$ which is the minimum of the potential energy subject to the constraint equation the potential energy is expanded as follows.

$$\Pi(x) = \Pi(x_o+\Delta x) = \Pi(x_o) + g^T(x_o)\Delta x + \frac{1}{2}\Delta x^T K(x_o)\Delta x + \cdots \qquad (9)$$

where $\Delta x = x - x_o$

with $g^T(x_o) = 0$ (Newton-Raphson solution) equation (9) can be arranged as

$$\Pi(x) = \frac{1}{2}x^T K_o x - (x_o^T K_o)x + \left[\Pi(x_o) + \frac{1}{2}x_o^T K_o x_o\right] + \cdots \qquad (10)$$

The term in the brackets in (10) is constant and the first two terms are in a form which can be used with the quadratic programming algorithms.

To continue the discussion the contact problem is considered for the buckled elastica. The vector of nodal unknowns is given by $X^T$ and the vector involved in the contact constraint is $u^T$ where

$$X^T = (X_1\ \dot{X}_1\ Y_1\ \dot{Y}_1\ X_2\ \dot{X}_2 \cdots Y_N\ \dot{Y}_N) \qquad (11)$$

and $$u^T = (X_1 Y_1 X_2 Y_2 \cdots X_N Y_N) \qquad (12)$$

Using symmetry to solve the buckled elastica problem as indicated in Figure 4 the constraint equation is

$$u \geq 0 \qquad (13)$$

Then, given $x_o$ an approximate contact solution can be found by solving the following quadratic programming problem.

Find $$\min \varphi(x) = \min_{\forall x}\left[\frac{1}{2}x^T K_o x - \rho_o^T x\right] \qquad (14)$$

with $u \geq 0$

where $$K_o = \left.\frac{\partial \Pi}{\partial X^2}\right|_{X_o}$$

and $$P_o^T = X_o^T K_o$$

To obtain the optimal contact soution the quadratic programming problem is repeated by recomputing Ko and Po at each solution of (14) until the region of contact no longer changes. One last Newton-Raphson solution obtained while enforcing the appropriate region of contact yields the solution to the nonlinear contact problem.

The method described above was used to obtain post contact solutions to the buckled elastica. The location of the contact point with respect to the loaded end of the elastica was computed as a function of the loading for a uniform mesh of 40 elements and is shown in Figure 5. At some loads, two contact points were indicated. In all such cases, the actual contact point was located between them. That is, when two contact points were computed they were never on one side of the actual point contact solution given by Flaherty and Keller[4]. Some solutions were obtained for a uniform mesh of 80 elements. The resulting contact point locations were approximately twice as close to the analytical solution.

NONLINEAR CONTACT USING THE PENALTY METHODS: The use of gap elements to treat contact was described by Chan and Tuba[5]. This method is similar to the penalty method[6]. The penalty method is easily adapted to the Lagrangian finite element formulation being used here and provides another finite element method to compare with the analytical solution. The contact constraint as stated in equation (13) can be obtained in a least squares sense by adding the square of the nodal value (i.e., $x_n$ or $y_n$) for any nodal variable that fails to satisfy equation (13). On an element level this is done by adding the product of the squared terms and the penalty parameters as follows (only $X_n$'s are included here).

$$\Pi_e' = \Pi_e + \sum_{X_n<0} \gamma_{n\rho} X_n^2 \qquad (15)$$

where $\gamma_{n\rho} = C_n K_{nn} 10^{\rho}$ = penalty parameters or gap element stiffness

$\rho$ = a variable for convergence studies

$C_n = \begin{cases} 1/2 & \text{if one or no adjacent nodes have } x_n < 0 \\ 1 & \text{if both adjacent nodes have } x_n < 0 \end{cases}$

$K_{nn}$ = diagonal term from the tangent stiffness matrix associated with the $x_n$ displacement.

The selection of the penalty parameter chosen here reflects the ideas presented by Chan and Tuba[5] in which gap element stiffnesses are selected based on the current global tangent stiffness matrix and the portion of the element for which the constraint is violated (thus, $k_{nn}$ and $c_n$ as described above).

The computation of the element gradient and tangent stiffness is straight forward and proceeds as follows.

The element gradient is computed as

$$g'_e = \frac{\partial \Pi'_e}{\partial u_e} = g_e + \frac{\partial}{\partial u_e}\left[\sum \gamma_{np} X_n^2\right] \tag{16}$$

or

$$g'_e = g_e + g_{ep}$$

and the tangent matrix is found from

$$K_e = \frac{\partial^2 \Pi_e}{\partial u_e^2} = K_e + \frac{\partial^2}{\partial u_e^2}\left[\sum \gamma_{np} X_n^2\right]$$

or

$$K'_e = K_e + K_{ep} \tag{17}$$

The assembly of the global gradient and tangent matrices and the introduction of displacement boundary conditions are again similar to the methods used in linear analysis. Then, the Newton-Raphson method is applied to minimize the global form of equation (15). As an example when $X_{n_1}$ and $X_{n_2}$ are two variables which violate the constraint equation (13) the gradient is

$$g' = g + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 2\gamma_{n_1 p} X_{n_1} \\ 0 \\ \vdots \\ 2\gamma_{n_2 p} X_{n_2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{18}$$

and the tangent stiffness is

$$K' = K + \begin{bmatrix} 0 & 0 & \cdots & & & & \\ & & & 2\gamma_{n_1 p} & 0 & \cdots & \\ & & & & & & 2\gamma_{n_2 p} \\ 0 & \cdots & & & & & 0 \end{bmatrix} \tag{19}$$

The buckled elastica contact problem was solved for an end load of 80.0 and the range of penalty parameters indicated in Table 1. The location of the loaded end of the elastica appears to be converged to 6 or 7 decimal places when $\gamma_n = C_n K_{mm} 10^4$. Convergence with respect to mesh size was also considered and the results are shown in Table 2. After the penalty parameter was selected, convergence with respect to mesh size was numerically determined. The meshes chosen for the convergence study in Table 2 all have an identical contact point location so there was no effect from a shifting contact point due to a finer mesh. A uniform mesh with 80 elements was required to obtain convergence to about five significant digits in the displacements. This may be due to the fact that there is a large curvature along the short portion of the elastica between the loaded end and the contact point which is difficult to approximate with a coarse mesh.

The location of the nodes in contact was determined as a function of the load for a uniform mesh of 40 elements. The results are shown in Figure 6. They agree well with Flaherty and Keller's results in that the location of contact is always known within the length of an element. A more sophisticated choice of the penalty parameter ( $C_n$ in particular) would improve the results.

CONCLUSION: The use of quadratic programming and penalty methods for the analysis of geometrically nonlinear contact problems were presented. Of the two methods, the penalty method was easier to adapt to the total Lagrangian finite element formulation being used. The solutions obtained using the simplex method implied point contact which is in agreement with theory. The penalty method as used here was unable to predict point contact as clearly as the simplex method. For a given mesh, the penalty method requires that a convergence study be made with respect to the penalty parameter, whereas, the simplex method determines the results in one analysis. Both methods gave satisfactory results when the location of contact was determined as a function of the load. Additional work could include modifying the methods in such a way that regional and point contact can be easily determined numerically.

REFERENCES:


1. D.R. Westbrook, Contact Problems for the Elastic Beam, Comput. Structures, 15, 473-479 (1982).

2. M.H. Rusin, A Revised Simplex Method for Quadratic Programming, SIAM J. Appl. Math., 20, No. 2, 143-160, (1971).

3. I. Fried, Nonlinear Finite Element Computation of the Equilibrium, Stability and Motion of the Extensional Beam and Ring, Comp. Meth. Appl. Mech. Eng., 38, 29-44 (1983).

4. J.E. Flaherty and J.B. Keller, Contact Problems Involving a Buckled Elastica. SIAM J. Appl. Math. 24, No. 2, 215-225 (1973).

5. S.M.Chan and I.S. Tuba, A Finite Element Method for Contact Problems of Solid Bodies. Int. J. Mech. Sci. 13, 615-639 (1971).

6. N. Kikuchi, Penalty/Finite Element Approximations of a Class of Unilateral Contact Problems. Penalty-Finite Element Methods in Mechanics, AMD-Vol 51, ASME, 1982, Ed. J.N. Reddy.

Table 1. Effect of penalty parameter on X(1/2)

Penalty parameter = $C_n k_{nn} 10^P$ Load = 80

| P | X(1/2) |
|---|---|
| -2 | 0.071186117 |
| -1 | 0.073314913 |
| 0 | 0.073536903 |
| 1 | 0.073559195 |
| 2 | 0.073561426 |
| 3 | 0.073561649 |
| 4 | 0.073561671 |

Note: $N_e = 20$ $\|g\| < 10^{-10}$

Table 2. Effect of mesh size on X(1/2)

Penalty parameter = $C_n k_{nn} 10^4$ Load = 80

| $N_e$ | X(1/2) |
|---|---|
| 20 | 0.073561671 |
| 40 | 0.073554789 |
| 80 | 0.073554449 |

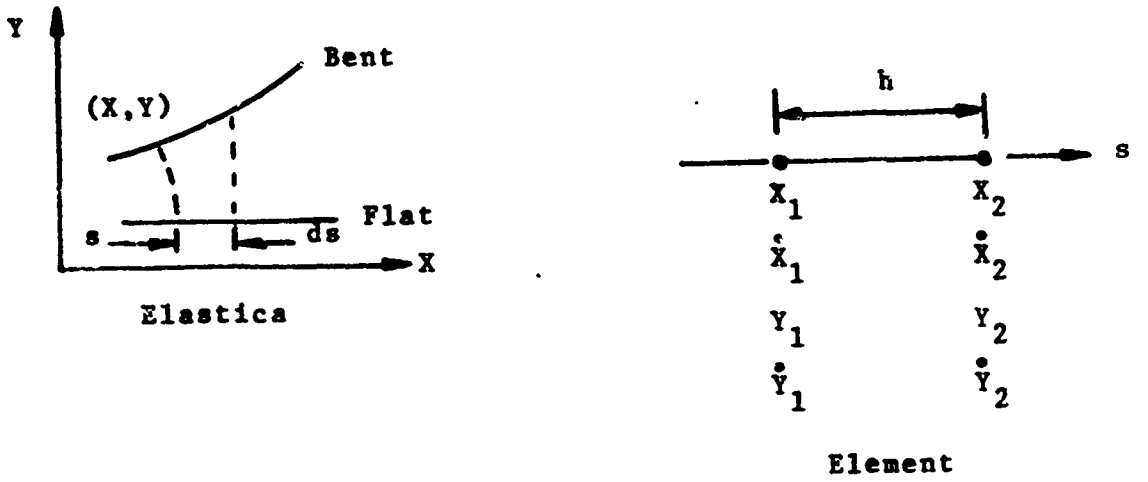



Figure 1. Coordinate systems for elastica and element.

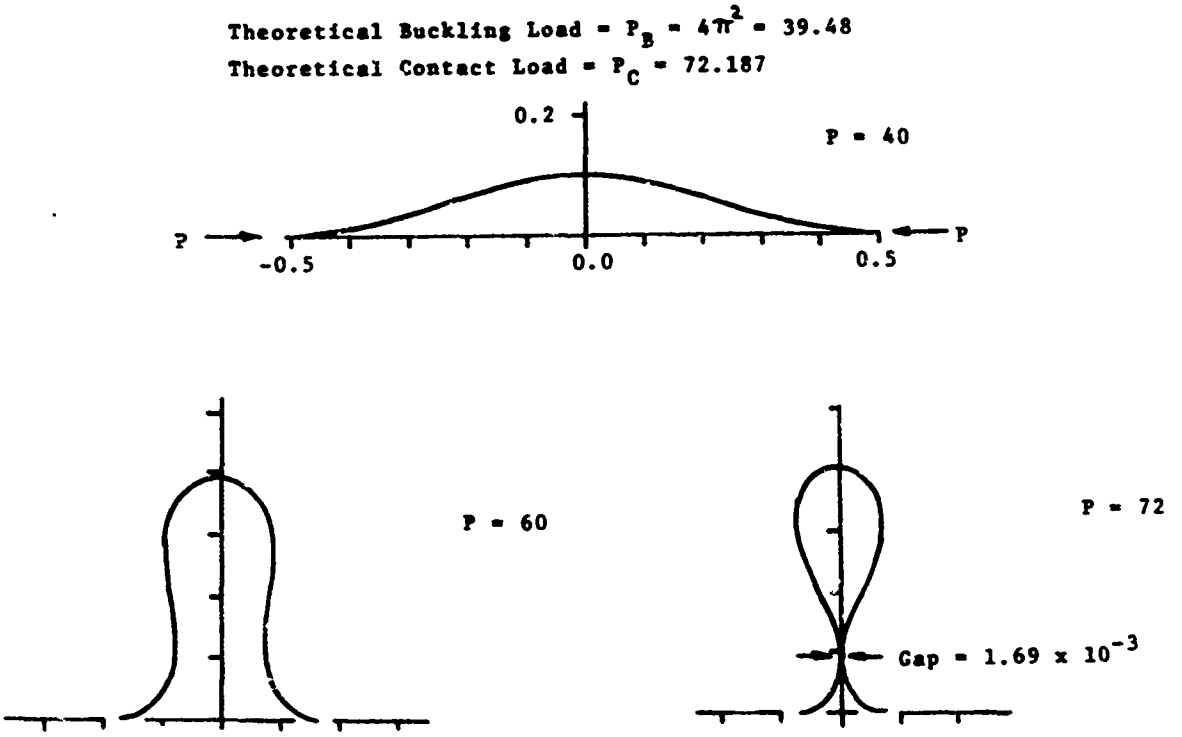



Figure 2. Deformed shape of elastica for loads of 40, 60 and 72.

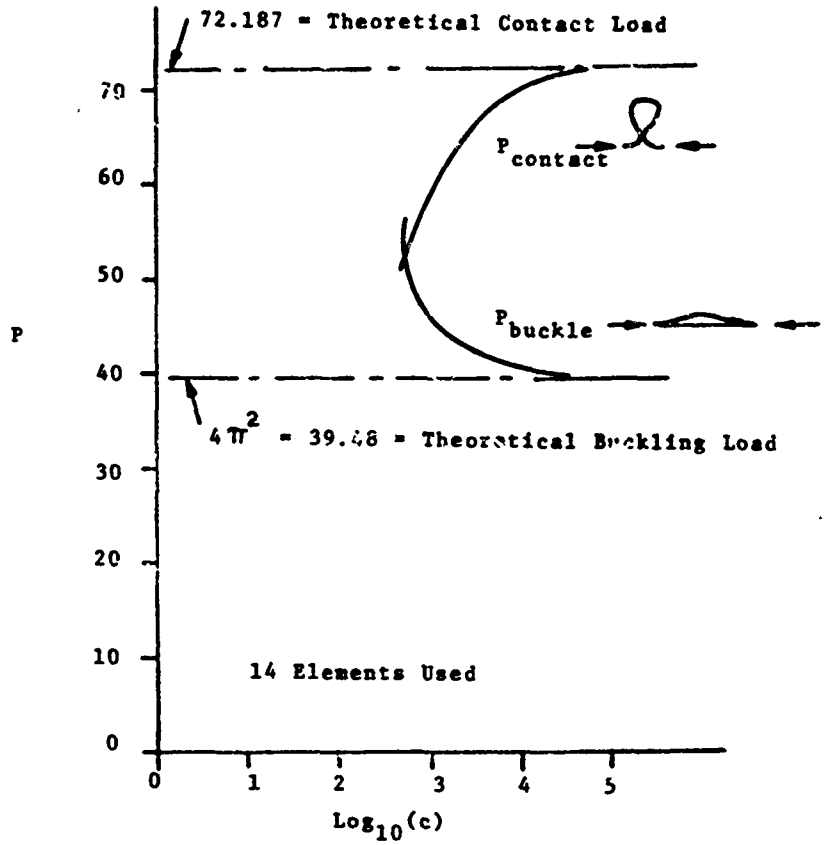



Figure 3. Effect of the magnitude of c on the buckling and contact loads.

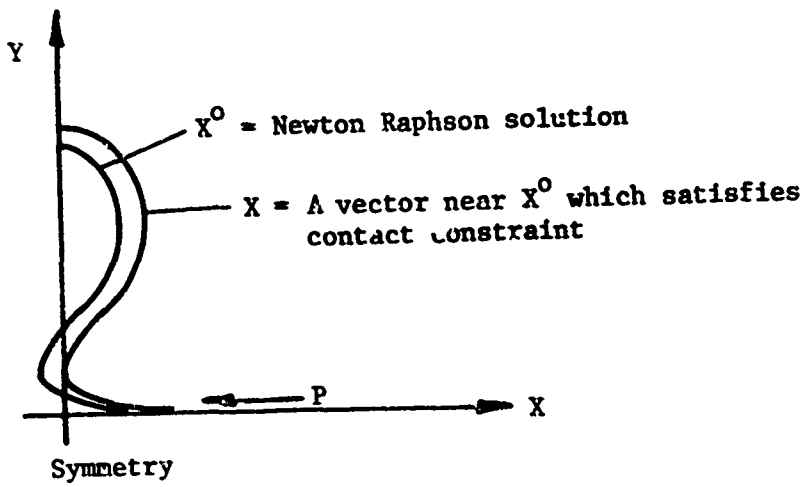



Figure 4. Newton-Raphson solution and near-by contact solution.

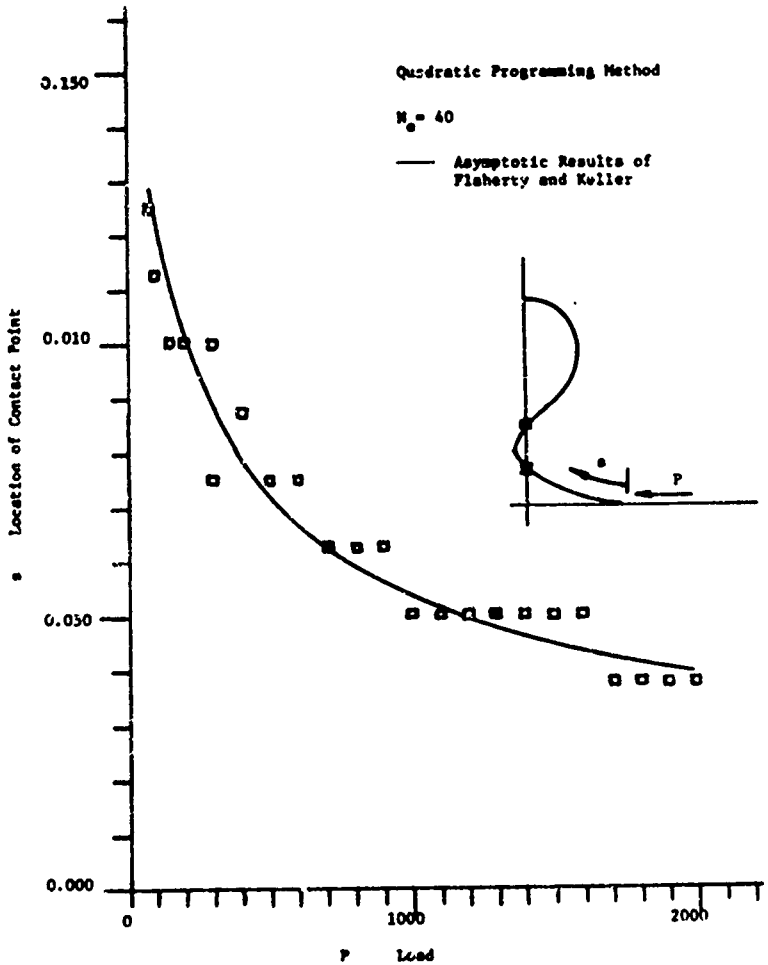



Figure 5. Location of contact point as a function of end load - simplex method.

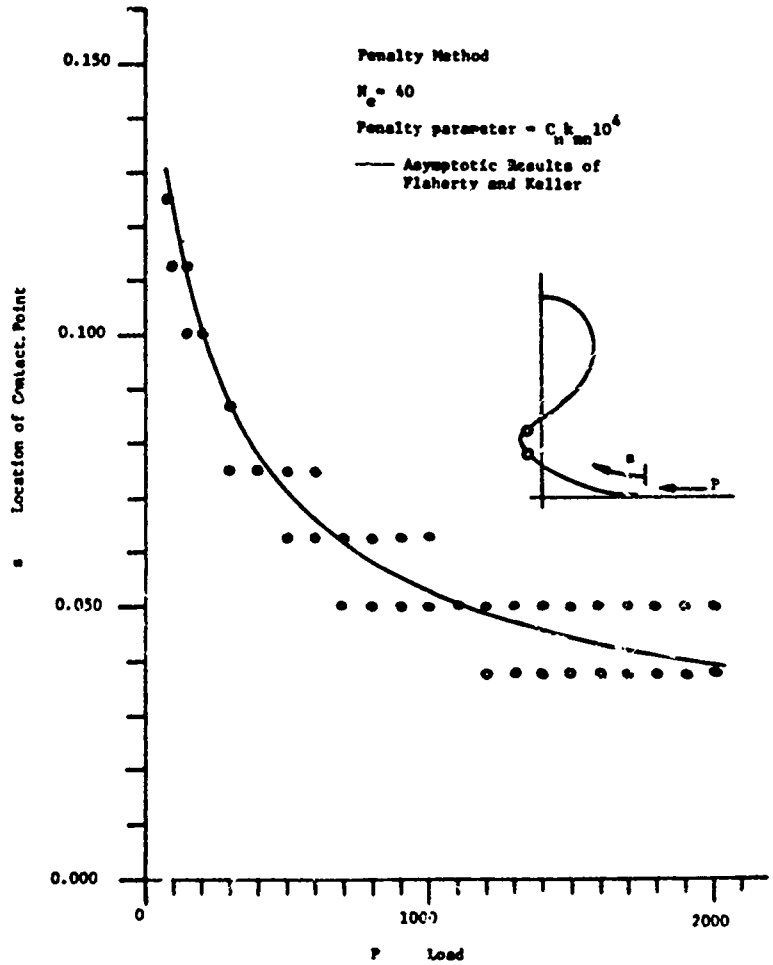



Figure 6. Location of contact point as a function of end load - penalty method.

# OPTIMAL CONTROL TECHNIQUES FOR COMPUTING STATIONARY FLOWS OF VISCOELASTIC FLUIDS WITH MEMORY

Patrick Le Tallec
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53705

ABSTRACT. We consider the problem of computing the stationary flows of a viscoelastic fluid with memory flowing through a given domain. The proposed numerical technique is based on optimal control techniques, which replace the original equations of the problem by a minimization problem to be solved by a conjugate gradient algorithm. Such techniques are very powerful and can handle equations which change type, provided that, as done here, one uses an adequate preconditioning operator and that one computes efficiently the gradient of the function to be minimized.

I. EQUATIONS OF THE PROBLEM. Let us consider a viscoelastic fluid with memory (of upper-convected Maxwell type), flowing viscously through a given domain $\Omega$ (Fig. 1). We suppose that no slip occurs along the solid walls and that the fluid velocity at the entrance and at the exit of the domain is given.

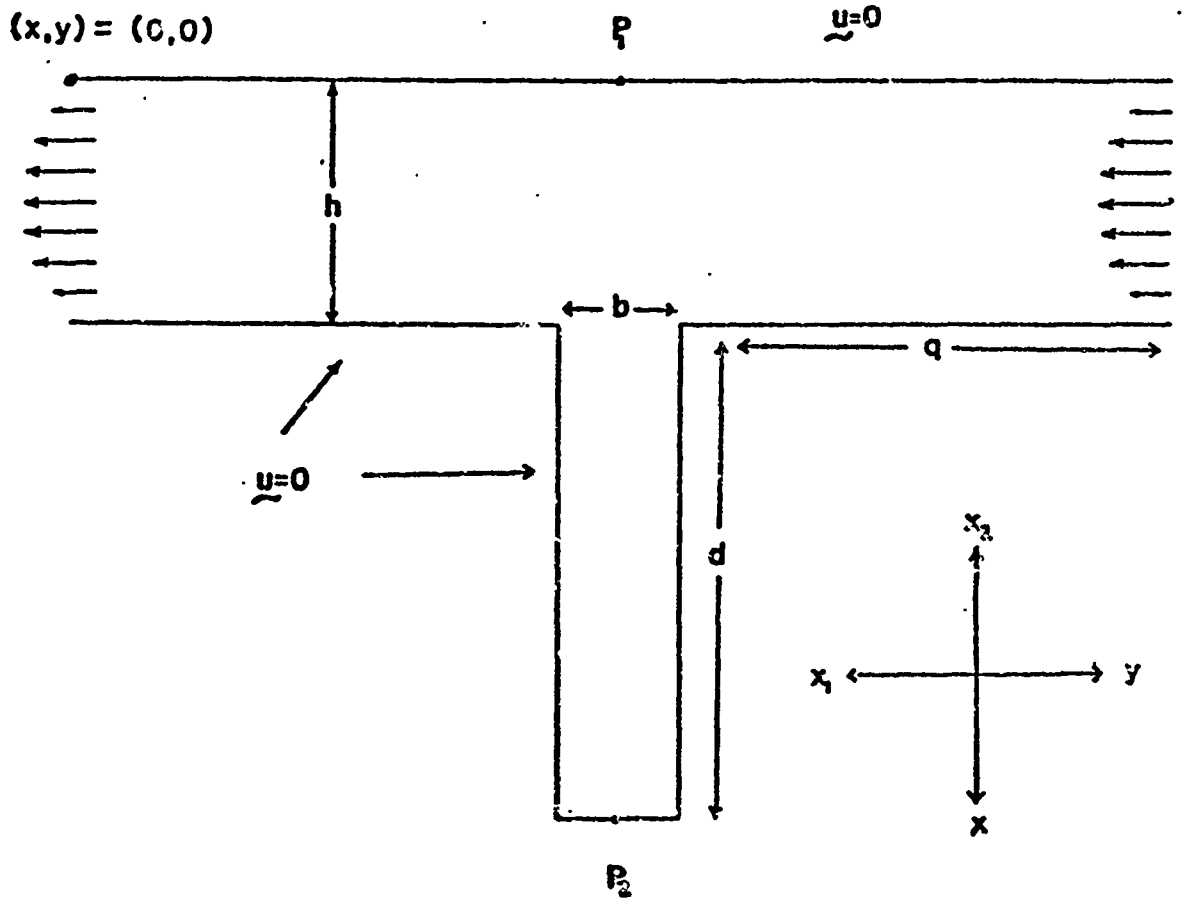



Fig. 1: The physical problem
(out of Malkus [1984])

PREVIOUS PAGE IS BLANK

For examples, such situations arise while studying plane flows over slots, such as those studied experimentally by Bird and al. [1982]. The equations governing such situations are simply:

(1) EQUILIBRIUM

$$-\mathrm{div}(\underset{\sim}{\sigma}) + \rho(\underset{\sim}{u}\cdot\underset{\sim}{\nabla})\underset{\sim}{u} = \underset{\sim}{f} \quad \text{in } \Omega ,$$

(2) CONSTITUTIVE LAW (upper-convected Maxwell)

$$\begin{cases} \underset{\sim}{\sigma} = \underset{\sim}{\sigma}_D - p\underset{\sim}{1} , \\ \underset{\sim}{\sigma}_D(\underset{\sim}{x},t) = \displaystyle\int_{-\infty}^{t} \frac{\mu}{\lambda^2} \exp\left(\frac{\tau - t}{\lambda}\right)\left[(\underset{\sim}{F}_t^T\underset{\sim}{F}_t)^{-1} - \underset{\sim}{1}\right]d\tau , \\ \underset{\sim}{F}_t(\underset{\sim}{x},\tau) = \dfrac{\partial \underset{\sim}{X}_t(\underset{\sim}{x},\tau)}{\partial \underset{\sim}{x}} , \\ X_t(\underset{\sim}{x},\tau) = \text{position at } \tau \text{ of the particle which is in } \underset{\sim}{x} \text{ at time } t \text{ and which is subjected to the velocity field } \underset{\sim}{u}. \end{cases}$$

(3) KINEMATIC RESTRICTIONS

$$\mathrm{div}\ \underset{\sim}{u} = 0, \quad \underset{\sim}{u} = \underset{\sim}{u}_0 \quad \text{on } \Gamma .$$

Above, $\underset{\sim}{u}$ represents the fluid velocity, $\underset{\sim}{\sigma}$ the Cauchy stress tensor, $p$ a hydrostatic pressure, $\rho$ the fluid density, $\mu$ the fluid viscosity and $\lambda$ the relaxation time. Observe that, as an extra boundary condition, the constitutive law (2) requires the knowledge of what happened to the fluid before it enters the domain.

It has been observed in Joseph, Renardy, Saut [1984], that these equations change type when the viscoelastic Mach number $U/\sqrt{\mu/\lambda\rho}$ reaches 1 ($U$ is a characteristic velocity of the considered flow). Real characteristics then appear along which the vorticity can be discontinuous. However, most numerical methods employed for solving (1)-(3) (such as the classical fixed point method solving iteratively for the velocities and then for the stresses) cannot handle this change of type.

The idea of this paper is to employ for the numerical solution of (1)-(3) optimal control techniques in $\underset{\sim}{H}^{-1}(\Omega)$, which were used with success in transonic flow computations (Glowinski [1984]), where a similar change of type occurs. To be applied, such techniques require the definition of an appropriate isomorphism between $\underset{\sim}{H}_0^1(\Omega)$ and $\underset{\sim}{H}^{-1}(\Omega)$ (the preconditioning operator) and the computation of the transpose of the linearized constitutive equations. Here, for Maxwell viscoelasticity, this turns out to be very natural: the Stokes operator acts as a very good preconditioning operator and the transpose of the constitutive laws leads to equations strongly related to a lower-convected Maxwell model with opposite velocities.

## 2. $\underset{\sim}{H}^{-1}$ LEAST-SQUARES FORMULATION OF THE PROBLEM.

2.1 A one-dimensional model problem. Let $f : R \to R$ be differentiable and consider the problem of solving numerically the nonlinear equation

$$f(x) = 0 .$$

If it has a solution in $R$, then this equation is equivalent to

$$\underline{\text{Minimize}} \ \frac{1}{a} |f(x)|^2 \ \underline{\text{over}} \ R, \quad (a > 0) ,$$

problem which can be numerically solved by the gradient algorithm

$$+ \ x_0 = \text{given} ,$$

$$+ \ \text{for} \ \ n = 0 \ \ \text{until satisfied set}$$

$$x_{n+1} = x_n - \frac{2}{a} f(x_n) f'(x_n) .$$

This algorithm can be a very efficient method for solving $f(x) = 0$, provided that $a$ is properly chosen and that $f(x)f'(x)$ is easy to compute. It will be the basis of the numerical technique that we will use to solve (1)-(3).

2.2 Maxwell viscoelasticity. Let $V$ be the space

$$V = \{\underset{\sim}{v} \in \underset{\sim}{H}_0^1(\Omega), \ \text{div} \ \underset{\sim}{v} = 0\} ,$$

let $V^*$ be its topological dual, denote by $\langle \cdot, \cdot \rangle$ the duality pairing between $V$ and $V^*$ and introduce the following operators

$$A : V \to V^*, \quad A(\underset{\sim}{v}) = -\text{div}[\mu(\underset{\sim}{\nabla v} + \underset{\sim}{\nabla v}^T)] , \quad \text{(Stokes)}$$

$$L : V \to R, \quad L(\underset{\sim}{v}) = \int_\Omega \underset{\sim}{f} \cdot \underset{\sim}{v} dx ,$$

$$\begin{cases} T : V \to V^*, \quad T(\underset{\sim}{v}) = \rho(\underset{\sim}{u} \cdot \underset{\sim}{\nabla})\underset{\sim}{u} - \text{div}(\underset{\sim}{\sigma}_D(\underset{\sim}{u})) - A(\underset{\sim}{v}) \ \text{ with} \\ \underset{\sim}{u} = \underset{\sim}{v} + \underset{\sim}{u}_0, \ \ \underset{\sim}{\sigma}_D(\underset{\sim}{u}) \ \text{ being given by the constitutive relation (2).} \end{cases}$$

With these new notations, Equations (1) to (3) take the form

$$(4) \qquad \begin{cases} A(\underset{\sim}{u} - \underset{\sim}{u}_0) + T(\underset{\sim}{u} - \underset{\sim}{u}_0) - L = 0 \ \text{ in } \ V^* , \\ \underset{\sim}{u} - \underset{\sim}{u}_0 \in V . \end{cases}$$

If (4) has a solution, then (4) is obviously equivalent to the $\underset{\sim}{H}^{-1}$ least-squares formulation:

(5)

$$\boxed{\begin{array}{l}\underline{\text{MINIMIZE}}\ \ J(\underset{\sim}{v}) = \frac{1}{2}\langle A\underset{\sim}{y}(\underset{\sim}{v}),\ \underset{\sim}{y}(\underset{\sim}{v})\rangle\ \ \underline{\text{OVER}}\ \ V\ \ \underline{\text{WHERE}} \\ \underset{\sim}{y}(\underset{\sim}{v}) \in V\ \ \underline{\text{IS THE SOLUTION OF THE LINEAR PROBLEM}} \\ A\underset{\sim}{y}(\underset{\sim}{v}) = A\underset{\sim}{v} + T(\underset{\sim}{v}) - L\ \ \text{in}\ \ V^* \ .\end{array}}$$

Indeed, if (4) has a solution $\bar{\underset{\sim}{v}} = \underset{\sim}{u} - \underset{\sim}{u}_0$, and if we take $\underset{\sim}{v} = \bar{\underset{\sim}{v}}$ in (5), then the right hand side of (5) is equal to zero, thus the associated state vector $\underset{\sim}{y}(\bar{\underset{\sim}{v}})$ is also equal to zero, and therefore $J(\bar{\underset{\sim}{v}})$ is equal to zero. Since $J(\underset{\sim}{v})$ is positive by construction (A is a monotone operator on V), this implies that $\bar{\underset{\sim}{v}}$ is a minimizer of J over V.

Conversely, let $\bar{\underset{\sim}{w}}$ be a minimizer of J over V. As seen above, since (4) has a solution, J attains the value 0 on V. Thus, $J(\bar{\underset{\sim}{w}})$ must be equal to zero, therefore $\underset{\sim}{y}(\bar{w})$ must be equal to zero (the Stokes operator A is strictly monotone on V). By construction of $y(\bar{\underset{\sim}{w}})$, this implies that the right-hand side of (5) is equal to zero when we take $\underset{\sim}{v} = \bar{\underset{\sim}{w}}$, which means precisely that $\bar{\underset{\sim}{w}}$ is a solution of (4).

In summary, if we assume the existence of a solution to our original problem (1)-(3), we can replace these equations by the equivalent minimization problem just written above. This minimization formulation is the one which will be used in our numerical techniques. It reduces our initial problem to an optimal control problem, if we identify $\underset{\sim}{v}$ to a control variable, $\underset{\sim}{y}$ to a state vector, (5) to a state equation and $J(\cdot)$ to a cost function.

## 3. CONJUGATE GRADIENT METHOD.

The minimization formulation of §2 is interesting because it can be solved numerically by a conjugate gradient algorithm which has superlinear convergence properties. This algorithm is:

INITIALIZATION

$$\begin{cases} + \text{ Take } \underset{\sim}{u}^0 \in V + \underset{\sim}{u}_0 , \\ + \text{ Solve } \langle A(\underset{\sim}{g}^0), \underset{\sim}{w} \rangle = \langle J'(\underset{\sim}{u}^0 - \underset{\sim}{u}_0), \underset{\sim}{w} \rangle, \quad \forall \underset{\sim}{w} \in V , \\ + \text{ Set } \underset{\sim}{z}^0 = \underset{\sim}{g}^0 . \end{cases}$$

LOOP

For $n = 0$ until satisfied, do

$$\begin{cases} + \rho_n = \text{Argmin } J(\underset{\sim}{u}^n - \underset{\sim}{u}_0 - \rho \underset{\sim}{z}^n), \\ \quad \text{(to be achieved by quadratic interpolation of } J(\rho)) \\ + \underset{\sim}{u}^{n+1} = \underset{\sim}{u}^n - \rho_n \underset{\sim}{z}^n , \\ + \text{ Solve } \langle A\underset{\sim}{g}^{n+1}, \underset{\sim}{w} \rangle = \langle J'(\underset{\sim}{u}^{n+1} - \underset{\sim}{u}_0), \underset{\sim}{w} \rangle, \quad \forall \underset{\sim}{w} \in V , \\ \quad \text{(computation of the gradient)} \\ + \gamma_n = (\langle A\underset{\sim}{g}^{n+1}, \underset{\sim}{g}^{n+1} - \underset{\sim}{g}^n \rangle)/(\langle A\underset{\sim}{g}^n, \underset{\sim}{g}^n \rangle) , \\ + \underset{\sim}{z}^{n+1} = \underset{\sim}{g}^{n+1} + \gamma_n \underset{\sim}{z}^n . \quad \text{(conjugate gradient)} \end{cases}$$

## 4. COMPUTATION OF THE GRADIENT $J'(\underset{\sim}{v})$

Obviously, the difficult part of the algorithm of §3 is the computation of the gradient $J'(\underset{\sim}{v})$ of the cost function. Here, the cost function is a quadratic function of the state vector, itself depending on the added stress $\underset{\sim}{\sigma}_D$, the latter being the image of a nonlinear integral operator acting on $(\underset{\sim}{v} + \underset{\sim}{u}_0)$. But unlike a classical Newton method which would require $0(N^3)$ operations ($N = \dim V$) to compute this gradient, our computation of $J'(\underset{\sim}{v})$ only requires $0(N^2)$ operations, because we introduce an adjoint state vector and reduce the computation of all terms $\langle J'(\underset{\sim}{v}), \underset{\sim}{w} \rangle$ to the explicit computation of local integrals defined on the support of $\underset{\sim}{w}$.

To see that, let us introduce the adjoint state $\underset{\sim}{H}(\underset{\sim}{v})$ defined as

$$(6) \qquad \underset{\sim}{H}(\underset{\sim}{v})(\underset{\sim}{x},t) = \frac{1}{\lambda} \int_{-\infty}^{0} \exp\left(\frac{\tau - t}{\lambda}\right) (\underset{\sim}{F}_t^-(\underset{\sim}{x},\tau))^T \underset{\sim}{D}(X_t^-(\underset{\sim}{x},\tau)) \underset{\sim}{F}_t^-(\underset{\sim}{x},\tau) d\tau ,$$

with

$$
(7)\qquad
\begin{cases}
\underset{\sim}{D} = \dfrac{1}{2}\left(\nabla\underset{\sim}{y}(\underset{\sim}{y}) + (\nabla\underset{\sim}{y}(\underset{\sim}{y}))^T\right) & \text{if } \underset{\sim}{x} \in \Omega\ , \\
\underset{\sim}{D} = 0 \quad \text{if } \underset{\sim}{x} \notin \Omega\ , \\
\underset{\sim}{F}_t^-(\underset{\sim}{x},\tau) = \dfrac{\partial \underset{\sim}{x}_t^-}{\partial \underset{\sim}{x}}(\underset{\sim}{x},\tau)\ , \\
\underset{\sim}{x}_t^-(\underset{\sim}{x},\tau) = \text{position at time } \tau \text{ of the particle which is in } \underset{\sim}{x} \text{ at time } t \text{ and which is subjected to the velocity field } \underset{\sim}{u}^-(x) = -\underset{\sim}{u}(\underset{\sim}{x}).
\end{cases}
$$

This adjoint state can be computed by an explicit integration along the trajectories of $\underset{\sim}{u} = \underset{\sim}{y} + \underset{\sim}{u}_0$. Then, we can prove

THEOREM: The term $\langle J'(\underset{\sim}{v}),\underset{\sim}{w}\rangle$ is equal to

$$
(8)\qquad
\begin{aligned}
\langle J'(\underset{\sim}{v}),\underset{\sim}{w}\rangle = &\int_\Omega \rho\{(\underset{\sim}{w}\cdot\nabla)\underset{\sim}{u} + (\underset{\sim}{u}\cdot\nabla)\underset{\sim}{w}\}\cdot\underset{\sim}{\chi}(\underset{\sim}{v})d\underset{\sim}{x} \\
&+ \int_\Omega \{\mu(\nabla\underset{\sim}{w} + \nabla\underset{\sim}{w}^T) + \lambda(\nabla\underset{\sim}{w})\underset{\sim}{\sigma}_D + \lambda\underset{\sim}{\sigma}_D(\nabla\underset{\sim}{w})^T - \lambda(\underset{\sim}{w}\cdot\nabla)\underset{\sim}{\sigma}_D\}\cdot\underset{\sim}{H}(\underset{\sim}{v})d\underset{\sim}{x}\ .
\end{aligned}
$$

Proof: By definition of the gradient, we have

$$
\langle J'(\underset{\sim}{v}),\underset{\sim}{w}\rangle = \lim_{t\to 0}\frac{1}{t}\left[J(\underset{\sim}{y} + t\underset{\sim}{w}) - J(\underset{\sim}{v})\right] = \langle A\delta\underset{\sim}{\chi},\underset{\sim}{\chi}\rangle\ ,
$$

where $\underset{\sim}{\chi} = \underset{\sim}{\chi}(\underset{\sim}{y})$ is the solution of the state equation (5) and where $\delta\underset{\sim}{\chi}$ is obtained from $\underset{\sim}{w}$ by differentiation of (5), that is by solving

$$
A\delta\underset{\sim}{\chi} = A\underset{\sim}{w} + T'(\underset{\sim}{y})\cdot\underset{\sim}{w} \quad \text{in } V^*\ .
$$

Substituting this definition of $\delta\underset{\sim}{\chi}$ in the expression of the gradient, we get

$$
\langle J'(\underset{\sim}{v}),\underset{\sim}{w}\rangle = \langle A\underset{\sim}{w} + T'(\underset{\sim}{y})\cdot\underset{\sim}{w},\underset{\sim}{\chi}\rangle\ .
$$

Using the definition of $T(\cdot)$, and integrating by parts the term in $\mathrm{div}(\underset{\sim}{\sigma}_D'(\underset{\sim}{u})\cdot\underset{\sim}{w})$, this gives

$$
(9)\qquad \langle J'(\underset{\sim}{v}),\underset{\sim}{w}\rangle = \int_\Omega \{\rho(\underset{\sim}{w}\cdot\nabla)\underset{\sim}{u} + \rho(\underset{\sim}{u}\cdot\nabla)\underset{\sim}{w}\}\cdot\underset{\sim}{\chi}d\underset{\sim}{x} + \int_\Omega [\underset{\sim}{\sigma}_D'(\underset{\sim}{u})\cdot\underset{\sim}{w}]\cdot\underset{\sim}{D}(\underset{\sim}{\chi})d\underset{\sim}{x}\ ,
$$

with $\underset{\sim}{D}(\underset{\sim}{\chi}) = \frac{1}{2}(\nabla\underset{\sim}{\chi} + \nabla\underset{\sim}{\chi}^T)$. In (9), to compute the action of the derivative of $\underset{\sim}{\sigma}_D(\underset{\sim}{u})$ on $\underset{\sim}{w}$, we differentiate the constitutive law (2), first with respect to time, then with respect to the velocity $\underset{\sim}{u}$. We obtain

(10)
$$\begin{cases} \underset{\sim}{\tau} = \underset{\sim}{\sigma}_D'(\underset{\sim}{u})\cdot\underset{\sim}{w} \ \underline{\text{satisfies the differential equation}} \\ \lambda(\underset{\sim}{u}\cdot\nabla)\underset{\sim}{\tau} - \lambda(\nabla\underset{\sim}{u})\underset{\sim}{\tau} - \lambda\underset{\sim}{\tau}(\nabla\underset{\sim}{u})^T + \underset{\sim}{\tau} \\ \qquad = 2\mu\underset{\sim}{D}(\underset{\sim}{w}) - \lambda(\underset{\sim}{w}\cdot\nabla)\underset{\sim}{\sigma}_D(\underset{\sim}{u}) + \lambda(\nabla\underset{\sim}{w})\underset{\sim}{\sigma}_D(\underset{\sim}{u}) + \lambda\underset{\sim}{\sigma}_D(\underset{\sim}{u})(\nabla\underset{\sim}{w})^T , \\ \underset{\sim}{\tau} = 0 \quad \text{on} \quad \Gamma_1 \quad (= \text{part of} \quad \Gamma \quad \text{with} \quad \underset{\sim}{u}_0 \cdot \underset{\sim}{n} < 0) . \end{cases}$$

On the other hand, by differentiating the adjoint state equation (6) with respect to time, we have

(11)
$$\int_\Omega \underset{\sim}{\tau}\cdot\underset{\sim}{D}(\chi)d\underset{\sim}{x} = \int_\Omega \underset{\sim}{\tau}\cdot[-\lambda(\underset{\sim}{u}\cdot\nabla)\underset{\sim}{H} - \lambda(\nabla\underset{\sim}{u})^T\underset{\sim}{H} - \lambda\underset{\sim}{H}(\nabla\underset{\sim}{u}) + \underset{\sim}{H}]d\underset{\sim}{x} .$$

Integrating (11) by parts, and taking the incompressibility constraint div $\underset{\sim}{u} = 0$ into account, (11) yields

$$\int_\Omega \underset{\sim}{\tau}\cdot\underset{\sim}{D}(\chi)d\underset{\sim}{x} = \int_\Omega \underset{\sim}{H}\cdot\{\lambda(\underset{\sim}{u}\cdot\nabla)\underset{\sim}{\tau} - \lambda\underset{\sim}{\tau}(\nabla\underset{\sim}{u})^T - \lambda(\nabla\underset{\sim}{u})\underset{\sim}{\tau} + \underset{\sim}{\tau}\}d\underset{\sim}{x} ,$$

which, from (10), is equivalent to

(12)
$$\int_\Omega \underset{\sim}{\tau}\cdot\underset{\sim}{D}(\chi)d\underset{\sim}{x} = \int_\Omega \underset{\sim}{H}\cdot\{2\mu\underset{\sim}{D}(\underset{\sim}{w}) - \lambda(\underset{\sim}{w}\cdot\nabla)\underset{\sim}{\sigma}_D + \lambda(\nabla\underset{\sim}{w})\underset{\sim}{\sigma}_D + \lambda\underset{\sim}{\sigma}_D(\nabla\underset{\sim}{w})^T\}d\underset{\sim}{x} .$$

Plugging (12) back in (9) finally gives (8) and our proof is complete.

## 5. FINAL FLOW CHART.

With the gradient given by (8), the final flow chart corresponding to our conjugate gradient method of §3 is

## INITIALIZATION

+ $\underset{\sim}{u}^0$ given;
+ compute $\underset{\sim}{\sigma}_D(\underset{\sim}{u}^0 - \underset{\sim}{u}_0)$ by INTEGRATION OF (2) ALONG TRAJECTORIES;
+ compute the right-hand side $\underset{\sim}{r}^0$ of the state equation (5);
+ compute the state vector $\underset{\sim}{y}^0$ by SOLVING THE LINEAR STOKES PROBLEM (5);
+ compute the adjoint state $\underset{\sim}{H}^0$ by INTEGRATION OF (6) ALONG TRAJECTORIES;
+ compute the right-hand side $\underset{\sim}{j}^0$ of the gradient equation by (8);
+ compute the gradient $\underset{\sim}{g}^0$ by SOLVING A LINEAR STOKES PROBLEM;
+ set $\underset{\sim}{z}^0 = \underset{\sim}{g}^0$.

## LOOP ON $n \geq 0$

+ compute $\rho_n = \text{Argmin}\,(J(\underset{\sim}{u}^n - \underset{\sim}{u}_0 - \rho\underset{\sim}{z}^n))$ BY POLYNOMIAL INTERPOLATION;
+ set $\underset{\sim}{u}^{n+1} = \underset{\sim}{u}^n - \rho_n\underset{\sim}{z}^n$;
+ compute $\underset{\sim}{\sigma}_D(\underset{\sim}{u}^{n+1} - \underset{\sim}{u}_0)$ BY INTEGRATION OF (2) ALONG TRAJECTORIES OF $\underset{\sim}{u}^{n+1} + \underset{\sim}{u}_0$;
+ compute the right-hand side $\underset{\sim}{r}^{n+1}$ of the state equation (5);
+ compute the state vector $\underset{\sim}{y}^{n+1}$ by SOLVING THE LINEAR STOKES PROBLEM (5);
+ compute the adjoint state $\underset{\sim}{H}^{n+1}$ by INTEGRATION OF (6) ALONG TRAJECTORIES;
+ compute the right-hand side $\underset{\sim}{j}^{n+1}$ of the gradient equation by (8);
+ compute the gradient $\underset{\sim}{g}^{n+1}$ by SOLVING A LINEAR STOKES PROBLEM;
+ compute $\underset{\sim}{z}^{n+1} = \underset{\sim}{g}^{n+1} + \underset{\sim}{z}^n \langle A\underset{\sim}{g}^{n+1}, \underset{\sim}{g}^{n+1} - \underset{\sim}{g}^n\rangle / \langle A\underset{\sim}{g}^n, \underset{\sim}{g}^n\rangle$ .
+ end loop.

Here, the minimization of $J(\underset{\sim}{u}^n - \underset{\sim}{u}_0 - \rho\underset{\sim}{z}^n)$ by quadratic polynomial interpolation is achieved by

+ computing $J(0) = \frac{1}{2}\langle A\underset{\sim}{y}^n, \underset{\sim}{y}^n\rangle = \frac{1}{2}\,\underset{\sim}{r}^n \cdot \underset{\sim}{y}^n$;
+ computing $J'(0) = -\langle A\underset{\sim}{g}^n, \underset{\sim}{z}^n\rangle = -\underset{\sim}{j}^n \cdot \underset{\sim}{z}^n$;
+ computing $\underset{\sim}{v}_1 = \underset{\sim}{u}^n - \rho_{n-1}\underset{\sim}{z}^n$;
+ computing $\underset{\sim}{y}_1 = \underset{\sim}{y}_1(\underset{\sim}{v}_1)$ by inverting (5) with $\underset{\sim}{v} = \underset{\sim}{v}_1$;
+ computing $J(\rho_{n-1}) = \frac{1}{2}\langle A\underset{\sim}{y}_1, \underset{\sim}{y}_1\rangle$;
+ setting $\rho_n$ equal to the minimizer of the parabola which agrees with $J$ twice at 0 and once at $\rho_{n-1}$.

## 6. NUMERICAL IMPLEMENTATION.

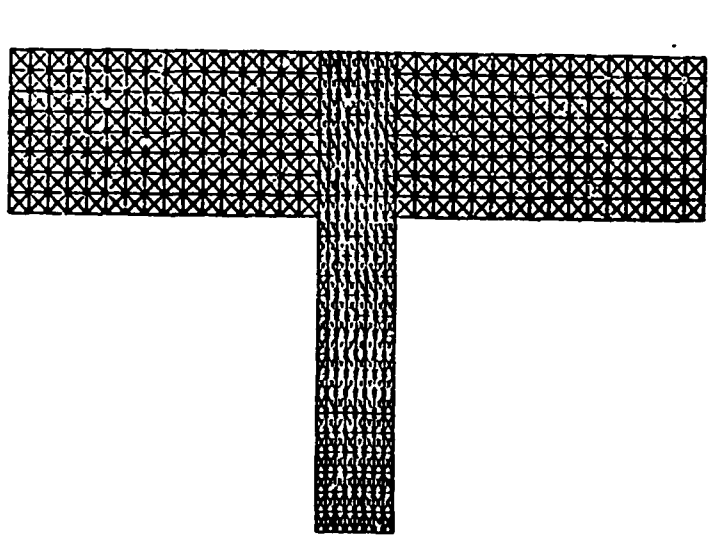
The practical implementation on a computer of the above flow chart requires the solution of two numerical problems:

(i) what type of approximation can be used for the numerical solution of the Stokes problems?

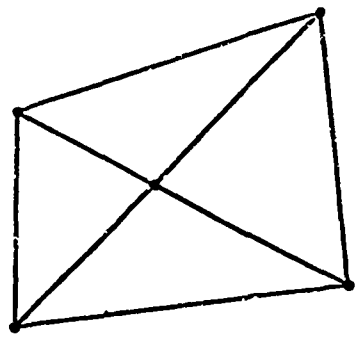
(ii) what numerical integration techniques can be used for the integration along the trajectories, while respecting the mechanical objectivity of the process?

Those problems are strongly interconnected since, for example, the finite element which is used determines the aspect of the computed trajectories. D. Malkus [1984] proposes answers which are very attractive because they respect the physical structure of the problem. His technique decomposes as follows:

(i) choice of an exactly incompressible piecewise linear finite element (such as the linear crossed triangle) for approximating the velocity field;



mesh of the domain



the original finite element

Fig. 2: Linear Crossed Triangle

(ii) exact computation of the trajectories incoming at the center of each finite element through a piecewise analytical solution of the ordinary differential equation

$$\dot{X}_t(\underset{\sim}{x},\tau) = \underset{\sim}{u}[X_t(\underset{\sim}{x},\tau)], \quad X_t(\underset{\sim}{x},t) = \underset{\sim}{x} ;$$

(iii) computation of the deformation gradient history by solving analytically the equation

$$\dot{\underset{\sim}{F}}_t(X_t(\underset{\sim}{x},\tau)) = \underset{\sim}{\underset{\sim}{\nabla u}}[X_t(\underset{\sim}{x},\tau)]\underset{\sim}{F}_t(X_t(\underset{\sim}{x},\tau)), \quad \underset{\sim}{F}_t(\underset{\sim}{x},t) = \underset{\sim}{1} ;$$

(iv) computation of the added stresses $g_D$ by a Laguerre type numerical quadrature

$$g_D(\underset{\sim}{x}) = \int_{-\infty}^{t} m_1(\tau)\ [\underset{\sim}{F}_t(\underset{\sim}{x},\tau)]\,d\tau \approx \sum_{i=1}^{NT} W(\tau_i) m_1(\tau_i)\ (\underset{\sim}{F}_t(\underset{\sim}{x},\tau_i)) \ .$$

The numerical quadrature of (iv) slightly changes the constitutive law but respects its objectivity since the trajectories and deformation gradients are exactly computed.

If we use D. Malkus' ideas, our numerical technique for the solution of (1)-(3) finally reduces to

a) the transformation of the original equations (1)-(3) into an equivalent minimization formulation (§2),

b) the solution of this minimization problem by the conjugate gradient algorithm of §5,

c) the solution of each Stokes problem involved in this algorithm by a finite element method using linear crossed triangles for the approximation of the velocity field,

d) the numerical integration of the integrals along trajectories by a Laguerre numerical quadrature, the trajectories being computed analytically.

If needed, an upwinding scheme can be added to this method, simply by replacing in the state equation (5) the term $(\underset{\sim}{u}\cdot\underset{\sim}{\nabla})\underset{\sim}{u}$ by

$$\frac{1}{\varepsilon}\ [\underset{\sim}{u}(\underset{\sim}{x}) - \underset{\sim}{u}(X_t(\underset{\sim}{x},t-\varepsilon))] \ .$$

This replacement renders our method fully deterministic and corresponds, in the case of a transonic flow computation, to the usual entropy condition.

## 7. CONCLUSIONS

As described, the proposed numerical technique can be expected to be cheap, because it only involves fixed positive definite sparse matrices (namely those associated to the linear Stokes problem), and to be able to handle hyperbolic situations, because it is based on a least-squares formulation. Moreover, it respects the physical objectivity of the problem and uses, in a critical way, the special structure of the adjoint problem.

Here, we have restricted ourselves to the case of Maxwell viscoelasticity. Of course, the proposed techniques can be generalized to any viscoelastic constitutive law, provided that we can easily compute the term

(13) $$\int_{\Omega} (\underset{\sim}{g}_D'(\underset{\sim}{v})\cdot\underset{\sim}{w})\cdot\underset{\sim}{D}(\underset{\sim}{y})\,d\underset{\sim}{x} \ .$$

In the case of constitutive laws which reduce to a differential equation, this can usually be done by the techniques of §4, through the introduction of an adjoint differential equation which defines an adjoint state. However, for the general case of a constitutive law given under an integral form, an efficient procedure for computing (13) is still to be found.

## REFERENCES


R. B. Bird, H. H. Saab, C. Curtiss [1982, J. Chem. Phys, 77, p. 4747.

D. Joseph, M. Renardy, J. C. Saut [1984], "Hyperbolicity and change of type in the flow of viscoelastic fluids". MRC Report, March 84.

R. Glowinski [1984], "Numerical methods for nonlinear variational problems", Springer Verlag.

D. Malkus, B. Bernstein [1984], "Flow of a Curtiss-Bird fluid over a transverse slot using the finite element Drift function method", to appear in J. Non-Newtonian Fluid Mech.

# Programming with Binary Relations and an Associated Algebra of Programs

*Paul Broome*

Ballistic Research Laboratory
SECAD
Aberdeen Proving Ground, MD 21078

and

Department of Computer and Information Science
University of Delaware
Newark, DE 19711

ABSTRACT

Mathematical logic is a particularly fruitful vehicle for expressing programs in a declarative style and seems well suited for artificial intelligence applications. On the other hand, Horn clause logic limited to conjuction and disjunction is somewhat primitive and low level. When programs are written as arbitrary n-ary relations, as in Prolog, hierarchically constructing more complex programs from simpler ones is often quite awkward.

John Backus, in his 1977 Turing award lecture, described a well structured and expressive functional programming style along with a useful algebra of programs. However, programs as functions are less general than programs as relations.

This work restricts logic programs to only binary relations and shows how to combine the backtracking of logic programming and the higher order operations of functional programming. This combination allows us to define a richer set of program forming operations that includes program inversion. We include new optimization rules and point out which FP-like rules are invalid in a relational context. As an example, we show how the algebra is used by synthesizing, from specifications, an efficient program for the N queens problem.

## Introduction

The purpose of this paper is both to show the advantages of describing programs as binary relations, or more accurately "set valued functions", and to give laws relating program forming operations whose arguments are relations. As an exhibition of the power in an algebra of binary relations we synthesize a program for the N queens problem from a general specification.

Warren [13] described how to define higher order operators in Prolog and argued that extending the Prolog language definition to include them was, in large part, unnecessary. As he says, they do not extend the power of the language.

One of Warren's examples is an accumulation operator, called iterate. He says that if predicate variables are used in more than small doses the program becomes excessively abstract and therefore hard to understand.



PREVIOUS PAGE IS BLANK

```
iterate([],R,Identity,Identity).
iterate([First|Rest],R,Identity,Result) :-
        iterate(Rest,R,Identity,MidValue),
        R(First,MidValue,Result).
```

We can think of the last argument as the result of accumulating the relation R, with identity 'Identity', over a sequence. The first clause says that the result of accumulating the relation R over an empty sequence is the identity. The second clause breaks the sequence up into a head and a tail, accumulates over the tail of the sequence to find a middle value, and finally applies R to the head of the sequence and the middle value to give the result. Some examples of uses of this operator are

```
iterate([1,2,3,4],(+),0,10).

iterate([1,2,3,4],(*),1,24).

iterate([[1,2],[],[3,4]],append,[],[1,2,3,4]).
```

These definitions are difficult to read partly because variables that represent predicates and variables that represent objects (numbers) are mixed in at the same level. This straightforward mixture of higher order operators and Prolog clauses leads to a difficulty of combining programs defined with operators.

We argue that higher order operators allow us to lift our level of programming. We can write programs that operate in bigger chunks. Operators act as recursion pattern templates and save the programmer from always relying on assertions based on induction to define his recursions. We can also give concise optimization rules between different operators. So we are claiming that higher order operators can play an important role in logic programming. On the other hand, it is commendable that such operators are so easily defined in Prolog.

A program forming operation is a very useful tool for accomplishing sophisticated data abstraction. The solution that is being proposed here depends heavily on composition of relations as a program forming operation (PFO) and allows other PFOs to name only their relational arguments. A PFO with its arguments is a term that is to be rewritten in a forward direction. The final result of this term rewriting is also a relation. Backtracking occurs at the lower, object level.

An example of a PFO is map. The term map(F) is a new relation between inputs and outputs (unnamed) that is formed from the relation F. If F is a relation of type A x A, i.e. it is a set of ordered pairs of type A, then the relation map(F) is of type *A x *A, a set of ordered pairs of sequences of type A. Thus map is a function, a PFO, from relations into relations; its type is AxA -> *Ax*A.

We will distinguish two levels of program definition. The first describes how to relate objects and allows definition of new nondeterministic relations. The second level allows definition of new relations only in terms of program forming operations or previously defined relations. New PFOs can be defined at the second level. No new nondeterminism can be introduced at the second level.

Our solution includes restricting our attention to binary relations and depending heavily on an infix functor, ';', for composition of relations.

In first level definitions new relations are created by asserting set memberships. An infix operator 'in', meaning 'element of', will represent a test of membership. Thus '(X,Y) in R' can be read as "R is applied to (is true of) the pair (X,Y)" or as "(X,Y) is an element of the relation R." This, along with a representation of infix composition of relations that doesn't require us to talk about arguments, gives us a powerful and expressive notation.

Therefore a first level definition for map(F) would be

```
([],[]) in map(F).
([A|B],[C|D]) in map(F) :-
        (A,C) in F,
        (B,D) in map(F).
```

and reduction can be defined as

```
([],Z) in accum(F,Z).
([A|B],D) in accum(F,Z) :-
        ([A,Z],X) in F,
        (B,D) in accum(F,X).
```

A third operator is 'gen(P,F)' which generates successive applications of F until P holds.

```
(X,[X]) in gen(P,F) :-
        (X,true) in P.
(X,[X|Y]) in gen(P,F) :-
        (X,false) in P,
        (X,Z) in F,
        (Z,Y) in gen(P,F).
```

For example, it holds that (5,[5,4,3,2,1]) in gen(eq1,sub1), ([3,2,1],[9,4,1]) in map(sqr), and ([5,4,3,2,1],15) in accum(+,0).

We interpret F;G to mean that F;G is true of (A,B) if F is true of (A,C) and G is true of (C,B).

```
(A,B) in F;G :-
        (A,C) in F,
        (C,B) in G.
```

As an example we can compute the sum of squares of a few integers as

```
([1,2,3,4],Y) in map(sqr);accum(+,0).
```

The second level definitions state an equivalence between relations. The left hand side is the more abstract version; the right hand side the more desirable from an efficiency standpoint. For example, we may define inner product of two vectors as

```
ip => transpose;map(*);accum(+,0).
```

Where 'transpose' transposes a matrix.

Second level definitions are not restricted to just new relations but can also be used to give optimization rules. For example a loop merging optimization rule for map can be stated as

```
map(F);map(G) => map(F;G).
```

with the understanding that '=>' means 'is equivalent to and is to be rewritten as', i.e.

```
F => G iff ((X,Y) in F iff (X,Y) in G).
```

Now we can painlessly add new higher order operators and write second level definitions that are like "logic programs in the style of APL." Because composition will be from left to right it is more suggestive of FQL [6]. The arguments given by Backus [1] apply to programs of this sort.

We write definitions at the second level as rewrite rules without bringing ourselves down to the object level. Even if the objects in question are nondeterministically specified we can still use rewrite rule theory to reason about these program forming operations. Since they are deterministic, and if we are careful, the rules can be designed to obey Church-Rosser properties. As another example, we may assert a definition for the inner product of two vectors, i.e.

```
ip => transpose;map(*);accum(+,0).
```

Where transpose transposes a matrix. The symbol 'ip' represents an abstraction of the computation 'transpose;map(*);accum(+,0)'. An example of the reasoning that would go along with such a definition is

```
([[1,2,3],[4,5,6]],Y) in ip
=> ([[1,2,3],[4,5,6]],Y) in transpose;map(*);accum(+,0).
=> ([[1,4],[2,5],[3,6]],Y) in map(*);accum(+,0).
=> ([4,10,18],Y) in accum(+,0).
=> Y=32.
```

**Selectors, Constructors and Nondeterminism**

A convenient way of sharing an input to more than one relation is with Backus' constructor functional. We read [F|G] as a sequence of relations whose head is F and whose tail is G. The relation [F|G] is true of an input/output pair if F is true of the input and the head of the output and G is true of the input and the tail of the output. More formally,

```
(X,[]) in ||.
(X,[U|V]) in |F|G| :-
        (X,U) in F,
        (X,V) in G.
```

Here it is understood that since [U|V] is in the object position then U and V represent components of a sequence of objects. For example, (3,[9,27]) in [sqr,cube].

In like fashion we define selectors that disassemble what the constructors build. We capture the Kth component of a sequence with #K. The relation, tl, gives us the rest of a sequence. The identity is id.

```
(X,X) in id.
([X|Y],X) in #1.
([X|Y],Z) in #K :-
        J is K-1,
        (Y,Z) in #J.
([X|Y],Y) in tl.
```

We try to localize instances of nondeterminism to easily transform programs that do not use nondeterminism. In this paper we describe all nondeterminism with a 'select' relation A sequence is the input to select. The value returned from it is a pair: one (any) element of the sequence and a subsequence of all other elements. We can describe the set of ordered pairs as

```
([X|Y], [X,Y]) in select.
([X|Y], [U,[X|V]]) in select :-
                (Y, [U,V]) in select.
```

Some sample pairs in the relation select are ([1,2,3],[1,[2,3]]), ([1,2,3],[2,[1,3]]) and ([1,2,3],[3,[1,2]]). These are all the ways of selecting an element from the sequence [1,2,3].

**Guards**

In our representation of computations as binary relations we often must package a collection of inputs into a single structured input. So for example, the Prolog clause append(A,B,C) must be written as ([A,B],C) in append. Likewise, we will often need to extend predicates of single arguments so that they can return Boolean values, for example (3,true) in odd. These extensions are important because we can use higher order operators with Boolean valued relations, e.g.

every(P) => map(P);accum(and,true).

Thus 'every(odd)' can check every element of a sequence for oddness, e.g. ([1,3,17],true) in every(odd).

A guard is any Boolean valued relation. Backtracking choices are created with 'select' and pruned with guards. These definitions are patterned after Dijkstra's guarded command language [5] except that our nondetermism is uncommitted. In (P -> F), P is a precondition for the input to F and in (F <- P), P is a postcondition for the values returned from F. More precisely,

```
(A,B) in (P -> F) :-
                (A,true) in P
                (A,B) in F.
```

Postconditions are appropriate for the 'generate and test' paradigm. This form tests the output of a relation with a guard.

```
(A,B) in (F <- P) :-
        (A,B) in F,
        (B,true) in P.
```

The first clause, (A,B) in F, may generate several feasible solutions for B, only a few of which may satisfy P For example an inefficient way to sort a sequence is to permute the sequence and then check that it is ordered.

sort => (perm <- ordered).

**Inversions with respect to composition**

An unusual program forming operation, one that can play a major role in program transformations, is the inverse. The program inv(F) is formed from F by running F backwards,i.e. with known output but unknown input. The following is an obvious definition but is often useless in a sytem where goals are solved in depth first order.

```
(X,Y) in inv(F) :-
        (Y,X) in F.
```

For example the expression ([],Y) in inv(accum(append,[])) cannot be solved for Y with a depthfirst search.

However, the concept turns out to be useful. It allows us to retain a left to right order of computation and make effective use of previously defined relations. By assuming that left hand arguments (inputs) are known and that outputs are to be determined, we can control the extent of our searches. This is done by ordering our subgoals so that the computation tree is not bushy but stringy. Unfortunately, the above definition for inverse is not an efficient computation rule and may even lead to a nonterminating computation.

A logic programming system with the completeness property (all clauses with a solution can be solved) cannot carry out all searches in depth first order. However, a totally breadth first search is overkill because, although such a method is complete, it is inefficient. We want to do most searches, those that are assured of termination, in depth first fashion and carry out the rest breadth first. The problems caused by the interaction of inverses with the depth first solution strategy are being examined in another paper [2]. It includes a collection of operator directed transformation rules that 'solve' the program for an inverse program.

**Limitations of the algebra**

If true functions are evaluated deterministically there is still a practical concern. It is often difficult to prove termination even for deterministic computations. Our approach is to interpret the constructor function under a lazy, rather than strict, semantics. In fact, lazy evaluation is

required in order to talk about potentially unbounded structures which arise when interacting with external physical devices.

The following logic program fails if g(X,V) fails. But failure may not be the desired behavior if [U|V] represents a potentially unbounded sequence of which we only need the head.

```
h(X,U) :- f(X,U),g(X,V).
```

Equivalently, the rule

```
[F,G];#1 => F.
```

does not hold under a strict semantics for at least two reasons. As stated, if G fails either by returning an undefined value or by not terminating then the left hand expression fails whereas the right hand side may still succeed.

A less obvious discrepancy involves domain elements with uninstantiated variables. If, for example, G succeeds only when the input is guarded, then G may instantiate an undefined variable to that value. Therefore F in the left hand expression will be solved under the assumption that G succeeds and would be more defined than what the right hand side would find as a solution.

More specifically, consider the following relation

```
[id,(eq4 -> id)];#1.
```

It is not equivalent to 'id.' The difference is noticed if the relation is solved with both unspecified domain and range elements.

```
(X,Y) in [id,(eq4 -> id)];#1
        =>
          (X,U) in id,
          (X,V) in (eq4 -> id),
          [U,V]=Z, #1:(Z,Y).
        =>
          X=U,
          (X,V) in (eq4 -> id),
          [U,V]=Z,
          U=Y.
        =>
          X=Y,
          (X,true) in eq4,
          X=V.
        =>
          X=Y,
          (4,true) in eq4,
          X=4.
        =>
          X=4,
          Y=4.
```

This solution is more specific than the solution of

```
(X,Y) in id.
```

This point must be considered when optimizing programs whose domains include 'difference lists.' The term difference list is often used to refer to an incompletely defined data structure that has become increasingly instantiated during a computation. However, the method could also apply to data structures other than lists such as trees. Operations such as the append of two difference lists can be performed in constant time because they can involve just the instantiation of variables instead of copying of entire structures. Another example is the process of

maintaining a queue; insertion of a new element can be done in constant time, if access to the uninstantiated variable at the end of the queue is shared and immediate.

Although there are methods for accessing and destructively updating recursive data structures ([9]) in functional languages, they involve more complex concerns and are more restrictive. In particular the paths of [9] do not allow sharing. Therefore only trees and not DAGs or graphs can be treated.

We define conditions under which we can apply these rules. One constraint that would assure applicability is to require that every domain element be completely defined and that G terminate on this value. This requirement is so strict that it rules out both difference lists and streams and is thus unsatisfactory.

Our solution is to isolate on a particular kind of difference list. Instead of creating two values with the constructor functional, we return a pair of suspensions, i.e. promises to compute the values. We will read the clause, 'X suchthat P' as 'a value X that has property P.' So we replace our previous definition of the constructor with

```
(X,|(U suchthat (X,U) in F)|(V suchthat (X,V) in G)|)
                                    in [F|G].
(X,||) in ||.
```

and replace our previous definition of composition with

```
(X,Y) in F;G :-
        (Z suchthat (X,Z) in F),Y) in G.
```

We must also include rules that recognize suspensions if their values are needed.

```
(X suchthat P,Y) in F :-
        var(X),
        P,
        (X,Y) in F.
(X suchthat P,Y) in F :-
        nonvar(X),
        (X,Y) in F.
```

These are equivalent to the rules that describe 'lazy evaluation' or 'call by need' [12], [8], [7] although our definition is much simpler. Vuillemin has shown that, as an argument evaluation rule, this mechanism is optimal. If values are never requested from subcomputations, then they never get computed. Lazy evaluation allows a more expressive style of programming because terminating conditions can often be ignored. One can conceptually build and manipulate infinite structures or streams.

The greatest disadvantage is that there is a constant overhead associated with making every structure a stream. The work of Mycroft [11] and Hudak [10] might be adjusted to optimize some lazy computations into equivalent strict computations. The greatest advantage that lazy evaluation holds for us is that algebraic laws are valid without other conditions such as success of irrelevant subcomputations. The computation carried out with lazily evaluated arguments agrees with what a mathematician would expect if he did the algebraic manipulation by hand. Thus the rule

```
[F,G];#1 => F
```

holds independently of whether G succeeds or fails. This is proved by showing that [F,G];#1 is a subrelation of G and vice versa.

The following is a simple example of stream oriented programming. It is a program that returns the infinite list of powers of two.

```
powersoftwo ==> @1;doubleall.
doubleall ==> [id|times2;doubleall].
times2 ==> [id,@2];(*).
```

The relation @1 is a constant function. It ignores all inputs and returns 1 as an output. The constructor in 'doubleall' must be evaluated lazily.

There are concerns to be voiced about interactions with backtracking and streams. A program that returns an infinite structure cannot be inverted. We must be careful about constructing streams because we cannot backtrack over a partially built structure that has, for example, already been printed. This point is relevant to the N queens problem that we consider in a later section.

**Towards an algebra of nondeterministic programs**

Before we look at the example we first consider the rules we will be using. Some of the rules from functional programming do not hold when combining relations. In particular, expressions that contain nondeterminism cannot be replicated or removed without care. We cannot, for example, distribute a nondeterministic expression whose results are shared ([4]). The rule from Backus' FP algebra [1] that distributes shared input to a constructor functional holds only if the program whose result is shared is a function, i.e.

```
F;[G,H] ==> [F;G, F;H] :-
                deterministic(F).
```

is valid only if F is deterministic. Also

```
[F,G];#1 ==> F :-
                deterministic(G).
```

requires that the number of replications of [F,G] be determined only by the nondeterminism in F, not in G. This is because our relations are built on a multiset model instead of on sets. In general, we must be careful about applying any rules that change the number of occurrences of an expression.

Finally, if we maintain a lazy constructor semantics we cannot propagate guards into the components of a constructor. For example,

```
([F,G] <- #1;P)
```

is a subrelation of

```
[(F <- P),G].
```

Some of the following rules will be used in the derivation of the N queens program.

```
(F <- G;P) ==> (F;G <- P);inv(G)
                if G;inv(G) ==> id.

map(F);every(P) ==> every(F;P).

(F;G <- P) ==> (F <- G;P);G
                if G is deterministic

gen(P,F);tl ==> F;gen(P,F).

gen(P,F);last ==> loopuntil(P,F).

(gen(P,F) <- every(Q)) ==> (gen(P,(F <- Q)) <- defined)
```

The last rule shows one problematical interaction between streams and backtracking. The expression gen(P,(P <- Q)) generates, until P becomes true, a stream whose elements satisfy Q. If somewhere later in the stream Q fails, then the entire stream becomes invalid as an answer. In operational terms we should "back up the line printer, erasing invalid characters."

We avoid this problem by adding the postcondition 'defined' to strictify the stream. Therefore the stream is not passed as valid until the entire structure has been created and checked.

### The N Queens Problem

The problem is to place N queens, on an N by N chessboard, in a manner such that no queen is attacking any other. This problem is a common example of backtracking, however the N queens problem is simpler to solve if we don't think of it as a traditional backtracking problem. Instead, we just consider applying a restricting condition to a set of possible answers. In other words, we view it as just another instance of the generate and test paradigm.

We start with a simple, understandable version of the N queens problem that generates every possible board position and then tests each for validity. Through the successive application of rewrite rules, we transform this program to a more efficient one. However, we do not prove the rules in this paper.

The N queens problem is first solved by generating a feasible set and then testing each to see if all queens are safe. The original feasible set is just the set of all permutations of placements of the N queens in N columns. So initially there are N factorial elements in the set. The relation that generates this set is just

```
iota;perm.
```

One elemental pair in this relation is (4,[3,2,1,4]). This is a placement of queens by columns: first column, third row; second column, second row and so on. This is only a tentative solution and not a solution. A solution is (4,[2,4,1,3]).

We use a random selector relation that picks an item from within a sequence. This is the source of all nondeterminism in the N queens problem. A logic programming definition for select would be

```
([X|Y],[X,Y]) in select.
([X|Y],[U,[X|V]]) in select :-
        (Y,[U,V]) in select.
```

These permutations are screened according to the condition that there must be no diagonal conflicts. The representation makes it impossible to have row or column conflicts

Our specification of the N queens program is reminiscent of Clark and Darlington's specification of sort[3]. That specification generates all permutations of the sequence and accepts only those that are ordered. Several efficient sort algorithms are synthesized from this specification. Correspondingly, we generate all permutations of 1..N and accept only those boards where all queens are safe. Our straight forward but inefficient program for N queens is as follows.

```
nqueens => (iota;perm <- nonesamediag).
```

where

```
iota => gen(eql,sub1);rev.

perm => [[],id];
        gen(#2;null, [#1,#2;select];
                     [[#2;#1|#1], #2;#2]);
        tl;map(#1);last.

nonesamediag => splitoff;
                every(noconflict).

noconflict => [#1,tl];
              [distl;map(absdiff), #2;len;iota];
              trans;every(ne).

splitoff => gen(tl;null,tl).

every(P) => map(P);accum(and,true).
```

Our objective is rewrite this inefficient, but straightforward, definition into a more efficient, but possibly less understandable, program.

Before we continue along these lines there is a particular kind of sequence of sequences that needs attention. We define a fading permutation as a sequence with these properties:

1. The head of the sequence is a permutation.

2. The i+1th element of a sequence is the tail of the ith element.

3. The last element is itself a sequence but with one element.

Given a permutation the following program is one way to create a fading permutation.

```
gen(tl;null, tl).
```

We must recognize the fact that the program

```
#1;gen(tl;null, tl)
```

is an identity on fading permutations. We return to the original problem.

We can rewrite nqueens as

```
nqueens => (iota;perm
                <-
            gen(tl;null,tl);every(noconflict)).
```

Noting that #1 is a right inverse for splitoff we have

```
nqueens => (iota;perm;gen(tl;null,u)
                <-
            every(noconflict));
           #1.
```

We concentrate our attention on perm;splitoff and substitute rev;#1 for last.

```
perm;gen(tl;null,tl)

 =>   [[],id];gen(#2;null, [#1,#2;select];
                            [[#2;#1[#1], #2;#2]);
                tl;map(#1);rev;#1;
                gen(tl;null,tl).

 =>   [[],id];gen(#2;null, [#1, #2;select];
                            [[#2;#1[#1], #2;#2]);
                tl;map(#1);rev.
```

as long as we know that this final expression generates fading permutations.

To show that this generates fading permutations we note that the first element of a resulting sequence (after being reversed, i.e. originally the last) is a permutation. This is because the first element of this expression is equivalent to the definition of perm. Next we note that successive elements are sequences that differ only by a missing first element. Thus the i+1th element is the tail of the ith element. Finally we note that the last element of the fading permutation has only one element. From these three facts we claim that our use of the simplification rule is justified.

Now we note that since 'and' is associative and commutative we can say

```
rev;every(P) => every(P).
```

Also we know that

```
(F;G <- P) => (F <- G;P);G.
```

where G is deterministic. Using these two rules we now have the expression

```
nqueens =>
          (iota;[[],id];
           gen(#2;null, [#1, #2;select];
                        [[#2;#1[#1], #2;#2]);tl
               <-
           map(#1);every(noconflict));
          map(#1);last.
```

We can move tl out of the way with the following rule

```
gen(P,F);tl => F;gen(P,F)
```

and we can merge map(#1) with every(noconflict) with

```
map(F);every(P) => every(F;P).
```

Finally the constraint that for every new queen there be no conflict can be propagated into the gen loop with the following rule

```
gen(P,F) <- every(Q) => (gen(P,(F <- Q)) <- defined).
```

So we now have the expression

```
nqueens =>
        iota;
        [[],id];
        [#1,#2;select];
        [[#2;#1|#1],#2;#2];
        gen(#2;null, ([#1, #2;select];
                      [[#2;#1|#1], #2;#2]
                        <- #1;noconflict));
        map(#1);
        last.
```

or

```
nqueens =>
        [[],iota;select];
        [[#2;#1|#1],#2;#2];
        loopuntil(#2;null, ([#1, #2;select];
                            [[#2;#1|#1], #2;#2]
                              <- #1;noconflict));
        #1.
```

This final program is more obscure but much more efficient than the original.

### Further Work

A system that can do these transformations automatically would have to include a fairly large collection of rules, although the operator direction helps to reduce the number. We are working on rules that can derive inverses of binary relations. They are important for this work and should be more generally useful to other kinds of program transformations such as the Burstall-Darlington unfold/fold method.

Even the synthesis shown in this paper has some weak spots that need filling. The most direct definition of 'perm'

```
perm =>
        select;
        gen(#2;null, #2;select);
        map(#1).
```

does not easily lead to the program given here although this does not appear to be a serious problem. We also should be able to transform the definition of 'noconflict' into a more efficient form such as

```
noconflict =>
        [@true, @1, #1, tl];
        loopuntil(#2;tl;null,
                  [[#1,[#2,[#3,#4;#1];absdiff];ne];and,
                   #2;add1,
                   #3,
                   #4;tl]);
        #1.
```

Transforming 'noconflict' into this program seems to require a different collection of rules than given in this paper.

## Discussion of Applications

As the kinds of programs we write become more complex and more ambitious, they become more difficult to write efficiently. In particular, when separate programs are combined, their loops should often be combined. The gap between efficiency and generality is becoming more difficult to bridge and the need for program transformation systems is becoming apparent. This work views such a program transformation system as a kind of computer algebra package that manipulates programs as expressions.

A second important application includes the formal modeling of the nondeterminism inherent in message passing in networks. We hope that properties of a model can be proved by algebraic simplification and transformation in place of a simulation.

## Conclusions

We have exhibited an operator directed method for reasoning about nondeterministic programs. The notation is a combined form of functional programming and logic programming. This method does not depend on recursion and its associated induction arguments but instead represents patterns of recursion with program forming operations and uses rules relating them.

The most important result in this paper is the synthesis, from a formal specification, of an efficient backtracking program for the N queens problem. We used the method of propagation of constraints so that nonsolutions are rejected before they are constructed. We showed that streams and backtracking may interact in unexpected ways and pointed to the occasional need to 'strictify' streams.


## Acknowledgements

This work greatly benefited from many hours of discussion with Thomas J Myers. We thank Shayla Moody for comments on the organization of this paper and the suggestions of Ken Cowen and Madhu Murthy are gratefully acknowledged. Thanks also go to Steve Wolff, Harry Reed, and R.J. Eicheiberger for providing an environment conducive to research.


## References


1. J Backus, Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs, *Comm. ACM 21*,8 (Aug. 1978), 613-641.
2. P. Broome, Program Transformation of Relations with Higher Order Operators and Streams, in preparation.
3. K. L. Clark and J. Darlington, Algorithm Classification Through Synthesis, *The Computer Journal 23*,1 (1980), 61-65.
4. A. T. Cohen and T. J. Myers, Toward an Algebra of Nondeterministic Programs, *Conf. Record of the ACM Symposium on LISP and FP*, 1982, 235-242.
5. E. W. Dijkstra, Guarded Commands, Nondeterminacy and Formal Derivation of Programs, *Comm. ACM 18*(1975), 453-457.
6. R. E. Frankel, FQL -- The Design and Implementation of a Functional Database Query Language, Contract N00014-75C-0462 (Rept 79-05-03).
7. D. P. Friedman and D. S. Wise, CONS Should Not Evaluate Its Arguments, in *Automata, Languages and Programming*, S. Michaelson and R. Milner (editors), Edinburgh University Press, Edinburgh.
8. P. Henderson and J. H. Morris, A Lazy Evaluator, *Conf. Record ., ie 3rd ACM Symposium on Principles of Programming Languages*, 1976, 95-103.
9. R. T. Hood, R. Cartwright and P. Matthews, Paths: An Abstract Alternative to Pointers, *Conf. Record of the 8th ACM Symp. on Principles of Programming Languages*, , 14-27.

10. P. Hudak and D. Kranz, A combinator-based compiler for a functional language, *Conf. Record of the 11th ACM Symp. on Principles of Programming Languages*, , 122-132.

11. A. Mycroft, The Theory and Practice of Transforming Call-by-Need into Call-by-Value, in *Proceedings of the Fourth International Symposium on Programming*, vol. LNCS 83, 1980, 269-281.

12. J. Vuillemin, Correct and Optimal Implementations of Recursion in a Simple Programming Language, *J. Computer and System Sciences* 9(1974), 322-354.

13. D. H. D. Warren, Higher-order extensions to Prolog: are they needed?, in *Machine Intelligence*, vol. 10, Hayes, Michie and Pao (editors), N.Y., 1982, John Wiley and Sons.

# CODE ITERATION FOR NOISY CHANNELS

A. Brinton Cooper, III

System Engineering and Concept Analysis Division

US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

## ABSTRACT

Error control codes having modest error correction capabilities and additional error detection features often are used beyond their correction abilities to force retransmission of messages under severe channel conditions. High noise or interference levels can cause a significant fraction of messages to experience multiple transmissions in order to provide successful delivery.

To improve the reliability of message delivery, a second level of error control is examined. A nonbinary Reed-Solomon (RS) code treats each detected character error as an erasure: the location of the error is known, but the identity of the transmitted symbol is not. The decoder then fills in symbols which the channel has "erased."

The performance of such an error control scheme in a typical application is studied, and the assumed scheme is shown to provide significant improvement for a modest change.

Concern about the numbers of message retransmissions often required by the occurrence of detectable but not correctable error patterns on degraded communication channels has prompted corrective suggestions including transmitting each message twice and eliminating acknowledgement messages. This suggestion probably will not play well in Peoria because the military are not likely to give up the unambiguous report which they feel is provided by receipt of ACK.

Typically, an error correcting code with modest error correction capabilities and the ability to detect somewhat more severe error conditions is used to correct some channel-induced errors and to indicate the existence of others. When such an error detection condition occurs, the datalink protocol [TANE81] suppresses acknowledgement of receipt of the message. After a timeout period, the source of the unacknowledged message will retransmit it. High noise or interference channels can cause a significant fraction of all messages to require multiple transmissions in order to achieve successful receipt.

To improve the reliability of message reception, a second level of error detection and correction is examined. A nonbinary Reed-Solomon (RS) code can treat each detected character error as an erasure: the location of the error is known, but the identity of the transmitted symbol is not. Decoders for error correcting codes not only correct received symbol errors but also fill in symbols which the channel has erased.

In what follows, the performance of error detection and correction in a typical system is determined for a representative communication channel, and augmentation of the assumed scheme is shown to provide significant improvement for a modest change.

## II. ERROR DETECTION AND CORRECTION IN MESSAGE COMMUNICATIONS

### A. INTRODUCTION

This note is concerned with the use of algebraic or block error correcting codes to improve message reception in message handling systems using very noisy communication channels. The performance of such codes is outlined in this section.

Figure 1 shows a model of the process by which information from a binary source (e.g., a message device) is conveyed to a destination in an accurate and timely manner over a noisy communication channel.




Figure 1. Model of Linear Block Codes

To each block of k information bits produced by the source, the encoder appends (n - k) redundant bits, each computed as a modulo-2 sum (linear combination) of at least some of the information digits. That is, they transmit no additional information but represent a form of controlled redundancy which is exploited at the decoder in order to recover the information actually transmitted. It is said that each redundant bit is a "parity check" on the information bits which constitute its sum.

The structure thus produced is called an (n,k) linear or block code. (Mathematically, a linear block code is a k-dimensional sub space of the vector space of n-tuples over a finite field of q elements, where q is an integer power of a prime number and, in this paper, is 2.)[PETE72]. The value of k is known as the dimension of the code; n is its length.

Rules for selecting the subsets of information digits to be checked by a parity digit are constructed so as to make the codewords pairwise as different as possible. If they are as different as possible, correct decoding can often be unambiguously accomplished by selecting as the transmitted codeword that which is "nearest" to the received n-tuple. Thus, while encoding produces a unique n-tuple for every block of k information digits, decoding must map many n-tuples into one k-tuple.

## B. THE BINARY SYMMETRIC CHANNEL

The channel model is shown in Figure 2.1. The random binary symbol generator produces a binary ONE with probability p and a ZERO with probability (1 - p). Information bits from the source and noise bits from the random symbol generator are added modulo-2 in the channel to produce output bits. Thus, an information bit will be inverted if and only if the noise bit is a ONE, and we say that the channel bit error probability is p. It is always assumed that $p < 0.5$ because if $p = 0.5$ the channel can transmit no information [GALL68] and if $p > 0.5$ symbol redefinition will make $p < 0.5$.

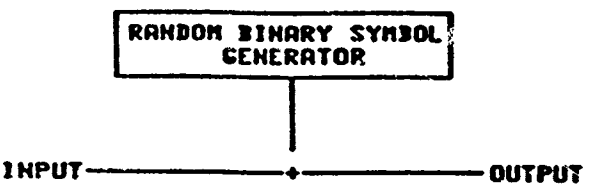



Figure 2.1 The Binary Symmetric Channel

This channel is known as the binary symmetric channel with transition probability p, BSC(p), and its behavior is represented by the state transition diagram of Figure 2.2.

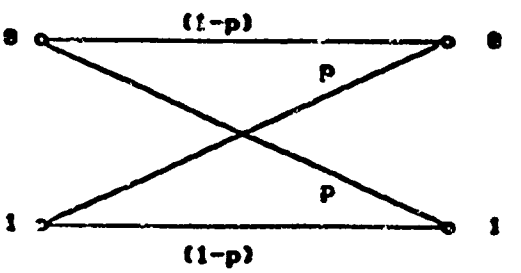



Figure 2.2 State Transitions in the Binary Symmetric Channel

The BSC provides a convenient vehicle for the comparison of error detection and correction techniques. No claim is made as to its fidelity in representing actual channels although it is a valid model for FM radios carrying binary data over long distances in the absence of external interference.

## C. HAMMING CODES AND SHORTENED HAMMING CODES

One aim of code design is to make the codewords as different as possible so that, when corrupted by channel noise, a received word tends to be nearer the word transmitted than to any other word. For

transmission over BSC(p), code word difference is expressed as Hamming distance: the number of positions in which the two words differ. It can be shown [PETE72] that the minimum distance d between two words over a given code guarantees that the code can correct any error pattern of t or fewer errors provided

$$t \leqslant [(d-1)/2]$$

where the notation indicates the integer part of the argument. For linear block codes, the minimum distance between two codewords is equal to the Hamming weight (number of non-zero positions) of the minimum weight, non-zero codeword.

Detailed structure of the codewords can be encapsulated in a (k x n) code generator matrix, G. A binary k-tuple is encoded by postmultiplying it by G to produce a length n codeword. Hence, all codewords are linear combinations of the rows of G. Each linear code

$$(v_1, v_2, \cdots v_n) = (a_1, a_2, \cdots a_k)G \tag{2-1}$$

also has an associated parity check matrix, H, with the property that the product of any codeword with the transpose of H gives zero:

$$vH^T = 0 \tag{2-2}$$

The relation between H and G can be seen by ordering the columns of H so that it assumes the form:

$$H = [Q|I_{n-k}] \tag{2-3}$$

The orthogonality property of (2-2) then causes the code generator to have the form [LIN&83]

$$G = [I_k|Q^T] \tag{2-4}$$

where $I_j$ is the jth order identity matrix and T indicates matrix transposition.

Hamming codes [HAMM50] are block codes having the capability to correct exactly one error per codeword. If an additional parity check is computed on the entire codeword, the Hamming decoder can detect any combination of two errors in a received word as well.

Codewords have length and dimension as shown in (2-5). All the non-zero m-tuples are the columns of the code's parity check matrix.

$$\begin{aligned} n &= 2^m - 1 \\ k &= n - m \end{aligned} \tag{2-5}$$

Thus, there is one Hamming code for each value of m.

The (15,11) single error correcting Hamming code has the parity check matrix, (2-6), produced by writing as columns all the binary 4-tuples, ordered numerically.

$$H = \begin{bmatrix} 000000011111111 \\ 000111100001111 \\ 011001100110011 \\ 101010101010101 \end{bmatrix} \tag{2-6}$$

To obtain the generator matrix of this code, the columns of the parity check matrix are reordered as in (2-3):

$$H1 = \begin{bmatrix} 1000000011111111 \\ 0100011110001111 \\ 0010101101100011 \\ 0001110110101011 \end{bmatrix} \tag{2-7}$$

The generator matrix, then, can be written according to (2-4) as:

$$G = \begin{bmatrix} 1000000000000111 \\ 0100000000001011 \\ 0010000000001101 \\ 0001000000001110 \\ 0000100000010011 \\ 0000010000010101 \\ 0000001000010110 \\ 0000000100011001 \\ 0000000010011010 \\ 0000000001011100 \\ 0000000000111111 \end{bmatrix} \tag{2-8}$$

The 16th column, an even parity check on the entire row, has been added for double error detection.

Any error correcting block code of length n can be shortened to length (n - s) by setting to zero s positions in the information vector. If the first s positions are those set to zero, then all codewords will begin with s zeros which need not be transmitted. This results in a code of length (n - s) and dimension (k - s). For example, to shorten the (16,11) code of (2-8), we can set the first four information positions to zero, resulting in the (12,7) shortened Hamming code of (2-9).

$$G = \begin{bmatrix} 100000010011 \\ 010000010101 \\ 001000010110 \\ 000100011001 \\ 000010011010 \\ 000001011100 \\ 000000111111 \end{bmatrix} \tag{2-9}$$

## D. PERFORMANCE OF (12,7) SHORTENED HAMMING CODE ON BSC(p)

It is useful to demonstrate the performance of this code at this point; the results will be needed later, as well. For this note, the binary symmetric channel will be assumed. Later analyses will deal with errors which occur in bursts.

According to the TACFIRE datalink protocol [TACF80], the occurrence of two bit errors in one character, which causes an error detection condition, prevents acknowledgement of receipt of the message; therefore, the source perceives failure of message receipt and retransmits the message. The probability of such an event is

$$P_{ce} = \binom{12}{2} p^2 (1-p)^{12-2} \tag{2-10}$$

Table 2.1 shows the detected character error probability (and, hence, the probability of a non-acknowledged message) as a function of the channel bit error probability for values of the latter from 0.003 to 0.10.

| p | $P_{ce}$ |
|---|---|
| 0.10 | 0.2301 |
| 0.08 | 0.1835 |
| 0.07 | 0.1565 |
| 0.06 | 0.1279 |
| 0.05 | 0.09879 |
| 0.04 | 0.07206 |
| 0.03 | 0.04380 |
| 0.02 | 0.02157 |
| 0.01 | $5.968 \times 10^{-3}$ |
| 0.008 | $3.898 \times 10^{-3}$ |
| 0.006 | $2.237 \times 10^{-3}$ |
| 0.005 | $1.569 \times 10^{-3}$ |
| 0.004 | $1.015 \times 10^{-3}$ |
| 0.003 | $5.764 \times 10^{-4}$ |

Table 2.1. Probability of detected error patterns on the BSC

These data are plotted in Figure 3.1.

## III. ERASURE CHANNELS AND REED-SOLOMON CODES

### A. INTRODUCTION

An iteration of encoding and decoding can be added to the scheme so far described. Essentially, the output of the original encoder can be further encoded according to the rules for another suitably chosen error correcting code. At the channel output, the original code can first be decoded as before and the result submitted to a second decoder for further processing [COOP78]. It is useful to postulate a different kind of channel when introducing the additional coding.

### B. THE ERASURE CHANNEL

In a received message, a character position where a detectable but uncorrectable error pattern has occurred in the channel can be considered as an erasure, i.e., a location where the decoder knows that an error pattern has occurred which it cannot correct. Linear block codes can handle erasures more handily than they can handle errors whose positions are unknown. For example, a code with minimum distance d can correct (fill in) e erasures in a received word where

$$d \geq e + 1 \tag{3-1a}$$

whereas

$$d \geq 2t + 1 \tag{3-1b}$$

where t = the number of errors correctable by the same code.

We now take a modified viewpoint and consider a noisy channel transmitting characters (binary m-tuples) rather than individual bits. [GORE73]. Characters either are received correctly from this channel or they are erased. The probability of a character erasure is the probability of any double weight error pattern. For the (12,7) shortened Hamming code discussed in Section II, this is given by (2-10).

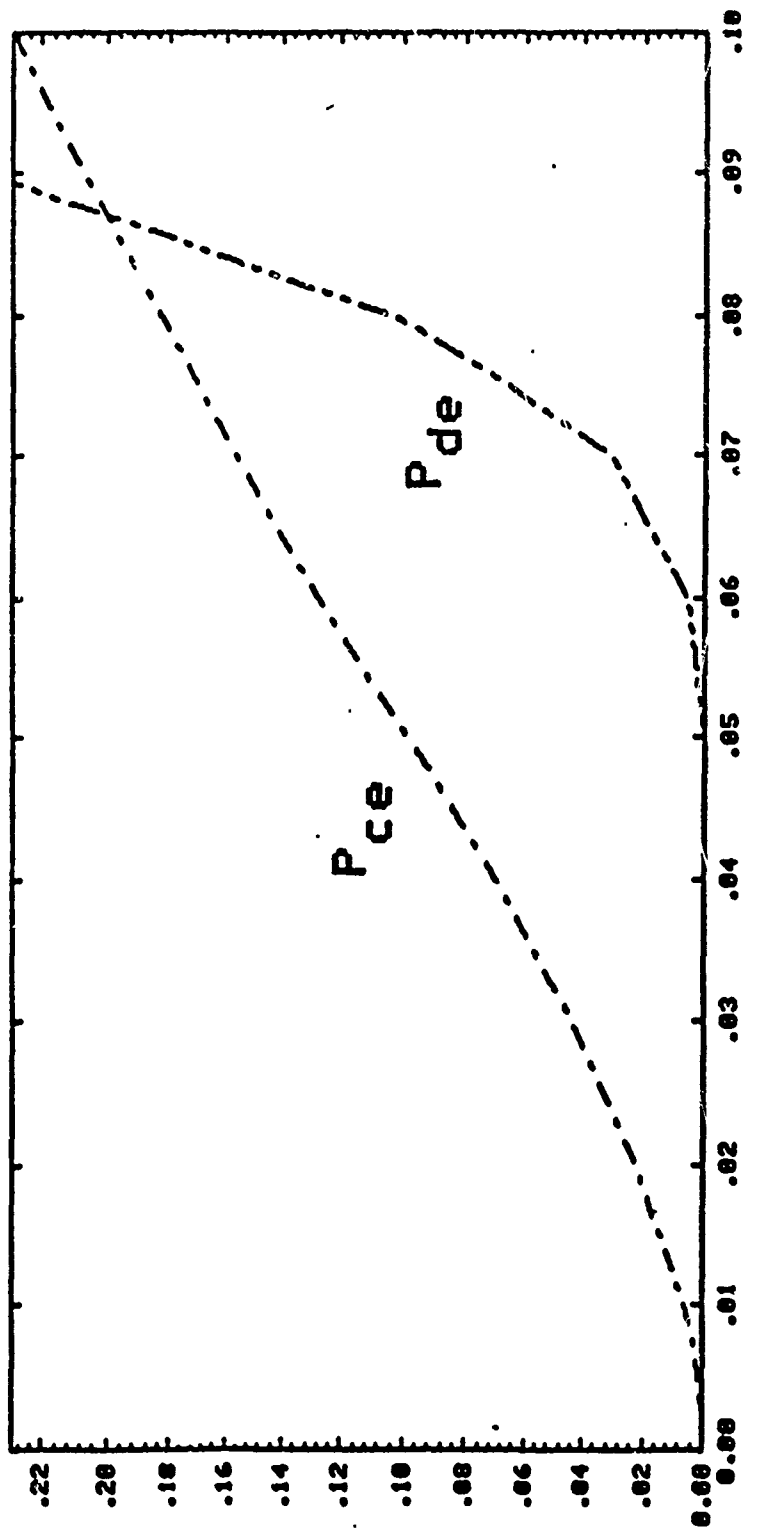



*Figure 3.1. Character Error Detection and RS Decoding Failure Probabilities*

So the channel under consideration accepts characters represented by binary m-tuples and presents to the destination a character erasure with probability given by (2-10). In what follows, an error detecting and correcting scheme to make this channel quite reliable is described.

## C. EXTENSION FIELDS AND THE BINARY SYMMETRIC CHANNEL

If binary symbols are manipulated m at a time, modern algebra permits all the ordinary arithmetic operations (addition, multiplication, inverses, and identities) customarily performed upon real numbers [BERL78]. We say that such a set of elements and operations is a Galois field of $2^m$ elements, GF($2^m$). (When m = 1, we have the familiar binary field.) In GF($2^m$), we can construct linear block codes as we did in the finite case: to every block of k information symbols from GF($2^m$) append (n - k) parity check symbols as linear combinations of the information. Note that this arithmetic is performed in GF($2^m$).

An interesting class of codes for these fields is that of the Reed-Solomon (RS) codes, which have the largest minimum distance possible for a given length and dimension (n,k) [BERL68]. A RS code is a linear block code with symbols from GF($2^m$), length $n = 2^m - 1$, and minimum distance $d = n - k + 1$ [REED60]. For a suitable choice of m, we will select a RS code for use on the erasure channel previously described.

## D. PERFORMANCE OF THE COMBINED CODING SCHEME

Before actually computing the performance of this coding scheme, its mappings will be carefully presented. A specific example will be considered.

The character set is a 49-element subset of full 7-bit ASCII. Each character is formed according to the standard 7-bit patterns. However, since 49 < 64, only 6 bits are actually needed in order to have a unique binary pattern for each symbol to be transmitted--a fact which we shall now use.

As 7-bit ASCII characters are produced, they are encoded by the (12,7) Hamming encoder as at present. Simultaneously, however, these 7 bit characters are mapped into 6-bit patterns. When 48 of these have been buffered, 15 parity checks will be computed on them according to the generator matrix of a (63,48) Reed-Solomon code over GF($2^6$). These parity check symbols are, of course, binary 6-tuples. These 6-tuples will be mapped back into binary 7-tuples according to the inverse of the 7 to 6 mapping. (A rule or a table can be used, so long as the transformation is reversible.) These 15 7-bit characters are now encoded by the (12,7) Hamming encoder and are concatenated with the 48 message characters previously encoded.

Decoding is accomplished in two stages. Received information is processed first by the decoder for the (12,7) Hamming code. Any double weight error pattern will force this decoder to output an "erasure" condition. Correctly decoded characters are presented as binary 7-tuples. These are converted to 6-tuples using the same map as at the encoder and are presented to the decoder for the (63,48) RS code.

Since character erasures are easily sensed by the RS decoder, it should not try to decode when more than 15 erasures have occurred. In the context of this note, the received message should not be acknowledged. The probability that a received message is not acknowledged is, therefore,

$$P_{dc} = \sum_{j=16}^{63} \binom{63}{j} P_{ce}^{j} (1-P_{ce})^{63-j} \qquad (3\text{-}2)$$

where $P_{ce}$ is the probability of a character erasure at the output of the Hamming decoder. Values of $P_{ce}$ for the BSC are obtained from Table 2.1 and the final results are shown in Table 3.1 and plotted in Figure 3.1.

| p | $P_{cr}$ | $P_{dr}$ |
|---|---|---|
| 0.10 | 0.2301 | 0.372 |
| 0.08 | 0.1835 | 0.103 |
| 0.07 | 0.1565 | 0.0307 |
| 0.06 | 0.1279 | $4.94 \times 10^{-3}$ |
| 0.05 | 0.09879 | $3.21 \times 10^{-4}$ |
| 0.04 | 0.07206 | $5.25 \times 10^{-6}$ |
| 0.03 | 0.04380 | $9.36 \times 10^{-9}$ |
| 0.02 | 0.02157 | $3.09 \times 10^{-13}$ |
| 0.01 | $5.968 \times 10^{-3}$ | $7.30 \times 10^{-22}$ |
| 0.008 | $3.968 \times 10^{-3}$ | $8.72 \times 10^{-25}$ |
| 0.006 | $2.237 \times 10^{-3}$ | $1.30 \times 10^{-28}$ |
| 0.005 | $1.569 \times 10^{-3}$ | $4.61 \times 10^{-31}$ |
| 0.004 | $1.015 \times 10^{-3}$ | $4.45 \times 10^{-34}$ |
| 0.003 | $5.764 \times 10^{-4}$ | $5.29 \times 10^{-38}$ |

Table 3.1 Probability of decoding failure for RS codes.

# IV. CONCLUSIONS

## A. DISCUSSION OF RESULTS

Improvement of the message rejection rate by an order of magnitude has been demonstrated. The original coding technique employed only a shortened Hamming code; when used on channels with coherent frequency shift keying (FSK) (a common method of impressing digital information onto FM radio signals) with values of signal to noise ratio of approximately 4 to 8 dB, it produced message rejection rates ranging from $6 \times 10^{-4}$ to 0.1. With the concatenation of RS codes, these rates dropped to a range of $5 \times 10^{-38}$ to $4 \times 10^{-4}$.

The penalty to be paid for this improvement is twofold. First, messages have been lengthened from 48 to 63 characters, an increase of 31% with no corresponding increase in the amount of information transmitted. Second, an additional stage of encoding and, more significantly, of decoding must be added. While many efficient decoding algorithms for RS codes are known, e.g. [BLAH79], the evaluation of the added complexity must be the subject of another report.

## B. FURTHER WORK

Undetectable error patterns are more insidious than those considered in this note. For example, 38 error patterns of weight four are codewords. Since the sum of two codewords is a codeword, the received vector will be one also. In such cases, no error condition can be detected. This behavior will be examined in a forthcoming report.

In addition to determining the impact of adding RS decoders to existing message processing systems, other factors must be studied before this effort is complete:

For this analysis, the binary symmetric channel with values of p from 0.003 to 0.1 was used. As asserted, the BSC is a valid model for certain quiet channels which are limited by the noise generated in the rf amplifier of the receiver. Conditions under which such a model is valid must be determined. Further, more realistic noise and interference models (e.g. noise bursts) must be considered, and coding techniques such as the one studied here must be evaluated against them.

Another consideration is that channel models such as the BSC assume that the world is discrete when, in fact, it is not. Such channel models are realized in practice by examining the received waveform (signal + noise) and making a statistical decision as to whether a binary 0 or a 1 was received. In the process, information about the reliability of that decision is discarded. To use that information in order to improve character and message reception reliability, "soft decision" detection and decoding techniques [FARR79] are under investigation. Significant improvements in message communication have been claimed for these methods, and they should be investigated.

Finally, the Hamming and RS codes were chosen for this part of the investigation because of their popularity among coding theorists and communication system designers. The technique studied above is related to "concatenated codes" [FORN66] which can be constructed from a variety of sets of constituent codes [COOP78]. Research is needed to select, for this application, codes which are optimum in terms of performance and decoding complexity.

## REFERENCES


[BERL68] Berlekamp, Elwyn R., ALGEBRAIC CODING THEORY, McGraw-Hill, New York, 1968.

[BLAH79] Blahut, R. E., "Transform Techniques for Error Control Codes," IBM. J. R&D, v23, No 3, May 1979.

[COOP78] Cooper, A. Brinton, III, "Algebraic Codes Constructed from other Algebraic Codes: A Short Survey and some Recent Results," in COMMUNICATION SYSTEMS AND RANDOM PROCESS THEORY, Sijthoff & Noordhoff, 1978 The Netherlands.

[FARR79] Farrell, P.G., "Soft Decision Techniques," in ALGEBRAIC CODING THEORY AND APPLICATIONS, G. Longo ed., Springer-Verlag, New York, 1979.

[FORN66] Forney, G. D. CONCATENATED CODES, MIT Press, Cambridge, 1966.

[GALL68] Gallager, R. G., INFORMATION THEORY AND RELIABLE COMMUNICATIONS, Wiley, New York, 1968.

[GORE73] Gore, W. C., "Transmitting Binary Symbols with Reed-Solomon Codes," 1973 Princeton Conference on Information Sciences and Systems.

[HAMM50] Hamming, R. W., "Error Detecting and Error Correcting Codes," BSTJ, vol XXVI, 1950.

[LIN&83] Lin, S. and D. J. Costello, Jr, ERROR CONTROL CODING: Fundamentals and Limitations, Prentice-Hall, Inc., Englewood Cliffs, 1983.

[PETE72] Peterson, W. W. and E. J. Weldon, Jr. ERROR CORRECTING CODES, MIT Press, Cambridge, 1972.

[REED60] Reed, I. S. and G. Solomon, "Polynomial Codes over certain Finite Fields," J. SIAM, v8, June 1960.

[TANE81] Tanenbaum, A. S., COMPUTER NETWORKS, Prentice-Hall, Englewood Cliffs, 1981.

# APPLICATION OF MACSYMA TO KINEMATICS AND MECHANICAL SYSTEMS

M.A. Hussain
General Electric Company
Corporate Research and Development

B. Noble
Mathematics Research Center
University of Wisconsin

## ABSTRACT

The objective of this paper is to illustrate that symbol manipulation systems can readily handle many of the typical symbolic calculations arising in the formulation of problems in kinematics and mechanical systems.

The paper consists of two parts. First, we discuss the use of MACSYMA in connection with the algebraic manipulations involved in transferring a body from one position to another in space, with particular reference to Rodrigues and Euler parameters and successive rotations, and an example involving quaternions. Second, we indicate how MACSYMA can be used to set up dynamical equations for the Stanford manipulator arm, and a spacecraft problem.

## INTRODUCTION

Kinematics is a basic tool for the analysis of mechanisms and mechanical systems. Until recently, the most common approach has been to use vectors and Euler angles. More recently, other approaches have been gaining in popularity because of computers. We illustrate by several examples that these approaches are particularly amenable to symbolic manipulation. The immediate objective is limited, namely to indicate that several methods of representing rotations including Rodrigues and Euler parameters, and quaternions can be handled by MACSYMA by a unified approach that would seem to have some elements of novelty. But also it should be clear that our examples suggest a different approach to dynamical problems such as those considered by Branets and Shmyglevskiy [3] using quaternions and Dimentberg [4] using the screw calculus. The nearest connected account of the type of approach we have in mind is the mss. [12] by Nikravesh et al., but a systematic use of computer symbolic manipulation would certainly affect the detailed treatment. This is the first part of the paper.

It is clear that the complexity of mechanical systems is increasing to the point where symbol manipulation must play an important part in their formulation and solution. We illustrate by two dynamical examples, one involving a robot arm, the other a spacecraft problem. The main reason for choosing these particular examples is that

This paper will also be published in the Proceedings of the Third MACSYMA Users' Conference 1984, supported by DOD.

the equations have been formulated and published in quite a detailed form already. By comparing our treatment with those already published, the reader will be able to make a judgment for himself concerning the usefulness of MACSYMA, and also how thinking in terms of symbol manipulation does change one's approach to the formulation of the equations. This is the second part of the paper.

We give the MACSYMA programs in detail in Appendices for two reasons. Experienced MACSYMA users may be able to suggest improvements. Readers familiar with other symbol manipulation systems may care to write programs for the same problems, and compare their programs with those in the Appendices. For the benefit of new and old users we encourage authors to publish their programs in detail, as done here.

## KINEMATICS EXAMPLES

### 1. The Representation of Rotation by Orthogonal Matrices

We remind the reader of some standard results. We work in terms of matrices (this can be converted into vector interpretations as appropriate) using upper case for matrices.

A rotation of a body with a fixed point by an angle $\phi$ around an axis defined by the unit column matrix $\mathbf{n} = [n_1,n_2,n_3]^T$ transfers a point $\mathbf{r} = [x,y,z]^T$ into a point $\mathbf{r}' = [x',y',z']^T$ by (cf. Bottema and Roth [1] p. 59) $\mathbf{r}' = \mathbf{A}\mathbf{r}$ where (see Figure 1)

$$\mathbf{A} = [\cos\phi\mathbf{I} + (1 - \cos\phi)\mathbf{n}\mathbf{n}^T + \sin\phi\mathbf{N}] \ , \tag{1}$$

$$\mathbf{N} = \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & -n_1 \\ -n_2 & n_1 & 0 \end{bmatrix} \tag{2}$$

(Note that $\mathbf{N}\mathbf{r}$ corresponds to the vector $\mathbf{n} \times \mathbf{r}$, and $\mathbf{I}$ is the identity matrix).

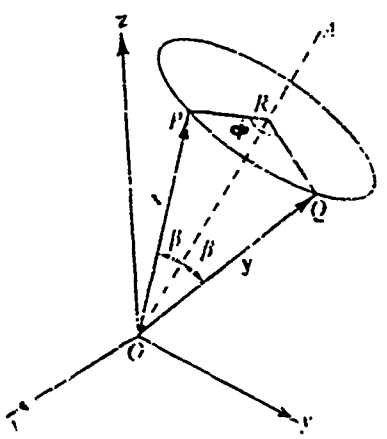



**Figure 1. Rotation of a body with a fixed point.**

The matrix $\mathbf{A}$ is orthogonal. We discuss three different ways of proving this using MACSYMA:

a) The simplest and most direct way is to express (1) in component form and simply check by brute force that $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, using TELLSIMP to impose side-conditions.

b) Alternatively we could use MACSYMA interactively as follows. It is easily checked that

$$\mathbf{n}^T\mathbf{n} = 1,\ \mathbf{N}^T = -\mathbf{N},\ \mathbf{n}^T\mathbf{N} = 0,\ \mathbf{Nn} = 0,\ \mathbf{N}^2 = \mathbf{nn}^T - \mathbf{I} \tag{3}$$

We use MACSYMA to form $\mathbf{A}^T\mathbf{A}$, which will give nine terms involving $\mathbf{n}^T\mathbf{N}$, $\mathbf{nn}^T\mathbf{nn}^T$, $\mathbf{N}^2$ etc., and we use SUBST to simplify and finally derive $\mathbf{A}^T\mathbf{A} = \mathbf{I}$.

c) We can use TELLSIMP to build the rules (3) into MACSYMA. Then a MACSYMA program can be written to produce the result I for $\mathbf{A}^T\mathbf{A}$.

Method a) is clearly simplest. Method c) is surprisingly tricky in MACSYMA because in addition to (3) we have to distinguish between scalars and matrices, and set proper switches. For verifying that $\mathbf{AA}^T = \mathbf{I}$, the simplest method is to use a) not c), but for more complicated problems, method a) soon produces algebraic expressions of horrendous complexity. As problem size increases, method c) will become preferable. In this paper, we have used the component form but further developments may require the more abstract approach.

## 2. Rodrigues Parameters

We introduce these by stating the result that any $3 \times 3$ orthogonal matrix A can be expressed in the following product form by the Cayley-Klein decomposition which says that there exists a skew-symmetric $3 \times 3$ matrix **B** such that (cf. Bottema and Roth [1], p. 10):

$$\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} + \mathbf{B}) \tag{4}$$

This tells us immediately that $\mathbf{B} = (\mathbf{A} - \mathbf{I})(\mathbf{A} + \mathbf{I})^{-1}$. MACSYMA gives us directly (Appendix I):

$$b_i = n_i \tan\frac{1}{2}\phi \qquad i = 1,2,3 \tag{5}$$

The $b_i$ $(i = 1,2,3)$ are the Rodrigues parameters.

We first express A in terms of the Rodrigues parameters. We find (Appendix II) cf. Bottema and Roth [1] p. 148:

$$\mathbf{A} = \frac{1}{\Delta}\begin{bmatrix} 1 + b_1^2 - b_2^2 - b_3^2 & 2(b_1b_2 - b_3) & 2(b_1b_3 + b_2) \\ 2(b_2b_1 + b_3) & 1 - b_1^2 + b_2^2 - b_3^2 & 2(b_2b_3 - b_1) \\ 2(b_3b_1 - b_2) & 2(b_3b_2 + b_1) & 1 - b_1^2 - b_2^2 + b_3^2 \end{bmatrix} \tag{6}$$

where $\Delta = 1 + b_1^2 + b_2^2 + b_3^2$. Using the notation $\mathbf{A} = [a_{ij}]$, it is clear from this result that:

$$\begin{aligned} b_1 &= (a_{32} - a_{23})/d \\ b_2 &= (a_{13} - a_{31})/d \\ b_3 &= (a_{21} - a_{12})/d \end{aligned} \tag{7}$$

with $d = 1 + a_{11} + a_{22} + a_{33}$. Having established the necessary background, we

derive typical basic results by means of MACSYMA. The reader should compare our derivation with those of, for example, Bottema and Roth [1], Gibbs [5], and Dimentberg [4].

Consider the result of first rotating a body round an axis $\mathbf{n}$ by angle $\phi$, then around a second axis $\mathbf{n}'$ by an angle $\phi'$. Euler's theorem tells us that the result is equivalent to a rotation by some angle $\phi''$ round some axis $\mathbf{n}''$. In matrices, if the matrices corresponding to these three rotations are $\mathbf{A}$, $\mathbf{A}'$, $\mathbf{A}''$ and we start with a point $\mathbf{r}$, this is first transformed into $\mathbf{r}' = \mathbf{Ar}$, and then $\mathbf{r}'$ is transformed into $\mathbf{r}'' = \mathbf{A}'\mathbf{r}'$. We also have $\mathbf{r}'' = \mathbf{A}''\mathbf{r}$ so that

$$\mathbf{A}'' = \mathbf{A}'\mathbf{A}$$

The Rodrigues parameters corresponding to $\mathbf{n}''$, $\phi''$ are given by (7) where $a_{ij}$ are the elements of $\mathbf{A}''$. But these are given in terms of the first two rotations by the the corresponding elemnts of $\mathbf{A}'\mathbf{A}$ These matrix relations are carried out by MACSYMA in Appendix III, giving the result:

$$\mathbf{b}'' = \frac{\mathbf{b} + \mathbf{b}' - \mathbf{B}\mathbf{b}'}{1 - \mathbf{b}^T\mathbf{b}'} \tag{8}$$

where $\mathbf{B}$ is related to $\mathbf{b}$ as $\mathbf{N}$ was to $\mathbf{n}$ in (2).

Note that this is a straightforward derivation that would be laborious to carry out by hand, as compared with derivations carried out in the literature designed for the ease of hand computation.

## 3. Euler Parameters

Instead of using Rodrigues parameter $b_i$, it is often convenient to use Euler parameters $c_i$ related to $b_i$ by (Bottema and Roth [1] p. 150)

$$b_i = c_i/c_0 \ , \ c_0^2 + c_1^2 + c_2^2 + c_3^2 = 1 \tag{9}$$

The relation (5) then gives

$$\mathbf{A} = \begin{bmatrix} c_0^2 + c_1^2 - c_2^2 - c_3^2 & 2(-c_0c_3 + c_1c_2) & 2(c_0c_2 + c_1c_3) \\ 2(c_0c_3 + c_2c_1) & c_0^2 - c_1^2 + c_2^2 - c_3^2 & 2(-c_0c_1 + c_2c_3) \\ 2(-c_0c_2 + c_3c_1) & 2(c_0c_1 + c_3c_2) & c_0^2 - c_1^2 - c_2^2 + c_3^2 \end{bmatrix} \tag{10}$$

Although it would seem that the Euler parameters are straightforward homogeneous forms of the Rodrigues parameters, it turns out that some relations are expressed much more simply in terms of the Euler parameters.

One example is the Euler parameter analog of (8) for two successive rotations. To derive this, substitute $\mathbf{b} = \mathbf{c}/c_0$ , $\mathbf{b}' = \mathbf{c}'/c_0$ in (8) which gives:

$$\mathbf{b}'' = \frac{c_0'\mathbf{c} + c_0\mathbf{c}' - \mathbf{C}\mathbf{c}}{c_0c_0' - \mathbf{c}^T\mathbf{c}'} \tag{11}$$

When this is written out in detail we find that by introducing

$$c_0'' = c_0'c_0 - c_1'c_1 - c_2'c_2 - c_3'c_3 \tag{12}$$

$$
\begin{aligned}
c_0'' &= c_0'c_0 - c_1'c_1 - c_2'c_2 - c_3'c_3 \\
c_1'' &= c_1'c_0 + c_0'c_1 - c_3'c_2 + c_2'c_3 \\
c_2'' &= c_2'c_0 + c_3'c_1 + c_0'c_2 - c_1'c_3 \\
c_3'' &= c_3'c_0 - c_2'c_1 + c_1'c_2 + c_0'c_3
\end{aligned}
\tag{12}
$$

equation (11) can be written in the simple form

$$
\mathbf{b}'' = \mathbf{c}''/c_0'' \tag{13}
$$

In Appendix IV we ch: ; by MACSYMA that if $c_0^2 + \mathbf{c}^T\mathbf{c} = 1$, $(c_0')^2 + (\mathbf{c}')^T\mathbf{c}' = 1$, then $(c_0'')^2 + (\mathbf{c}'')^T\mathbf{c}'' =$ which is a well-known result due to Euler. This result and (12) mean that $c_0''$, $c_1''$, $c_2''$, $c_3''$ are the Euler parameters corresponding to the total rotation.

In the literature, the result (12) is often derived via quaternions (e.g. Bottema and Roth [1], p. 518,520). It is of some interest to express this approach in the present context of Euler parameters and matrices which can be done without mentioning quaternions explicitly. Introduce $\gamma$ and $\Gamma$ defined as follows:

$$
\gamma = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}, \quad \Gamma = \begin{bmatrix} c_0 & -c_1 & -c_2 & -c_3 \\ c_1 & c_0 & -c_3 & c_2 \\ c_2 & c_3 & c_0 & -c_1 \\ c_3 & -c_2 & c_1 & c_0 \end{bmatrix}
$$

If $\gamma',\Gamma'$ are the corresponding matrices with $\mathbf{c}'$ in place of $\mathbf{c}$, and similarly for $\gamma'',\Gamma''$, we define the product $\gamma'\gamma$ by (compare the remark following (2)):

$$
\gamma'' = \gamma'\gamma = \Gamma'\gamma \tag{14a}
$$

which says exactly the same as (12). We first note that if we define $\gamma^{-1} = [c_0,-c_1,-c_2,-c_3]^T$ then $\gamma\gamma^{-1} = \gamma^{-1}\gamma = [1,0,0,0]^T$. It can be verified (e.g. by the MACSYMA program in Appendix V) that introducing $\boldsymbol{\rho} = [r_0,r_1,r_2,r_3]^T$, $\mathbf{r} = [r_1,r_2,r_3]^T$ and $\boldsymbol{\rho}',\mathbf{r}'$ correspondingly, then if we form $\boldsymbol{\gamma\rho\gamma}^{-1}$, and denote the result by $\boldsymbol{\rho}'$, then

$$
\begin{bmatrix} r_0' \\ \mathbf{r}' \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{A} \end{bmatrix} \begin{bmatrix} r_0 \\ \mathbf{r} \end{bmatrix}
$$

Where $\mathbf{A}$ is precisely the matrix that appeared in (10), i.e., $\boldsymbol{\gamma\rho\gamma}^{-1}$ represents a rotation. This is our version of the standard quaternion theorem on rotation, derived of course from a completely different point of view (cf. Brand [2], p. 417). A second rotation would give $\boldsymbol{\rho}'' = \boldsymbol{\gamma}'\boldsymbol{\rho}'(\boldsymbol{\gamma}')^{-1}$, and combining the rotations leads to $\boldsymbol{\rho}'' = \boldsymbol{\gamma}'\boldsymbol{\gamma\rho}(\boldsymbol{\gamma}')^{-1}\boldsymbol{\gamma}^{-1}$, i.e., if $\boldsymbol{\gamma}''$ represents the combined rotation then $\boldsymbol{\gamma}'' = \boldsymbol{\gamma}'\boldsymbol{\gamma}$, which is identical with (12).

Still another way of obtaining (12) is suggested by the discussion of Cayley-Klein parameters in Bottema and Roth ([1] p. 529), namely that a result corresponding to equation (9.8) in that reference should hold for Euler parameters. We introduce the notations:

$$
\mathbf{V} = c_0\mathbf{I} + \mathbf{S}
$$

$$
\mathbf{V}^{-1} = c_0\mathbf{I} - \mathbf{S}
$$

$$S = \begin{bmatrix} 0 & -c_1 & -c_2 & -c_3 \\ c_1 & 0 & -c_3 & c_2 \\ c_2 & c_2 & 0 & -c_1 \\ c_3 & -c_2 & c_1 & 0 \end{bmatrix}, \quad q = \begin{bmatrix} y & z & 0 & -x \\ z & -y & x & 0 \\ 0 & x & y & z \\ -x & 0 & z & -y \end{bmatrix}$$

$$Q = \begin{bmatrix} Y & Z & 0 & -X \\ Z & -Y & X & 0 \\ 0 & X & Y & Z \\ -X & 0 & Z & -Y \end{bmatrix}$$

The MACSYMA program in Appendix VI does the following. We form $VqV^{-1}$ and equate this to Q. This gives 16 equations. However, it is easily checked by MACSYMA that, in fact, there are only *three* independent relations involving x,y,z and X,Y,Z which can be written in the form

$$Aq = Q$$

where A is exactly the A given in (10). The implication of this, in connection with repeated rotations, is that if q corresponds to r and Q to r′ defined in the second paragraph of Section 1 and the corresponding A is denoted by V, then

$$VrV^{-1} = r'$$

Similarly, the second rotation gives $V'r'V'^{-1} = r''$ and the rotation from the initial position to the final position gives $V''rV''^{-1} = r''$. Eliminating r′ we have $V''r(V'')^{-1} = V'VrV^{-1}V'^{-1}$ so that finally

$$V'' = V'V \tag{14b}$$

and this is precisely equation (12) cf. (14a).

## 4. An Example Involving Dual Quaternions

The discussion in the last two sections was concerned with the rotation of a body with a fixed point and involved only three independent parameters. The general motion of a body involves displacement, as well as rotation, and requires six independent parameters. Rather than extending the methods of the last two sections, we illustrate how MACSYMA deals with a rather different approach to kinematics, namely via quaternions, by considering a calculation in a classic paper by Yang and Freudenstein ([14], 1964) dealing with a spatial four-bar mechanism.

In Figure 2, MA and NB are two nonparallel and nonintersecting lines. MN is the common perpendicular. Let a,b denote unit vectors in the direction of MA, NB respectively, and let $r_a, r_b$ denote the vectors $\vec{OM}$, $\vec{ON}$. We introduce the quaternions

$$\hat{a} = a + \epsilon(r_a \times a) \quad , \quad \hat{b} = b + \epsilon(r_b \times b)$$

where $\epsilon$ is a symbol with the property that $\epsilon^2 = 0$. Note that this implies, for example, that if $\hat{\theta} = \theta + \epsilon s$ then

$$\sin\hat{\theta} = \sin\theta + \epsilon s\cos\theta \quad , \quad \cos\hat{\theta} = \cos\theta - \epsilon s\sin\theta \tag{15}$$

As discussed by Yang et al. [14], the relative shift between $\hat{a}$ and $\hat{b}$ can be expressed as

$$\hat{b} = Q\hat{a} \quad , \quad \hat{a} = \hat{b}Q$$

where Q is a dual quaternion (see [14], (22, 23)). Successive application of formulae of this type gives rise to a loop closure equation for the mechanism of the form:

$$A(\hat{\theta}_1)\sin\hat{\theta}_4 + B(\hat{\theta}_1)\cos\hat{\theta}_4 = C(\hat{\theta}_1) \tag{16}$$

where

$$A(\hat{\theta}_i) = \sin\hat{\alpha}_{12}\sin\hat{\alpha}_{34}\sin\hat{\theta}_1$$

$$B(\hat{\theta}_1) = -\sin\hat{\alpha}_{34}\,(\sin\hat{\alpha}_{41}\cos\hat{\alpha}_{12} + \cos\hat{\alpha}_{41}\sin\hat{\alpha}_{12}\cos\hat{\theta}_1)$$

$$C(\hat{\theta}_1) = \cos\hat{\alpha}_{23} - \cos\alpha_{34}\,(\cos\hat{\alpha}_{41}\cos\hat{\alpha}_{12} - \sin\hat{\alpha}_{41}\sin\hat{\alpha}_{12}\cos\hat{\theta}_1)$$

Here

$$\hat{\alpha}_{12} = \alpha_{12} + \epsilon a_{12} \quad , \quad \hat{\theta}_1 = \theta_1 + \epsilon s_{11}$$

$$\hat{\alpha}_{23} = \alpha_{23} + \epsilon a_{23} \quad , \quad \hat{\theta}_2 = \theta_2 + \epsilon s_2$$

$$\hat{\alpha}_{34} = \alpha_{34} + \epsilon a_{34} \quad , \quad \hat{\theta}_3 = \theta_3 + \epsilon s_3$$

$$\hat{\alpha}_{41} = \alpha_{41} + \epsilon a_{41} \quad , \quad \hat{\theta}_4 = \theta_4 + \epsilon s_4$$

It is then clear that (15) can be reduced to the form

$$P + \epsilon Q = R + \epsilon S$$

where P, Q, R, and S are independent of $\epsilon$. It is required to find the explicit form of P, Q, R, and S. To calculate this by hand is extremely laborious, but straightforward in MACSYMA. The program is given in Appendix VII.

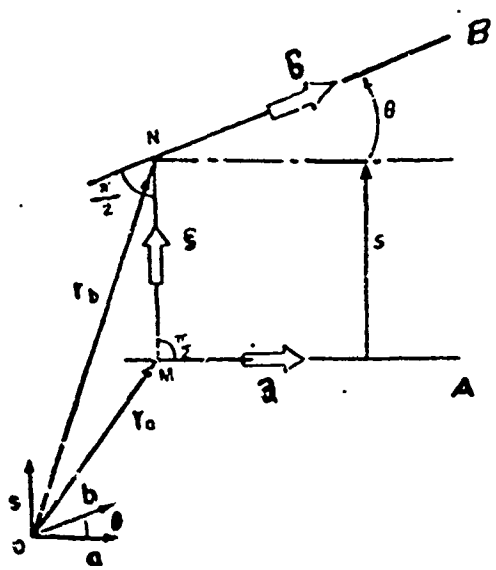



**Figure 2. Relative position of two line vectors.**

## TWO EXAMPLES IN DYNAMICS

### 5. Equations of Motion for the Stanford Manipulator Arm

There are a number of ways to set up dynamical equations for robot manipulator arms (see Paul [13]). Kane-Levinson [9] have given an example of setting up dynamical equations for the Stanford manipulator. Our objective is to reproduce these equations from an algorithmic point of view, without having to do by hand the kind of extensive manipulation given in that paper. The method can help us to set up similar sets of equations for any manipulator automatically, thereby reducing the labor. We also show that MACSYMA can simplify the Kane-Levinson end-result, reducing the numbers of arithmetic operations required to obtain numerical results.

We consider the Stanford manipulator arm (Paul [13]), a six-element, six-degree-of-freedom manipulator. A schematic representation of this arm is given in Figure 3, from Kane-Levinson [9], where more details can be found. The six bodies are designated A, ..., F. Body A can be rotated about a vertical axis fixed in space. A supports B which can be rotated about a horizontal axis fixed relative to A. The figure should now be self-explanatory, the joint connecting B and C being translational, and the remaining joints rotational.

$q_1$, ..., $q_6$ are generalized coordinates characterizing the instantaneous configuration of the arms, the first five being rotational and $q_6$ translational. For the plane configuration of the arms as drawn in Figure 3, it is assumed that $q_1, \cdots q_5$ are zero.

We choose coordinate axes as follows. $n_1$, $n_2$, $n_3$ are unit vectors fixed in space as indicated in Figure 3, $n_1$, $n_2$ lying in the plane of the paper. $a_1$, $a_2$, $a_3$ are unit vectors fixed in the arm A which coincide with $n_1$, $n_2$, $n_3$ when the arm is in the configuration of Figure 3. Similarly, $b_1$, $b_2$, $b_3$ are unit vectors attached to the arm B and similarly for C, D, E, and F.

We give a mathematical description of an algorithm for setting up the dynamical equations. This is essentially the algorithm described by Kane-Levinson [9], but organized in a somewhat different way in order to facilitate implementation on MACSYMA. The stages and details of the MACSYMA program which are in Appendix VIII, parallel the mathematical description that follows:

*Stage 1*: Set up angular velocities:

Rotations about x,y,z axes can be described by orthogonal matrices of simple form as discussed in detail by Paul [13], Chapter 1. For instance, rotation by an angle $\theta$ about the x-axis involves ([13], p. 15)

$$\mathrm{Rot}_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$$

Let $R_1$, ... $R_5$ denote matrices corresponding to rotations $\theta_1$, ..., $\theta_5$ about axes, y,x,y,x,y respectively in the local coordinates fixed relative to arms A, B, D, E, F. Let $\dot{q}_1$, ..., $\dot{q}_5$ denote angular velocities around y,x,y,x,y axes respectively. These are

vector quantities represented by matrices that we denote by $\omega_1$, ..., $\omega_5$. For instance, $\omega_1 = [0,\dot{q}_1,0]$ etc. Similarly, for the linear velocity $\dot{q}_6$.

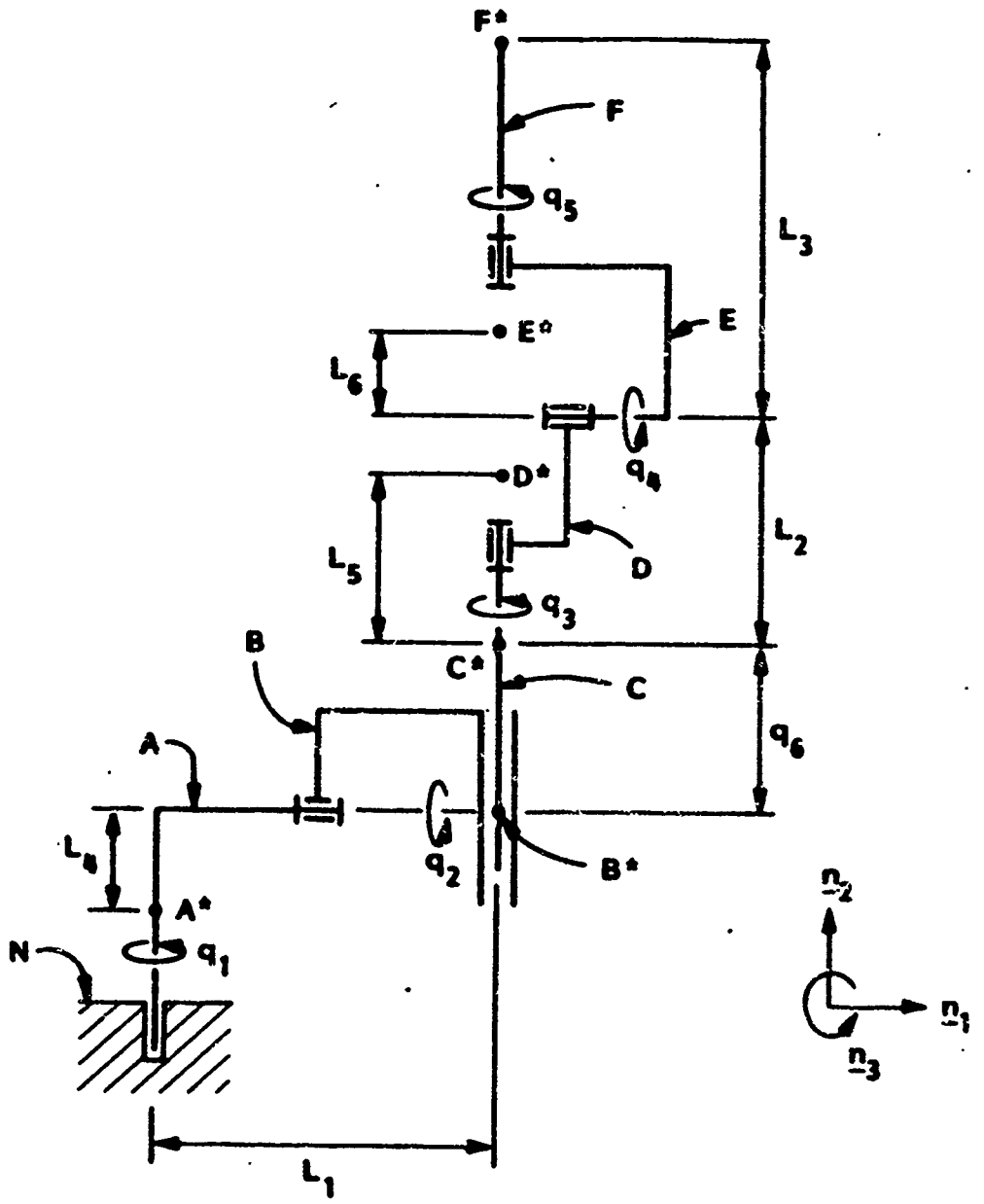



**Figure 3. A schematic representation of Stanford manipulator arm.**

Next introduce $\omega^A$, ..., $\omega^F$, the angular velocities of A, ..., F in our Newtonian frame of reference, but with components expressed in the local coordinate frame of reference. For example:

$$\omega^D = [u_1,u_2,u_3] \text{ means: } \omega^D = u_1\underline{d}_1 + u_2\underline{d}_2 + u_3\underline{d}_3 \tag{17}$$

The algorithm for computing $\omega^A$, ..., $\omega^F$ is given by:

$$\omega^A = \omega_1 R_1$$

$$\omega^B = (\omega^A + \omega_2)R_2$$

$$\omega^C = \omega^B$$

$$\omega^D = (\omega^C + \omega_3)R_3$$

$$\omega^E = (\omega^D + \omega_4)R_4$$

$$\omega^F = (\omega^E + \omega_5)R_5$$

If these formulae are used as they stand, the expression for $\omega^F$ in terms of $\dot{q}_i$ will be complicated. The complexity can be reduced using a method due to Kane-Levinson [9]. The $u_i$ that occur in (16) can be expressed in terms of $\dot{q}_1$, $\dot{q}_2$, $\dot{q}_3$ as follows

$$u_1 = \dot{q}_1 \sin q_2 \sin q_3 + \dot{q}_2 \cos q_3$$

$$u_2 = \dot{q}_1 \cos q_2 + \dot{q}_3$$

$$u_3 = -\dot{q}_1 \sin q_2 \cos q_3 + q_2 \sin q_1$$

$$u_i = \dot{q}_i \quad i = 4,5,6$$

*Stage 2*: Set up linear velocities:

In stage 1, the angular velocities were always expressed in local coordinates corresponding to the arm being considered. This is not necessarily the case for the way in which Kane-Levinson [9] formulate the linear velocities (see paragraph preceeding (28) in the paper). Because we wish our results to be comparable to those in [9], we state the formulae we use, which will lead to results that are the same as those in equations (28-43) in [9]. (Note that the stars in the following refer to the velocities of the centers of mass of the corresponding arms.)

$$v^{A*} = 0$$

$$v^{B*} = \omega^A \times R^B$$

$$v^{C*} = \omega^B \times R^C + \dot{\bar{q}}_6$$

$$v^{D*} = \omega^B \times R^D + \dot{\bar{q}}_6$$

The expressions for $v^{E*}$, $v^{F*}$ correspond to those in equation (40) and (42) in the Kane-Levinson paper [9]. The exact form we use can be found from the expressions for VE and VF in the MACSYMA program given in Appendix VIII.

The remaining stages are relatively straightforward.

*Stage 3*: Find the partial angular velocities.

*Stage 4*: Find the partial linear velocities.

These are explained in the Kane-Levinson paper [9] and the MACSYMA implementation in Appendix VIII is self-explanatory.

*Stage 5*: Find the angular accelerations.

*Stage 6*: Find the linear accelerations.

These are obtained by simple differentiation of the corresponding angular and linear velocities as given in the MACSYMA program in Appendix VIII.

*Stage 7*: Define moments of inertia.

We next have to consider forces.

*Stage 8*: Define torques.

*Stage 9*: Set up generalized forces.

*Stage 10*: Set up active forces.

*Stage 11*: Set up Kane's equations.

These steps are straightforward; the MACSYMA program is given in Appendix VIII.

Finally, Figure 4 gives a comparison of some numerical results obtained from MACSYMA and Kane-Levinson [9].

It is of some interest to compare the mathematical equations in the Kane-Levinson paper with the corresponding MACSYMA expressions. For example, consider:

| Kane-Levinson [9] (underlined quantities are vectors) | | MACSYMA |
|---|---|---|
| $\underline{\omega}^A = \dot{q}_1 \underline{a}_2$ | (13) | WA: EXPAND(W1.R1) |
| But | | $\underline{\omega}^A \equiv$ WA , $\dot{q}_1 \underline{a}_2 \equiv$ W1.R1 |
| $\dot{q}_1 = \dfrac{u_1 s_3 - u_3 c_3}{s_2}$ | (8) | Here $\omega^A$, $\dot{q}_1$, $\underline{a}_2$ are vectors; WA, W1.R1 are matrices |
| Introduce | | |
| $Z_4 = \dfrac{s_3}{s_2}$, $Z_5 = -\dfrac{c_3}{c_2}$ | | |
| Then (13) becomes: | | |
| $\underline{\omega}^A = (Z_4 u_1 + Z_5 u_3)\underline{a}_2$ | (15) | |
| Similarly, | | |
| $\underline{\omega}^B = Z_2 \underline{b}_1 + Z_{10} \underline{b}_2 + Z_{11} \underline{b}_3$ | (16) | WB: EXPAND ((WA+W2).R2) |
| where | | |
| $Z_{10} = Z_6 u_1 + Z_7 u_3$, $Z_{11} = Z_8 u_1 + Z_9 u_3$ | | |

One point here is that because Kane-Levinson [9] are carrying out the algebra by hand, it is convenient for them to introduce intermediate symbols $Z_1, Z_2 \cdots$ going up to $Z_{196}$, and similarly, 36 X's and 31 W's. MACSYMA has no difficulty in generating the end result in explicit form. These end results are no more complex than

the complexity of the equations given in [9]. At the time of writing this paper a preliminary number count on additions and multiplication for $X_{ij}$, the coefficients of equations of motion, obtained by MACSYMA, as compared to those in [9], shows a reduction by approximately a factor of two.

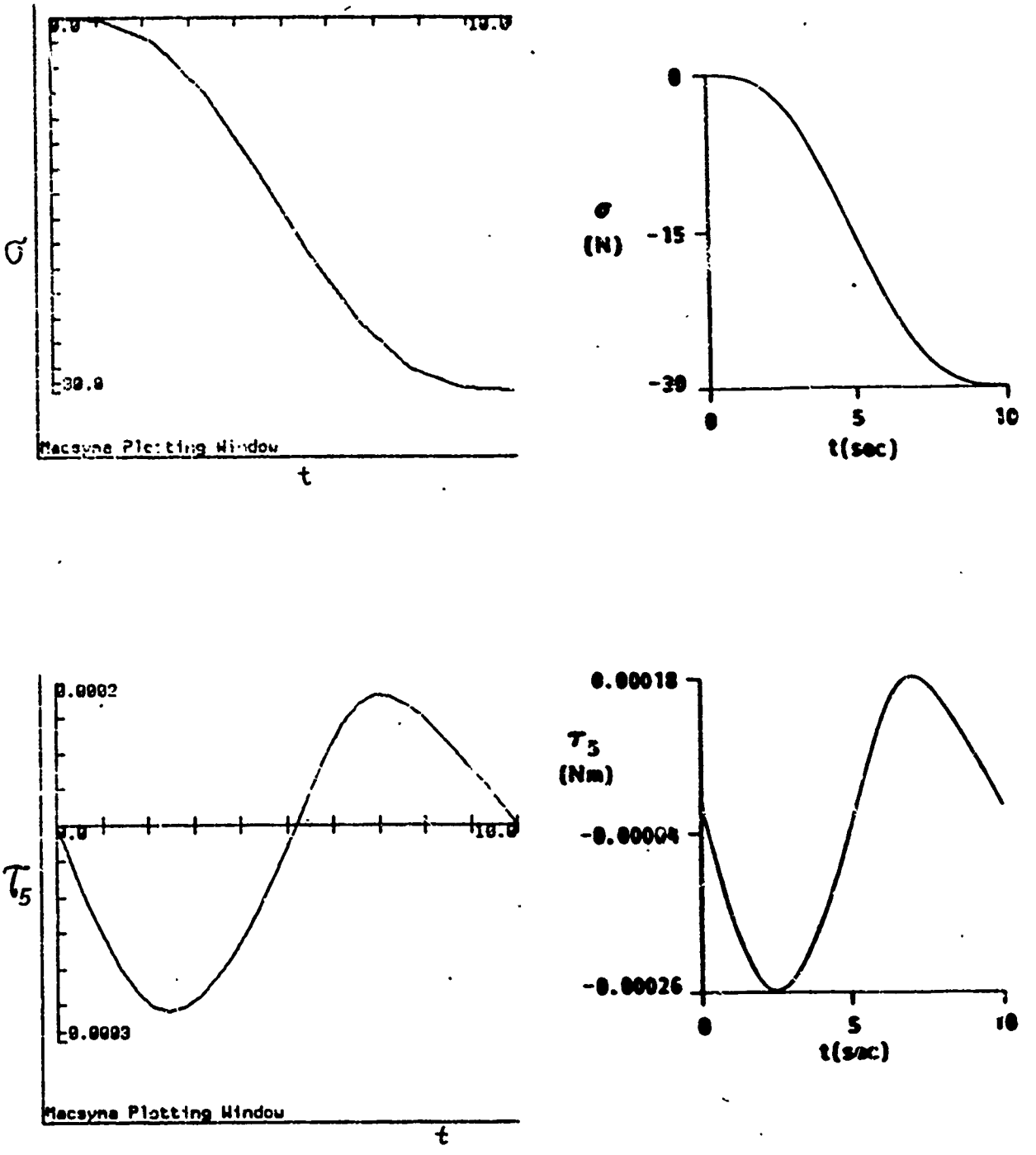



**Figure 4a. Comparison of numerical results for $\sigma$, $\tau_5$ obtained by MACSYMA and Reference 9.**

In conclusion, we note that Paul [13] sets up the dynamical equation of the Stanford manipulator arm using the Lagrangian equation approach. See also [6]. Some applications of the Lagrange method using MACSYMA are discussed in [9].

Various methods of setting up dynamical equations that could be carried out by MACSYMA are illustrated in [8].

## 6. A Spacecraft Problem

Levinson [11] has described in detail an application of the symbolic language FORMAC to formulate the spacecraft problem shown in Figure 5, consisting of two rigid bodies with a common axis of rotation *b*. (See also [10], pp.279-285).

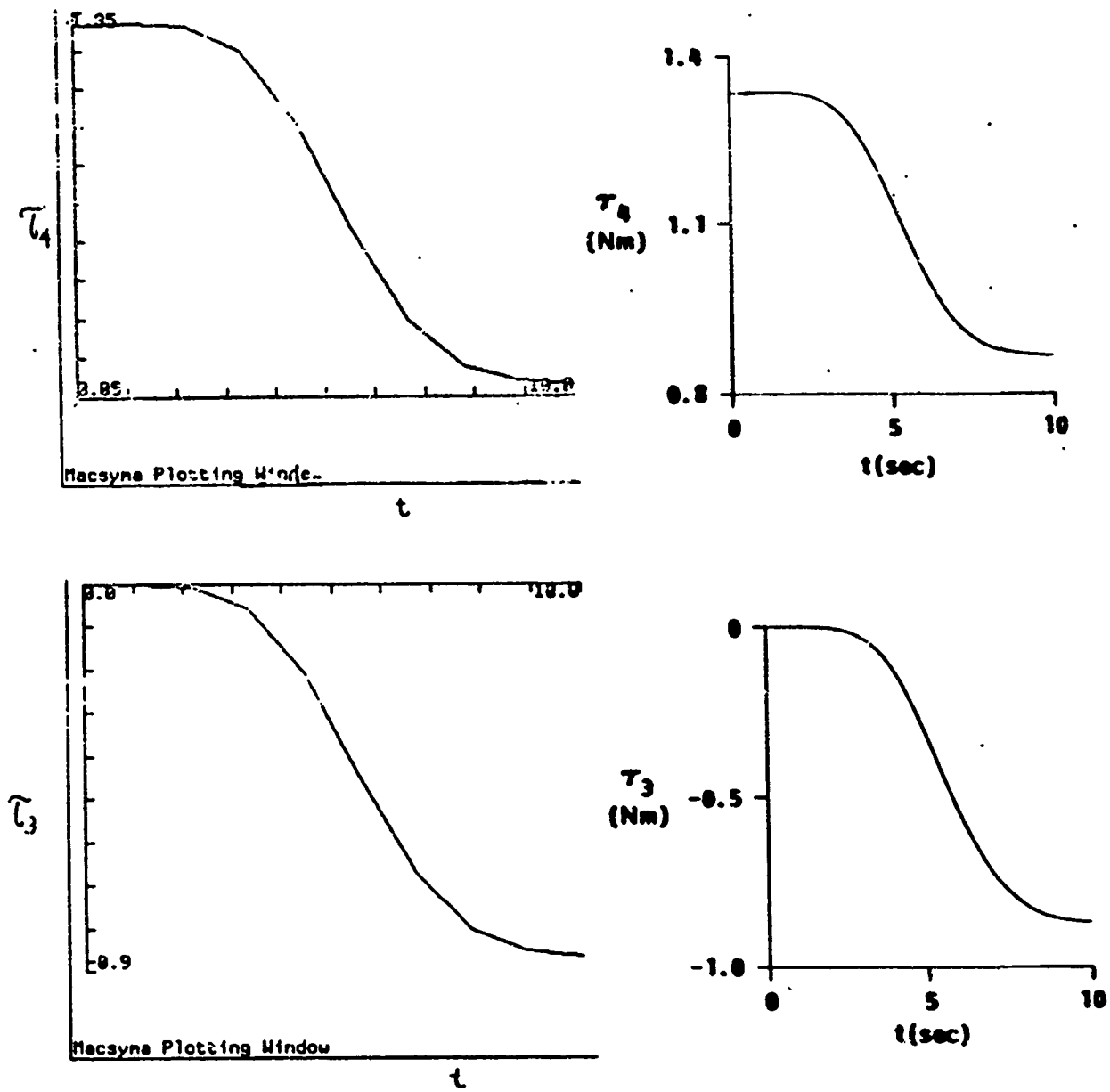



**Figure 4b. Comparison of numerical results for $\tau_4$ and $\tau_3$ obtained by MACSYMA and Reference 9.**

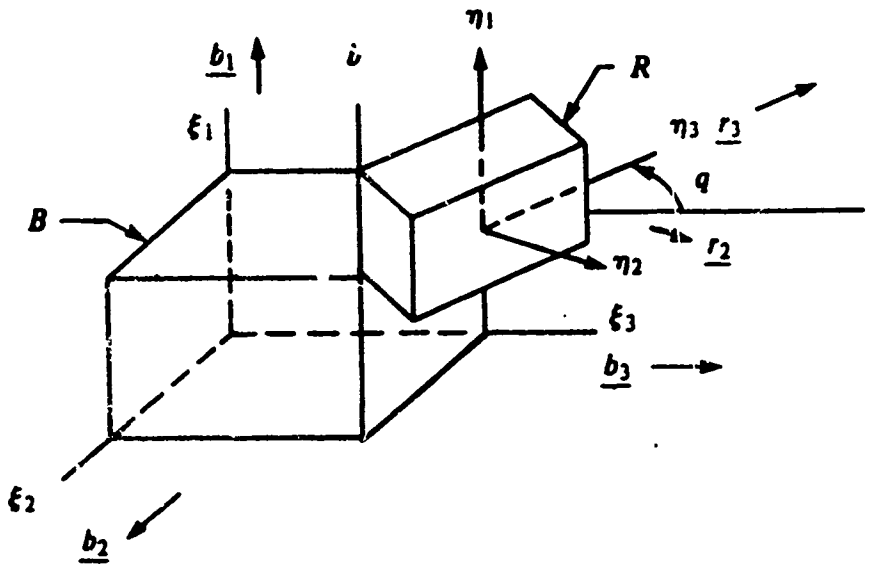



**Figure 5. Two rigid bodies with a common axis of rotation.**

The equations are given in complete detail in Ref. [11], and translated into MACSYMA in Appendix IX. In the example in the last section, we wrote the MACSYMA program in terms of matrices. In Appendix IX, the present example is written in terms of vectors, by writing BLOCK functions to perform the dot and cross products. To illustrate the comparison of the vector equation with the corresponding MACSYMA expressions:

| Equations from Ref. [11] | | | | MACSYMA |
|---|---|---|---|---|
| $\underline{r}_2$ | = | $\cos q\, \underline{b}_2 + \sin q\, \underline{b}_3$ | (1) | R[2]:COS(Q)*B[2]+SIN(Q)*B[3]; |
| $\omega$ | = | $u_1\underline{b}_1 + u_2\,\underline{b}_2 + u_3\,\underline{b}_3$ | (3) | WB:U[1]*B[1]+U[2]*B[2]+U[3]*B[3]; |
| $\mu_4$ | = | $\dot{q}$ | | U[4]:DIFF(Q,T); |
| $\alpha^R$ | = | $\frac{d}{dt}(\omega^R) + \omega^B \times \omega^R$ | (7) | ALPR:DIFF(WR,T)+CROSS(WB,WR); |

We discuss only one other correspondence. Equation (27) in Ref. [11] is

$$F_r = \frac{\partial v^{B^*}}{\partial \mu_r} \cdot (F)_B + \frac{\partial \omega^B}{\partial \mu_r} \cdot (T)_B \quad (r=1, \cdots, 7)$$

which becomes in MACSYMA

```
F[R]:=DOT(DIFF(VBS,U[R]),FB) + DOT(DIFF(WB,U[R]),TB);
```

The complete set of equations given in Ref. [11] is generated by Appendix IX. The reader should compare the corresponding FORMAC program given in Levinson [11].

## CONCLUDING REMARKS

It should be clear from the examples given that symbolic manipulation by computer can carry out many of the laborious and routine calculations involved in the analysis of mechanical systems. But, potentially even more important, is the influence that

symbolic manipulation is likely to have on the methods used to formulate problems. The reader should compare, for example, the algorithmic approach we have adopted to the Stanford manipulator arm problem with the approach in [9]. As another example, if symbolic manipulation methods are used, this will influence whether we formulate problems in terms of Euler angles, Euler parameters, Rodrigues parameters, or quaternions, etc. In addition, one can visualize the production of standard MACSYMA software to produce equations corresponding to those of Kane-Levinson [9] for *any* given combination of rotating and sliding joints.

## REFERENCES


1. O. Bottema and B. Roth, *Theoretical Kinematics*, North-Holland, 1979.
2. L. Brand, *Vector and Tensor Analysis*, Wiley, 1957.
3. V. N. Branets and I. P. Shmyglevskiy, "Application of Quarternions to Rigid Body Rotation Problems," NASA Tech. Transl. TTF-15, 414, 1974 (1973 Russian original).
4. F. M. Dimentberg, "The Screw Calculus and its Applications in Mechanics," NTIS Transl. FTD-HT-23-1632-67, 1968 (1965 Russian original).
5. J. W. Gibbs, *Vector Analysis*, Dover reprint, 1960 (original published in 1909).
6. J. M. Hollerbach, "A Recursive Lagrangian Formulation of Manipulator Dynamics and a Comparative Study of Dynamics Formulation Complexity," *IEEE Trans. on Syst., Man and Cyb. SMC-10*, 1980, 730-736.
7. M. A. Hussain and B. Noble, "Application of Symbolic Computation to the Analysis of Mechanical Systems, Including Robot Arms," General Electric Technical Report 84CRD062, 1984. Also to be published in the Proceedings of the NATO Conference on Mechanisms, E. Haug, ed., University of Iowa, 1984.
8. T. R. Kane and D. A. Levinson, "Formulation of Equations of Motion for Complex Spacecraft," *J. Guidance and Control*, 1980, 99-112.
9. T. R. Kane and D. A. Levinson, "The Use of Kane's Dynamical Equations in Robotics", *Int. J. Robotics Research 2*, 1983, 3-21.
10. T. R. Kane, P. W. Likins, and D. A. Levinson, *Spacecraft Dynamics*, McGraw-Hill, 1983.
11. D. A. Levinson, "Equations of Motion for Multiple-Rigid-Body Systems via Symbol Manipulation," *J. Spacecraft and Rockets 14*, 1977, 479-487.
12. P. E. Nikravesh, R. A. Wehage, and E. J. Haug, *Computer-Aided Analysis of Mechanical Systems*, to be published.
13. R. P. Paul, *Robot Manipulators - Mathematics, Programming, and Control*, M. I. T. Press, 1981.
14. A. T. Yang and F. Freudenstein, "Application of Dual-Number Quaternion Algebra in the Analysis of Spatial Mechanisms," *Trans ASME J. Appl. Mech.*, 1966, 300-308.

## APPENDIX I

```
/*PROVE THE IDENTITY OF EQUATION 5 */
/* TRIGNOMETRIC SIMPLIFICATION */
MATCHDECLARE(A,TRUE),
TFELLSIMP(SIN(A)^2,1-COS(A)^2);
/* DEFINE CROSS PRODUCT MATRIX OR ALTERNATING TENSOR */
ALT(N):=MATRIX([0,-N[3,1],N[2,1]],[N[3,1],0,-N[1,1]],
[-N[2,1],N[1,1],0]);
N:MATRIX([N1],[N2],[N3]);
NN:ALT(N);
I:IDENT(3);
AA:COS(ALPHA)*I+(1-COS(ALPHA))*(N . TRANSPOSE(N))+NN*SIN(ALPHA);
AAP:I+AA;
/* WORK WITH HALF ANGLES */
ALPHA BTA*2;
EV(AA);
TRIGEXPAND(%);
AA:%$
/* ADD IDENTITY MATRIX AND INVERT */
AAP:AA+I$
IAAP:AAP^^(-1)$
/*SUBTRACT IDENTITY MATRIX AND FORM MATRIX PRODUCT AS ANSWER*/
AAM:AA-I$
ANSWER:AAM . IAAP$
/*USE IDENTITY THAT N1^2+N2^2+N3^2=1 */
NN3:1-N1**2-N2**2;
ANSWER.RATSUBST(NN3,N3^2,ANSWER);
ANSWER.RATSIMP(%);
```

## APPENDIX II

```
/* CAYLEY'S DECOMPOSITION OF ORTHOGONAL MATRIX
A =(I-B)^-1(I+B),WHERE B1,B2,B3 ARE RODRIGUES PARAMETERS*/
/* DEFINE CROSS PRODUCT OR ALTERNATING TENSOR MATRIX*/
ALT(N):=MATRIX([0,-N[3,1],N[2,1]],[N[3,1],0,-N[1,1]],
[-N[2,1],N[1,1],0]);
B MATRIX([B1],[B2],[B3]);
BB:ALT(B);
I:IDENT(3);
INBB.(I-BB)^^-1;
A.INBB.(I+BB);
ANSWER.RATSIMP(%);
/*SOLVE ABOVE FOR B1 B2 B3,FOLLOWIN IS A CROSS CHECK */
DEL:RATSIMP(1+A[1,1]+A[2,2]+A[3,3]);
BB1:RATSIMP(1/DEL*(A[3,2]-A[2,3]));
BB2.RATSIMP(1/DEL*(A[1,3]-A[3,1]));
BB3:RATSIMP(1/DEL*(A[2,1]-A[1,2]));
```

## APPENDIX III

```
/*TWO SUCCESIVE ROTATIONS IN TERMS OF RODRIGUES PARAMETER*/
ALT(N):=MATRIX([0,-N[3,1],N[2,1]],[N[3,1],0,-N[1,1]],[-N[2,1],N[1,1],0]);
B.MATRIX([B1],[B2],[B3]);
BB:ALT(B);
I:IDENT(3);
INBB.(I-BB)^^-1;
A.INBB.(I+BB)$
A:RATSIMP(%);
BP:MATRIX([BP1],[BP2],[BP3]);
BBP:ALT(BP);
INBBP.(I-BBP)^^-1;
AP.INBBP.(I+BBP)$
AP.RATSIMP(%);
APP:AP.A;
/*SOLVE ABOVE FOR BPP1 BPP2 BPP3 */
DEL:RATSIMP(1+APP[1,1]+APP[2,2]+APP[3,3]);
BPP1:RATSIMP(1/DEL*(APP[3,2]-APP[2,3]));
BPP2 RATSIMP(1/DEL*(APP[1,3]-APP[3,1]));
BPP3:RATSIMP(1/DEL*(APP[2,1]-APP[1,2]));
/*THE ABOVE RESULTS ARE SAME AS EQUATION (11) */
```

## APPENDIX IV

```
/* DERIVE EULER IDENTITY SEE ALSO  BRAND REF. [2] P.408*/
S:MATRIX(
[0, CC1, CC2, CC3],
[CC1,0,-CC3,CC2],
[CC2,CC3,0,-CC1],
[CC3,-CC2,CC1,0]);
SP.MATRIX(
[0,-CP1,-CP2,-CP3],
[CP1,0,-CP3,CP2],
[CP2,CP3,0,-CP1],
[CP3,-CP2,CP1,0]);
I:IDENT(4);
V:CC0*I+S;
VP:CP0*I+SP;
MAT1:V.VP;
/* NOW TAKE THE FIRST COLUMN OF THE ABOVE MATRIX AND SQUARE IT*/
MAT2:SUBMATRIX(%,2,3,4);
ANSWER:%.%;
ANSWER:FACTOR(ANSWER);
/* NOTE ABOVE IS A COMPLETE SQUARE */
```

## APPENDIX V

```
/*QUATERNION MULTIPLICATION EXAMPLE */
/*ANALOG OF CAYLEY-KLEIN RESULT */
I:IDENT(4);
/* NOW WE DEFINE AN OPERATION SS ON A COLUMN MATIX BASED ON ANALOG
OF CAYLEY KLEIN DECOMPOSITION */
SS(CC):=MATRIX([CC[1,1],-CC[2,1],-CC[3,1],-CC[4,1]],
[CC[2,1],CC[1,1],-CC[4,1],CC[3,1]],
[CC[3,1],CC[4,1],CC[1,1],-CC[2,1]],
[CC[4,1],-CC[3,1],CC[2,1],CC[1,1]]);
/* DEFINE AN INVERSE OPERATION */
INV(CC):=1/(CC.CC)*MATRIX([CC[1,1]],[-CC[2,1]],[-CC[3,1]],[-CC[4,1]]);
/* NOW THE BRANDS'S THEOREM ON QUATERNION FORMULATED IN MATRIX FORM *
RHO:MATRIX([R0],[R1],[R2],[R3]);
GAM:MATRIX([Q0],[Q1],[Q2],[Q3]);
/*NOW DEFINE  QUATERNION PRODUCT */
APROD(R,Q):=SS(R).Q;
A:MATRIX([A0],[A1],[A2],[A3]);
RATSIMP(APROD(INV(A),A));
ANSWER:RATSIMP(APROD(GAM,APROD(RHO,INV(GAM))));
EQ1:ANSWER[1,1];
EQ2:ANSWER[2,1];
EQ3.ANSWER[3,1];
EQ4:ANSWER[4,1];
/* NOW GENERATE COEFFICIENT MATRIX FOR RHO */
COEFMATRIX([EQ1,EQ2,EQ3,EQ4],[R0,R1,R2,R3]);
/* THE ABOVE IS SAME AS EXTENDED EULER PARAMETER MATRIX */
```

## APPENDIX VI

```
/* THE BASIC DECOMPOSITION FOR EULER PARAMETER */
/*TEST OUT (C0*I+S)X(C0*I-S)      */
II:IDENT(4);
SS:MATRIX([0,-CC1,-CC2,-CC3],
[CC1,0,-CC3,CC2],
[CC2,CC3,0,-CC1],
[CC3,-CC2,CC1,0]);
Q.MATRIX([Y,Z,0,-X],
[Z,-Y,X,0],
[0,X,Y,Z],
[-X,0,Z,-Y]);
EQ1:EXPAND((CC0*II+SS).Q.(CC0*II-SS));
T1:EQ1[2,3];
T2:EQ1[1,1];
T1:EQ1[4,3];
ANSWER COEFMATRIX([T1,T2,T3],[X,Y,Z]);
/*ABOVE IS SAME A S EQUATION 11 */
```

## APPENDIX VII

```
/*..... ALGEBRA FOR QUATERNIONS FROM YANG'S PAPER....*/
NNPRED(N):=IS(N>=2);
MATCHDECLARE(NN,NNPRED);
TELLSIMPAFTER(EP^NN,0),
/*ABOVE WILL ELIMINATE EP**2 TERMS */
AL12H:AL12+EP*A12;
AL23H:AL23+EP*A23;
AL34H:AL34+EP*A34;
AL41H:AL41+EP*A41;
TH1H:TH1+EP*S1;
TH2H:TH2+EP*S2;
TH3H:TH3+EP*S3;
TH4H:TH4+EP*S4;
SAL12H: EXPAND(TAYLOR(SIN(AL12H),EP,0,1));
SAL23H: EXPAND(TAYLOR(SIN(AL23H),EP,0,1));
SAL34H: EXPAND(TAYLOR(SIN(AL34H),EP,0,1));
SAL41H: EXPAND(TAYLOR(SIN(AL41H),EP,0,1));
STH1H:  EXPAND(TAYLOR(SIN(TH1H),EP,0,1));
STH2H:  EXPAND(TAYLOR(SIN(TH2H),EP,0,1));
STH3H:  EXPAND(TAYLOR(SIN(TH3H),EP,0,1));
STH4H:  EXPAND(TAYLOR(SIN(TH4H),EP,0,1));
CAL12H:EXPAND(TAYLOR(COS(AL12H),EP,0,1));
CAL23H EXPAND(TAYLOR(COS(AL23H),EP,0,1));
CAL34H:EXPAND(TAYLOR(COS(AL34H)),EP,0,1));
CAL41H:EXPAND(TAYLOR(COS(AL41H),EP,0,1));
CTH1H: EXPAND(TAYLOR(COS(TH1H)),EP,0,1));
CTH2H: EXPAND(TAYLOR(COS(TH2H)),EP,0,1));
CTH3H: EXPAND(TAYLOR(COS(TH3H)),EP,0,1));
CTH4H: EXPAND(TAYLOR(COS(TH4H)),EP,0,1));
AATHHH:SAL12H*SAL34H*STH1H;
BBTHHH:-SAL34H*(SAL41H*CAL12H+CAL41H*SAL12H*CTH1H);
CCTHHH:CAL23H-CAL34H*(CAL41H*CAL12H-SAL41H*SAL12H*CTH1H);
EQ1:AATHHH*STH4H+BBTH1H*CTH4H-CCTHHH;
PRIMARY:EV(EQ1,EP=0);
DUAL:RATCOEFF(EQ1,EP);
A:RATCOEFF(PRIMARY,SIN(TH4));
B RATCOEFF(PRIMARY,COS(TH4));
C:EXPAND(PRIMARY-A*SIN(TH4)-B*COS(TH4));
DUAL1:DUAL-S4*(A*COS(TH4)-B*SIN(TH4));
A0 RATCOEFF(DUAL1,SIN(TH4));
B0 RATCOEFF(DUAL1,COS(TH4));
CC0.EXPAND(DUAL1-A0*SIN(TH4)-B0*COS(TH4));
CC0.RATSIMP(CC0);
```

## APPENDIX VIII

```
/*DYNAMICAL EQUATIONS FOR STANFORD MANIPULATOR*/
MATCHDECLARE(A,TRUE);
DEPENDS([Q1,Q2,Q3,Q4,Q5,Q6],T);
DEPENDS(U,T);
/*TRIGNOMETRIC SIMPLIFICATIONS */
TELLSIMP(SIN(A)^2,1-COS(A)^2);
S1:SIN(Q1);
CC1:COS(Q1);
S2:SIN(Q2);
CC2:COS(Q2);
S3:SIN(Q3);
CC3:COS(Q3);
/* EXPRESS LOCAL ANGULAR VELOCITIES IN TERMS OF GENERALIZED ONES*/
QD1:1/S2*(U[1]*S3-U[3]*CC3);
QD2.U[1]*CC3+U[3]*S3;
QD3.U[2]+(U[3]*CC3-U[1]*S3)*CC2/S2;
QD4 U[4],
QD5 U[5];
QD6 U[6];
GRADEF(Q1,T,QD1);
GRADEF(Q2,T,QD2);
GRADEF(Q3,T,QD3);
GRADEF(Q4,T,QD4);
GRADEF(Q5,T,QD5);
GRADEF(Q6,T,QD6);
/*DEFINE ROTATIONS */
ROTX(Q):=MATRIX([1,0,0],[0,COS(Q),-SIN(Q)],[0,SIN(Q),COS(Q)]);
ROTY(Q):=MATRIX([COS(Q),0,SIN(Q)],[0,1,0],[-SIN(Q),0,COS(Q)]);
ROTZ(Q):=MATRIX([COS(Q),-SIN(Q),0],[SIN(Q),COS(Q),0],[0,0,1]);
W1:MATRIX([0,QD1,0]);
W2:MATRIX([QD2,0,0]);
W3:MATRIX([0,QD3,0]);
W4.MATRIX([QD4,0,0]);
W5:MATRIX([0,QD5,0]);
W6.MATRIX([0,QD6,0]);
/*SET UP ROTATION MATRICES */
R1.ROTY(Q1);
R2:ROTX(Q2);
R3:ROTY(Q3);
R4:ROTX(Q4);
R5.ROTY(Q5);
/*STAGE 1. SET UP ANGULAR VELOCITIES */
WA:EXPAND(W1 . R1);
WB:EXPAND(W1 . R1 . R2+W2 . R2);
WC:WB;
WD:EXPAND(W1 . R1 . R2 . R3+W2 . R2 . R3+W3 . R3);
WE:EXPAND(W1 . R1 . R2 . R3 . R4+W2 . R2 . R3 . R4+W3 . R3 . R4+W4 . R4);
WF:EXPAND(W1.R1.R2.R3.R4. R5+W2.R2.R3.R4.R5+W3.R3 R4.R5+W4.R4.R5+W5 R5);
/* SET UP BASE VECTORS AND CROSS PRODUCT */
AA:MATRIX([AA1,AA2,AA3]);
BB.MATRIX([BB1,BB2,BB3]);
CC:MATRIX([CC1,CC2,CC3]);
DD.MATRIX([DD1,DD2,DD3]);
EE:MATRIX([EE1,EE2,EE3]);
FF.MATRIX([FF1,FF2,FF3]);
CROSS(A,B,BASE):=BLOCK([],MATRIX([A[1,2]*B[1,3]-A[1,3]*B[1,2],
-(A[1,1]*B[1,3]-A[1,3]*B[1,1]),A[1,1]*B[1,2]-A[1,2]*B[1,1]]));
/* LENTHE VECTORS FOR VELOCITIES */
VECL1:MATRIX([L1,0,0]);
VECQ6:MATRIX([0,Q6,0]);
VECL5:MATRIX([0,L5,0]);
VECL2:MATRIX([0,L2,0]);
VECL6:MATRIX([0,L6,0]);
VECL3:MATRIX([0,L3,0]);
VECL4:MATRIX([0,L4,0]);
/*STAGE 2. SET UP LINEAR VELOCITIES */
VA:MATRIX([0,0,0]);
RB:MATRIX([L1,L4,0]);
VB:CROSS(WA,RB,AA);
RC:VECQ6+VECL1 . R2;
VC:CROSS(WB,RC,CC);
/*ADD LINEAR COMPONENT */
VC:VC+W6;
RD:VECL1 . R2+VECQ6+VECL5;
VD:CROSS(WB,RD,CC);
/*ADD LINEAR COMPONENT */
VD:VD+W6;
/*FOR VF START WITH VELOCITY OF C */
VE:EXPAND(VC . R3 . R4)+CROSS(WE-W4,VECL2 . R3 . R4,EE)+CROSS(WE,VECL6,EE);
/*REPLACE L6 BY L3 IN ABOVE FOR VELOCITY OF F*/
VF:RATSUBST(L3,L6,%);
/*STAGE 3. SET UP PARTIAL ANGULAR VELOCITIES */
FOR I THRU 6 DO LDISPLAY(WAR[I]:RATCOEF(WA,U[I]));
FOR I THRU 6 DO WBR[I]:RATCOEF(WB,U[I]);
FOR I THRU 6 DO WCR[I]:RATCOEF(WC,U[I]);
FOR I THRU 6 DO WDR[I]:RATCOEF(WD,U[I]);
FOR I THRU 6 DO WER[I]:RATCOEF(WE,U[I]);
FOR I THRU 6 DO WFR[I]:RATCOEF(WF,U[I]);
/*STAGE 4. SET UP PARTIAL LINEAR VELOCITIES */
FOR I THRU 6 DO VAR[I]:RATCOEF(VA,U[I]);
FOR I THRU 6 DO VBR[I].RATCOEF(VB,U[I]);
FOR I THRU 6 DO VCR[I]:RATCOEF(VC,U[I]);
FOR I THRU 6 DO VDR[I]:RATCOEF(VD,U[I]);
FOR I THRU 6 DO VER[I]:RATCOEF(VE,U[I]);
FOR I THRU 6 DO VFR[I]:RATCOEF(VF,U[I]);
/*STAGE 5. FIND THE ANGULAR ACCELERATIONS */
ALPHAA:DIFF(WA,T)$
ALPHAB:DIFF(WB,T)$
ALPHAC:DIFF(WC,T)$
ALPHAD:DIFF(WD,T)$
ALPHAE.DIFF(WE,T)$
ALPHAF:DIFF(WF,T)$
/*STAGE 6. FIND THE LINEAR ACCELERATION */
ACCA:DIFF(VA,T);
ACCB.DIFF(VB,T)+CROSS(WA,VB,AA)$
ACCC:DIFF(VC,T)+CROSS(WB,VC,BB)$
ACCD:DIFF(VD,T)+CROSS(WC,VD,CC)$
ACCE:DIFF(VE,T)+CROSS(WE,VE,EE)$
ACCF.DIFF(VF,T)+CROSS(WF,VF,EE)$
/*STAGE 7. MOMENTS OF INERTIA */
```

```
IA:MATRIX([IA1,IA2,IA3]);
IB:MATRIX([IB1,IB2,IB3]);
IC:MATRIX([IC1,IC2,IC3]);
ID:MATRIX([ID1,ID2,ID3]);
IE:MATRIX([IE1,IE2,IE3]);
IF:MATRIX([IIF1,IIF2,IIF3]);
/*STAGE 8,9,10. DEFINE TORQUES,REACTIONS,AND GENERALIZED FORCES FOR A,B,C,D,E,F*/
TAS:-ALPHAA*IA-CROSS(WA,IA*WA,AA)$
RAS:-MA*ACCA$
FOR I THRU 6 DO LDISPLAY(KAS[I]:WAR[I] . TAS+VAR[I] . RAS)$
TBS:-ALPHAB*IB-CROSS(WB,IB*WB,BB)$
RBS:-MB*ACCB$
FOR I THRU 6 DO KBS[I]:WBR[I] . TBS+VBR[I] . RBS$
TCS:-ALPHAC*IC-CROSS(WC,IC*WC,CC)$
RCS:-MC*ACCC$
FOR I THRU 6 DO KCS[I]:WCR[I] . TCS+VCR[I] . RCS$
TDS:-ALPHAD*ID-CROSS(WD,ID*WD,DD)$
RDS:-MD*ACCD$
FOR I THRU 6 DO KDS[I]:WDR[I] . TDS+VDR[I] . RDS$
TES:-ALPHAE*IE-CROSS(WE,IE*WE,EE)$
RES:-ME*ACCE$
FOR I THRU 6 DO KES[I]:WER[I] . TES+VER[I] . RES$
TFS:-ALPHAF*IF-CROSS(WF,IF*WF,FF)$
RFS:-MF*ACCF$
FOR I THRU 6 DO KFS[I]:WFR[I] . TFS+VFR[I] . RFS$
/*SUM ALL CORRESPONDING GENERALIZED FORCES */
KK1:KAS[1]+KBS[1]+KCS[1]+KDS[1]+KES[1]+KFS[1]$
KK2:KAS[2]+KBS[2]+KCS[2]+KDS[2]+KES[2]+KFS[2]$
KK3:KAS[3]+KBS[3]+KCS[3]+KDS[3]+KES[3]+KFS[3]$
KK4:KAS[4]+KBS[4]+KCS[4]+KDS[4]+KES[4]+KFS[4]$
KK5:KAS[5]+KBS[5]+KCS[5]+KDS[5]+KES[5]+KFS[5]$
KK6:KAS[6]+KBS[6]+KCS[6]+KDS[6]+KES[6]+KFS[6]$
/*STAGE 10. SET UP ACTIVE FORCES */
GA:MATRIX([0,-G*MA,0]);
GB:MATRIX([0,-G*MB,0]);
GC:-G*MC*MATRIX([0,CC2,-S2]);
GD:-G*MD*MATRIX([0,CC2,-S2]);
GE:-G*ME*MATRIX([0,1,0]) . R1 . R2 . R3 . R4;
GF:-G*MF*MATRIX([0,1,0]) . R1 . R2 . R3 . R4;
TNA:MATRIX([0,TAU1,0]);
TBA:MATRIX([TAU2,0,0]);
TCB:MATRIX([0,-SIGMA,0]);
TDC:MATRIX([0,TAU3,0]);
TED:MATRIX([TAU4,0,0]);
TFE:MATRIX([0,TAU5,0]);
RNA:MATRIX([0,0,0]);
/*SET UP GENERALIZED ACTIVE FORCES */
SPECIAL2[R]:=BLOCK(IF R = 6 THEN -SIGMA ELSE 0);
KTOTALR[R]:=SPECIAL2[R]+WAR[R] . TNA+(WAR[R] . R2-WBR[R]) . TBA
+(WCR[R] . R3-WDR[R]) . TDC . R3+(WDR[R] . R4-WER[R]) . TED . R4
+(WER[R] . R5-WFR[R]) . TFE . R5+VBR[R] . GB+VCR[R] . GC+VDR[R] . GD+VER[R]
KEEPFLOAT:TRUE;
/*NUMERICAL EXAMPLE WITH VALUES GIVEN IN REF. [9]*/
G:9.8; L1:.1; L2:.6; L3:.2; L4:.1; L5:0.7; L6:0.06; MA:9; MB:6; MC:4; MD:1;
ME:0.6; MF:0.5; IA1:0.01; IA2:0.02; IA3:0.01; IB1:0.06; IB2:0.01; IB3:0.05;
IC1:0.4; IC2:0.01; IC3:0.4, ID1:0.0005; ID2:0.001; ID3:0.001; IE1:0.0005;
IE2:0.0002; IF3:0.0005; IIF1:0.001; IIF2:0.002; IIF3:0.003;
EV(FT:(T-10/(2*%PI)*SIN(2*%PI*T/10))*(%PI/180),NUMER);
TQ1:60/10*FT;
TQ2:%PI/2+(60-90)/10*FT;
TQ3:TQ1;
TQ4:TQ1;
TQ5:TQ1;
TQ6:1/10;
U[1]:DIFF(TQ1,T)*SIN(TQ2)*SIN(TQ3)+DIFF(TQ2,T)*COS(TQ3);
U[2]:DIFF(TQ1,T)*COS(TQ2)+DIFF(TQ3,T);
U[3]: DIFF(TQ1,T)*SIN(TQ2)*COS(TQ3)+DIFF(TQ2,T)*SIN(TQ3);
U[4]:DIFF(TQ4,T);
U[5]:DIFF(TQ5,T);
U[6]:DIFF(TQ6,T);
Q1:TQ1;
Q2:TQ2;
Q3:TQ3;
Q4:TQ4;
Q5:TQ5;
Q6:TQ6;
/* NOW WE PLOT RESULTS AND COMPARE WITH REF. [9] */
FINAL6:KTOTALR[6]+SIGMA+KK6;
FINAL6:EV(FINAL6,DIFF)$
EQUALSCALE:FALSE;
PLOTNUM:10;
PLOT(FINAL6,T,0,10,"PLOT OF SIGMA");
FINAL5:EV(KK5,DIFF)$
PLOT(FINAL5,T,0,10,"PLOT OF TAU 5");
FINAL4:EV(KK4+KTOTALR[4]+TAU4,DIFF)$
PLOT(FINAL4,T,0,10,"PLOT OF TAU 4");
FINAL3:EV(KK2+KTOTALR[2]+TAU3,DIFF)$
PLOT(FINAL3,T,0,10,"PLOT OF TAU3");
/*TRY TO SIMPLIFY AND COLLECT TERMS X I J IN EQUATION OF MOTION*/
/*FIRST DELETE NUMERICAL VALUES */
FOR I:1 QRU 6 DO ( FOR J:1 QRU 6 DO XXA[I,J]:RATCOEFF(KAS[I],DIFF(U[J],T)));
FOR I:1 QRU 6 DO ( FOR J:1 QRU 6 DO XXB[I,J]:RATCOEFF(KBS[I],DIFF(U[J],T)));
FOR I:1 QRU 6 DO ( FOR J:1 QRU 6 DO XXC[I,J]:RATCOEFF(KCS[I],DIFF(U[J],T)));
FOR I:1 QRU 6 DO ( FOR J:1 QRU 6 DO XXD[I,J]:RATCOEFF(KDS[I],DIFF(U[J],T)));
FOR I:1 QRU 6 DO ( FOR J:1 QRU 6 DO XXE[I,J]:RATCOEFF(KES[I],DIFF(U[J],T)));
FOR I:1 QRU 6 DO ( FOR J:1 QRU 6 DO XXF[I,J]:RATCOEFF(KFS[I],DIFF(U[J],T)));
FOR I:1 QRU 6 DO ((FOR J:1 QRU 6 DO XXX[I,J]:RATSIMP(XXA[I,J]+XXB[I,J]+
XXC[I,J]+XXD[I,J]+XXE[I,J]+XXF[I,J])));
FOR I:1 THRU 6 DO ( FOR J:1 THRU 6 DO LDISPLAY (XXX[I,J]));
/*COMPARE ABOVE X,I,J WITH THOSE OF REF. [9]*/
```

## APPENDIX IX

```
/*SPACECRAFT EXAMPLE*/
        /*.. .....CARTESIAN DIV AND CURL DEFINITION.. .....
        .....UNIT VECTORS ARE B1 B2 B3..... SEE LEVINSON      */
        /*... ..... ..DEFINE DOT AND CROSS PRODUCTS. .*/
        DOT(V1,V2):=BLOCK([P,PP],
        FOR I:1 THRU 3 DO P[I]:RATCOEFF(V1,B[I]),
        FOR I:1 THRU 3 DO PP[I]:RATCOEFF(V2,B[I]),
        P[4]:SUM(P[I]*PP[I],I,1,3),
        RETURN(P[4]))$
        CROSS(V1,V2):=BLOCK([P,PP,PPP],
        FOR I:1 THRU 3 DO P[I]:RATCOEFF(V1,B[I]),
        FOR I:1 THRU 3 DO PP[I]:RATCOEFF(V2,B[I]),
        PPP[1]:(P[2]*PP[3]-P[3]*PP[2]),
        PPP[2]:(-P[1]*PP[3]+P[3]*PP[1]),
        PPP[3]:(P[1]*PP[2]-P[2]*PP[1]),
        PPP[4]:B[1]*PPP[1]+B[2]*PPP[2]+B[3]*PPP[3],
        RETURN(PPP[4]))$
        /*.....NOW WE INPUT EQUATIONS FROM LEVINSON'S PAPER */
        DEPENDS(U,T);
        DEPENDS(Q,T),
        R[2]:COS(Q)*B[2]+SIN(Q)*B[3];
        R[3]:-SIN(Q)*B[2]+COS(Q)*B[3];
        WB:U[1]*B[1]+U[2]*R[2]+U[3]*B[3];
        DERIVABBREV:TRUE;
        U[4]:DIFF(Q,T);
        WR:U[1]*U[4])*B[1]+ U[2]*B[2]+U[3]*B[3];
        ALPB DIFF(U[1],T)*B[1]+DIFF(U[2],T)*B[2]+DIFF(U[3],T)*B[3];
        ALFR:DIFF(WR,T)+CROSS(WB,WR);
        PRS:R1*B[1]+R2*R[2]+R3*B[3];
        PP2S:R1*B[1]+R2*R[2]+R3*R[3];
        PRSBS:PP2S-PPRS;
        VBS:U[5]*B[1]+U[6]*B[2]+U[7]*B[3];
        VRS.VBS +DIFF(PRSPS,T) +  .(OSS(WB PRSBS);
        ABS:DIFF(VBS,T)+CROSS(WB,VBS);
        ARS:DIFF(VRS,T)+CROSS(WR,VRS);
        IBBSWB:BET1*B[1]*DOT(B[1],WB. +BET2*B[2]*DOT(B[2],WB)+BET3*B[3]*DOT(B[3],WB);
        IRRSWR:RHO1*B[1]*DOT(B[1],WR.+RHO2*B[2]*DOT(B[2],WR)+RHO3*B[3]*DOT(B[3],WR);
        IBBSALPB:BET1*B[1]*DOT(B[1],ALPB)+ BET2*B[2]*DOT(B[2],ALPB)+BET3*B[3]*DOT(B[3],ALPB);
        IRRSALPR:RHO1*B[1]*DOT(B[1],ALPR)+RHO2*B[2]*DOT(B[2],ALPR)+RHO3*B[3]*DOT(B[3],ALPR);
        FSB:-MD*ARS;
        FSR:-MR*ARS;
        TSB:CROSS(IBBSWB,WB)-IBBSALPB;
        TSR:CROSS(IRRSWR,WR)-IRRSALPR;
        FB.F1*B[1]+F2*B[2]+F3*B[3];
        TB.T1*B[1]+T2*B[2]+T3*B[3];
        F[R]:=DOT(DIFF(VBS,U[R]),FB)+DOT(DIFF(WB,U[R]),TB);
        FS[R]:=DOT(DIFF(VBS,U[R]),FSB)+DOT(DIFF(VRS,U[R]),FSR)
        +DOT(DIFF(WB,U[R]),TSB)+DOT(D. F(WP,U[R]),TSR);
        EQ[R]:=F[R]+FS[R];
        EQ[1];
        FOR I:1 THRU 7 DO LDISPLAY ( X[I,1]:RATCOEFF(EQ[I],DIFF(U[I],T)));
```

# Dynamic Instability of the Flexible Coupler of a Four-Bar Mechanism

Iradj G. Tadjbakhsh*

Abstract

Dynamic behavior of flexible components of mechanisms is prone to instabilities which create resonant speed barriers. By considering small deformations superimposed on the steady dynamic state equations governing evolution of disturbances can be obtained. For the case of mechanisms driven by periodic inputs these equations reduce to a system of coupled Mathieu-Hill equations for the amplitudes of modes of vibrations. Application of the Floquet theory determines the critical conditions of speed, geometry and material properties causing dynamic instability.

## 1. Introduction

A primary cause of instability in linkages is the flexibility of its members. Under steady operations periodic time dependent axial forces act on these members that may lead to instability characterized by unbounded growth of small disturbances. The situation is similar to dynamic buckling of columns under the impact of periodic time dependent loads [1-2]. Some recent investigations into dynamic stability of flexible mechanisms is reported in references [3-5].

In this paper we consider the dynamic instability of the coupler of a four-bar mechanism with equal input and output crank lengths. This

PREVIOUS PAGE IS BLANK

*Professor of Civil Engineering, Rensselaer Polytechnic Institute, Troy, New York.

type of mechanism with its simple kinematics is often encountered and affords the opportunity for analyses which can be generalized to more complex situations.

Our method of analysis consists of considering small deformations of the elastic coupler superimposed on its undeformed straight configuration. Utilization of Galarkin's method in the resulting equations of motions leads to a system of coupled Mathieu-Hill equations determining the behavior of the amplitude of various modes of vibrations. Finally we employ known results from the theory of such systems to determine conditions of parametric resonance.

## 2. Formulation and Analysis

The accelerations $\underset{\sim}{a}$ of points along the coupler AB, of the four-bar mechanism, shown in Fig. 1, is given by

$$\underset{\sim}{a} = -R(\dot{\beta}^2\cos\beta+\ddot{\beta}\sin\beta)\ \underset{\sim}{i} + (-R\dot{\beta}^2\sin\beta+R\ddot{\beta}\cos\beta+\ddot{v})\ \underset{\sim}{j} \qquad (1)$$

Here $\beta(t)$ is the angular displacement of the wheels, $v(s,t)$ the transverse displacement of the coupler, s the arc length along the central line of the couler, R the radius of the wheels and $\underset{\sim}{i}$ and $\underset{\sim}{j}$ unit vectors in the horizontal and vertical directions. Assumption of small displacements is employed.

Dynamic equilibrium of the coupler and the wheels, Fig. 2, are expressed by

$$\underset{\sim}{F_1} + \underset{\sim}{F_2} = \gamma \int_0^{\ell} \underset{\sim}{a}\, ds \qquad (2)$$

$$\ell\ \underset{\sim}{i} \times \underset{\sim}{F_2} = \gamma \int_0^{\ell} (s\ \underset{\sim}{i} + v\ \underset{\sim}{j}) \times \underset{\sim}{a}\, ds \qquad (3)$$

$$\underset{\sim}{M}_1 - R\,\underset{\sim}{e} \times \underset{\sim}{F}_1 = J_1\,\ddot{\beta}\,\underset{\sim}{k} \tag{4}$$

$$\underset{\sim}{M}_2 - R\,\underset{\sim}{e} \times \underset{\sim}{F}_2 = J_2\,\ddot{\beta}\,\underset{\sim}{k} \tag{5}$$

in which $\underset{\sim}{F}_1$ and $\underset{\sim}{F}_2$ are the end forces in the coupler, $\underset{\sim}{k} = \underset{\sim}{i} \times \underset{\sim}{j}$, and $\underset{\sim}{M}_1$ ($= M_1\,\underset{\sim}{k}$) and $\underset{\sim}{M}_2$ ($= M_2\,\underset{\sim}{k}$) the externally applied couples on the wheels. Also $J_1$, $J_2$ stand for the mass moments of inertia of the wheels, $\gamma$ ($=\rho A$) and $\ell$ the mass per unit length and the total length of the coupler and $\underset{\sim}{e} = \cos\beta \cdot \underset{\sim}{i} + \sin\beta \cdot \underset{\sim}{j}$ the unit vector along OA.

Eqs. (2)-(4) constitute five scalar equations for the determination of the four unknowns comprised of the components of $\underset{\sim}{F}_1$ and $\underset{\sim}{F}_2$. Thus we find for $\underset{\sim}{F}_1$

$$\underset{\sim}{F}_1 = \left(\cot\beta \cdot \psi - \frac{M_1 - J_1\,\ddot{\beta}}{R \sin\beta}\right) \underset{\sim}{i} + \psi\,\underset{\sim}{j} \tag{6}$$

where

$$\psi = \frac{1}{2} R\,\ell\,\gamma(-\dot{\beta}^2\sin\beta + \ddot{\beta}\cos\beta) - \frac{\gamma R}{\ell}\left(\dot{\beta}^2\cos\beta + \ddot{\beta}\sin\beta\right) \int_0^\ell v\,ds + \gamma \int_0^\ell \left(1 - \frac{s}{\ell}\right) \ddot{v}\,ds \tag{7}$$

This result is valid provided

$$(J_1 + J_2 + R\,\ell\,\gamma)\,\ddot{\beta} + \gamma\cos\beta \int_0^\ell \ddot{v}\,ds = M_1 + M_2 \tag{8}$$

which is also a consequence of (2)-(5).

the stress resultant $\underset{\sim}{f}$ the axial force T and the transverse shear N in the coupler, at a point s, are given by

$$\underset{\sim}{f} = -\underset{\sim}{F}_1 + \gamma \int_0^s \underset{\sim}{a}\,ds \tag{9}$$

$$T = (\underset{\sim}{i} + v' \underset{\sim}{j}) \cdot \underset{\sim}{f} \tag{10}$$

$$N = (- v' \underset{\sim}{i} + \underset{\sim}{j}) \cdot \underset{\sim}{f} \tag{11}$$

in which primes denote differentiation with respect to s.

Carrying out the indicated operations in (9)-(11) and retaining only linear terms in y, one obtains

$$T = \gamma \cot\beta[\frac{R}{\ell}(\dot\beta^2\cos\beta+\ddot\beta\sin\beta)\int_0^\ell v\,ds + \int_0^\ell (\frac{s}{\ell} - 1)\,\ddot v\,ds]$$

$$+ R\gamma[(\dot\beta^2\sin\beta-\ddot\beta\cos\beta)\,(\frac{\ell}{2} - s)\,v' + \dot\beta^2\cos\beta.(\frac{\ell}{2} - s)$$

$$- \frac{\ell}{2}\ddot\beta\cot\beta.\cos\beta - s\ddot\beta.\sin\beta] + \frac{M_1 - J_1\ddot\beta}{R\sin\beta} \tag{12}$$

and

$$N = \gamma[- \int_0^\ell \ddot v\,ds + \frac{R}{\ell}(\dot\beta^2\cos\beta+\ddot\beta\sin\beta)\int_0^\ell v\,ds + \frac{1}{\ell}\int_0^\ell s\,\ddot v\,ds]$$

$$- \{R\gamma[\dot\beta^2\cos\beta.(\frac{\ell}{2} - s) - \frac{1}{2}\ell\,\ddot\beta\cos\beta\cot\beta - s\ddot\beta\sin\beta] + \frac{M_1 - J_1\ddot\beta}{R\sin\beta}\}v'$$

$$+ R\gamma\,(\frac{\ell}{2} - s)\,(\dot\beta^2\sin\beta-\ddot\beta\cos\beta) \tag{13}$$

Conservation of linear and angular momenta is expressed by (9) and

$$EI\,v''' - \rho I\ddot v' + N = 0 \tag{14}$$

respectively. In (14) it is assumed that bending moment-curvature relationship is $M = EIv''$. Substitution for N, from (13) into (14), results in an inhomogeneous linear equation for y.

Restricting subsequent considerations to the steady case, $\ddot{\beta} = 0$, $\beta = \omega t$, we select $\beta$ as the independent variable replacing t. Then (14) becomes

$$EIv''' - \rho I\omega^2 v'_{\beta\beta} + \gamma\omega^2\{\frac{R}{\ell}\cos\beta \int_0^{\ell} v\, ds + \frac{1}{\ell}\int_0^{\ell} s\, v_{\beta\beta}\, ds - \int_0^{\ell} v_{\beta\beta}\, ds$$

$$-[R(\frac{\ell}{2} - s)\cos\beta + \frac{M_1}{R\sin\beta}]v'\} = -\gamma\omega^2 R(\frac{\ell}{2} - s)\cos\beta \qquad (15)$$

The boundary conditions that accompany (15) are

$$v(0,\beta) = v(\ell,\beta) = 0$$

$$v''(0,\beta) = v''(\ell,\beta) = 0 \qquad (16)$$

We shall attempt to obtain the solution for y for a given constant input moment $M_1$. The output moment $M_2$ is then obtained from (8).

3. Approximate Solution

Employing Galerkin procedure we assume

$$v = \sum_{m=1}^{N} g_m(\beta)\sin\frac{m\pi s}{\ell} \qquad (17)$$

which satisfies (16). Substitution into (15) and minimizing the error leads to the solution of the system

$$\underset{\sim}{C}\, \underset{\sim}{g}'' + \frac{1}{\lambda}(\underset{\sim}{D} + \frac{m_1}{\sin\beta}\underset{\sim}{I})\, \underset{\sim}{g} + 2\mu r^2\cos\beta\, \underset{\sim}{E}\, \underset{\sim}{g} = 2\mu r^2 \sin\beta . \underset{\sim}{f} \qquad (18)$$

Here $\underset{\sim}{g} = (g_1, g_2, \ldots g_N)^T$, $\mu = R/\ell$, $m_1 = M_1\ell^2/REI$, $\lambda = \rho\ell^2\omega^2/E$ and $r = \ell(A/I)^{1/2}$ is the parameter defining the slenderness of the coupler. The matrices C and D are diagonal with elements $C_{mm} = 1 + r^2(m\pi)^{-2}$,

$D_{mm} = (m\pi)^2$. The matrix $\underset{\sim}{I}$ is the unit matrix and $\underset{\sim}{E}$ is a symmetric zero-diagonal matrix with elements

$$E_{mn} = \frac{m^2 + n^2}{\pi^2(m^2-n^2)^2}\,[1 - (-1)^{m+n}] \tag{19}$$

Finally, components of the vector $\underset{\sim}{f}$ are given by

$$f_m = \frac{1-(-1)^m}{(m\pi)^2}\ , \qquad m = 1,2,3 \ \ldots\ N \tag{20}$$

Limits of stability of solutions $\underset{\sim}{g}$ of (18) are determined by the behavior of the solutions to the homogeneous form of (18) which is written in the form

$$\sin\beta\ \underset{\sim}{C}\ \underset{\sim}{g}'' + \lambda^{-1}(\sin\beta.\underset{\sim}{D} + m_1\underset{\sim}{I})\underset{\sim}{g} + \mu r^2 \sin 2\beta\ \underset{\sim}{E}\ \underset{\sim}{g} = 0 \tag{21}$$

It can be shown that system of differential equations with periodic coefficients of period $2\pi$ admit of solutions of periods $2\pi$ and $4\pi$ [6]. More precisely, in the case of (21), the regions of dynamic instability in the four-dimensional parameter space of $\lambda$, $m_1$, r and $\mu$ are confined by surfaces $S(\lambda, m_1, r, \mu) = 0$ for which periodic solutions of the same period exist. Regions bounded by surfaces which correspond to solutions of different periods contain parameter values of stable solutions [2].

With a view to constructing solutions of period $2\pi$ we set

$$\underset{\sim}{g} = \frac{1}{2}\,\underset{\sim}{A}_o + \sum_{k=1}^{N} \underset{\sim}{A}_k \cos k\theta + \underset{\sim}{B}_k \sin k\theta \tag{22}$$

The constants $\underset{\sim}{A}_o$, $\underset{\sim}{A}_k$, $\underset{\sim}{B}_k$, k = 1,2,3, ..., can be determined by substituting (22) into (21) and, in the resulting equations, equating coefficients of various Fourier components to zero. In this way a system of algebraic equations are obtained the first three of which have special forms and

thereafter may be given with a general formula. These are:

$$m_1 \underset{\sim}{A}_0 + (\underset{\sim}{D} - \lambda \underset{\sim}{C})\, \underset{\sim}{B}_1 + \mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{B}_2 = 0$$

$$2m_1 \underset{\sim}{A}_1 + \mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{B}_1 + (\underset{\sim}{D} - 4\lambda \underset{\sim}{C})\, \underset{\sim}{B}_2 + \mu r^2 \underset{\sim}{E}\, \underset{\sim}{B}_3 = 0$$

$$\underset{\sim}{D}\, \underset{\sim}{A}_0 + \mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{A}_1 + 2m_1 \underset{\sim}{B}_1 + (4\lambda \underset{\sim}{C} - \underset{\sim}{D})\, \underset{\sim}{A}_2 - \mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{A}_3 = 0$$

$$- \mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{B}_{k-2} + [(k-1)^2 \lambda \underset{\sim}{C} - \underset{\sim}{D}]\, \underset{\sim}{B}_{k-1} + 2m_1 \underset{\sim}{A}_k +$$

$$[\underset{\sim}{D} - (k+1)^2 \lambda \underset{\sim}{C}]\, \underset{\sim}{B}_{k+1} + \mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{B}_{k+2} = 0, \quad k = 2,3,\ldots.N$$

$$\mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{A}_{k-2} + [\underset{\sim}{D} - (k-1)^2 \lambda \underset{\sim}{C}]\, \underset{\sim}{A}_{k-1} + 2m_1 \underset{\sim}{B}_k + \qquad (23)$$

$$[(k+1)^2 \lambda \underset{\sim}{C} - \underset{\sim}{D}]\, \underset{\sim}{A}_{k+1} - \mu r^2 \lambda \underset{\sim}{E}\, \underset{\sim}{A}_{k+2} = 0, \quad k = 2,3,\ldots.N$$

Existance of a solution to the homogeneous system (23) requires vanishing of an infinite determinant. In order to study the problem approximately but in a systematic way, we use a hierarchy of subsystem of increasing order in the number of unknowns. As the smallest subsystem we consider the case when the only non-zero elements are $\underset{\sim}{A}_0 = [A_{01}]$, $\underset{\sim}{A}_1 = [A_{11}]$, $\underset{\sim}{B}_1 = [B11]$. Then by considering only the three leading equations in the set (23) we obtain $A_{11} = 0$ and

$$m_1 A_{01} + (D_{11} - \lambda C_{11})\, B_{11} = 0$$

$$D_{11} A_{01} + 2m_1 B_{11} = 0 \qquad (24)$$

This implies

$$2m_1^2 = \pi^4 - (\pi^2 + r^2)\, \lambda \qquad (25)$$

which is a parabola in the $\lambda$–$m_1$ plane. The next larger and more accurate

first mode system consists of non-zero constants $A_{01}$, $A_{11}$, $B_{11}$, $A_{21}$, $B_{21}$. The resulting 5x5 system decomposes into

$$\begin{aligned} 2m_1 A_{11} + (D_{11} - 4\lambda C_{11}) B_{21} &= 0 \\ (D_{11} - \lambda C_{11}) A_{11} + 2m_1 B_{21} &= 0 \end{aligned} \tag{26}$$

and

$$\begin{aligned} m_1 A_{01} + (D_{11} - \lambda C_{11}) B_{11} &= 0 \\ D_{11} A_{01} + 2m_1 B_{11} - (D_{11} - 4\lambda C_{11}) A_{21} &= 0 \\ -(D_{11} - \lambda C_{11}) B_{11} + 2m_1 A_{21} &= 0 \end{aligned} \tag{27}$$

Existence of a solution to (26) and (27) implies

$$\begin{aligned} 4m_1^2 &= (D_{11} - \lambda C_{11})(D_{11} - 4\lambda C_{11}) = \\ &= [\pi^2 - \lambda(1 + r^2/\pi^2)][\pi^2 - 4\lambda(1 + r^2/\pi^2)] \end{aligned} \tag{28}$$

and

$$\begin{aligned} 4m^2 &= (D_{11} - \lambda C_{11})(3D_{11} - 4\lambda C_{11}) = \\ &= [\pi^2 - \lambda(1 + r^2/\pi^2)][3\pi^2 - 4\lambda(1 + r^2/\pi^2)] \end{aligned} \tag{29}$$

respectively.

A complete picture is otained when $4\pi$-periodic solutions are also considered. Such solutions can be represented by

$$g = \sum_{k=1,3,5,\ldots}^{N} \underset{\sim}{A}_k^* \cos \tfrac{1}{2} k\theta + \underset{\sim}{B}_k^* \sin \tfrac{1}{2} k\theta \tag{30}$$

The same procedure following (22) is employed resulting in an algebraic system for $\underset{\sim}{A}_k^*$, $\underset{\sim}{B}_k^*$; k = 1,3,5, ... . The first four vector equations

have a special form and thereafter may be given by general formulas. These are:

$$8m_1 \underset{\sim}{A}_1 + (4\underset{\sim}{D} - \lambda\underset{\sim}{C})\underset{\sim}{B}_1 + [(4\mu r^2 \underset{\sim}{E} - 9\underset{\sim}{C})\lambda + 4\underset{\sim}{D}]\underset{\sim}{B}_3 + 4\mu r^2 \lambda \underset{\sim}{E}\underset{\sim}{B}_5 = 0$$

$$(4\underset{\sim}{D} - \lambda\underset{\sim}{C})\underset{\sim}{A}_1 + 8m_1 \underset{\sim}{B}_1 + [(9\underset{\sim}{C} + 4\mu r^2 \underset{\sim}{E})\lambda - 4\underset{\sim}{D}]\underset{\sim}{A}_3 - 4\mu r^2 \underset{\sim}{E}\underset{\sim}{A}_5 = 0$$

$$[(\underset{\sim}{C} + 4\mu r^2 \underset{\sim}{E})\lambda - 4\underset{\sim}{D}]\underset{\sim}{B}_1 + 8m_1 \underset{\sim}{A}_3 + (4\underset{\sim}{D} - 25\lambda \underset{\sim}{C})\underset{\sim}{B}_5 + 4\mu r^2 \lambda \underset{\sim}{E}\underset{\sim}{B}_7 = 0$$

$$[(4\underset{\sim}{D} - \lambda\underset{\sim}{C}) + 4\mu r^2 \lambda \underset{\sim}{E}]\underset{\sim}{A}_1 + 8m_1 \underset{\sim}{B}_3 + (25\lambda \underset{\sim}{C} - 4\underset{\sim}{D})\underset{\sim}{A}_5 - 4\mu r^2 \lambda \underset{\sim}{E}\underset{\sim}{A}_7 = 0$$

$$- 4\mu r^2 \lambda \underset{\sim}{E}\underset{\sim}{B}_{k-4} + [\lambda(k-2)^2 \underset{\sim}{C} - 4\underset{\sim}{D}]\underset{\sim}{B}_{k-2} + 8m_1 \underset{\sim}{A}_k + [4\underset{\sim}{D} - \lambda(k+2)^2 \underset{\sim}{C}]\underset{\sim}{B}_{k-2} + 4\mu r^2 \lambda \underset{\sim}{E}\underset{\sim}{B}_{k+4} = 0, \quad k = 5,7,9, \ldots \tag{31}$$

$$4\mu r^2 \lambda \underset{\sim}{E}\underset{\sim}{A}_{k-4} + [4\underset{\sim}{D} - \lambda(k-2)^2 \underset{\sim}{C}]\underset{\sim}{A}_{k-2} + 8m_1 \underset{\sim}{B}_k + [(k+2)^2 \lambda \underset{\sim}{C} - 4\underset{\sim}{D}]\underset{\sim}{A}_{k+2} - 4\mu r^2 \lambda \underset{\sim}{E}\underset{\sim}{A}_{k+4} = 0, \quad k = 5,7,9, \ldots$$

Corresponding to a subsystem consisting of only $A_{11}$, $B_{11}$ $A_{31}$, $B_{31}$ with all others being zero, one obtains the characteristic equation

$$64\, m_1^2 - (4D_{11} - 9\lambda C_{11})(4D_{11} - \lambda C_{11}) = \pm 8m_1(4D_{11} - \lambda C_{11}) \tag{31}$$

Fig. 3 shows the results of the analysis in the $\lambda$–$m_1$ plane. These results are based on formulas (28)–(29) and (31). Boundaries of the regions of instability, shown as shaded, are solid lines when based upon $2\pi$-periodic solution and dashed lines when based upon $4\pi$-periodic solutions. The values of $\mu = 0.4$ and $r = 100$ were used.

Acknowledgement

The investigations reported here were supported by the U.S. Army Research Office under grant to the Rensselaer Polytechnic Institute. The author wishes to acknowledge the assistance of his graduate student, C. Younis, who carried out the numerical calculations and checked all the manipulations.

References

1. S. Lubkin and J.J. Stoker, "Stability of Columns and Strings under Periodically Varying Forces," Quart. Appl. Math., 1, 3 (1943) 215-236.

2. V.V. Bolotin, The Dynamic Stability of Elstic Systems, Holden-Day, 1964, pp. 216-220.

3. M. Badlani and W. Kleinhenz, "Dynamic Stability of Elastic Mechanisms," J. Mech. Des. Trans. ASME 101, 1, (Jan. 1979) 149-153.

4. P.W. Jasinski, H.C. Lee and G.N. Sandor, "Vibrations of Elastic Connecting Rod of a High Speed Slider-Crank Mechanism," J. Engineering for Industry, Trans. ASME, 1971, pp. 636-644.

5. S. Kalayeioglu and C. Bagci, "Determination of the Critical Operating Speeds of Planar Mechanisms by the Finite Element Method Using Planar Actual Line Elements and Lumped Mass Systems," ASME Journal of Mechanical Design, 101, 2 (1979) 210-223.

6. E.A. Coddington and N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill 1955, pp. 78-81.

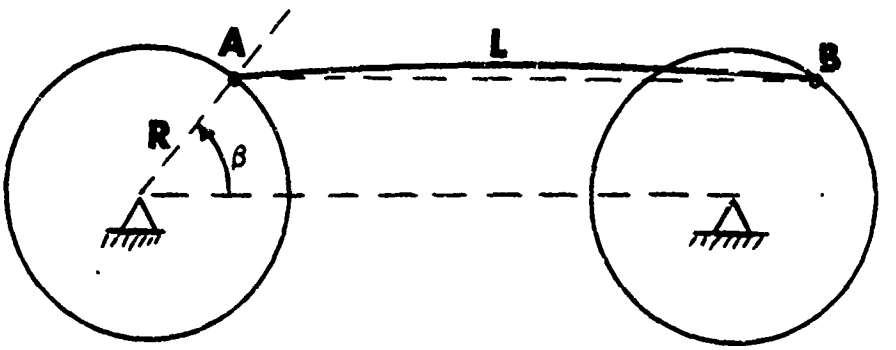



Fig. 1

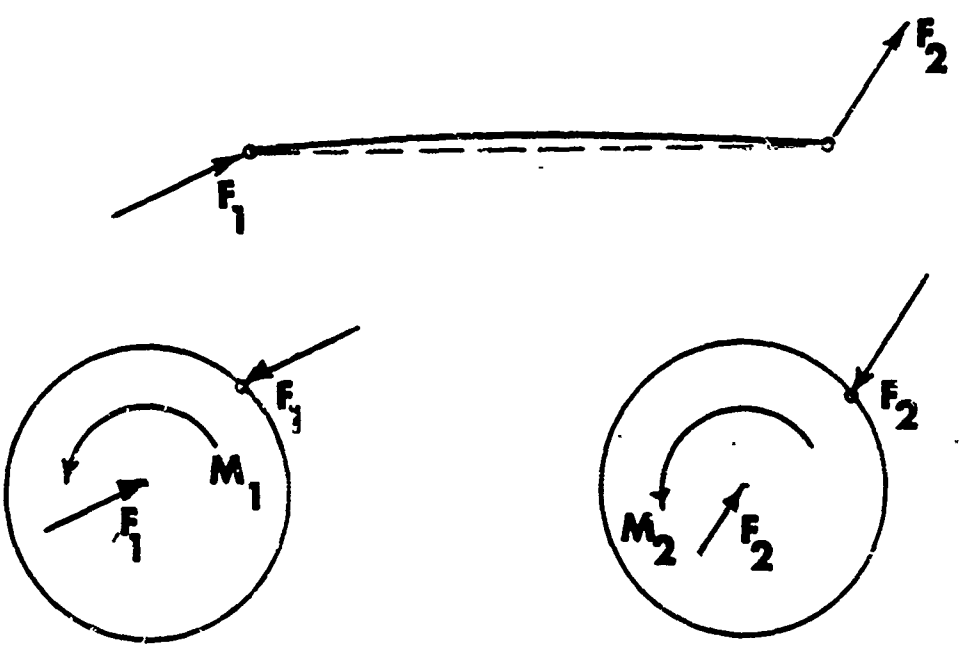



Fig. 2

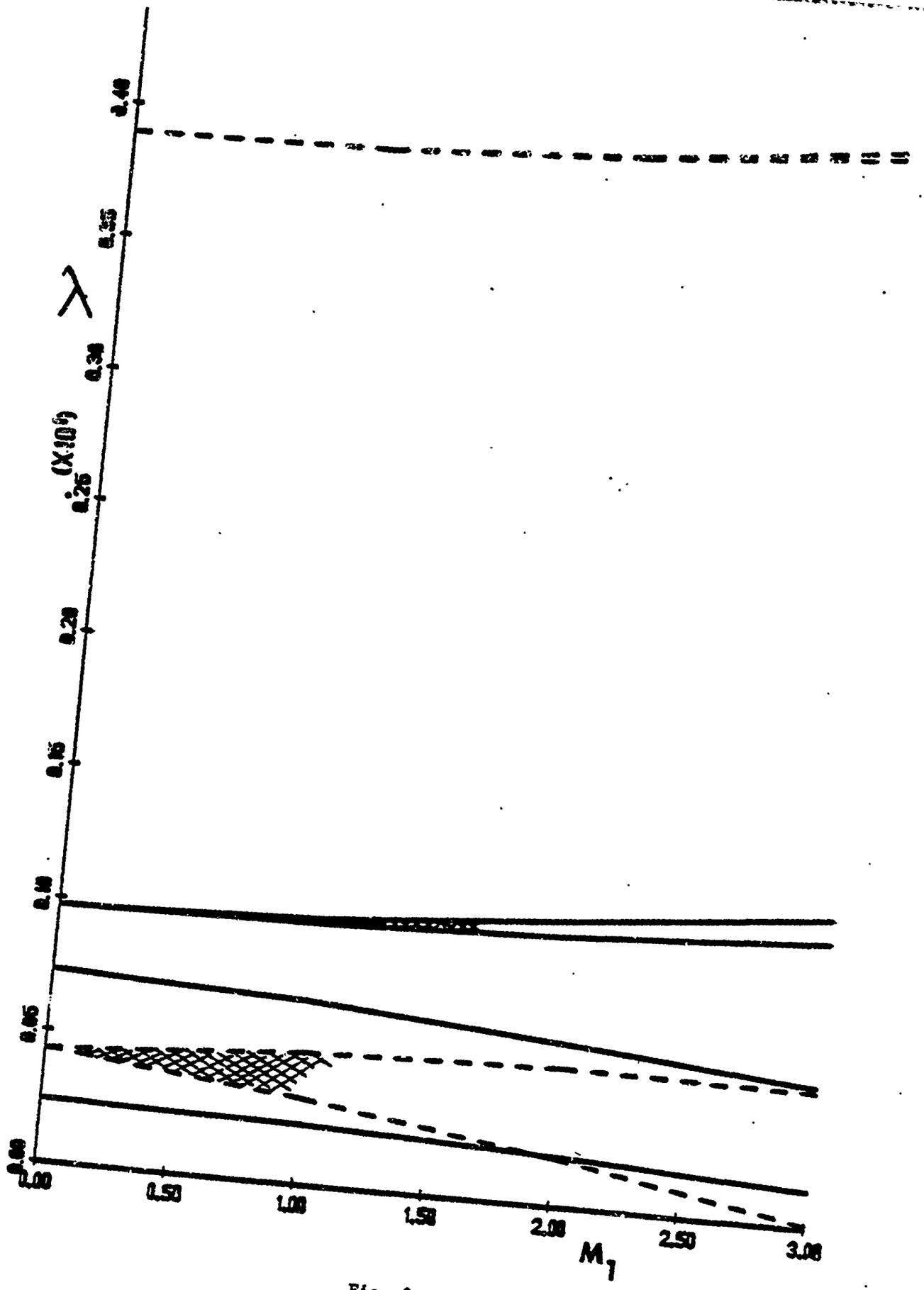

λ
(X10⁶)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.00
0.50
1.00
1.50
2.00
2.50
3.00
M₁

Fig. 3

# AN INTRODUCTION TO THE SCIENTIFIC COMPUTING LANGUAGE PASCAL-SC

L. B. Rall
Mathematics Research Center
University of Wisconsin-Madison

ABSTRACT. Microcomputers are now being widely used for small-scale scientific, engineering, and statistical computing. Pascal-SC (Pascal for Scientific Computing) is a language which has been developed specifically for this application. Its most important features are: (i) accurate floating-point arithmetic for real, complex, and interval numbers, vectors, and matrices, with controlled rounding if desired; (ii) the convenience of operator notation for numerical data types, which makes programs easier to write, read, and document, together with the ability to accept user-defined operators for nonstandard data types; and (iii) compatibility with ordinary Pascal, so that Pascal programming techniques and programs already written in Pascal can be used immediately. In Pascal-SC, solutions of linear systems of equations, inverses of matrices, and eigenvalues and eigenvectors are computed with guaranteed error bounds, and scalar products of vectors and sums of arbitrary length of floating-point numbers are computed to the closest floating-point number, or rounded as desired. These basic features of Pascal-SC will be described, together with applications to research on numerical methods which have been carried out on a microcomputer.

1. OBJECTIVE. A brief description of the language Pascal-SC (Pascal for Scientific Computation) will be given to explain features of this extension of ordinary Pascal [10] which are particularly useful for scientific, engineering, and statistical calculations, particularly on microcomputers. The reader is assumed to be familiar with Pascal programming on at least an introductory level, and to have had some experience with numerical computation of the kind which arises in scientific and engineering problems. With this background, it should be easy to appreciate the advantages of the additional features of Pascal-SC.

In order to keep the discussion short and to the point, many details will be omitted, and a formal description of the language will not be given. For operational and programming details, the reader should consult [3], [19], [28]; formalities are given in [5]. Here, simple examples will be used to illustrate ideas as they are introduced.

2. WHY PASCAL-SC? As pointed out by Wirth [10], the introduction of a new computer language requires careful justification. The same applies to an extension or modification of an existing language, particularly a language which is as successful and widely used as Pascal. The most important additional features of Pascal-SC are:

Sponsored by the U. S. Army under Contract No. DAAG29-80-C-0041.

(i) **Accurate floating-point arithmetic with controllable rounding;**

(ii) **User-defined operators to facilitate programming and documentation.**

Furthermore, it is of considerable importance to note that:

(iii) **Pascal-SC retains the features of ordinary Pascal.**

Thus, none of the investment in learning to program in Pascal or in programs already written in Pascal is lost in going from Pascal to Pascal-SC. The Pascal-SC compiler can translate programs written in ordinary Pascal. Moreover, programming in Pascal-SC will come very naturally to the Pascal programmer, as will be seen from the examples given below.

With regard to (i), the floating-point arithmetic provided by Pascal-SC is implemented not only for real floating-point numbers (type REAL), but also for complex numbers, intervals, and vectors and matrices of these types. This allows convenient and accurate computation with the kinds of numerical data most frequently encountered in scientific and engineering problems. The floating-point arithmetic of Pascal-SC [28] is constructed on the basis of the theory of computer arithmetic given by Kulisch and Miranker [14], which guarantees accuracy, controllability and reliability of the results. In order to keep the compiler small enough to be convenient to use on microcomputers and still provide these additional features, extensive use is made of external libraries of pretranslated code which the compiler can link to the user's program, or source code which can be compiled as part of it.

The second important additional capability of Pascal-SC is that it allows user-defined operators to permit the manipulation of nonstandard data types in ordinary mathematical notation. For example, if A is a matrix, and x, b, c are vectors, then the programmer can write the statement

(2.1) `c := A*x + b;`

in Pascal-SC to perform the indicated calculations. This notation follows ordinary mathematical usage, and thus has the advantages of clarity and simplicity compared to the calling of procedures and functions to obtain the same result in ordinary Pascal. In order for the Pascal-SC compiler to accept (2.1), the heading of the program has to contain a definition of the binary operators * to perform matrix by vector multiplication and + to perform vector addition. (Source code for these operators is included in the Pascal-SC package for vector and matrix arithmetic.) In addition to this "overloading" of standard operator symbols, Pascal-SC permits the user to give operators arbitrary names (for example, XOR for "exclusive or"), and assign priorities to such operators. One particular convenience of Pascal-SC is that the operator ** can be defined to perform exponentiation, which makes the writing of polynomials and other functions containing powers simpler than in ordinary Pascal. In allowing user-defined operators, Pascal-SC is similar to Algol 68 and Ada.

These points will be discussed in more detail in the following sections.

3. FLOATING-POINT REAL ARITHMETIC. This is the "built-in" arithmetic of Pascal-SC for floating-point numbers (type REAL). Since this arithmetic is based on the general theory of computer arithmetic given in [14], it is accurate, controllable, and reliable. Before going on to details, a precise statement of the meaning of these terms is necessary because the related

concepts of "accuracy" (the exactness with which results are calculated) and "precision" (the number of digits used in the representation of floating-point numbers) are often confused. For example, the result 32.0 - 31.0 = 1.00 is calculated with low precision (3 decimal digits), but high accuracy (exactly). By contrast, UNVAC 1100 floating-point hardware gives

$$134217728.0 - 134217727.0 = 2.00000000,$$

which is done with higher precision (9 significant digits), but no accuracy, since the correct answer is 1.00000000 [21].

It is possible to discuss the idea of accuracy independently of the particular precision used, since each floating-point system contains only a finite set of numbers. In a given system S, two floating-point numbers u,v with $u < v$ will be said to be **adjacent** if there is no floating-point number w such that $u < w < v$. For $x,y \in S$, the exact result $x \circ y$ of an arithmetic operation $\circ$, where $\circ \in \{+, -, *, /\}$, will either be a floating-point number, or a real number w such that $u < w < v$, where u,v are adjacent floating-point numbers. In order to produce a floating-point number in the latter case, w is "rounded" to an element of S, which should be either u or v. This is the basic requirement for **reliability** of a floating-point arithmetic operation. The floating-point software supplied for some microcomputers does not meet this simple requirement, and neither does the floating-point hardware of some mainframe computers. This unhealthy situation has lead the IEEE to undertake the promulgation of standards for floating-point arithmetic (see the SIGNUM Newsletter for October, 1979).

An **accurate** rounding R of the floating-point operation $\circ$ selects $R(x \circ y)$ equal to u or v to minimize the **roundoff error**

$$\varepsilon(x \circ y) = |R(x \circ y) - (x \circ y)|. \tag{3.1}$$

This is the **best possible answer (BPA)** rounding [29]. In case of a tie between u and v, a suitable rule is invoked. In Pascal-SC, this rounding is to the one of u,v which is furthest from zero, in order to satisfy the condition of **antisymmetry** of R, $R(-x) = -R(x)$, which is required by the general theory [14].

The distance $|u - v|$ between u and v will depend on the precision of the floating-point numbers being used; this determines the maximum roundoff error of a reliable calculation. Of course, if the arithmetic of the machine being used is not reliable, then roundoff error is not related in a simple way to precision, and the attempt to "buy" more accuracy by using increased precision can be futile, as well as expensive.

There are other ways in which rounding to a reliable result can be carried out, including:

(i) x is rounded upward to v;

(ii) x is rounded downward to u.

The **directed** roundings (i) and (ii) are necessary to support interval arithmetic [17], [16], [14], among other things. Rounding in a floating-point arithmetic is said to be **controlled** if the user can choose the method desired for the result of a given floating-point arithmetic operation. Pascal-SC gives the user the choice of BPA and the upward and downward directed roundings, which means that a total of twelve operators are provided for the four basic

arithmetic operations +, -, *, /, and the three roundings listed above:

```
                +   -   *   /          { BPA rounding }
(3.2)           +>  ->  *>  />         { Upward rounding }
                +<  -<  *<  /<         { Downward rounding }
```

Thus, the Pascal-SC programmer can control the direction of rounding if desired, for example, to obtain guaranteed lower or upper bounds for the values of arithmetic expressions [6]. It also follows from the reliability of Pascal-SC arithmetic that the addition and multiplication operators (3.2) are commutative, which is not true for the kind of floating-point arithmetic ordinarily encountered.

For the microcomputer implementations of Pascal-SC, **decimal** arithmetic is used, and the precision of floating-point numbers is twelve decimal digits in scientific notation, with an exponent range in powers of 10 from -99 to +99. The smallest and largest positive numbers are thus MINREAL = 1.00000000000E-99 (= $10^{-99}$) and MAXREAL = 9.99999999999E+99, respectively. Zero is represented by 0 = 0.00000000000E+00, as usual [28]. The use of decimal arithmetic avoids the errors introduced by conversion between binary and decimal upon input and output. Decimal values are represented internally by two BCD digits per byte. The precision of the Pascal-SC floating-point number system for microcomputers is thus adequate for the representation of most numerical quantities of interest in scientific and engineering computation. Furthermore, since Pascal-SC floating-point arithmetic is reliable, accurate, and controllable, there is seldom any need for more than 12 decimal digits of precision in a given computation.

To illustrate the features of Pascal-SC arithmetic, consider the product

(3.3) $$A = (13.4565432278)(0.00045378239245).$$

The **exact** value of A is not a 12-digit floating-point number, so rounding will take place in the corresponding floating-point multiplication operations. The floating-point operation * gives

(3.4) B = 1.34565432278E+01 * 4.53782392145E-04 = 6.10634237591E-03,

as the BPA for A, which is in error by at most $\frac{\delta}{2} = 5.0\times10^{-15}$, since Pascal-SC arithmetic is reliable and the distance between B and its two neighboring floating-point numbers is $\delta = 10^{-14}$. More precisely, (3.4) establishes that A belongs to the half-open interval $[B - \frac{\delta}{2}, B + \frac{\delta}{2})$, since ties are rounded away from 0. The operations *< and *> with directed rounding give

```
        C = 1.34565432278E+01 *< 4.53782392145E-04 = 6.10634237591E-03,
(3.5)
        D = 1.34565432278E+01 *> 4.53782392145E-04 = 6.10634237592E-03,
```

respectively. Since C = B, the result (3.5) shows that the exact answer A belongs to the interval [C,D] = [B,D] = [B, B + δ]; therefore, on the basis of

(3.4), A belongs to the half-open interval $[B - \frac{\delta}{2}, B + \frac{\delta}{2}) \cap [B, B + \delta] = [B, B + \frac{\delta}{2})$ (see Figure 3.1). Thus, this calculation proves that $A \geq B$ and $A - B < 5.0 \times 10^{-15}$. This gives a more accurate location for A than the BPA answer (3.4).

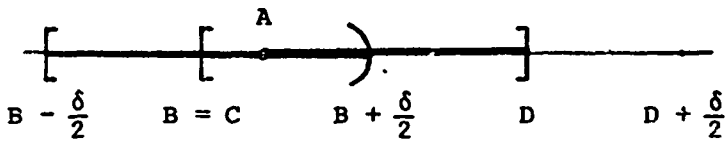



Figure 3.1. A Detail of the REAL Floating-Point Screen.

In addition to the twelve rounded arithmetic operations listed above, Pascal-SC provides the most frequently used **standard functions**, computed to BPA accuracy [28]. In order to keep the microcomputer version of the Pascal-SC system small, extensive use is made of external libraries, so the compiler will bring in code only for functions, procedures, and operators actually needed by the program.

5. FLOATING-POINT COMPLEX ARITHMETIC. In Pascal-SC, manipulation of complex numbers is accomplished by subroutines for operators, functions, and procedures which are stored in external libraries. For this reason, the declaration of complex numbers has the stereotyped form

```
TYPE COMPLEX = RECORD RE,IM: REAL END;
```

Thus, the representation of a complex number Z in Pascal-SC is in Cartesian coordinates. For input and output of Z, the standard format is (Z.RE, Z.IM).

In ordinary Pascal, addition and other arithmetic operations with complex numbers have to be done by procedures and functions. In Pascal-SC, however, operator overloading simplifies notation in the program considerably. To illustrate this, consider as a simple example the source code to enable the operator + to add complex numbers:

```
OPERATOR + (A,B: COMPLEX) RES: COMPLEX;
   VAR U: COMPLEX;
   BEGIN
     U.RE := A.RE + B.RE;
     U.IM := A.IM + B.IM;
    RES := U
   END;
```

The actual coding for this operator is essentially the same as for a function or procedure for the same purpose. However, to add the two complex numbers V,W and assign the result to the complex number Z, one writes only

```
(5.1)                    Z := V + W;
```

in the subsequent program. Addition operators have to be defined for all pairs of operands of types INTEGER, REAL, and COMPLEX, since these would occur naturally in expressions being evaluated. If K, R, C denote generic variables of types INTEGER, REAL, and COMPLEX, respectively, then six addition operators are needed:

(5.2) +C, K + C, C + K, R + C, C + R, C + C.

Similarly, six subtraction operators

(5.3) -C, K - C, C - K, R - C, C - R, C - C,

are required, as well as five multiplication and five division operators:

(5.4)
K*C, C*K, R*C, C*R, C*C,

K/C, C/K, R/C, C/R, C/C.

All 22 of the operators (5.2)-(5.4) are provided in an external library in the form of pretranslated code, and the corresponding declarations, for example,

```
OPERATOR + (RA: REAL; B: COMPLEX) RES: COMPLEX;
  EXTERNAL 155;
```

are available to the programmer in an external text file. Actual coding is also simplified considerably the fact that the programmer can direct the compiler to refer to external libraries for type declarations and definitions of operators and other needed functions and procedures [19], [28]. The use of this feature makes the source code for a Pascal-SC program more compact and readable. Examples of programs using such directives are given in Appendices A and B.

All complex floating-point operations in Pascal-SC calculate the real and imaginary parts of the result to BPA accuracy. For addition and subtraction, operators similar to the one given above for complex addition are satisfactory; however, multiplication and division require special algorithms to attain this accuracy [14], [28]. For example, consider the function CDIV which does complex division by the usual formula:

(5.5)
```
FUNCTION CDIV(A,B : COMPLEX) : COMPLEX;
   VAR DENOM: REAL;
           U: COMPLEX;
   BEGIN
     DENOM := B.RE * B.RE + B.IM * B.IM;
      U.RE := (A.RE * B.RE + A.IM * B.IM) / DENOM;
      U.IM := (A.IM * B.RE - A.RE * B.IM) / DENOM;
      CDIV := U
   END;
```

The Pascal-SC operator for complex division,

(5.6)
```
OPERATOR / (A,B: COMPLEX) RES: COMPLEX;
  EXTERNAL 182;
```

has a number of advantages over the function CDIV. Among these are:

a. Accuracy. The results of Pascal-SC complex operations are calculated to BPA accuracy in the sense that their real and imaginary parts are given to BPA accuracy. Ordinary Pascal functions and procedures for COMPLEX multiplication or division, such as CDIV given above, cannot attain this accuracy because of the number of roundoff errors which occur. In fact, catastrophic cancellations can happen which render the results almost meaningless when the ordinary formulas in (5.5) are used. For example, for

```
                V = (1.23456789, 1.23456789),
(5.7)
                W = (1.0000123E-05, 1.0000321E-05),
```

one gets

```
                V/W = (1.23454048308E+05, -1.22216794612E+00),
(5.8)
          CDIV(V,W) = (1.23454048308E+05, -1.22216773503E+00),
```

where the incorrect digits of CDIV(U,V) are indicated in boldface. Even in this fairly harmless-looking case, the algorithm (5.5) has lost **five** significant digits in the imaginary part of the quotient in a single division, while all the digits of the Pascal-SC result for V/W are correct. To be sure, roundoff error will also increase with repeated use of the Pascal-SC division operator (5.6), but at a slower and more predictable rate.

b. Deferred overflow. In the Pascal-SC algorithms for complex multiplication and division, overflow does not result unless the real or imaginary part of the result is > MAXREAL in absolute value, whereas overflow can occur in (5.6) in the calculation of the intermediate values DENOM, U.RE, U.IM, even though the actual result has real and imaginary parts which are representable by floating-point numbers. For example, for

```
(5.9)          V = (3.0E+99, -1.0E+99),  W = (1.0E+99, -1.0E+99),
```

Pascal-SC complex division gives the result

```
(5.10)           V/W = (2.00000000000E+00, 1.00000000000E+00),
```

while the subroutine (5.6) for CDIV overflows when trying to compute DENOM.

c. Ease of use. For VAR U,V,W: COMPLEX, the use of (5.6) allows one to write

```
(5.11)                U := V / W;
```

in the source code for the program instead of

```
(5.12)                U := CDIV(V,W);
```

as in ordinary Pascal. In the case of complicated expressions involving complex numbers, the gain in programming ease using ordinary mathematical notation with operators instead of function and procedure calls is significant. The source code is more likely to be correct in the first place, and also will be easier to document and read later.

d. Compilation time. The function CDIV has to be compiled from the source code (5.5) for each ordinary Pascal program which uses complex division. The operator (5.6), on the other hand, is given by pretranslated code which is automatically linked to the user's program in the last stage of the compilation. This saves a considerable amount of compilation time.

The accurate complex divison in Pascal-SC turns out to be slower than the function CDIV. According to the table given in [4], p. 267, a typical time for the complex division U/V is 100 milliseconds for a 2.5MHz Z80 processor. The

function CDIV(U,V) uses six real multiplications, two real divisions, and three real addition/subtractions. The typical times for these operations given in [4] total 60.4 milliseconds. It will be seen later that some Pascal-SC operations are actually **faster** than their inaccurate real simulations; however, in the case of complex multiplication and division, one pays a little for guaranteed, reliable accuracy.

In addition to the arithmetic operators +, -, *, / for type COMPLEX, a number of additional operators, functions, and procedures are provided in the Pascal-SC complex library for convenience. For details on these, including the domains and ranges of the standard functions, see [28].

Rounded complex operations +<, +>, -<, ->, *<, *>, /<, /> are also included in an external library in the form of pretranslated code [28]. Here, rounding is carried out componentwise. Each complex number z = (x,y) with $|x|, |y| <$ MAXREAL will belong to a rectangle with corners which are the floating-point complex numbers A = (u,v), B = (u + δ, v), C = (u + δ, v + η), D = (u, v + η), and which contains no other floating-point complex numbers (see Figure 5.1). The result of rounding z downward will be $\nabla z$ = A = (u,v), while z is rounded upward to $\Delta z$ = C = (u + δ, v + η), where δ and η are the spacings in the floating-point screen in the horizontal and vertical directions in the complex plane, respectively. The BPA rounding of z will be (BPA(x), BPA(y)), and thus could be any one of the four points A,B,C,D. For example, suppose

(5.13) $$z = (100 - 4i)/(565 + 789i),$$

which is not a complex floating-point number. Pascal-SC operations give

(5.14)
A = (100,-4)/<(565,789) = (5.66437234668E-02,-8.61803501157E-02),
C = (100,-4)/>(565,789) = (5.66437234669E-02,-8.61803501156E-02),

while the BPA for (5.13) is

(5.15) D = (100,-4)/(565,789) = (5.66437234668E-02,-8.61803501156E-02).

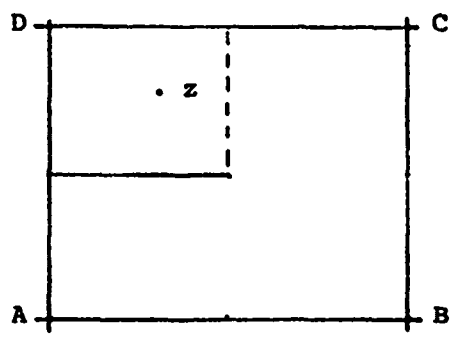



Figure 5.1. A Detail of the COMPLEX Floating-Point Screen.

In this case, $\delta = \eta = 1.0 \times 10^{-13}$, and the results (5.14) and (5.15) locate z in the rectangular complex interval A + $[0, 5.0\times10^{-14})$ + i$[5.0\times10^{-14}, 1.0\times10^{-13}]$, according to the rules for Pascal-SC rounding (see Figure 5.1).

Complex arithmetic with directed rounding can be used as the basis for complex interval arithmetic [14].

6. FLOATING-POINT INTERVAL ARITHMETIC. Interval arithmetic [1] [17], [16] is based on the use of closed, finite intervals [a,b] of real numbers as its basic elements. Interval arithmetic has a number of significant applications in scientific, engineering, and statistical computation; however, its use has not been widespread up to now because of the limitations of conventional computer arithmetic units [18] and ordinary programming languages. In Pascal-SC, interval arithmetic has been implemented efficiently, and is just as convenient to use as real or complex arithmetic.

In the general theory of computer arithmetic [14], interval arithmetic is regarded as a special case of arithmetic on subsets of real numbers. Here, if X, Y are subsets of R, and $o \in \{+, -, *, /\}$, then

$$Z = X \circ Y = \{xoy \mid x \in X, y \in Y\}, \tag{6.1}$$

by definition. For division, of course, $0 \notin Y$ is excluded. If X = [a,b] and Y = [c,d] are intervals, then Z = [r,s] is an interval if defined, and the endpoints r,s of Z can be calculated from the endpoints of X,Y [14], [17], [16]. It is assumed, of course, that the intervals X = [a,b] considered are finite and **proper**, that is, $a \leq b$.

The most basic application of interval arithmetic is the following: If the value of a function, for example,

$$w = 9x^4 - y^4 + 2y^2, \tag{6.2}$$

is calculated in interval arithmetic for intervals X,Y, then the resulting interval W will contain **all** values w of the function (6.2) for **all** values of $x \in X$, $y \in Y$. This property of interval arithmetic allows one to bound the ranges of functions without detailed analysis of maxima and minima. Since rounding of interval operations is outward to the smallest floating-point interval which contains the exact results [28], automatic bounds for round-off error can be obtained conveniently, by taking the input intervals X,Y to be single points, that is, X = [x,x], Y = [y,y], [1], [17], [16]. For a simple application of this principle, consider the evaluation of (6.2) by the expressions

```
(6.3)   (a)   w := 9*x*x*x*x - y*y*y*y + 2*y*y;
        (b)   w := 9*(x**4) - y**4 + 2*(y**2);
        (c)   w := (3*(x**2) - y**2)*(3*(x**2) + y**2) + 2*(y**2);
```

in REAL arithmetic, and

```
(6.4)   (a)   W := 9*X*X*X*X - Y*Y*Y*Y + 2*Y*Y;
        (b)   W := 9*(X**4) - Y**4 + 2*(Y**2);
        (c)   W := (3*(X**2) - Y**2)*(3*(X**2) + Y**2) + 2*(Y**2);
```

in INTERVAL arithmetic, respectively. The power operators ** in (6.3) and (6.4) are given in Appendix A in the source code for the program WEVAL which was used to calculate the following results.

For x = 10864, y = 18817, the statements (6.3) give

For x = 10864, y = 18817, the statements (6.3) give

$$
\begin{aligned}
&\text{(a)} \quad w = \ \ 1.58978 \times 10^{5}, \quad W = [-1.841022 \times 10^{6},\ 1.158978 \times 10^{6}],\\
(6.5)\quad &\text{(b)} \quad w = -8.41022 \times 10^{5}, \quad W = [-8.410220 \times 10^{5},\ 1.158978 \times 10^{6}],\\
&\text{(c)} \quad w = \ \ 1.00000000000, \quad W = [\ 1.00000000000\ ,\ 1.00000000000\ ].
\end{aligned}
$$

The enormous width of W in (6.5)(a) and (6.5)(b) indicates that the values given for w can be subject to large roundoff error, and hence are untrustworthy. On the other hand, evaluation of (6.2) by the statements (6.3)(c) and (6.4)(c), respectively, shows that this formulation is highly accurate, and in fact proves that (6.3)(c) yields the exact value w = 1 of the function (6.2) for the given values of x and y. This example also shows that the way in which arithmetic expressions are written can be crucial for accuracy. The techniques for evaluation of arithmetic expressions with maximum accuracy [6] can be automated, and are incorporated in a Pascal-SC demonstration program [26].

In Pascal-SC, floating-point intervals are declared in the stereotyped way:

```
TYPE INTERVAL = RECORD INF,SUP: REAL END;
```

As in the case of complex numbers, the operators, functions, and procedures in an external library manipulate intervals declared in this form. In order to preserve the inclusion property of interval arithmetic, all rounding in floating-point INTERVAL arithmetic is outward: The result given for XoY is the **smallest** floating-point interval Z which contains the actual result. This can be implemented by the use of directed rounding, for example, addition can be performed by the operator

(6.6)
```
OPERATOR + (A,B: INTERVAL) RES: INTERVAL;
  VAR C: INTERVAL;
  BEGIN
    C.INF := A.INF +< B.INF;
    C.SUP := A.SUP +> B.SUP;
    RES := C;
  END;
```

This operator is available as a pretranslated subroutine, declared by

```
OPERATOR + (A,B: INTERVAL) RES: INTERVAL
  EXTERNAL 68;
```

If I denotes a generic variable of type INTERVAL, and K of type INTEGER, the following 15 arithmetic operations are provided for interval arithmetic:

$$
\begin{gathered}
\pm I,\ K \pm I,\ I \pm K,\ I \pm I,\\
K * I,\ I * K,\ I * K,\\
K / I,\ I / K,\ I / I.
\end{gathered}
$$

Because REAL floating-point expressions do not necessarily yield the exact values of their real results, operators between types REAL and INTERVAL are not included, since the resulting interval might not contain the true outcome of a computation. In addition to interval versions of standard functions and various utility procedures [28], the interval library includes some operators which work with intervals as sets of real numbers. There are the "lattice operators"

| | | |
|---|---|---|
| | ** | Intersection, |
| (6.7) | | |
| | ++ | Interval Hull, |

and the relational operators

| | | |
|---|---|---|
| | <= | Subinterval, |
| | >= | Superinterval, |
| (6.8) | | |
| | >< | Disjointness, |
| | IN | Point Inclusion, |

[28]. The intersection operator ** will generate an error interrupt if its operands are disjoint intervals, otherwise, their intersection $I \cap J$ will be computed. If I = [a,b] and J = [c,d], then I ++ J = [min{a,c},max{b,d}] is the smallest interval which contains both I and J. The relation I <= J is TRUE if $I \subset J$, otherwise FALSE; similarly, I >= J is TRUE if $I \supset J$. The result of I >< J is TRUE if I and J are disjoint intervals. This test can be used to avoid I ** J in this case. If R is a floating-point number (type REAL), R IN I is TRUE if $R \in I$ as a real number.

With regard to the efficiency of the implementation of interval arithmetic in Pascal-SC, Bohlender and Grüner [4] give the following typical times in milliseconds for a microcomputer with a 2.5MHz Z80 processor:

| Operation | + | - | * | / |
|---|---|---|---|---|
| REAL | 2.2 | 2.2 | 6.0 | 10.0 |
| INTERVAL | 5.4 | 5.4 | 23.0 | 31.0 |

Considering the fact that each interval operation has to determine two real numbers, and that interval multiplication and division require the calculation of four real products in one case [14], the above indicates an almost optimal implementation in software. By contrast, factors of 100 or 200 between REAL and INTERVAL arithmetic speed have been noted on conventional computers, such as the UNIVAC 1100 series. [18]. In addition, while a REAL computation provides only a floating-point number which approximates the result, an INTERVAL computation on the other hand provides an interval which is guaranteed to contain the true result. An important application of this property of interval calculations will be described below in connection with the solution of linear systems of equations with guaranteed error bounds.

7. REAL FLOATING-POINT VECTOR AND MATRIX ARITHMETIC. Calculations with real vectors and matrices are among the most commonly encountered tasks in scientific, engineering, and statistical computation. Pascal-SC offers the user the same reliability, accuracy, controllability, and convenience when calculating with real n-dimensional floating-point vectors and matrices as it does for the scalar types REAL, COMPLEX, and INTERVAL. This is accomplished with the aid of an external source code library of operators, functions, and procedures for vector and matrix manipulation, and a built-in function SCALP for the calculation of scalar products of floating-point vectors to BPA accuracy or with directed rounding at the option of the user.

7.1. Convenience. In order to illustrate the convenience of Pascal-SC for vector and matrix calculations, suppose that A, B are n×n matrices, and x, y, z are n-dimensional vectors. To evaluate

(7.1) $$z = 5.5ABx + 3y,$$

the corresponding expression in Pascal-SC is

(7.2)
```
z := 5.5*A*B*x + 3*y;
```

which uses ordinary operator notation instead of the function and procedure calls which would be required in ordinary Pascal and most other languages. In order to make use of the software provided in the corresponding external library, a stereotyped declaration of floating-point vector and matrix data types is expected:

```
CONST     DIM = #;  {The actual dimension replaces #}
 TYPE DIMTYPE = 1..DIM;
      RVECTOR = ARRAY [DIMTYPE] OF REAL;
      RMATRIX = ARRAY [DIMTYPE] OF RVECTOR;
```

The operators +, -, *, and the operators +<, +>, -<, ->, *<, *> with directed rounding are available for various permissible combinations of operands, for example, multiplication of an RVECTOR by an INTEGER or REAL, and so on [28].

7.2. Reliability, accuracy, and controllability: The scalar product SCALP. The general theory of computer arithmetic [14] requires that each component of the result of a vector or matrix operation be rounded to the BPA for the actual real result, or downward or upward to the closest neighboring floating-point number if desired. Addition and subtraction of vectors and matrices present no problems from the standpoint of this requirement, since the desired results can be calculated componentwise with the aid of the six REAL arithmetic operators ±, ±<, ±> described in §3. Calculation of the scalar products of vectors, which is an inherent component of matrix and matrix by vector multiplication, is a different matter. Ordinarily, this calculation is simulated by a FOR loop of real operations, such as in the following function:

(7.3)
```
FUNCTION SPROD(A,B: RVECTOR): REAL;
  VAR I: DIMTYPE; S: REAL;
  BEGIN
    S := 0;
    FOR I:=1 TO DIM DO
      S := S + A[I] * B[I];
    SPROD := S
  END;
```

This is an example of what Kulisch calls the "vertical" definition of computer arithmetic [141, [12]. Of course, there is no hope that the result of SPROD will be accurate in general. For this reason, the internal calculations in a function of this kind are often done in higher precision than the external calculation. While this is a often a great help in some cases, it still does not solve the accuracy problem. On the other hand, the Pascal-SC function

(7.4)
```
SCALP(A,B: RVECTOR; ROUND: INTEGER);
```

which is a built-in feature of the compiler, will calculate the value of the **exact** scalar product of A and B to an **adjacent** floating-point number, with rounding downward, upward, or to the BPA controlled by the value of the parameter ROUND [28]. This reliability, accuracy, and controllability is required by the general theory of computer vector and matrix arithmetic [14]. It can be achieved by special algorithms [14], or the provision of a sufficiently long accumulator. In the microcomputer version of Pascal-SC, this "long accumulator" is implemented in software [4], but the same thing can be done in hardware [13], and can be expected to be a feature of future advanced mainframe computers.

In order to allow the accumulation of several scalar products, the parameter ROUND can also inhibit the clearing of the long accumulator before the product is calculated [28]. The corresponding values are given in the following table:

| Rounding | Clear Long Accumulator | Inhibit Clearing |
|---|---|---|
| BPA | ROUND = 0 | ROUND = 4 |
| Downward | ROUND = -1 | ROUND = 3 |
| Upward | ROUND = 1 | ROUND = 5 |

If the long accumulator is not being cleared, other arithmetic operations are not permitted between successive calls of SCALP [28].

The use of SCALP makes it possible to calculate the results of matrix and matrix by vector multiplications to the closest floating-point numbers, or rounded to the closest larger or smaller neighboring values if desired. This reliability, accuracy, and controllability distinguishes Pascal-SC vector and matrix arithmetic from traditional packages.

Some of the important properties of SCALP are illustrated by the following examples.

a. Accuracy. For

$$A = (10^{99},\ 10^{-99},\ -10^{99}),\qquad B = (1,\ 1,\ 1), \tag{7.5}$$

the value of SPROD(A,B) is 0, of course, while SCALP(A,B,0) gives the correct answer $10^{-99}$. Persons who believe that multiple precision will solve all accuracy problems should determine how much precision is required on their machine for SPROD(A,B) to duplicate this result. Once satisfied, they can then try (7.6) below. Although these examples are extreme, examples can be given for which SPROD is highly inaccurate for vectors of the types one can expect to encounter in actual problems.

b. Speed. It turns out that the accurate scalar product function SCALP is **faster** than the corresponding FOR loop in SPROD if DIM > 1. The typical times given in [4] in milliseconds for a 2.5MHz Z80 processor are:

SCALP 8 + 5.5(DIM),

FOR LOOP 1 + 9.6(DIM).

Since all the matrix and matrix by vector multiplication routines are based on SCALP, they can be expected to execute faster than their inaccurate simulations in REAL arithmetic. Furthermore, the current implementation of Pascal-SC

handles access to elements of arrays very efficiently by means of an "array descriptor" [11], from which addresses of elements can be calculated quickly.

c. Deferred overflow. Under ordinary circumstances, SCALP will not indicate an overflow unless the actual real result is outside the range of representable floating-point numbers. For example, for

(7.6) $$A = (10^{99},\ 10^{-99},\ -10^{99}), \qquad B = (10^{99},\ 1,\ 10^{99}),$$

SCALP(A,B,0) will compute the correct result $10^{-99}$, while SPROD(A,B) will overflow for I = 1.

d. Ease of use. The function SCALP is just as easy to use as SPROD, and requires no additional source code in the program, since it is part of the Pascal-SC system. Furthermore, SCALP is more versatile, since its arguments can be arbitrary one-dimensional arrays of floating-point numbers of the same length, of which RVECTOR is only a special case. When called, SCALP consults the array descriptors of its arguments [11] to determine if they are in fact of equal length, and then calculates their scalar product if this is true. Thus, SCALP can be used to calculate scalar products of vectors of various dimensions in the same program.

8. COMPLEX AND INTERVAL FLOATING-POINT VECTOR AND MATRIX ARITHMETIC. The basic ideas here are generally the same as for real vector and matrix arithmetic: Convenience, based on the use of operator notation for expressions, and accuracy. The subroutines in MCLIB and MILIB, respectively, expect the type declarations

```
TYPE CVECTOR = ARRAY [DIMTYPE] OF COMPLEX;
     CMATRIX = ARRAY [DIMTYPE] OF CVECTOR;
```

and

```
TYPE IVECTOR = ARRAY [DIMTYPE] OF INTERVAL;
     IMATRIX = ARRAY [DIMTYPE] OF IVECTOR;
```

in the corresponding cases.

For accurate scalar products, the respective functions

```
FUNCTION CSCALP (VAR A,B: CVECTOR; AKDIM: INTEGER): COMPLEX;
  EXTERNAL 188;
```

and

```
FUNCTION ISCALP (VAR A,B: IVECTOR; AKDIM: INTEGER): INTERVAL;
  EXTERNAL 88;
```

are provided. The functions form the basis of the subroutines for accurate matrix and matrix by vector multiplication. The products are computed from the first AKDIM components of each vector argument. This permits the flexibility of using vectors of various dimensions AKDIM $\leq$ DIM in the same program. CSCALP computes the BPA for the scalar product of complex vectors; directed rounding is not provided for the complex scalar product or complex matrix and matrix by vector multiplications. ISCALP computes the smallest interval which contains the exact result, as in the case of other interval operations [28].

9. SOLUTION OF LINEAR SYSTEMS OF EQUATIONS AND MATRIX INVERSION. These are problems which arise time after time in scientific, engineering, and statistical computation. The Pascal-SC system subroutines for these purposes, which use the accurate scalar product and interval arithmetic, yield results which are far more exact than can be obtained by ordinary floating-point arithmetic. Furthermore, **guaranteed** error bounds are given for results, so the reliability of the computation is immediately determinable. These subroutines will accept either real or interval vectors and matrices.

The basic procedure in the system library LGLLIB for the solution of linear systems of equations is LGLP (an acronym for the German words for "linear equations solution program"). This procedure is declared by

```
PROCEDURE LGLP(DIM,AKDIM: INTEGER; VAR A: RMATRIX, VAR B: RVECTOR;
               VAR Y: IVECTOR);
  EXTERNAL 524;
```

[28]. The purpose of this procedure is to solve the linear system

$$Ax = B \tag{9.1}$$

with coefficient matrix A and right-hand side B. Instead of a floating-point approximation to the solution x, LGLP calculates an interval vector Y which, if proper, contains the **exact** solution x of (9.1), and proves that the floating-point matrix A is a nonsingular real matrix [27]. This allows one to determine not only an approximate value for x, but also guaranteed error bounds for it [23]. The parameter AKDIM in the formal parameter list allows one to solve systems of size smaller than DIM if desired; only the first AKDIM rows and columns of A and components of B are involved in the calculation.

Failure of LGLP to return a proper interval vector Y indicates that A is singular or **extremely** ill-conditioned. In this case, the components of Y will be set equal to the **improper** interval [+1,-1]. A test should be made for this condition immediately on return from LGLP, since all interval subroutines expect proper intervals as data [28].

Thus, LGLP either gives a solution with guaranteed accuracy or an error indication. In practice, LGLP has been observed to succeed for well-known examples of badly conditioned matrices, such as Hilbert matrices [26], [27]. In Appendix B, a simple program to solve linear equations of order up to 20 is given, together with its application to a system of five equations in five unknowns in which the coefficient matrix has a condition number larger than $4 \times 10^{18}$. Inspection of the resulting interval vector Y shows that LGLP was able to solve this system to an accuracy of one unit in the twelfth significant digit. The same 5 × 5 system defeated a standard linear equation solver on a VAX 11/780.

When one is solving several systems with the same matrix but different right sides, considerable time can be saved if the approximate LU-decomposition of A and other preliminary calculations are only done once. The procedure LGLPR is available for this purpose. Its declaration is

```
PROCEDURE LGLPR (DIM,AKDIM: INTEGER; VAR A: RMATRIX; VAR B: RVECTOR;
                 NRS: BOOLEAN; VAR R: RMATRIX; VAR MB: IMATRIX;
                 VAR Y: RVECTOR);
  EXTERNAL 522;
```

[28]. The first call of this procedure is with NRS = FALSE. Matrices needed to process subsequent right sides will be computed and stored as R and MB. Subsequently, LGLPR is called with NRS = TRUE for each new right side.

Similar procedures LGLI and LGLIR are available for the case that the components of the coefficient matrix A and right side B are intervals, that is, A is of type IMATRIX and B is of type IVECTOR. The interval vector Y in this case bounds all solutions of real systems with coefficient matrices belonging to A and right sides belonging to B. This can be helpful in case where the data are subject to uncertainty.

For matrix inversion, the subroutine

```
PROCEDURE INVP (DIM,AKDIM: INTEGER; VAR A: RMATRIX; VAR C: IMATRIX);
  EXTERNAL 526;
```

will, if successful, compute an interval matrix C which contains the inverse of the real (point) matrix A. Singularity or extreme ill-condition of A is reported in the same way as for LGLP, while successful calculation of C proves that A is nonsingular, as before. Finally,

```
PRODCEDURE INVI (DIM,AKDIM: INTEGER; VAR A,C: IMATRIX);
  EXTERNAL 527;
```

is used for inversion of interval-valued matrices. If C is computed successfully as an IMATRIX of proper intervals, then C contains the inverses of all real matrices contained in the interval matrix A. Return of C with all components equal to the improper interval [+1,-1] indicates that A contains at least one singular or very badly conditioned real matrix.

On a microcomputer with 64Kb of storage, LGLP and LGLPR are limited to about DIM = 25 or less, LGLI, LGLIR, and INVP to DIM = 20, and INVI to DIM = 15. For these relatively small systems, the time required for execution seems to be reasonable. As in the case of the rest of the Pascal-SC system, source code for declarations and pretranslated code for the above procedures can be found in an external library.

10. EIGENVALUES AND EIGENVECTORS. The Pascal-SC system provides the standard subroutine

```
PROCEDURE EIGEN (DIM,AKDIM: INTEGER; VAR A: RMATRIX; LAMBDA: REAL;
                 VAR X: RVECTOR; VAR ILAMBDA: INTERVAL; VAR Y: IVECTOR);
  EXTERNAL 534;
```

for the calculation of guaranteed interval bounds for real eigenvalues and vectors of real matrices A, in particular, symmetric matrices. This is another type of calculation which occurs often in engineering and other scientific computation. In addition to the actual dimension AKDIM and the matrix A, EIGEN expects floating-point approximations LAMBDA and X to the eigenvalue and eigenvector of interest, or at least values with which to start the calculation. If successful, the interval value ILAMBDA and interval vector Y returned include an exact real eigenvalue and eigenvector of the floating-point matrix A, and furthermore guarantee that the included eigenvalue is of multiplicity one. Hence, EIGEN will not succeed for multiple eigenvalues [s1], [28]. In case of failure, EIGEN will return improper intervals for ILAMBDA and the components of Y.

11. THE ACCURATE SUM OF n FLOATING-POINT NUMBERS. Statistical calculations, in particular, often require the computation of the sum of n floating-point numbers and perhaps also their squares,

$$(11.1) \qquad S = \sum_{i=1}^{n} a_i , \quad T = \sum_{i=1}^{n} a_i^2 .$$

In Pascal-SC, it is possible to compute the BPA for S and T, or round the result upward or downward to the closest floating point number by taking the $a_i$ as components of an RVECTOR A and using SCALP. For E = (1,1,1,...,1), one has

(11.2) S = SCALP(A,E,ROUND); and T = SCALP(A,A,ROUND);

with the desired best-possible result. However, in the case of the sum S, the standard Pascal-SC subroutine

```
FUNCTION SUM (VAR A: RVECTOR; AKLIM,ROUND: INTEGER): REAL;
  EXTERNAL 480;
```

is also provided. This function performs the addition of the first AKDIM elements of A to the BPA or result of directed rounding of the BPA. SUM, like SCALP, uses the long accumulator, and the values of ROUND have the same significance as given in §7 for SCALP. It is thus possible to call SUM again without clearing the long accumulator, to allow independent accumulation of partial sums without loss of accuracy. However, no other arithmetic operations are allowed between successive calls to SUM [28].

When computing sums of interval numbers, one can use

(11.3) IS := ISCALP(IA,IE,DIM);

where IE has components all equal to [1,1]. For the interval sum of squares, however, it is preferable to form the interval vector IQA with components equal to ISQR(A[I]) for I = 1..DIM, and then compute

(11.4) IT := ISCALP(IQA,IE,DIM);

rather than ISCALP(IA,IA,DIM), for the reason given in §6 about the preferability of ISQR(X) to X*X for intervals.

12. PROGRAMMING IN PASCAL-SC. The only new techniques in Pascal-SC for a Pascal programmer to acquire are the definition and use of operators. Except for these, there is no difference between Pascal and Pascal-SC programming. Therefore, the discussion here will focus on the operator concept [19]. The definition of an operator subroutine is headed by

```
OPERATOR <name> (<formal parameter list>) <result name>: <result type>;
```

The code following this heading is the same as for a function having the same purpose. The result must be assigned to <result name> in its entirety before leaving the subroutine. For example, if <result name> = RES is of type INTERVAL, one must calculate an interval U and make the assignment RES := U; before leaving the subroutine, rather than calculating RES.INF and RES.SUP separately. The formal parameter list consists of one or two identifiers and their types. Thus, operators are either unary or binary. Since operators occur

in expression strings ("infix" notation), their arguments have to be of expression type. That is, VAR A, etc., is not allowed in the formal parameter list. The examples of operators given in §5 and §6 and below can be used as models.

There are two ways to name an operator in Pascal-SC:

(i) By redefining ("overloading") one of the standard Pascal-SC operator symbols for a new data type or types.

(ii) By use of an arbitrary name selected by the user which conforms to the ordinary rules for identifiers in Pascal [10]. In this case (see below), the priority of the operator also has to be declared.

These two methods will be discussed separately.

12.1. Overloading standard operator symbols. This is the most common method used in scientific and engineering computing to name Pascal-SC operators, since one usually wishes to follow the ordinary mathematical notation encountered in the formulas being used. The standard operator symbols in Pascal-SC are, in order of decreasing priority:

Unary operators:

```
NOT, + (unary), - (unary)
```

Multiplicative (binary) operators:

```
*  /  DIV  MOD  AND  **  *>  *<  />  /<
```

Additive (binary) operators:

```
+  -  +>  +<  ->  -<  ++  OR
```

Relational (binary) operators:

```
=  <>  <=  >=  <  >  IN  ><
```

The fundamental distinction between a unary and a binary OPERATOR is that the formal parameter list for the operator contains exactly one parameter in the first case, and exactly two in the second, and these are the only possibilities. An overloaded operator will have the same priority as its symbol in the table above. In the case of + and -, the parameter list of the operator heading will specify whether they are unary (highest priority) or binary.

One convenience of Pascal-SC that is immediately apparent is that one can define ** to perform exponentiation on whatever numerical types are appropriate for the application at hand. However, this should be done with care. Some good methods are given in [7]. Source code for a simple, "repeated squaring" [22] implementation of

```
OPERATOR ** (R: REAL; K: INTEGER) RES: REAL;
```

to perform $R^K$ for integral powers of floating-point numbers is given in Appendix A. This operator makes it possible to write $x^3$, $x^4$, etc. as x**3, x**4, etc. in expressions to be evaluated, which is a more convenient way to represent these

simple powers than by a procedure or function call as in ordinary Pascal. (A function or procedure should be used to compute the result of raising an interval base to an interval power, since the operator ** is used to compute the intersection of INTERVAL variables (see (6.7)).)

The order of the operands in the formal parameter list determines the order in which the operator will be applied. The compiler distinguishes various uses of the same operator symbol by the type(s)of its operand(s), and their order if the operator is binary. Thus, in the same program, "+" can be used to denote addition of complex numbers, intervals, vectors, matrices, quaternions, polynomials, etc., in addition to its standard meaning for integers and floating-point numbers. All that is required is that the appropriate definition of OPERATOR + be given in the heading of the program for each meaning of "+" in the body of the program.

Of course, the compiler recognizes only the rules of arithmetic for user-defined data types which are provided to it by the programmer. For example, if one wishes to use expressions in which variables of both type INTEGER and type GRADIENT [22] appear, both

```
OPERATOR + (K: INTEGER; G: GRADIENT) RES: GRADIENT;
```

and

```
OPERATOR + (G: GRADIENT; K: INTEGER) RES: GRADIENT;
```

must be defined in the heading of the program so that the compiler can produce code for both K + G and G + K. Type GRADIENT consists of the value of a function together with its gradient vector, and is declared by

```
TYPE GRADIENT = RECORD F: REAL; DF: RVECTOR END;
```

in Pascal-SC [22]. In this case, both operators produce the same result, consisting of the alteration of the function value G.F of the GRADIENT variable G to K + G.F = G.F + K, respectively, with no change in the gradient vector G.DF [22]. However, it could happen that the user is working with quantities for which addition is not necessarily commutative. Pascal-SC allows the possibility of defining the result of "+" or any other binary operator to be dependent on the order of the operands.

12.2. Named operators. In Pascal-SC, the user can name operators according to the ordinary Pascal rules for identifiers [10]. For example, the factorial operator (a unary operator) could be called FAC. In this case, FAC 4 would have the value 4! = 24. Note that parentheses are not used unless the operation is applied to an expression. Similarly, a named binary operator is written between its operands ("infix" notation) in the same way as +, -, *, /, etc. The operator FAC can be defined in the program heading as follows:

```
OPERATOR FAC (A: INTEGER) RES: INTEGER;
  BEGIN
    IF A <= 1 THEN RES := 1
              ELSE RES := A * FAC (A - 1);
  END;  { Recursive definition of OPERATOR FAC }
```

In terms of FAC, the binomial coefficient C(N,K) could be computed by use of the statement

```
(12.1)              BINOM := FAC N DIV (FAC K * FAC (N - K));
```

This example is for only for illustration of the construction of a named operator and the possibility of recursion. On the microcomputer implementation of Pascal-SC, INTEGER arithmetic is implemented only for integers I such that $-32768 \leq I \leq 32767$ [28], and thus FAC N can be computed only for $N \leq 7$. Actual computation of factorials should be accomplished by a type conversion to REAL, and controlled by WHILE or UNTIL, in order to avoid stacking recursions too deeply.

Similarly, the binary Boolean operator XOR for "exclusive or" could be defined by

```
OPERATOR XOR (A,B: BOOLEAN) RES: BOOLEAN;
BEGIN
  RES := (A AND NOT B) OR (NOT A AND B);
END;
```

A typical program statement using XOR would be

```
(12.2)        IF OBS1 XOR OBS2 THEN PROB := 0.25 ELSE PROB := 0.75;
```

which would assign the value 0.25 to PROB if just one of OBS1,OBS2 is TRUE, or 0.75 otherwise.

In order for the Pascal-SC compiler to recognize FAC and XOR as the names of operators, and assign priorities to them, a PRIORITY declaration for each named operator must follow directly after the heading line of the program, which gives the name of the program and the list of files used. From highest to lowest priority, these **priority declarations** have the forms

```
PRIORITY <Operator name> = @;  { Unary operators }

PRIORITY <Operator name> = *;  { Multiplicative operators }

PRIORITY <Operator name> = +;  { Additive operators }

PRIORITY <Operator name> = =;  { Relational operators }
```

Thus, suppose one writes a program called CHANCE which uses the operators FAC and XOR to calculate probabilities of outcomes in some stochastic model, and only the standard files INPUT and OUTPUT are used for communication between the program and the outside world. If XOR is to have the same priority as OR, then the first three lines in the heading of the source code for the program would be

```
PROGRAM CHANCE (INPUT,OUTPUT);

  PRIORITY    FAC = @;
              XOR = +;
```

The standard sequence of definitions and declarations would then follow to complete the heading of the program, and then the body of the program consisting of the actual statements to be executed. The structure of a Pascal-SC program therefore differs only slightly from that of an ordinary Pascal program, as shown in the next section.

12.3. Structure of a Pascal-SC program. In Pascal-SC, the order of declarations in the heading is somewhat freer than in standard Pascal [19]. However, the general principle applies that everything must be declared or defined before use. Since overloading standard operator symbols is more common than using named operators, the headings of most Pascal-SC programs will look identical to Pascal programs except for the OPERATOR definitions. Programming is further simplified because Pascal-SC already provides operators for most of numerical data types commonly encountered in scientific and engineering computation, such as complex numbers, intervals, vectors, and matrices, as explained in the previous sections.

The difference between Pascal and Pascal-SC programs is most striking in the statements of the actual body of the program. Here, the power of operator notation makes is possible to write expressions clearly and compactly. This elimination of complicated sequences of function and procedure calls shortens programs and makes the source code much easier to read and understand. This facilitates documentation as well as use of the program.

With just two exceptions, noted below in **boldface** type, the sequence of a Pascal-SC program is identical to an ordinary Pascal program [10]:

```
PROGRAM <Name> ( <List of internal file names> );

   PRIORITY
   LABEL
   CONST
   TYPE
   VAR                  <Declarations and definitions of the program heading>
   PROCEDURE
   FUNCTION
   OPERATOR

 BEGIN
 ..                     <Statements comprising the body of the program>
 END.
```

The order in which procedures, functions, and operators are declared is arbitrary, as in Pascal. The implementation of the operator concept in Pascal-SC is based on the fact that the underlying virtual machine (the so-called KL/P machine) can stack operands of arbitray data types, so that functions with results of arbitrary type can be computed efficiently [11]. This refinement is permits considerable savings in the number of machine instructions actually needed, and hence leads to shorter execution times [11].

13. CONCLUSIONS. The accuracy of Pascal-SC arithmetic and the convenience of operator notation for manipulation of numerical and other data types make this language a valuable tool for scientific, engineering, and statistical computation. In addition to its usefulness for routine problems, such as solution of linear systems of equations, experience has shown that it is possible use this language to program and carry out some rather sophisticated computations, even on a microcomputer. Examples include numerical solution of a

nonlinear integral equation [25], the solution of nonlinear systems of equations by iterative methods [22], [9], and the solution of ordinary differential equations by real and interval Taylor series [8]. In these applications the operator concept of Pascal-SC was used to implement automatic evaluation of derivatives and Taylor series for functions defined by expressions in ordinary mathematical notation [8], [22], [24]. The microcomputer systems used in these investigations can best be described as minimal: Eight-bit machines with Z80 processors, 64Kb of main storage, and two disk drives. The Pascal-SC compiler used was developed by Profs. U. Kulisch and and H.-W. Wippermann and their associates at the Universities of Karlsruhe and Kaiserslautern in Germany, and is described in [3], [19], and [28]. Even these modest resources appear adequate for many of the day-to-day calculations needed by engineers, scientists, and statisticians, as well as for research on methods in numerical analysis which can be applied to larger problems. As "personal computers" grow in size and speed, the accuracy and convenience of Pascal-SC will provide the user with a more powerful tool, and its features will also be advantageous on forthcoming larger machines.

14. ACKNOWLEDGMENTS. The author would like to thank Professor George F. Corliss for reading the manuscript carefully, and a large number of valuable suggestions. Example (6.2) was provided by Prof. Dr. Ulrich Kulisch, and the system of equations solved by the Pascal-SC program in Appendix B was one of several similar systems solved for Dr. S. Takagi of the U. S. Army Cold Regions Research and Engineering Laboratory.

## REFERENCES


1. G. Alefeld and J. Herzberger. Introduction to Interval Computations. Tr. by Jon Rokne. Academic Press, New York, 1983.

2. U. Allendörfer. Gesamte Arithmetic des PASCAL-SC-Rechners. Benutzerhandbuch. Ergänzungen für den Betrieb unter CP/M [Supplement for Operation under CP/M to the User Handbook for the Complete Arithmetic of the Pascal-SC Computer]. Department of Computer Science, University of Kaiserslautern, 1982.

3. U. Allendörfer and R. Kirchner. PASCAL-SC Bedienungsanleitung [Pascal-SC User Guide]. Department of Computer Science, University of Kaiserslautern, 1982.

4. G. Bohlender and K. Grüner. Realization of an optimal computer arithmetic, [ 5], pp. 247-268. Academic Press, New York, 1983.

5. G. Bohlender, K. Grüner, E. Kaucher, R. Klatte, W. Krämer, U. W. Kulisch, S. M. Rump, Ch. Ullrich, J. Wolff v. Gudenberg, and W. L. Miranker. PASCAL-SC: A PASCAL for Contemporary Scientific Computation. Research Report RC 9009, IBM Thomas J. Watson Research Center, Yorktown Heights, N. Y., 1981.

6. H. Böhm. Evaluation of arithmetic expressions with maximum accuracy, [15], pp. 121-137. Academic Press, New York, 1983.

7. W. F. Cody, Jr. and W. Waite. Software Manual for the Elementary Functions. Prentice-Hall, Englewood Cliffs, N. J., 1980.

8. G. Corliss and L. B. Rall. Automatic Generation of Taylor Series in Pascal-SC: Basic Operations and Applications to Ordinary Differential Equations, MRC Technical Summary Report No. 2497, University of Wisconsin-Madison, 1983.

9. A. Cuyt and L. B. Rall. Computational Implementation of the Multivariate Halley Method, MRC Technical Summary Report No. 2481, University of Wisconsin-Madison, 1983.

10. K. Jensen and N. Wirth. Pascal User Manual and Report, 2nd ed. Springer-Verlag, Berlin-Heidelberg-New York, 1978.

11. R. Kirchner. Überblick über die vorliegende Implementierung der Pascal-Spracherweiterung [Overview of the Current Implementation of the Pascal Language Extension]. Department of Computer Science, University of Kaiserslautern, 1981.

12. U. Kulisch. A new arithmetic for scientific computation, [15], pp. 1-26. Academic Press, New York, 1983.

13. U. Kulisch and G. Bohlender. Features of a hardware implementation of an optimal arithmetic, [15], pp. 269-290. Academic Press, New York, 1983.

14. U. Kulisch and W. L. Miranker. Computer Arithmetic in Theory and Practice. Academic Press, New York, 1981.

15. U. Kulisch and W. L. Miranker (Eds.). A New Approach to Scientific Computation. Academic Press, New York, 1983.

16. R. E. Moore. Interval Analysis. Prentice-Hall, Englewood Cliffs, N. J., 1966.

17. R. E. Moore. Methods and Applications of Interval Analysis. SIAM Studies in Applied Mathematics, 2, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1979.

18. R. E. Moore. New results on nonlinear systems, [20], pp. 165-180. Academic Press, New York, 1980.

19. M. Neaga. PASCAL-SC Sprachbeschreibung und Programmieranleitung [PASCAL-SC Language Description and Programming Guide]. Department of Computer Science, University of Kaiserslautern, 1982.

20. K. L. E. Nickel (Ed.) Interval Mathematics 1980. Academic Press, New York, 1980.

21. L. B. Rall. Accurate computer arithmetic for scientific computation, Proceedings of the 1982 Army Numerical Analysis Conference, 343-356. U. S. Army Research Office, Research Triangle Park, N. C., 1982.

22. L. B. Rall. Differentiation in Pascal-SC: Type GRADIENT. MRC Technical Summary Report No. 2400. University of Wisconsin-Madison, 1982. To appear in ACM Trans. Math. Software, June, 1984.

23. L. B. Rall. Representation of intervals and optimal error bounds. Math. Comp. 41 (1983), 219-227.

24. L. B. Rall. Differentiation and generation of Taylor coefficients in Pascal-SC, [15], pp. 291-309. Academic Press, New York, 1983.

25. L. B. Rall. Interval Methods for Fixed-Point Problems. MRC Technical Summary Report No. 2583. University of Wisconsin-Madison, 1983.

26. S. M. Rump. Computer demonstration packages for standard problems of numerical mathematics, [15], pp. 27-49. Academic Press, New York, 1983.

27. S. M. Rump. Solving algebraic problems with high accuracy, [15], pp. 51-120. Academic Press, New York, 1983.

28. J. Wolff von Gudenberg. Gesamte Arithmetik des PASCAL-SC-Rechners. Benutzerhandbuch. [Complete Arithmetic of the PASCAL-SC Computer. User Handbook]. Institute for Applied Mathematics, University of Karlsruhe, 1981.

29. J. M. Yohe. Roundings in floating-point arithmetic. IEEE Trans. Computers C-22 (1973), 577-586.

APPENDIX A

Evaluation of (6.2) $w = 9x^4 - y^4 + 2y^2$ in Pascal-SC

1. Source code for the program WEVAL.

```
PROGRAM WEVAL(INPUT,OUTPUT);

(*  This program calculates w = 9*(x**4) - y**4 + 2*(y**2) in real and
    interval arithmetic.  *)

$USES INTERVAL;  (* DIRECTS COMPILER TO USE INTERVAL LIBRARY *)

    VAR C: CHAR; w,x,y: REAL; W,X,Y: INTERVAL;

    (* POWER OPERATORS *)

OPERATOR ** (R: REAL;K: INTEGER) RES: REAL;   (* R ** K *)

    VAR L: INTEGER;U: REAL;

    BEGIN    (* OPERATOR R ** K *)

        IF (R = 0) AND (K <= 0) THEN
            BEGIN   (* ERROR *)
                WRITELN('EXPONENTIATION ERROR, 0 ** K, K <= 0');
                SVR(0)   (* RETURN TO OPERATING SYSTEM *)
            END;    (* ERROR *)

        IF (K = 0) OR (R = 1) THEN U:=1

        ELSE IF K = 1 THEN U:=R

        ELSE    (* K <> 0,1 *)

            BEGIN                       (* REPEATED SQUARING *)

                L:=ABS(K);U:=1;
                REPEAT
                    IF L MOD 2 = 1 THEN U:=R*U;
                    L:=L DIV 2;
                    IF L <> 0 THEN R:=R*R
                UNTIL L = 0;
                IF K < 0 THEN U:=1/U   (* NEGATIVE EXPONENT *)

            END;                        (* REPEATED SQUARING *)

        RES:=U

    END;     (* OPERATOR R ** K *)
```

```
OPERATOR ** (I: INTERVAL;K: INTEGER) RES: INTERVAL;   (* I ** K *)

    VAR L: INTEGER;U: INTERVAL;

    BEGIN    (* OPERATOR I ** K *)

        IF (0 IN I) AND (K <= 0) THEN
            BEGIN   (* ERROR *)
                WRITELN('EXPONENTIATION ERROR, I ** K, 0 IN I');
                SVR(0)   (* RETURN TO OPERATING SYSTEM *)
            END;    (* ERROR *)

        IF K = 0 THEN
            BEGIN   (* K = 0 *)
                U.INF:=1;U.SUP:=1
            END     (* K = 0 *)

        ELSE IF K = 1 THEN U:=I

        ELSE    (* K <> 0,1 *)

            BEGIN                       (* REPEATED SQUARING *)

                L:=ABS(K);U.INF:=1;U.SUP:=1;
                REPEAT
                    IF L MOD 2 = 1 THEN U:=I*U;
                    L:=L DIV 2;
                    IF L <> 0 THEN I:=ISQR(I);
                UNTIL L = 0;
                IF K < 0 THEN U:=1/U   (* NEGATIVE EXPONENT *)

            END;                        (* REPEATED SQUARING *)

        RES:=U

    END;     (* OPERATOR R ** K *)

    (* END OF POWER OPERATORS *)

PROCEDURE WWRITE(w: REAL; W: INTERVAL);

    BEGIN   (* WRITE RESULTS OF REAL AND INTERVAL EXPRESSIONS *)

        WRITELN;WRITELN('   w = ',w);       (* OUTPUT OF RESULTS *)
        WRITELN;WRITELN('   W = [',W.INF,',',W.SUP,']');

    END;    (* WRITE RESULTS *)
```

```
BEGIN    (* MAIN PROGRAM *)

    C := 'Y'; WHILE C = 'Y' DO
    BEGIN   (* ACTUAL CALCULATION *)

        WRITELN('ENTER VALUES OF X AND Y');READ(x,y);
        X := INTPT(x); Y := INTPT(y);  (* CONVERT REAL VALUES TO INTERVALS *)

        w := 9*x*x*x*x - y*y*y*y + 2*y*y;  (* REAL RESULT *)

        W := 9*X*X*X*X - Y*Y*Y*Y + 2*Y*Y;  (* INTERVAL RESULT *)

        WRITELN;
        WRITELN('(a) w := 9*x*x*x*x - y*y*y*y + 2*y*y; gives:');
        WWRITE(w,W);                        (* OUTPUT RESULTS *)

        w := 9*(x**4) - y**4 + 2*(y**2);   (* REAL RESULT *)

        W := 9*(X**4) - Y**4 + 2*(Y**2);   (* INTERVAL RESULT *)

        WRITELN;
        WRITELN('(b) w := 9*(x**4) - y**4 + 2*(y**2); gives:');
        WWRITE(w,W);                        (* OUTPUT RESULTS *)

        w := (3*(x**2) - y**2)*(3*(x**2) + y**2) + 2*(y**2); (* REAL RESULT *)

        W := (3*(X**2) - Y**2)*(3*(X**2) + Y**2) + 2*(Y**2);
                                            (* INTERVAL RESULT *)

        WRITELN;
WRITELN('(c) w := (3*(x**2) - y**2)*(3*(x**2) + y**2) + 2*(y**2); gives:');
        WWRITE(w,W);                        (* OUTPUT RESULTS *)

        WRITELN;
        WRITELN('MORE VALUES (Y/N)?');READ(C,C)  (* CONTINUE . QUIT *)

    END    (* ACTUAL CALCULATION *)

END.    (* MAIN PROGRAM *)
```

2. Sample results using the program WEVAL.

```
B>XQPC WEVAL
ENTER VALUES OF X AND Y
*10864 18817

(a) w := 9*x*x*x*x - y*y*y*y + 2*y*y; gives:

   w =  1.58978000000E+05

   W = [-1.84102200000E+06, 1.15897800000E+06]

(b) w := 9*(x**4) - y**4 + 2*(y**2); gives:

   w = -8.41022000000E+05

   W = [-8.41022000000E+05, 1.15897800000E+06]

(c) w := (3*(x**2) - y**2)*(3*(x**2) + y**2) + 2*(y**2); gives:

   w =  1.00000000000E+00

   W = [ 1.00000000000E+00, 1.00000000000E+00]

MORE VALUES (Y/N)?
*N
KL/P-STOP
```

APPENDIX B

1. Source code for the program LINSYS.

```
PROGRAM LINSYS(INPUT,OUTPUT);

    (* This program computes an interval vector Y containing the solution
       of linear systems of equations AX = B up to order 20, or returns an
       error message if the coefficient matrix A is singular or extremely
       badly conditioned.  The matrix A is read by rows, followed by B. *)

$USES LGL,DIM=20;   (* SETS DIMENSION FOR EXTERNAL LIBRARY ROUTINES *)

    VAR     A: RMATRIX;       (* COEFFICIENT MATRIX *)
            B: RVECTOR;       (* RIGHT SIDE *)
            Y: IVECTOR;       (* INCLUSION OF SOLUTION OF AX = B *)

        AKDIM: INTEGER;       (* ACTUAL SIZE OF SYSTEM *)
          I,J: DIMTYPE;       (* INDEX VARIABLES *)
            C: CHAR;          (* CONTROL VARIABLE *)

BEGIN    (* MAIN PROGRAM *)

    C:='Y'; WHILE C = 'Y' DO

    BEGIN      (* SOLUTION OF SYSTEM *)

        WRITELN('INPUT DIMENSION');READ(AKDIM);WRITELN;
        WRITELN('INPUT MATRIX BY ROWS');
        FOR I:=1 TO AKDIM DO FOR J:=1 TO AKDIM DO READ(A[I,J]);
        WRITELN;WRITELN('INPUT RIGHT SIDE');
        FOR I:=1 TO AKDIM DO READ(B[I]);

        LGLP(DIM,AKDIM,A,B,Y);   (* SOLVE SYSTEM *)

        IF Y[1].INF <= Y[1].SUP THEN  (* Y IS PROPER *)

        BEGIN       (* OUTPUT OF SYSTEM AND RESULTS *)

            WRITELN;
            WRITELN('SOLVE AY = B WITH INTERVAL INCLUSION OF ANSWER');
            WRITELN;
            FOR J:=1 TO AKDIM DO     (* OUTPUT OF A BY COLUMNS *)
                BEGIN
                    FOR I:=1 TO AKDIM DO
                    WRITELN('A[',I:2,',',J:2,'] = ',A[I,J]);WRITELN;
                END;                 (* OUTPUT OF A *)
            WRITELN;
            FOR I:=1 TO AKDIM DO
            WRITELN('B[',I:2,'] = ',B[I]);   (* OUTPUT B *)
            WRITELN;FOR I:=1 TO AKDIM DO     (* OUTPUT Y *)
            WRITELN('Y[',I:2,'] = [',Y[I].INF,',',Y[I].SUP,']');

        END         (* OUTPUT OF SYSTEM AND RESULTS  *)
```

```
        ELSE        (* Y IS IMPROPER *)

        BEGIN       (* ERROR MESSAGE *)

            WRITELN;WRITELN('THE MATRIX IS SINGULAR OR BADLY CONDITIONED');
            FOR I:=1 TO AKDIM DO Y[I].SUP:=Y[I].INF  (* RESET Y *)

        END;        (* ERROR MESSAGE *)

        WRITELN;WRITELN('ENTER ANOTHER SYSTEM (Y/N)?');
        READ(C,C);

    END         (* SOLUTION OF SYSTEM *)

END.    (* MAIN PROGRAM *)
```

2. Sample results using LINSYS.

```
A>XQP LINSYS TAK.DAT CON:
INPUT DIMENSION

INPUT MATRIX BY ROWS

INPUT RIGHT SIDE

SOLVE AY = B WITH INTERVAL INCLUSION OF ANSWER

A[ 1, 1] = -1.91569421219E+11
A[ 2, 1] =  0.00000000000E+00
A[ 3, 1] =  8.54599767833E+08
A[ 4, 1] =  0.00000000000E+00
A[ 5, 1] =  0.00000000000E+00

A[ 1, 2] =  0.00000000000E+00
A[ 2, 2] =  6.12580328713E+11
A[ 3, 2] =  0.00000000000E+00
A[ 4, 2] =  4.67810493936E+07
A[ 5, 2] =  2.73290749688E+09

A[ 1, 3] =  1.00000000000E+00
A[ 2, 3] = -1.00000000000E+00
A[ 3, 3] =  1.12080226545E+02
A[ 4, 3] =  0.00000000000E+00
A[ 5, 3] =  1.12075040461E+02

A[ 1, 4] =  0.00000000000E+00
A[ 2, 4] =  5.21250000000E-01
A[ 3, 4] =  0.00000000000E+00
A[ 4, 4] =  1.00000000000E+00
A[ 5, 4] =  0.00000000000E+00

A[ 1, 5] = -5.03360090692E-03
A[ 2, 5] =  1.60975964126E-02
A[ 3, 5] =  2.24535604687E-05
A[ 4, 5] =  0.00000000000E+00
A[ 5, 5] =  7.18104363616E-05


B[ 1] =  1.85094646224E+04
   2] = -5.91833428873E+04
   3] = -8.25723911920E+01
   4] =  1.92589687381E+33
   5] =  2.15844970057E+35

Y[ 1] = [ 7.93624477937E+29, 7.93624477938E+29]
Y[ 2] = [ 7.93739034599E+29, 7.93739034600E+29]
Y[ 3] = [-4.20005760703E+32,-4.20005760702E+32]
Y[ 4] = [-3.71300190864E+37,-3.71300190863E+37]
Y[ 5] = [-3.02038610401E+43,-3.02038610400E+43]

ENTER ANOTHER SYSTEM (Y/N)?
KL/?-STOP
```

3. Contents of the data file TAK.DAT.

```
A>TYPE TAK.DAT
5
-1.91569421219E+11 0 1 0 -5.03360090692E-03
0 6.12580328713E+11 -1 5.2125E-01 1.60975964126E-02
8.54599767833E+08 0 1.12080226545E+02 0 2.24535604687E-05
0 4.67810493936E+07 0 1 0
0 2.73290749688E+09 1.12075040461E+02 0 7.18104363616E-05
1.85094646224E+04 -5.91833428873E+04 -8.25723911920E+01 1.92589687381E+33
2.15844970057E+35
N
```

# OPTIMAL CORRECTIONS OF A DAMPED LINEAR OSCILLATOR UNDER RANDOM PERTURBATIONS*

P. L. Chow and J. L. Menaldi

Department of Mathematics, Wayne State University

Detroit, Michigan 48202

ABSTRACT An additive control of a randomly excited linear damped oscillator is studied by the methods of dynamic programming and variational inequalities. The objective is to minimize the mean deviation from the rest position over a finite horizon, with a possible resource constraint. We shall present some analytical results on the optimal average cost function and the feedback control law. Numerical solution will be discussed briefly.

I. INTRODUCTION. The problem of controlling random vibrations occurs in a wide range of situations, from the vehicle dynamics to the stabilization of a skyscraper during the earthquake. In this article we study a relatively simple problem involving a unit mass attached to a spring and a dash-pot. If the system is excited by an external random force, one is interested in regulating the fluctuations about the rest position. The problem would have been classical if the controlling force were smooth and unlimited in its energy supply. However we will be mainly concerned with more realistic class of controls, where the impulsive forces are allowed and the total energy supply is fixed over a finite time-horizon. This makes the problem more meaningful but complicated. But it is essential to take the necessary first step in order to understand the more

* This work has been supported in part by the U.S. Army Research Office under Contract DAAG29-83-K-0014.

difficult problems with several degrees of freedom.

In a recent paper [1], we treated the problem of controlling a first-order Ito equation, which may be regarded as a special case of the present problem. It corresponds to the limiting case when the spring constant tends to zero. Our work is a generalization of the results by Gorbunov [2] in the sense that we admit a wider class of controls. Furthermore our techniques are rather different from his. For the general mathematical techniques involved, one is referred to the references [3] - [5].

The article summarizes some preliminary results concerning our investigation in the announced subject. In what follows, we shall first formulate the problem under study and describe the analytical results that we have obtained. Also the numerical solution, among others will be discussed.

II. FORMULATION OF PROBLEMS. Let us consider a damped linear oscillator excited by a white-noise and controlled by a regulating force:

$$\begin{cases} \ddot{x} + p\dot{x} + k^2 = r\dot{w}_t + \dot{v}_t \ , \ o < t \leq T, \\ x(o) = x_o \ , \ \dot{x}(o) = y_o \ , \end{cases} \tag{1}$$

where the dot means the time derivative and

$x(t)$ : the position of a unit mass at the time $t$,

$q$, $k$ : the damping and spring constants,

$x_o$, $y_o$ : the initial position and velocity,

$r\dot{w}_t$ : the white-noise with the intensity $r$,

$v_t$ : the momentum control at $t$,

$T$ : the horizon.

Since the system will be controlled by a finite source of energy, we introduce the class $V$ of controls consisting of all random processes $\{v_t, t \geq o\}$, adapted to the Wiener process $\{w_t, t \geq o\}$, with a bounded variation.

To give the equation (1) a mathematical meaning we interpret it in the Ito's sense:

$$\begin{cases} dx = ydt , \\ dy = -(px + qy)dt + rdw_t + dv_t, \\ x(o) = x_o \ , \ y(o) = y_o, \end{cases} \tag{2}$$

where $y = \dot{x}$ and $p = k^2$.

Let $v_t^+$, $v_t^-$ denote the positive and negative parts of $v_t$, so that

$$v_t = v_t^+ - v_t^- , \quad t \geq o . \tag{3}$$

Suppose the total supply of correctional energy is fixed. Then we have the constraint

$$v_t = v_t^+ + v_t^- \leq z, \quad t \geq o , \tag{4}$$

where $z > o$ is propotional to the total correctional energy supply and will be called simply the correctional energy. Let $V_{ad}$ be the class of admissible controls $v$ in $V$ satisfying (4).

For each control policy $v$, let $J(v)$ be the average cost defined by

$$J(x_o, y_o, v) = E_v \left\{ \int_o^T f[x(t), y(t)]\, dt + g[x(T), y(T)] \right\} \quad (5)$$

where $f$ and $g$ are some positive functions measuring the deviation from the rest state.

Our goal is to find an optimal policy $\hat{v}$ in $V_{ad}$, in the form of the feedback law, which minimizes $J$, i.e.

$$\hat{J} = J(\hat{v}) = \inf \{J(v) : v \in V_{ad}\}. \quad (6)$$

To this end we shall use the techniques of dynamic programming and variational inequalities.

III. PENALIZATION AND DYNAMIC PROGRAMMING Instead of (2.2), we consider the following system

$$\begin{cases} dx(s) = y(s)\, ds, \\ dy(s) = [px(s) + qy(s)]\, ds + r\, dv_s + dv_s, \\ dz(s) = -d|v_s| = -(dv_s^+ + dv_s^-), \quad o < s \leq (T-t), \\ dt(s) = ds, \\ x(o) = x, \quad y(o) = y, \quad z(o) = z \quad t(o) = t. \end{cases} \quad (7)$$

where $z(s)$ is the remaining amount of correctional energy. Corresponding to the cost function (5), the running cost from $t$ to $T$ is given by

$$J(x, y, z, t, v) = E_v \int_o^{T-t} f[x(s), y(s), y(s), y(s)]\, ds + g[x(T-t), y(T-t)]. \quad (8)$$

The optimal cost is given by

$$\hat{u}(x, y, z, t) = \inf \{J(x, y, z, t, v) : v \in V_{ad}\}. \quad (9)$$

In order to apply the dynamic programming principle to find $\hat{u}$, we introduce

a regularized version of $V_{ad}$ defined by:

$$V_{ad}^{\varepsilon} = \{v \in V_{ad} : \dot{v}_t = \nu, \ |\nu| \leq \frac{1}{\varepsilon}\}, \ \varepsilon > 0. \tag{10}$$

We set

$$\hat{u}_{\varepsilon}(x, y, z, t) = \inf \{J(x, y, z, t, v) : v \in V_{ad}^{\varepsilon}\}. \tag{11}$$

Then the standard dynamic programming argument yields the optimality equation for $\hat{u}_{\varepsilon}$:

$$\begin{cases} \inf_{\mathcal{L}(\nu)} u_{\varepsilon} = -f(x,y), \ x, y \in R, \ z > 0, \ 0 \leq t < T, \ |\nu| \leq \frac{1}{\varepsilon} \\ u_{\varepsilon}(x, y, z, T) = g(x, y), \\ u_{\varepsilon}(x, y, 0, t) = u^{0}(x, y, t). \end{cases} \tag{12}$$

Here we have put

$$\begin{cases} \mathcal{L}(\nu) = \mathcal{L}_0 + \mathcal{L}_1(\nu) \\ \mathcal{L}_0 = \partial_t + \frac{1}{2} r^2 \partial_{\dot{y}}^2 + \dot{y} \partial_x - (px + qy) \partial_y, \\ \mathcal{L}_1(\nu) = \nu \partial_{\dot{y}} - |\nu| \partial_z, \end{cases} \tag{13}$$

and $u^0$ satisfies

$$\begin{cases} \mathcal{L}_0 u^0 = -f \\ u^0(x, y, T) = g. \end{cases} \tag{14}$$

In (13) the partial derivatives $\partial_t$, $\partial_{\dot{}}$ ...., stand for the partial derivatives $\frac{\partial \cdot}{\partial t}$, $\frac{\partial \cdot}{\partial \dot{}}$ ....

It seems reasonable to expect that, as $\varepsilon \to 0$, the solution $u_{\varepsilon}$ of the penalized problem (12) approaches the optimal cost $\hat{u}$ giver by (9).

This fact among other results will be described in the next section.

IV. SOME RESULTS AND DISCUSSION. First let us present some preliminary results pertaining to the penalized problem (12) and its relation to the original one. To analyze (12) we need the following:

(R 1). If $\partial_z \varphi \leq 0$, then

$$\inf_{|\nu| \leq \frac{1}{\varepsilon}} \mathcal{L}_1(\nu)\, \varphi = \frac{1}{\varepsilon} \{ |\partial_y \varphi| + \partial_z \varphi \}^+, \quad (15)$$

where $\{x\}^+$ denotes the positive part of x.

Under appropriate conditions, we can show that

(R 2). Assume that the functions $f(x, y) = f(x) \geq 0$ and $g(x,y) = g(x) \geq 0$ are convex and even, and the functions together with their derivatives are uniformaly Lipschitz - continuous. Then the optimal cost function $\hat{u}_\varepsilon$ given by (11) has the following properties:

(a) For each $t > 0$, $\hat{u}_\varepsilon(x, y, z, t)$ is convex in $(x, y, z)$.

(b) $\hat{u}_\varepsilon(x, y, z, t) = \hat{u}_\varepsilon(-x, -y, z, t)$ for any $(x, y, z, t)$.

(c) $\hat{u}_\varepsilon$ is Lip-continuous in $(x, y, z, t)$, (uniformly in $\varepsilon$ ).

(d) $\partial_x \hat{u}_\varepsilon$ and $\partial_y \hat{u}_\varepsilon$ are uniformly Lip-continuous in x, y and $\varepsilon$ for each $(z, t)$.

(e) $\hat{u}_\varepsilon$ $(x, y, ., t)$ is a decreasing function of z for every fixed $(x, y, t)$.

Further $\hat{u}_\varepsilon$ is the solution of the Hamilton-Jacobi-Bellman equation (12).

Remarks: The fact that the optimal cost $\hat{u}_\varepsilon$ decreses with the total correction energy z is intuitively clear. It follows that $\varphi = \hat{u}_\varepsilon$ satisfies the condition

in (R1). The symmetry property (b) enables us to consider the problem either in the region $x \geq o$ or $y \geq o$.

Now we hope to recover the optimal cost $\hat{u}$ by taking the limit of $\hat{u}_\varepsilon$ as $\varepsilon \downarrow o$. This is indeed a valid procedure. In fact we can say much more, as indicated by the next result:

(R 3). Under the same assumptions in (R 2), we have

$$\hat{u}(x, y, z, t) = \lim_{\varepsilon \downarrow o} \hat{u}_\varepsilon (x, y, z, t),$$

where $\hat{u}$ is given by (9). Furthermore $\hat{u}$ preserves the properties (a), (b), (c), (e) in (R 2), and satisfies the variational inequalities:

$$\begin{cases} (Au - f) \leq 0 \quad \text{and} \quad M(u) \leq 0, \\ (Au - f) \cdot M(u) = 0, \quad z > o, \\ u|_{t=T} = g, \\ u|_{z=o} = u^o, \end{cases} \tag{16}$$

where $A = -\mathcal{L}_o$ with $\mathcal{L}_o$ given in (13); $u^o$ satisfies (14), and

$$M(\varphi) = |\partial_y \varphi| + \partial_z \varphi . \tag{17}$$

Remark: In the above analysis, we have neglected the control cost, i.e. f, g independent of $v$. A similar analysis has been carried out when the cost for control is not negligible.

(R 4) In the case of one-sided control ( $v = v^+$ ), let

$$w(x, y, t) = \lim_{z \to \infty} u(x, y, z, t) \tag{18}$$

Then the problem (16) reduces to

$$\begin{cases} (Aw - f) \leq o, \quad Bw = \partial_y w \leq o, \\ (Aw - f)(Bw) = o, \\ w|_{t=T} = g. \end{cases} \tag{19}$$

Finally we have the following decomposition.

(R 5). For a one-sided control, the problem (16) can be decomposed into the control-free problem (14) and the problem (19) with unlimited control energy. In particular, the solution to (16) can be expressed as :

$$u(x, y, z, t) = w(x, y, t) + w^{o}(x, y + z, t)$$
$$w^{o}(x, y, t) = u^{o}(x, y, t) - w(x, y, t). \tag{20}$$

With the aid of the above decomposition, the construction of solution to (16) is greatly simplified.

In spite of the above results, the following questions remain to be resolved:

(Q.1) Free Boundary The system (16) is a terminal and free-boundary value problem. For the reduced problem the free boundary $\Gamma : y = y^{*}(x,t)$, separates two regions in which $A w = f$, $Bw = o$, respectively. The smoothness and the shape of $\Gamma$ have yet to be determined.

(Q 2). Construction of the Optimal Control. In the one-dimensional case, we showed that the optimal control is a reflected diffusion from the free boundary [1]. It is conceivable that, in the present case, the optimal control is an oblique-reflected diffusion. But this has to be verified.

(Q 3) Numerical Approximation. To compute the optimal cost $\hat{u}_{\varepsilon}$, we may try to solve the variational inequalities (16) numerically. This can be done by an iterative scheme and the finite element method [6]. Another possibility is to apply (R2) and (R3) by solving the penalized problem for small $\varepsilon$ . The penalized problem (12), in view of (R1), is a nonlinear parabolic terminal-boundary value problem in (t, x, y, z):

$$
\begin{cases}
\mathcal{L}_0 u_\varepsilon - \frac{1}{\varepsilon}\{|\partial_y u_\varepsilon| + \partial_z u_\varepsilon\}^+ = f, \ o \le t \le T, \\
u_\varepsilon(x, y, z, T) = g(x, y), \\
u_\varepsilon(x, y, o, t) = u^o(x, y, t),
\end{cases} \tag{21}
$$

which is coupled to the problem (14) for $u^o$. Because of the high dimensionality, an efficient numerical algorithm for solving such problem is unknown to us.

The above questions and others are being investigated, and further results will be presented elsewhere.

## REFERENCES


[1] P.L. Chow, J.L. Menaldi and M. Robin, Additive Control of Stochastic Linear Systems with Finite Horizon, SIAM J. Control and Optim, to appear.

[2] V.K. Gorbunov, Minimax Impulsive Correction of Perturbations of a Linear Damped Oscillator, Appl. Math. Mech. (PMM), 40 (1976), pp. 252-259.

[3] A. Bensoussan and J.L. Lions, Applications des Inequations Variationnelles en Control Stochastique, Dunod, Paris, 1978.

[4] W.H. Fleming and R.W. Rishel, Deterministic and Stochastic Optimal Control, Springer-Verlag, New York, 1975.

[5] J.L. Menaldi and M. Robin, On some Cheap Control problems for Diffusion Processes, Trans. Amer. Math. Soc., 278 (1983), pp. 771-802.

[6] R. Glowinski, J.L. Lions and R. Tremolieres, Numerical Analysis of Variational Inequalities, North-Holland, Amsterdam, 1981.

# NON-PERIODIC CONDITIONS FOR CHAOS AND SNAP-BACK REPELLERS

Nam P. Bhatia
University of Maryland
Baltimore County, Maryland 21228
and
Walter O. Egerland
Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005-5066

## Introduction

Li and Yorke [1] introduced in their fundamental paper the term "chaotic" for a class of self-mappings of an interval. The real function f is chaotic if (a) there are points of arbitrarily large periods and (b) there is an uncountable set S such that for every $x_0, y_0 \in S$, $x_0 \neq y_0$, $\lim_{n\to\infty} \sup|f^n(x_0) - f^n(y_0)| > 0$ and $\lim_{n\to\infty} \inf|f^n(x_0) - f^n(y_0)| = 0$. Following the Li-Yorke result that "period three implies chaos", many authors worked on periodic conditions that allow the same conclusion. The best known of these conditions is "period $p \neq 2^n$ implies chaos". Such investigations are summarized in Targonski's monograph [2].

Li and Yorke also introduced four-point inequalities satisfied by a point and its three successors with respect to the given function f. They showed that these imply the existence of a three-period and hence chaos. Our investigations, begun under the US Army Summer Faculty Research and Engineering Program 1983, show that the Li-Yorke inequalities play a fundamental role in the theory of chaos. For this presentation we have singled out three theorems. The first is an addendum to the Li-Yorke theorem. The second establishes equivalent companion inequalities to the Li-Yorke inequalities. The third theorem, of a different character, is especially important in applications. We also introduce the elementary notion of a Newton function (our term) which we have found indispensable in the investigation of parameter families and which, we think, deserves to be better known. The quadratic mapping $g(y) = ay^2 + 2by + c$, $a \neq 0$, b, c, real constants, particular cases of which have received wide attention, is used to illustrate certain points. Finally, we state an open problem that is of interest in connection with "random number generators".

## Definitions and Notations

Let $f : R \to R$ be continuous. If $x_0 \in R$, the orbit of $x_0$ under $f$ is defined as the set $\{x : x = f^n(x_0),\ n = 0,1,\ldots\}$, where, for every positive integer $n$, $f^n$ is the $n$-th iterate of $f$ and $f^0(x_0) = x_0$. We shall write $x_n := f^n(x_0)$ for a given $x_0 \in R$ and call $x_1, x_2 \ldots$ the successors of $x_0$. A pre-orbit of a given $x_0 \in R$ is any (finite or infinite) sequence $x_0, x_{-1}, x_{-2}, \ldots$ such that $f(x_{-n}) = x_{-(n-1)}$ for all $n$ for which $x_{-n}$ is defined. The points $x_{-1}, x_{-2}, \ldots$ in any such sequence are called predecessors of $x_0$. A point $x_0$ is called critical if $f(x_0) = x_0$, i.e., a critical point of $f$ is a fixed point of $f$. A periodic point $x_0$ of period $p > 1$ ($p$ a positive integer) is a point for which the relations $f^p(x_0) = x_0$, $f^k(x_0) \neq x_0$, $1 \leq k < p$, hold.

The following fundamental results are now well-known.

Theorem. (Sarkovskii). For $m, n = 0,1,\ldots$ consider the total ordering of the positive integers:

$$3 < 5 < 7 < \cdots < 2\cdot 3 < 2\cdot 5 < 2\cdot 7 \cdots < 2^n\cdot 3 < 2^n\cdot 5 < 2^n\cdot 7$$
$$< \cdots < 2^m < 2^{m-1} < \cdots < 2^2 < 2 < 1.$$

If a continuous mapping $f : R \to R$ has a periodic point of period $p$, then it also has a periodic point of period $q$ for every $q$ [illegible] (in the above total order).

Theorem (Li-Yorke): Let $f : R \to R$ be continuous. If there is a point $x_0 \in R$ such that either $x_3 < x_0 < x_1 < x_2$ or $x_3 > x_0 > x_1 > x_2$, then $f$ has a point of period three. Furthermore, if $f$ has a three periodic point, there exists an uncountable set $S \subset R$ such that for every $x_0, y_0 \in S$, $x_0 \neq y_0$,

$$\lim_{n\to\infty} \sup|x_n - y_n| > 0$$

and

$$\lim_{n\to\infty} \inf|x_n - y_n| = 0.$$

Definition. A point $x_0 \in R$ satisfies a Li-Yorke inequality if

$$x_3 < x_0 < x_1 < x_2$$

or

$$x_3 > x_0 > x_1 > x_2.$$

An Example. The following example shows that the existence of a three periodic point does not guarantee the existence of a Li-Yorke inequality. The quadratic mapping $g(y) = ay^2 + 2by + c$, $a \neq 0$, b, c. real constants, may be brought in the form $f(x) = x^2 - r$ by setting $y = a^{-1}(x - b)$, $g = a^{-1}(f - b)$, and $r = b^2 - b - ac$. f has a three periodic orbit for $r = 7/4$, but no point $x_0 \in R$ satisfies a Li-Yorke inequality for $r \leq 7/4$.

Definition. Let $f : R \to R$ be continuous. A Newton function of f is defined as a nonnegative solution of the differential equation $N(x) = (x - f(x))N'(x)$.

It is clear from the definition that N is unique up to a multiplicative constant, piecewise monotone and differentiable. The application of the traditional method of Newton for solving iteratively the equation $N(x) = 0$ results in a "Newton sequence" that is identical with the iterates $x_n = f^n(x_0)$. Thus by determining the Newton function of a given f a familiar geometric picture for the evolution of the iterates becomes available that, in our opinion, is superior to others. Although the qualitative and quantitative characteristics of a Newton function that signal existence or nonexistence of chaos await exploration, its knowledge certainly settles the question of convergence at a glance and, in any case, provides an invaluable source of visual insight. For example, by considering the Newton function of $f(x) = x^2 - r$, it is apparent that (a) the iterates converge to $c_1$ if $x_0 \in (c_{-2}, c_2)$ and $-1/4 \leq r \leq 3/4$, where $c_1$ and $c_2$, $c_1 \leq c_2$, are the critical points of f, and (b) no orbit originating in the interval $(-c_2, c_2)$ can escape it for $3/4 \geq r \geq -1/4$.

Theorem 1. If $x_0 \in R$ satisfies a Li-Yorke inequality, then f has at least two distinct three periodic orbits.

The theorem extends the Li-Yorke theorem and may be illustrated by the example $f(x) = x^2 - 2$. The point $x_0 = \sqrt{2}$ satisfies a Li-Yorke inequality and $(2 \cos 2\pi/7, 2 \cos 4\pi/7, 2 \cos 8\pi/7)$, $(2 \cos 2\pi/9, 2 \cos 4\pi/9, 2 \cos 8\pi/9)$ are two distinct three-periodic orbits of f.

Theorem 2. The sets of points $A = \{x_0 : x_3 < x_0 < x_1 < x_2\}$, $B = \{x_0 : x_1 < x_2 < x_0 < x_3\}$, and $C = \{x_0 : x_2 < x_3 < x_0 < x_1\}$ are either all empty or all non-empty. The same statement holds for the sets with all inequalities reversed.

The theorem extends the Li-Yorke theorem by establishing equivalent companion inequalities to the Li-Yorke inequalities. Since examples show that other four-point inequalities between $x_0$, $x_1$, $x_2$, $x_3$ do not assure the existence of three-periodic orbits, the theorem lists the complete set of Li-Yorke four-point inequalities.

For the proofs of Theorem 1 and Theorem 2 the reader is referred to [4].

Theorem 3. If a critical point $c_0$ of f has predecessors satisfying

$$c_{-2} < c_0 < c_{-3} < c_{-1},$$

then f has a six-periodic point.

Proof. Since $f(c_{-2}) = c_{-1} > c_{-3}$ and $f(c_0) = c_0 < c_{-3}$, there exists $c_{-4} \in (c_{-2}, c_0)$, and $f(c_{-3}) = c_{-2} < c_{-4}$, $f(c_0) = c_0 > c_{-4}$ imply the existence of a $c_{-5} \in (c_0, c_{-3})$. Hence

$$c_{-2} < c_{-4} < c_0 < c_{-5} < c_{-3} < c_{-1}.$$

If we set $c_{-5} = x_0$, we obtain $x_6 < x_0 < x_2 < x_4$, i.e., $x_0$ satisfies a Li-Yorke inequality for the function $f^2$. Therefore, $f^2$ has a three-periodic point $x^*$ that is either a six-periodic or a three-periodic point of f. If $x^*$ is a six-periodic point of f, the proof is complete. If $x^*$ is a three-periodic point of f, the theorem of Sarkovskii ensures that f has also a six-periodic point. This completes the proof.

The theorem subsumes many propositions that ensure chaos via the concept of snap-back repeller, first introduced by Marotto [5]. (Given a critical point $c_0$, an interval $(a_0, b_0)$ containing $c_0$ is called a repelling interval if for every $x_0 \in (a_0, b_0)$, $x_0 \neq c_0$, there is an $n \geq 1$ such that $x_n \notin (a_0, b_0)$. A critical point $c_0$ is a repeller if it has a repelling interval. A repeller $c_0$ is called a snap-back repeller if every interval $(a_0, b_0)$ containing $c_0$ contains a point $x_0 \neq c_0$ such that $x_n = c_0$ for some positive integer $n \geq 2$.) For example, by adding the condition that the interval $(c_0, c_{-3})$ does not contain a critical point of $f^2$, one can conclude that $c_0$ is a snap-back repeller. We intend to explore the connection between our inequality condition and the snap-back repeller condition for chaos elsewhere.

Example. If $f(x) = x^2 - r$, then f has a six periodic point for all $r \geq r_0$, where $r_0 \approx 1.543689011$ is the real root of $r^3 - 2r^2 + 2r - 2 = 0$. And if we apply the theorem to the smaller of the two-periodic points of f, which is a critical point of $f^2$, we conclude the existence of a twelve-periodic orbit of f and hence chaos for all $r > 1.4305$. This is the best prediction to date that chaos has occured for $f(x) = x^2 - r$, for which the onset of chaos at $r \approx 1.402$ was determined by computer studies.

An Open Problem. Consider the sequence of random variables $x_{n+1} = x_n^2 - 2$ with $x_0$ uniform over $[-2,2]$. Then $x_n$ converges in distribution to the random variable x whose distribution is given by $F(x) = 1 - (1/\pi) \arccos x/2$, $-2 \leq x \leq 2$. What is the limit in distribution of $x_{n+1} = x_n^2 - p$, $3/4 \leq p \leq 2$, where $x_0$ is uniform over $[-1/2 - (p + 1/4)^{1/2}, 1/2 + (p + 1/4)^{1/2}]$?

Acknowledgement

The authors wish to thank Dr. S. S. Wolff of the BRL for his continued support and valuable discussions.

## References

[1] T. -Y. Li and J. A. Yorke, Period three implies chaos. Amer. Math. Monthly 82 (1975), pp. 985-992.

[2] György Targonski, Topics in Iteration Theory, Studia Mathematica, Skript 6, Vandenhoek and Ruprecht, Göttingen and Zürich (1981).

[3] A. N. Sarkovskii, Coexistence of Cycles of a Continuous Function of the Line into itself (Russian). Ukrain: Mat. Z. 16 (1964), pp. 61-71.

[4] N. P. Bhatia and W. O. Egerland, On the Existence of Li-Yorke Points in the Theory of Chaos. Mathematics Research Reports, No. 84-4 (1984), University of Maryland, Baltimore County.

[5] F. R. Marotto, Snap-Back Repellers imply Chaos in $R^n$, Journal of Mathematical Analysis and Applications 63 (1978), pp. 199-223.

# ON THE NUMERICAL SOLUTION OF SINGULARLY PERTURBED LINEAR TWO-POINT BOUNDARY-VALUE PROBLEMS

B.S. Ng
Department of Mathematical Sciences
Indiana University-Purdue University
Indianapolis, IN 46223

W.H. Reid
Department of Mathematics
The University of Chicago
Chicago, IL 60637

ABSTRACT We present an initial-value technique, based on the use of certain compound matrices, for the numerical solution of linear two-point boundary-value problems involving unstable ordinary differential equations of the singular perturbation type. We demonstrate the effectiveness of the method via certain examples which exhibit internal as well as end-point boundary-layers.

I. INTRODUCTION In two earlier papers [1, 2], we introduced the compound matrix method for the numerical solution of mathematically unstable linear eigenvalue and boundary-value problems with separated boundary conditions. The focus of these previous studies has been on fourth-order problems involving single differential equations of the Orr-Sommerfeld type. The basic ideas developed in [1,2] have since, however, been generalized in a variety of ways, and these include the use of compound matrices to deal with boundary-value problems involving higher-order systems of equations [3]. The method has also been successfully applied to a number of problems arising from the study of hydrodynamic stability [4, 5, 6, 7]. In this paper, we wish to demonstrate, via examples, that the method can also be effective in dealing with certain second-order problems of the singular perturbation type.

II. THE COMPOUND MATRIX METHOD For our purpose, it is convenient to first discuss the compound matrix method in terms of a third-order problem. Consider then a single third-order equation

$$L_3(\psi) = \psi''' - c_1\psi'' - c_2\psi' - c_3\psi = f, \qquad (1)$$

where $c_1$, $c_2$, $c_3$, and f are functions of x and $a \leq x \leq b$. We let $\underset{\sim}{\psi} = [\psi, \psi', \psi'']^T$; we shall also suppose that a single boundary condition is prescribed at x = a and is given by

$$\underset{\sim}{\underset{\sim}{P}}\underset{\sim}{\psi}(a) = p, \tag{2}$$

where $\underset{\sim}{P}$ is a 1 x 3 row vector and p is a constant. The remaining two boundary conditions at x = b are given by

$$\underset{\sim}{\underset{\sim}{Q}}\underset{\sim}{\psi}(b) = \underset{\sim}{q}, \tag{3}$$

where $\underset{\sim}{Q}$ is a 2 x 3 matrix of rank 2 and $\underset{\sim}{q}$ is a 2 x 1 constant vector. If we let $\underset{\sim}{\psi}_0$ denote the solution of (1) which satisfies (2) and let $\underset{\sim}{\psi}_1$ and $\underset{\sim}{\psi}_2$ denote two linearly independent solutions of the corresponding homogeneous problem (with $f \equiv 0$ and $p = 0$), the solution to (1) - (3) can be written in the form

$$\underset{\sim}{\psi} = \underset{\sim}{\psi}_0 + \alpha\underset{\sim}{\psi}_1 + \beta\underset{\sim}{\psi}_2, \tag{4}$$

for some constants $\alpha$ and $\beta$. In the usual application of the method of complementary functions, $\underset{\sim}{\psi}_0$, $\underset{\sim}{\psi}_1$ and $\underset{\sim}{\psi}_2$ must be computed as solutions to three separate initial-value problems. The boundary conditions at x = b then lead to a pair of linear equations from which, at least in principle, the constants $\alpha$ and $\beta$ in (4) can be determined. Nevertheless, it is well-known that the coefficient matrix of the linear equations for $\alpha$ and $\beta$ can be highly ill-conditioned if the homogeneous solutions of (1) exhibit inherent growth problems. In that case a further loss of accuracy will occur due to cancellation errors as $\underset{\sim}{\psi}$ must be obtained from (4) by superposition.

Alternatively, the compound matrix method is based on considering the solution matrices $\underset{\sim}{\Psi}_0$ and $\underset{\sim}{\Psi}$ where

$$\underset{\sim}{\Psi}_0 = \begin{bmatrix} \psi_0 & \psi_1 & \psi_2 \\ \psi_0' & \psi_1' & \psi_2' \\ \psi_0'' & \psi_1'' & \psi_2'' \end{bmatrix} \quad \text{and} \quad \underset{\sim}{\Psi} = \begin{bmatrix} \psi_1 & \psi_2 \\ \psi_1' & \psi_2' \\ \psi_1'' & \psi_2'' \end{bmatrix} \tag{5}$$

We note that the 2 x 2 minors of $\underset{\sim}{\Psi}$ are defined by

$$y_1 = \psi_1\psi_2' - \psi_1'\psi_2, \quad y_2 = \psi_1\psi_2'' - \psi_1''\psi_2, \quad y_3 = \psi_1'\psi_2'' - \psi_1''\psi_2', \tag{6}$$

and $\underset{\sim}{y} = [y_1, y_2, y_3]^T$ is called the second compound of $\underset{\sim}{\Psi}$. The only 3 x 3 minor of $\underset{\sim}{\Psi}$ is simply its determinant which can be expressed in the form

$$z = y_1\psi_0'' - y_2\psi_0' + y_3\psi_0, \tag{7}$$

and z is called the third compound of $\underset{\sim}{\Psi}_0$. On differentiating (6) and (7) and eliminating the third derivatives by the use of (1), we then obtain

$$
\begin{aligned}
y_1' &= y_2, \\
y_2' &= y_3 + c_1 y_2 + c_2 y_1, \\
y_3' &= c_1 y_3 - c_3 y_1, \\
z' &= c_1 z + y_1 f,
\end{aligned}
\tag{8}
$$

where the initial conditions for $y_1$, $y_2$, $y_3$ and $z$ can be derived from the corresponding conditions on $\underset{\sim}{\psi}_0$, $\underset{\sim}{\psi}_1$, and $\underset{\sim}{\psi}_2$ by using (6) and (7).

Assuming that we have obtained $y_1$, $y_2$, $y_3$ and $z$ by integrating (8) from $x = a$ to $b$ subject to the appropriate initial conditions, the solution $\psi$ can be determined in the following manner. We note that (4) can be rewritten as a homogeneous system of three linear equations of the form

$$
\begin{bmatrix}
(\psi - \psi_0) & \psi_1 & \psi_2 \\
(\psi - \psi_0)' & \psi_1' & \psi_2' \\
(\psi - \psi_0)'' & \psi_1'' & \psi_2''
\end{bmatrix}
\begin{bmatrix}
1 \\ -\alpha \\ -\beta
\end{bmatrix}
= 0
\tag{9}
$$

The existence of a non-trivial solution to (9) immediately implies that $\psi$ must satisfy the "auxiliary" equation

$$
y_1\psi'' - y_2\psi' + y_3\psi = z \ . \tag{10}
$$

Hence to determine $\psi$, we integrate (10) backwards from $x = b$ to $a$ subject to the prescribed boundary conditions at $x = b$. Moreover, it can easily be shown [2] that the solution $\psi$ thus obtained is indeed the solution of the boundary-value problem (1)-(3).

The foregoing analysis can now be adapted to deal with boundary-value problems involving a second-order equation of the form

$$
L_2(\phi) = \phi'' - c_1\phi' - c_2\phi = f \tag{11}
$$

together with separated boundary conditions at $x = a$ and $b$. If we now let $\phi = \psi'$, then Eq. (11) becomes

$$
\psi''' - c_1\psi'' - c_2\psi' = f \tag{12}
$$

and our discussion of third-order problems is then directly applicable on setting $c_3 \equiv 0$. Moreover, a further simplification is possible if we assume, as is often the case, that the boundary condition at $x = a$ is imposed on either $\phi$ or $\phi'$. In that case we have $y_3(a) = 0$ and, with $c_3 \equiv 0$, the third of Eqs. (8) shows that $y_3(x) \equiv 0$. Thus, it is sufficient to integrate the remaining three of Eqs. (8) from $x = a$ to $b$ to determine $y_1$, $y_2$, and $z$. The solution $\phi$ can then be obtained by integrating the first-order equation.

$$
y_1\phi' - y_2\phi = z \tag{13}
$$

from $x = b$ to $a$.

III. EXAMPLES AND DISCUSSION In order to demonstrate the effectiveness of the procedure just described, we have used it to compute the solutions for $-1 \leq x \leq 1$ of the equations [8]

$$\varepsilon^2\phi'' + x\phi' = 0, \tag{14}$$

$$\varepsilon^2\phi'' + (x \cos x)\phi' + (x + x^3)\phi = 0, \tag{15}$$

$$\varepsilon^2\phi'' + x\phi' - \phi = 0, \tag{16}$$

$$\varepsilon^2\phi'' + x\phi' - \frac{1}{2}\phi = 0, \tag{17}$$

$$\varepsilon^2\phi'' + |x|\phi' + (x - \frac{1}{2})^3\phi = 0, \tag{18}$$

and

$$\varepsilon^2\phi'' + (x^2 - \frac{1}{2})\phi' + x\phi = 0, \tag{19}$$

each subject to the boundary conditions

$$\phi(-1) = 1 \text{ and } \phi(1) = 2 . \tag{20}$$

Our results for $\varepsilon = 10^{-2}$ are shown in Figures 1 through 6 and they were found to be in good agreement with those of Pearson [8]. We note that the solutions of (14) and (15) exhibit internal boundary-layers, while those of (16) and (17) possess corner layers near $x = 0$. The solutions of (18) and (19) display boundary-layer behavior in the interior of the interval as well as at the left end-point.

All integrations were carried out in double-precision arithmetic on an IBM 3081 computer using a fourth-order Runge-Kutta routine with a constant stepsize $h = 0.0005$. The need to choose such a small value of h is the consequence of the use of a constant step integration scheme and the rapid variation of the solutions within the boundary-layer regions. We believe, however, this limitation can in part be overcome with the use of a more sophisticated variable step integration scheme, although this in turn would complicate somewhat the implementation of the method from a programming standpoint. It is of interest to note that the FORTRAN program used to obtain the present results contains fewer than 120 lines of codes. This then suggests that the compound matrix method can provide a simple alternative to other initial-value techniques such as orthonormalization at least for a certain class of problems of the singular perturbation type.

We also note that in constrast to other initial-value techniques, the compound matrix method first derives a lower-order auxiliary equation whose solutions automatically satisfy the boundary condition at one of the two end-points. The solution to the original boundary-value problem is then obtained by integrating the auxiliary equation subject to the boundary condition at the second end-point. This has the effect of pinning down the solution at both end-points. In this respect, the compound matrix method is

not unlike other boundary-value techniques such as finite-differences and indeed this is the chief reason for its effectiveness in dealing with unstable boundary-value problems.

## ACKNOWLEDGEMENTS

This work has been supported in part by the National Science Foundation under grants MCS81-01932 (B.S.N.) and MCS83-01125 (W.H.R.). The work of B.S.N. was done partly during his visit to the Rensselaer Polytechnic Institute in the Spring of 1984.

## REFERENCES

1. B.S. Ng and W.H. Reid, "An Initial Value Method for Eigenvalue Problems Using Compound Matrices," J. Computational Phys., Vol. 30 (1979), pp. 125-136.

2. B.S. Ng and W.H. Reid, "A Numerical Method for Linear Two-Point Boundary-Value Problems Using Compound Matrices," J. Computational Phys., Vol. 33 (1979), pp. 70-85.

3. B.S. Ng and W.H. Reid, "The Compound Matrix Method for Ordinary Differential Systems," J. Computational Phys. (to appear).

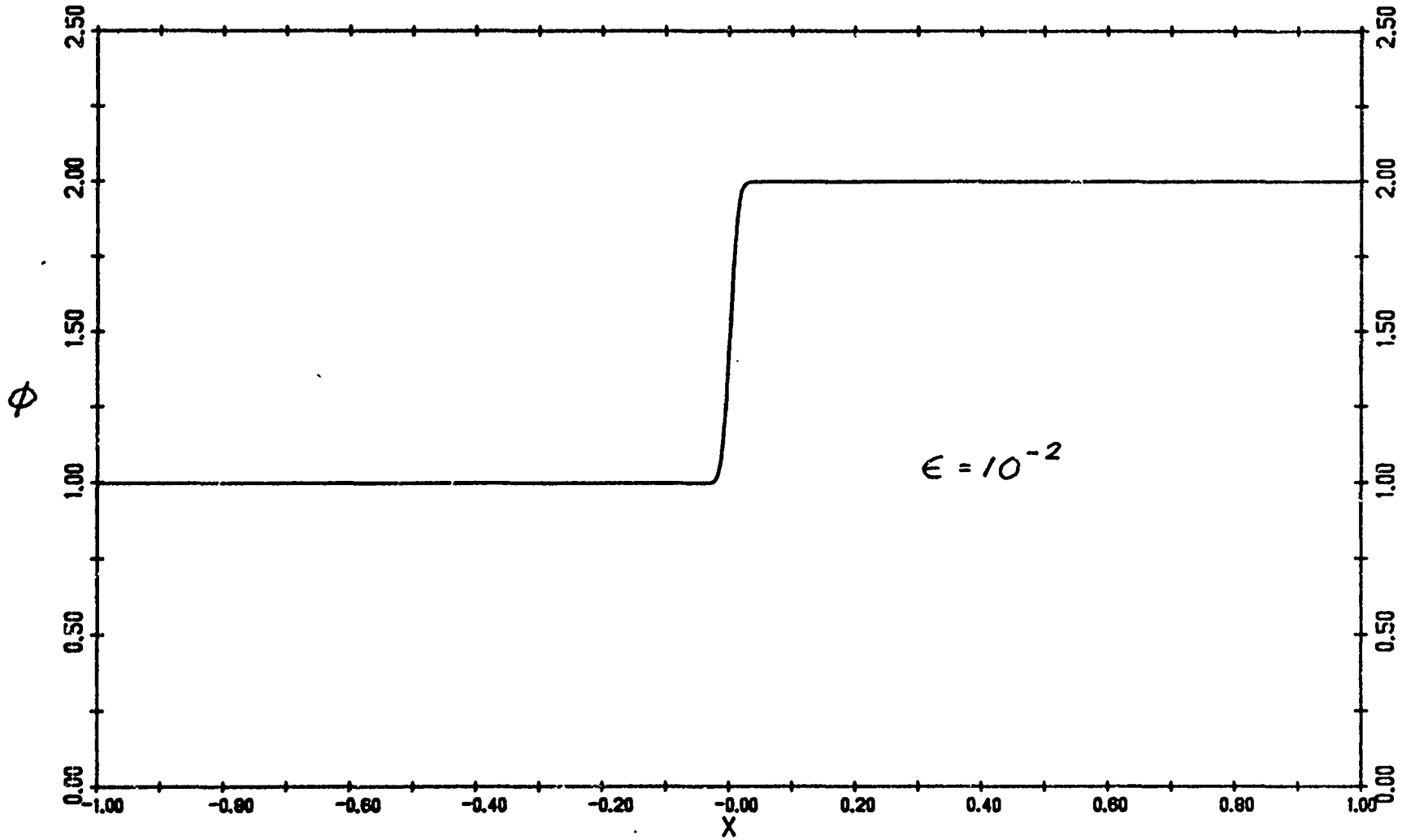
4. B.S. Ng and W.H. Reid, "On the Numerical Solution of the Orr-Sommerfeld Problem: Asymptotic Initial Conditions for Shooting Methods," J. Computational Phys., Vol. 38 (1980), pp. 275-293.

5. B.S. Ng and E.R. Turner, "On the Linear Stability of Spiral Flow Between Rotating Cylinders," Proc. R. Soc. Lond., Vol. A382 (1982), pp. 83-102.

6. R.J. Bodonyi and W.S. Ng, "On the Stability of the Similarity Solutions for Swirling Flow Above an Infinite Rotating Disk," J. Fluid Mech., Vol. 144 (1984), pp. 311-328.

7. R.C. DiPrima, P.M. Eagles and B.S. Ng, "The Effect of Radius Ratio on the Stability of Couette Flow and Taylor Vortex Flow," Phys. Fluids (to appear).

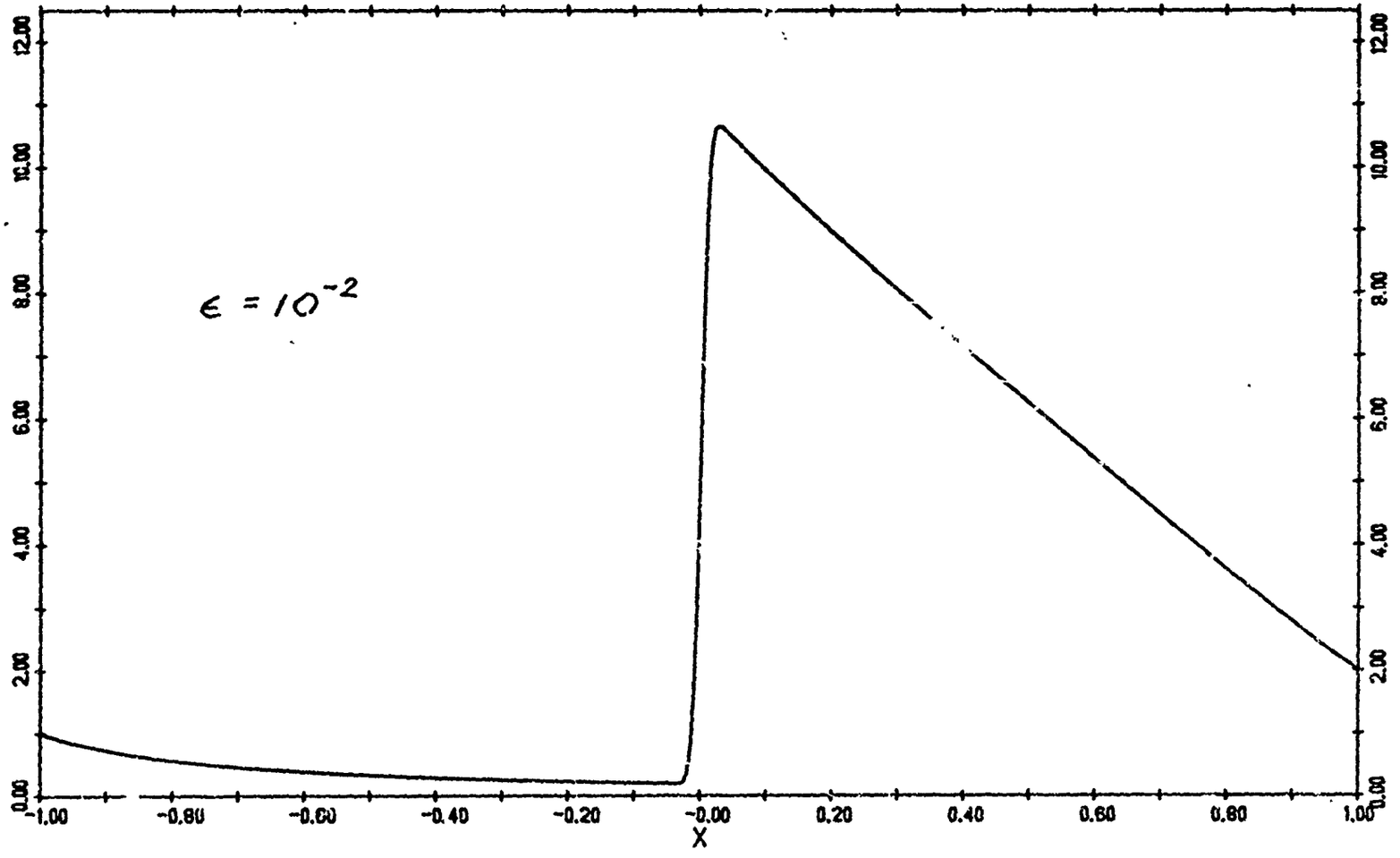
8. C.E. Pearson, "On a Differential Equation of Boundary Layer Type," J. Math. and Phys., Vol. 47 (1968), pp. 134-154.




$$\epsilon^2 \phi'' + x\phi' = 0$$

$$\phi(-1) = 1, \quad \phi(1) = 2$$

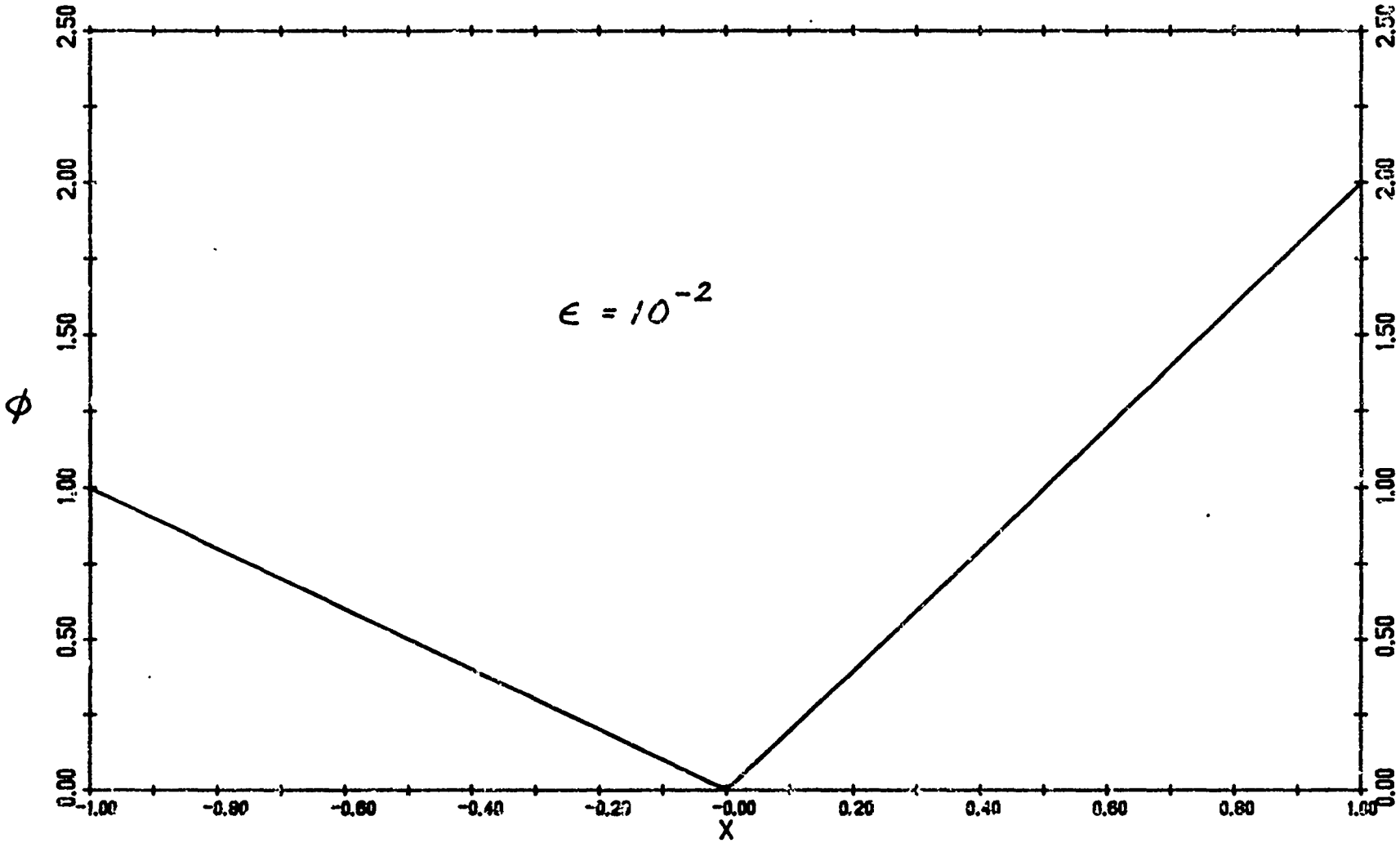
Figure 1




$$\epsilon^2\phi'' + (x\cos x)\phi' + (x+x^3)\phi = 0$$

$$\phi(-1) = 1, \quad \phi(1) = 2$$

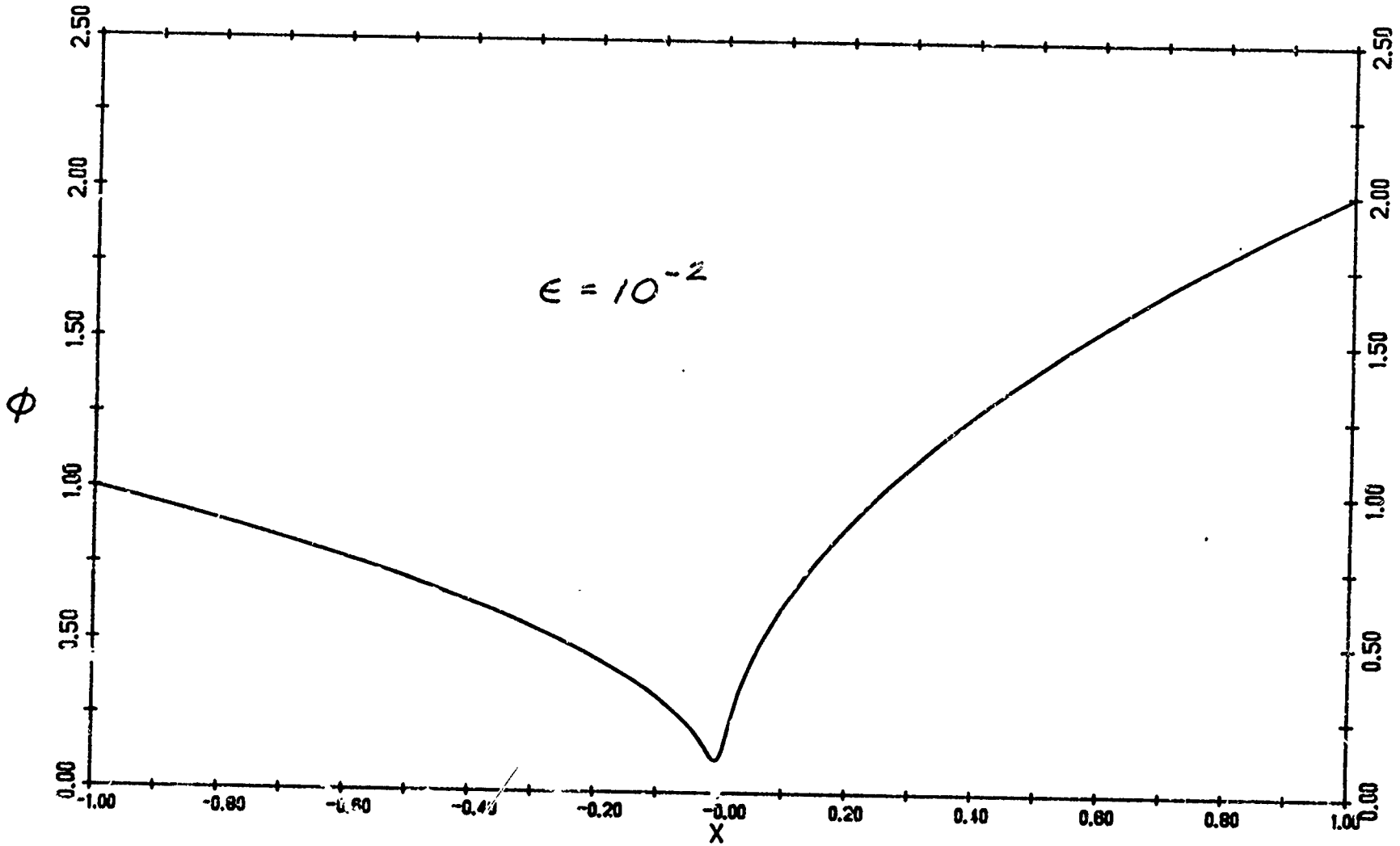
Figure 2




$$\epsilon^2 \phi'' + x\phi' - \phi = 0$$

$$\phi(-1) = 1, \quad \phi(1) = 2$$

Figure 3

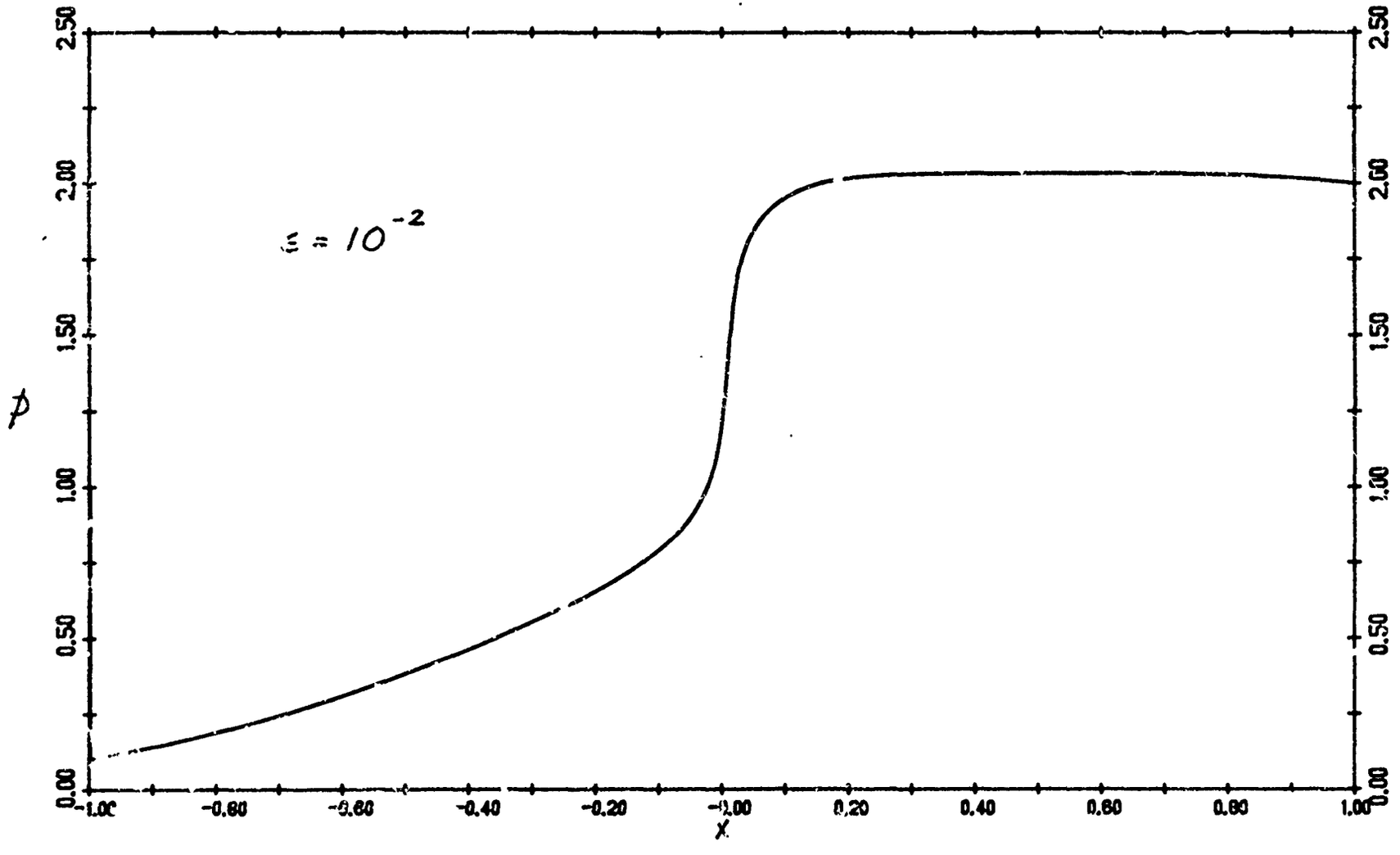





$$\epsilon^2 \phi'' + x\phi' - \frac{1}{2}\phi = 0$$

$$\phi(-1) = 1, \quad \phi(1) = 2$$

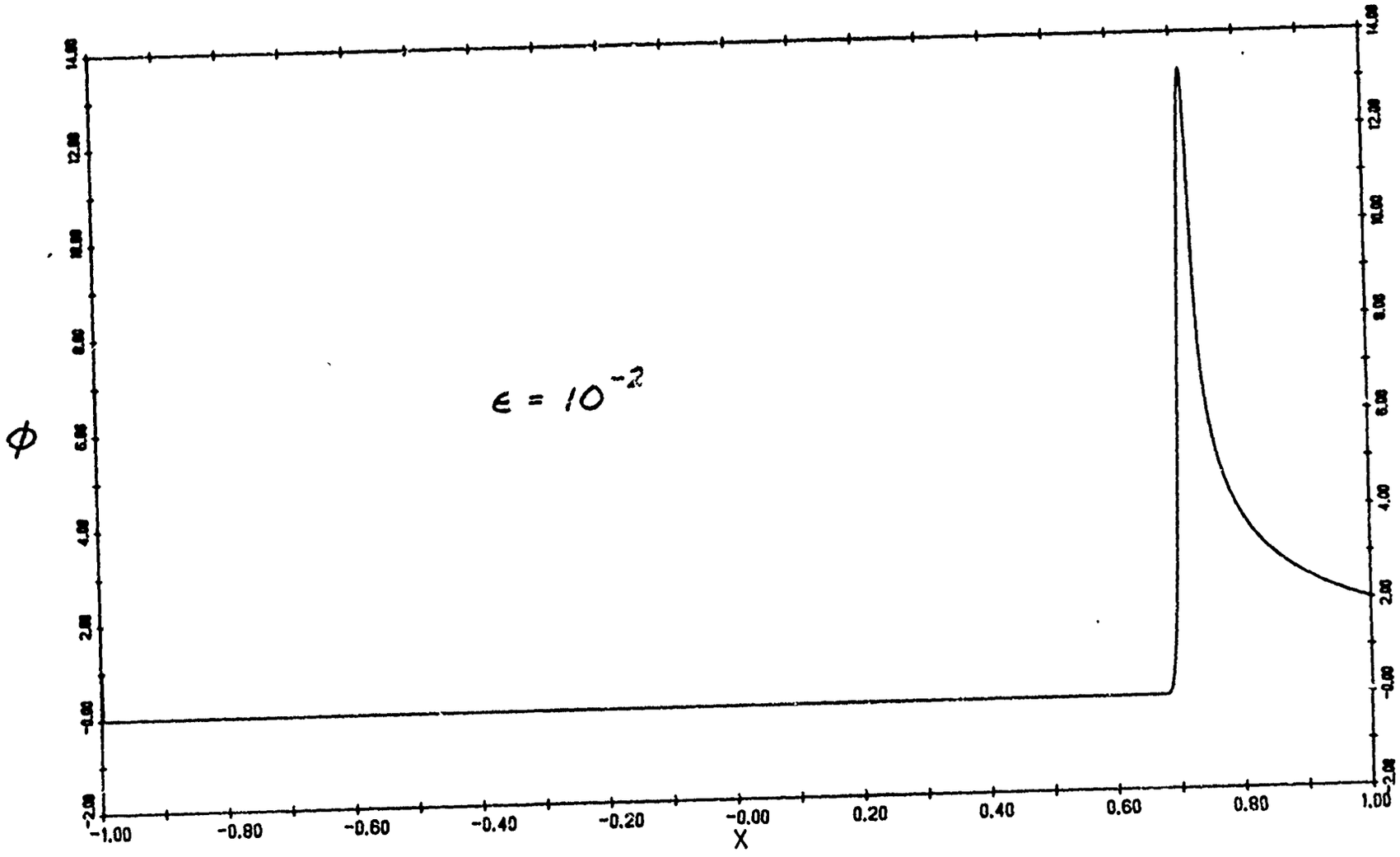
Figure 4




$$\epsilon^2\phi'' + |x|\phi' + (x-\tfrac{1}{2})^3\phi = 0$$

$$\phi(-1) = 1, \quad \phi(1) = 2$$

Figure 5




$$\epsilon^2\phi'' + (x^2 - \tfrac{1}{2})\phi' + x\phi = 0$$

$$\phi(-1) = 1, \quad \phi(1) = 2$$

Figure 6

# SELF SIMILAR SOLUTIONS FOR A DEGENERATE CAUCHY PROBLEM

Klaus Höllig and John A. Nohel
Mathematics Research Center
University of Wisconsin-Madison
Madison, WI 53705

ABSTRACT. We determine the self-similar solutions of the Cauchy problem

$$\begin{aligned} v_t &= \phi(v)_{xx}, \quad x \in R,\ t > 0 , \\ v(x,0) &= g(x) \end{aligned} \tag{P}$$

for the constitutive function $\phi(v) = \max(0,v)$ and the model datum

$$g(x) = \begin{cases} (p_+)x^\gamma, & x \geq 0 \\ -(p_-)|x|^\gamma, & x \leq 0 \end{cases} \tag{D}$$

where $\gamma, p_\pm > 0$. It is shown that the unique solution of (P), (D) is

$$v(x,t) = \begin{cases} -(p_-)|x|^\gamma, & x < -\kappa\sqrt{t} \\ t^{\gamma/2}\psi(\frac{x}{\sqrt{t}}), & x > -\kappa\sqrt{t} \end{cases}$$

where

$$\psi(\xi) = \left[b_1(\kappa)D_{-\gamma-1}(\frac{\xi}{\sqrt{2}}) + b_2(\kappa)D_\gamma(\frac{i\xi}{\sqrt{2}})\right]\exp(-\frac{\xi^2}{8}) ,$$

$$b_1(\kappa) = \frac{p_-\kappa^{\gamma+1}D_\gamma(-\frac{i\kappa}{\sqrt{2}})\exp(\frac{\kappa^2}{8})}{i\sqrt{2}\,\exp(\frac{\gamma+1}{2}\pi i)} ,$$

$$b_2(\kappa) = -\frac{p_-\kappa^{\gamma+1}D_{-\gamma-1}(-\frac{\kappa}{\sqrt{2}})\exp(\frac{\kappa^2}{8})}{i\sqrt{2}\,\exp(\pi i\,\frac{\gamma+1}{2})}$$

and $\kappa$ is implicitly determined by the equation

$$p := p_+/p_- = (\frac{\kappa}{\sqrt{2}})^{\gamma+1}D_{-\gamma-1}(-\frac{\kappa}{\sqrt{2}})\exp(\frac{\kappa^2}{8}) ;$$

$D_\nu(\cdot)$ denotes the parabolic cylinder function of index $\nu$.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.



PREVIOUS PAGE IS BLANK

1. INTRODUCTION AND RESULT. In this note we study certain aspects of the degenerate Cauchy problem

$$\begin{aligned} &v_t = \phi(v)_{xx}, \quad x \in R,\ t > 0 , \\ &v(x,0) = g(x) \end{aligned} \tag{P}$$

for the constitutive function $\phi(v) = \max(v,0)$. We assume that the initial data $g$ are smooth on $R \setminus \{0\}$ with at most polynomial growth at infinity and satisfy

$$xg(x) > 0, \quad x \neq 0 ,$$

$$g(0) = 0 .$$

Problems of this type arise as convexifications of diffusion equations with nonmonotone constitutive functions as has been discussed in [HN1]. The behavior of solutions for (P) is similar to the one phase Stefan problem where $g(x) \equiv -1$ for $x < 0$. Existence and uniqueness of weak solutions of (P) follows from nonlinear semigroup theory [BCP,E]. Moreover, using standard approximation arguments one can show the existence of a continuous monotone decreasing free boundary $t \to s(t)$ where $v(s(t)^+,t) = 0$.

The pair $(v,s)$ satisfies the free boundary problem

$$\begin{aligned} &v_t = v_{xx}, \quad x > s(t),\ t > 0 , \\ &v(x,0) = g(x), \quad x > 0 , \\ &v(s(t)^+,t) = 0 , \\ &g(s(t))s'(t) = v_x(s(t)^+,t) , \\ &s(0) = 0 . \end{aligned} \tag{$\bar{P}$}$$

Conversely, the solution $v$ of ($\bar{P}$) extended by $v(x,t) = g(x)$ for $x < s(t)$ is a weak solution of (P).

We are interested in the regularity and the qualitative behavior of the free boundary for small $t$. Here we consider only the model datum

$$g(x) = \begin{cases} p_+ x^\gamma, & x > 0 , \\ -p_- |x|^\gamma, & x < 0 , \end{cases} \tag{D}$$

where $p_\pm, \gamma > 0$ are given constants. For this case problem (P) has self-similar solutions which can be determined explicitly. The problem (P) with more general initial data will be included in a joint paper with J. Vazquez [HNV] where we use such self-similar solutions as comparison functions.

Proposition. For the model datum (D) problem ($\bar{P}$) has the unique self-similar solution

$$v(x,t) = t^{\gamma/2}\psi\left(\frac{x}{\sqrt{t}}\right), \quad x > s(t),\ t > 0$$

where $\psi(\cdot)$ is the unique solution of the ordinary differential equation (1) below, satisfying the initial conditions (2) and condition (3) at infinity. The free boundary is $s(t) = -\kappa\sqrt{t}$, $t > 0$. For $p := p_+/p_- > 0$ given, $\kappa > 0$ is the unique solution of the equation

$$p = \left(\frac{\kappa}{\sqrt{2}}\right)^{\gamma+1} D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right)\exp\left(\frac{\kappa^2}{8}\right), \tag{E}$$

where $D_{-\gamma-1}(\cdot)$ is the parabolic cylinder function of index $(-\gamma - 1)$.

In the forthcoming paper with Vazquez [HNV] we will use the above Proposition to analyse (P) for more general constitutive functions and more general data; we assume there that $\phi$ is smooth on $[0,\infty)$ with $0 < c \leq \phi' \leq C$, $\phi(v) \equiv 0$ for $v \leq 0$ and that the more general data satisfy

$$G(x) = g(x) + o(|x|^{\gamma}) \qquad (|x| \to 0),$$

where $g$ is the model datum. Then it will be shown that

$$s(t) = -\kappa\sqrt{\phi'(0^+)t} + o(\sqrt{t}) \qquad (t \searrow 0)$$

with $\kappa$ defined as before. For $\phi(v) = \max(0,v)$ and $\gamma = 1$ we proved in [HN2,3] the stronger result

$$s(t) = -\kappa\sqrt{t} + O(t^{1/2+\alpha}) \qquad (t \searrow 0)$$

for any $\alpha < 1/2$. In this case (E) reduces to

$$p = \frac{\kappa^2}{2} + \frac{\kappa^3}{4}\exp\left(\frac{\kappa^2}{4}\right)\int_{-\kappa}^{\infty}\exp(-y^2/4)dy\ .$$

Note, that a first order expansion of the equation $g(s(t))s'(t) = v_x(s(t),t)$ in ($\bar{P}$) formally yields a different result, namely

$$(p_-)s(t)s'(t) = (p_+) + \cdots,$$

which yields

$$s(t) = -\sqrt{2pt} + \cdots,$$

whereas, e.g. for $p = 1$, $\kappa = .9034\ldots \neq \sqrt{2}$. The reason for this apparent inconsistency is that all derivatives of $v$ become singular at $(x,t) = (0,0)$; in particular $v_x$ is not continuous at this point.

2. PROOF OF THE PROPOSITION. Substituting $v(x,t) = t^{\gamma/2}\psi(x/\sqrt{t})$ in (P) one sees that $\psi$ must satisfy the linear differential equation

$$2\psi''(\xi) + \xi\psi'(\xi) - \gamma\psi(\xi) = 0 \text{ for } \xi > -\kappa , \tag{1}$$

subject to the initial conditions

$$\psi(-\kappa) = 0, \ \psi'(-\kappa) = (p_-)\frac{\kappa^{\gamma+1}}{2} \tag{2}$$

and $\kappa$ is related to $p_\pm$ via

$$\lim_{\xi\to+\infty} \xi^{-\gamma}\psi(\xi) = p_+ . \tag{3}$$

The free boundary is given by

$$s(t) = -\kappa\sqrt{t} .$$

Equation (1) can be solved explicitly. Put $x = \xi/\sqrt{2}$ and $w(\xi) = \psi(x)$. Then (1) becomes

$$w''(x) + xw'(x) - \gamma w(x) = 0 . \tag{4}$$

Setting $w(x) =: y(x)\exp(-x^2/4)$ we obtain

$$y''(x) - (\frac{1}{2} + \gamma + \frac{x^2}{4})y(x) = 0 . \tag{5}$$

This differential equation has the general solution [B, p. 116-117]

$$y(x) = b_1 D_{-\gamma-1}(x) + b_2 D_\gamma(ix) \quad (-\infty < x < \infty, \ \gamma > 0) ,$$

where $D_\nu(\cdot)$ is the parabolic cylinder function of index $\nu$. Thus the general solution of (1) is

$$\psi(\xi) = [b_1 D_{-\gamma-1}(\frac{\xi}{\sqrt{2}}) + b_2 D_\gamma(\frac{i\xi}{\sqrt{2}})]\exp(\frac{-\xi^2}{8}) \tag{6}$$

for $-\infty < \xi < \infty$ and $\gamma > 0$. To impose the initial conditions (2) we need the formulae (above ref. p. 119)

$$\frac{d}{d\xi}\left[D_{-\gamma-1}(\frac{\xi}{\sqrt{2}})\exp(\frac{-\xi^2}{8})\right] = \frac{-1}{\sqrt{2}} D_{-\gamma}(\frac{\xi}{\sqrt{2}})\exp(\frac{-\xi^2}{8}) ,$$

$$\frac{d}{d\xi}\left[D_\gamma(\frac{i\xi}{\sqrt{2}})\exp(\frac{-\xi^2}{8})\right] = \frac{i\gamma}{\sqrt{2}} D_{\gamma-1}(\frac{i\xi}{\sqrt{2}})\exp(\frac{-\xi^2}{8}) .$$

Then the initial conditions (2) yield the pair of equations

$$b_1 D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right) + b_2 D_\gamma\left(-\frac{i\kappa}{\sqrt{2}}\right) = 0$$
$$-b_1 D_\gamma\left(-\frac{\kappa}{\sqrt{2}}\right) + i\gamma b_2 D_{\gamma-1}\left(-\frac{i\kappa}{\sqrt{2}}\right) = p_- \frac{\kappa^{\gamma+1}}{\sqrt{2}} \exp\left(\frac{\kappa^2}{8}\right) . \tag{7}$$

Because (5) is of self-adjoint form the Wronskian of $D_\gamma(\cdot)$, $D_{-\gamma-1}(\cdot)$ is constant,

$$W(D_{-\gamma-1}(\cdot), D_\gamma(\cdot)) \equiv -i\exp\left[\left(\frac{\gamma+1}{2}\right)\pi i\right] .$$

Thus

$$b_1(\kappa) = \frac{(p_-)\kappa^{\gamma+1} D_\gamma\left(-\frac{i\kappa}{\sqrt{2}}\right)\exp\left(\frac{\kappa^2}{8}\right)}{i\sqrt{2}\,\exp\left[\left(\frac{\gamma+1}{2}\right)\pi i\right]}$$
$$b_2(\kappa) = -\frac{(p_-)\kappa^{\gamma+1} D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right)\exp\left(\frac{\kappa^2}{8}\right)}{i\sqrt{2}\,\exp\left[\left(\frac{\gamma+1}{2}\right)\pi i\right]} , \tag{8}$$

and (6) with $b_1, b_2$ given by (8) is the solution of (1) satisfying the initial conditions (2). To compute the limit in (3) we use (see above ref. p. 122)

$$D_\nu(z) = z^\nu \exp\left(-\frac{z^2}{4}\right)[1 + O(|z|^{-2})] \quad \text{as} \quad |z| \to \infty , \tag{9}$$

which is valid for $-\frac{3\pi}{4} < \arg z < \frac{3\pi}{4}$. Thus for $\xi \in \mathbf{R}$, $\gamma > 0$,

$$D_{-\gamma-1}\left(\frac{\xi}{\sqrt{2}}\right) = \exp\left(-\frac{\xi^2}{8}\right)\left(\frac{\xi}{\sqrt{2}}\right)^{-\gamma-1}[1 + O(|\xi|^{-2})], \quad \xi \to +\infty ,$$
$$\left|D_\gamma\left(\frac{i\xi}{\sqrt{2}}\right)\right| = \exp\left(\frac{\xi^2}{8}\right)\left(\frac{\xi}{\sqrt{2}}\right)^{\gamma}\left|\exp\left(\frac{i\gamma\pi}{2}\right)\right|[1 + O(|\xi|^{-2})], \quad \xi \to +\infty . \tag{10}$$

Substitution of (10) and (8) into the general solution (6) yields

$$\psi(\xi) \sim b_2(\kappa)\exp\left(\frac{i\gamma\pi}{2}\right)\left(\frac{\xi}{\sqrt{2}}\right)^{\gamma}[1 + O(1)], \quad \xi \to +\infty . \tag{11}$$

From formula (8) we see that

$$b_2(k)\exp\left(\frac{i\gamma\pi}{2}\right) = \frac{p_-}{\sqrt{2}}\kappa^{\gamma+1} D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right)\exp\left(\frac{\kappa^2}{8}\right) . \tag{12}$$

Imposing the asymptotic condition (3) and using (11), (12) we finally obtain

$$\lim_{\xi \to +\infty} \frac{\psi(\xi)}{\xi^{\gamma}} = p_+ = p_-\left(\frac{\kappa}{\sqrt{2}}\right)^{\gamma+1} D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right)\exp\left(\frac{\kappa^2}{8}\right) \tag{13}$$

which yields the equation (E).

To complete the proof of the Proposition we have to show that given any $p > 0$, (E) is uniquely solvable for $\kappa$. From [BO, p. 573]

$$D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right) = \frac{\sqrt{\pi}}{2^{\frac{\gamma+1}{2}}\,\Gamma\left(1+\frac{\gamma}{2}\right)} \sum_{n=0}^{\infty} \frac{a_{2n}}{(2n)!}\left(\frac{\kappa}{\sqrt{2}}\right)^{2n} + \frac{\sqrt{\pi}}{2^{\frac{\gamma}{2}}\,\Gamma\left(\frac{1+\gamma}{2}\right)} \sum_{n=0}^{\infty} \frac{a_{2n+1}}{(2n+1)!}\left(\frac{\kappa}{\sqrt{2}}\right)^{2n+1}, \tag{14}$$

is an analytic function of $\kappa$, $-\infty < \kappa < \infty$, $\gamma > -1$; $a_0 = a_1 = 1$, $a_{n+2} = \left(\gamma + \frac{1}{2}\right)a_n + \frac{n}{4}(n-1)a_{n-2}$, and $D_{-\gamma-1}(0) = \frac{\sqrt{\pi}}{2^{(\gamma+1)/2}\Gamma\left(1+\frac{\gamma}{2}\right)}$. Since the coefficients $a_\nu$ are positive, $D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right)$ is a positive, strictly increasing function of $\kappa$ for $0 \leq \kappa < \infty$, and by (14) so is $p(\kappa)$. Moreover $p(0) = 0$. Also using (14) in (E)

$$p(\kappa) \sim \frac{\sqrt{\pi}}{2^{\frac{\gamma+1}{2}}\,\Gamma\left(1+\frac{\gamma}{2}\right)}\left(\frac{\kappa}{\sqrt{2}}\right)^{\gamma+1} \qquad (\kappa \to 0^+) .$$

Moreover, [BO, p. 574]

$$D_{-\gamma-1}\left(-\frac{\kappa}{\sqrt{2}}\right) \sim \frac{\sqrt{2\pi}}{\Gamma(1+\gamma)}\left(\frac{\kappa}{\sqrt{2}}\right)^{\gamma}\exp\left(\frac{\kappa^2}{8}\right) \qquad (\kappa \to +\infty) ,$$

and therefore, from (E)

$$p(\kappa) \sim \frac{\sqrt{2\pi}}{\Gamma(1+\gamma)}\left(\frac{\kappa}{\sqrt{2}}\right)^{2\gamma+1}\exp\left(\frac{\kappa^2}{4}\right) \qquad (\kappa \to +\infty) .$$

## REFERENCES


[B] Bateman Manuscript Project, A. Erdélyi, ed., McGraw Hill, 1954.

[BO] C. Bender and S. Orszag, Advanced Mathematical Methods for Scientists and Engineers, McGraw Hill, 1978.

[BCP] P. Benilan, M. G. Crandall and A. Pazy, M-Accretive operators, to appear.

[E] L. C. Evans, Application of nonlinear semigroup theory to certain partial differential equations, in: Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, 1978.

[H] K. Höllig, Existence of infinitely many solutions for a forward backward heat equation, Trans. Amer. Math. Soc. 278 (1983), 299-316.

[HN1] K. Höllig and J. A. Nohel, A diffusion equation with a nonmonotone constitutive function, Proceedings NATO/LONDON Math. Soc. Conf. on Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., Reidel Publishing Co. (1983), 409-422.

[HN2] K. Höllig and J. A. Nohel, A nonlinear integral equation occurring in a singular free boundary problem, Trans. Amer. Math. Soc., to appear.

[HN3] K. Höllig and J. A. Nohel, A singular free boundary problem, MRC Technical Summary Report #2582, Mathematics Research Center, University of Wisconsin-Madison.

[HNV] K. Höllig, J. A. Nohel and J. Vazquez, in preparation.

# VARIATIONAL PRINCIPLE FOR PENETRATOR DYNAMICS USING BILINEAR FUNCTIONAL AND ADJOINT FORMULATION

C. N. Shen
U.S. Army Armament, Munitions, and Chemical Command
Aramament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. The solution problems in both spatial and time domains using finite element method can be based on the variational principle employing bilinear functional and adjoint formulation. This principle is extended to coupling systems in matrix vector form such as penetration dynamics. The present hyperbolic type partial differential equation of interest has two dependent and two independent variables with the coupling in the spatial domain.

I. INTRODUCTION. The elastic-plastic stress-strain relations for a rod derived by T. Wright and the differential equations of the rod itself have been cast into variational form. The variational principle using bilinear functional and adjoint formulation has served as bases to numerical solutions by the finite element method. This principle has now been extended to coupling systems such as in the impact dynamics. The present hyperbolic type partial differential equation has two dependent and two independent variables, with coupling in the spatial domain. This new formulation is also ready to be used for the coupled impact problems which is given in the appendix of this paper.

II. THE VARIATIONAL PRINCIPLE. A dynamical system can be modeled by the matrix vector partial differential equation.

$$L(\zeta)\, y(\zeta) = -Q(\zeta) \tag{1}$$

with appropriate boundary and initial conditions. In the above equation, L is a matrix linear operator in both spatial and temporal domain, y is a vector dependent variable, Q is a vector forcing function, and $\zeta$ represents all independent variables, both spatial and temporal.

The inner produce $\langle\ \rangle$ of an adjoint forcing function $\bar{Q}$ and the solution $(y(\zeta))$ of Eq. (1) can be used for the purpose of estimation. This inner product is $\langle \bar{Q}, y \rangle$.

*The author is also employed by Rensselaer Polytechnic Institute, where he holds the title Professor in the Electrical Computer and Systems Engerineering Department.

PREVIOUS PAGE IS BLANK

An accurate estimation can be made by constructing a variational principle [1]. By using the adjoint variable $\bar{y}$ as a Lagrange multiply for Eq. (1) adding to $\langle \bar{Q}, y \rangle$, we have

$$J_1[y,\bar{y}] \overset{\Delta}{=} \langle \bar{Q}, y \rangle + \langle \bar{y}, (Q+Ly) \rangle = \langle \bar{Q}, y \rangle + \langle \bar{y}, Q \rangle + \langle \bar{y}, Ly \rangle \tag{2}$$

To keep the system symmetrical, let us define the adjoint system as

$$\bar{L}(\xi)\bar{y}(\xi) = -\bar{Q}(\xi) \tag{3}$$

By using the original variable y as a Lagrange multiply for Eq. (3) adding to $\langle Q, \bar{y} \rangle$, we have

$$J_2[y,\bar{y}] \overset{\Delta}{=} \langle \bar{Q}, y \rangle + \langle y, (\bar{Q}+\bar{L}\bar{y}) \rangle = \langle \bar{Q}, y \rangle + \langle y, \bar{Q} \rangle + \langle y, \bar{L}\bar{y} \rangle \tag{4}$$

The relationship of the adjoint system to the original system is

$$D \overset{\Delta}{=} \langle \bar{y}, Ly \rangle - \langle y, \bar{L}\bar{y} \rangle = D_e \tag{5}$$

where D is the bilinear concomitant [1]. Combining Eqs. (2), (4), and (5) one obtains

$$J_1 = J_2 + D_e \tag{6a}$$

In order to keep the functional symmetrical, we have

$$J_1 \overset{\Delta}{=} \frac{1}{2}\,[J_1 + J_2 + D_e] \tag{6b}$$

which is of the form

$$J_1 = \langle \bar{Q}, y \rangle + \langle \bar{y}, Q \rangle + \frac{1}{2}\langle \bar{y}, Ly \rangle + \frac{1}{2}\langle y, \bar{L}\bar{y} \rangle + \frac{D_e}{2} \tag{6c}$$

Similarly,

$$J_2 = \langle \bar{Q}, y \rangle + \langle \bar{y}, Q \rangle + \frac{1}{2}\langle \bar{y}, Ly \rangle + \frac{1}{2}\langle y, \bar{L}\bar{y} \rangle - \frac{D_e}{2} \tag{6d}$$

III. INTEGRAL OF BILINEAR EXPRESSION. The integral of a bilinear expression for a two-dimensional problem having a system of second order partial derivatives in time and in space can be written as

$$I = \int_{x_o}^{x_b} \int_{t_o}^{t_b} \bar{\Omega}[y(x,t), \bar{y}(x,t)]\,dt\,dx \tag{7}$$

where $\bar{\Omega}[y,\bar{y}]$ is a given bilinear expression in the form

$$\bar{\Omega}[y,\bar{y}] = \bar{y}_t^T \beta y_t - \bar{y}_x^T A y_x - \bar{y}^T P y + \bar{y}_x^T \Gamma y + \bar{y}^T N y_x \tag{8}$$

The subscripts t and x indicate the partial derivatives for the functions y and $\bar{y}$. The matrices A, B, and P are diagonal and $\Gamma$ and N are off-diagonal.

$$A = \begin{bmatrix} a_1^2 & 0 \\ 0 & a_2^2 \end{bmatrix}, \quad B = \begin{bmatrix} b_1^2 & 0 \\ 0 & b_2^2 \end{bmatrix}, \quad \text{and } P = \begin{bmatrix} p_1^2 & 0 \\ 0 & p_2^2 \end{bmatrix} \tag{9}$$

$$\Gamma = \begin{bmatrix} 0 & \gamma_1 \\ \gamma_2 & 0 \end{bmatrix}, \quad N = \begin{bmatrix} 0 & \eta_1 \\ \eta_2 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} w \\ u \end{bmatrix} \quad \text{and } \bar{y} = \begin{bmatrix} \bar{w} \\ \bar{u} \end{bmatrix} \tag{10}$$

Equation (8) can be integrated by parts. Two different forms of integration and end conditions are given. The first form of the integral is obtained by integrating by parts on the adjoint variable.

$$I_a = -\iint \{\bar{y}^T B y_{tt} - \bar{y}^T A y_{xx} + \bar{y}^T P y + \bar{y}^T P y_x - \bar{y}^T N y_x\} dt dx$$

$$+ \int \bar{v}^T B y_t dx - \int \bar{y}^T A y_{xx} dt + \int \bar{y}^T P y dt \tag{11}$$

which gives

$$I_a = \iint -\bar{w}(b^2{}_1 w_{tt} - a_1{}^2 w_{xx} + P_1{}^2 w + \gamma_1 u_x - \eta_1 u_x) dt dx$$

$$+ \iint -\bar{u}(b_2{}^2 u_{tt} - a_2{}^2 u_{xx} + P_2{}^2 u + \gamma_2 w_2 - \eta_2 w_x) dx dt$$

$$+ \int_{x_o}^{x_b} [\bar{w} b_1{}^2 w_t + \bar{u} b^2 u_t]_{t_o}^{t_b} dx +$$

$$\int_{x_o}^{x_b} [\bar{w} a_1{}^2 w_x + \bar{u} a_2{}^2 u_x]_{x_o}^{x_b} dt + \int_{t_o}^{t_b} [\bar{w}\gamma_1 u + \bar{u}\gamma_2 w]_{x_o}^{x_b} dt \tag{12}$$

$$I_a = -\langle \bar{y}, Ly \rangle + \int_{x_o}^{x_b} [b_1{}^2 \bar{w} w_t + b_2{}^2 \bar{u} u_t]_{t_o}^{t_b} dx$$

$$- \int_{t_o}^{t_b} [a_1{}^2 \bar{w} w_x + a_2{}^2 \bar{u} u_x]_{x_o}^{x_b} dt + \int_{t_o}^{t_b} [\gamma_1 \bar{w} u + \gamma_2 \bar{u} w]_{x_o}^{x_b} dt \tag{13}$$

On the other hand, we can perform integration on the original variable to give

$$I_b = -\iint \{y^T B \bar{y}_{tt} - y^T A \bar{y}_{xx} + y^T P \bar{y} - y^T \Gamma^T \bar{y}_x + y^T N^T \bar{y}_x\} dt dx$$

$$\int y^T B \bar{y}_t dx - \int y^T A \bar{y}_x dt + \int y^T N^T \bar{y} dt \tag{14}$$

which gives

$$I_b = \iint(-w)(b_1{}^2\bar{w}_{tt}-a_1{}^2\bar{w}_{xx}+p_1{}^2\bar{w}-\gamma_2\bar{u}_x+\eta_2\bar{u}_x)dtdx$$
$$+ \iint(-u)(b_1{}^2\bar{u}_{tt}-a_2{}^2\bar{u}_{xx}+p_2{}^2\bar{u}-\gamma_1\bar{w}_x+\eta_1\bar{w}_x)dtdx$$

$$+ \int_{x_o}^{x_b}[wb_1{}^2\bar{w}_t+ub_2{}^2\bar{u}_t]_{t_o}^{t_b}dx - \int_{t_o}^{t_b}[wa_1{}^2\bar{w}_x+ua^2\bar{u}_x]_{x_o}^{x_b}dt + \int_{t_o}^{t_b}[w\eta_2\bar{u} + u\eta_1\bar{w}]_{x_o}^{x_b}dt \quad (15)$$

$$I_b = -\langle y,\bar{\bar{L}}y\rangle + \int_{x_o}^{x_b}[b_1{}^2\bar{w}\bar{w}_t + b_2{}^2\bar{u}\bar{u}_t]_{t_o}^{t_b}dx - \int_{t_o}^{t_b}[a_1{}^2\bar{w}\bar{w}_x + a_2{}^2\bar{u}\bar{u}_x]_{x_o}^{x_b}dt +$$
$$+ \int_{t_o}^{t_b}[\eta_2\bar{w}\bar{u} + \eta_1\bar{u}\bar{w}]_{x_o}^{x_b}dt \quad (16)$$

To keep the form symmetrical, we take the average of the above two expressions

$$I = \frac{1}{2}I_a + \frac{1}{2}I_b = -\int_{x_o}^{x_b}\int_{t_o}^{t_b}\frac{1}{2}(\bar{y}L\bar{y}+y\bar{\bar{L}}y)dtdx +$$

$$+ \int(\bar{y}^TBy_t+y^T\bar{B}\bar{y}_t)\Big|_{t_o}^{t_b}dx - \frac{1}{2}\int(\bar{y}^TAy_x+y^T\bar{A}\bar{y}_x)\Big|_{x_o}^{x_b}dt + \frac{1}{2}\int(\bar{y}^TPy+y^T\bar{N}^T\bar{y})dt \quad (17)$$

which gives

$$I = -\frac{1}{2}\langle\bar{y},L\bar{y}\rangle - \frac{1}{2}\langle y,\bar{\bar{L}}y\rangle + \frac{1}{2}\int_{x_o}^{x_b}[b_1{}^2(\bar{w}\bar{w}_t+w\bar{w}_t) + b_2{}^2(\bar{u}\bar{u}_t+u\bar{u}_t)]_{t_o}^{t_b}dx$$
$$- \frac{1}{2}\int_{t_o}^{t_b}[a_1{}^2(\bar{w}\bar{w}_x+w\bar{w}_x) + b_2{}^2(\bar{u}\bar{u}_x+u\bar{u}_x)]_{x_o}^{x_b}dt$$
$$+ \frac{1}{2}\int_{t_o}^{t_b}[(\gamma_1+\eta_1)\bar{w}u + (\gamma_2+\eta_2)\bar{u}w]_{x_o}^{x_b}dt \quad (18)$$

where

$$L = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}, \quad \bar{L} = \begin{bmatrix} \bar{L}_{11} & \bar{L}_{12} \\ \bar{L}_{21} & \bar{L}_{22} \end{bmatrix} \quad (19)$$

$$L_{11} = \bar{L}_{11} = b_1{}^2\frac{\partial^2}{\partial t^2} + p_1{}^2 - a_1{}^2\frac{\partial^2}{\partial x^2}, \quad L_{22} = \bar{L}_{22} = b_2{}^2\frac{\partial^2}{\partial t^2} + p_2{}^2 - a_2{}^2\frac{\partial^2}{\partial x^2} \quad (20)$$

$$L_{12} = \bar{L}_{12} = -\frac{\partial}{\partial x}, \quad L_{21} = \bar{L}_{21} = \frac{\partial}{\partial x} \quad (21)$$

It is noted that we have used the following relationship in obtaining Eq. (21)

$$\eta_1 - \gamma_1 = 1 \quad \text{and} \quad \eta_2 - \gamma_2 = -1 \tag{22}$$

by comparing Eqs. (10), (21), and the last equation in the Appendix.

For a spatial and temporal partial system, Eq. (5) becomes

$$D = \int_{x_o}^{x_b} \int_{t_o}^{t_b} \bar{y}L\bar{y}dtdx - \int_{x_o}^{x_b} \int_{t_o}^{t_b} \bar{\bar{y}}Lydtdx \tag{23}$$

By equating Eqs. (11) and (14) and solving for D in Eq. (23), we are converting the double integral into single integrals in terms of the boundary conditions.

We can express the quantity $D_e$ as the sum of three parts for end conditions $D_1$, $D_2$, and $D_3$ as

$$D_e = D_1 + D_2 + D_3 \tag{24}$$

The terms in $D_1$ involve the initial conditions of y and $\bar{y}$ as

$$D_1 = \int_{x_o}^{x_b} (\bar{y}^T B y_t - y^T B \bar{y}_t)\Big|_{t_o}^{t_b} dx$$

$$= \int_{x_o}^{x_b} (b_1{}^2(\bar{w}w_t - w\bar{w}_t) + b_2{}^2(\bar{u}u_t - u\bar{u}_t)]_{t_o}^{t_b} dx \tag{25}$$

The terms in $D_2$ involve the boundary conditions from the second partials of y and $\bar{y}$ as

$$D_2 = -\int_{t_o}^{t_b} (\bar{y}^T A y_x - y^T A \bar{y}_x)\Big|_{x_o}^{x_b} dt$$

$$= -\int_{t_o}^{t_b} [a_1{}^2(\bar{w}w_x - w\bar{w}_x) + a_2{}^2(\bar{u}u_x - u\bar{u}_x)]_{x_o}^{x_b} dt \tag{26}$$

The terms in $D_3$ involve the boundary conditions from the coupling terms.

$$D_3 = \int_{t_o}^{t_b} (\bar{y}^T \Gamma y - y^T N^T \bar{y})dt = \int_{t_o}^{t_b} [(\gamma_1 - \eta_1)\bar{w}u + (\gamma_2 - \eta_2)\bar{u}w]_{x_o}^{x_b} dt \tag{27}$$

IV. THE SYMMETRICAL ADJOINT SYSTEM. The adjoint independent variable $\tau$ in Figure 1 can be expressed as

$$\frac{\tau_b - \tau}{\tau_b - \tau_o} = \frac{t - t_o}{t_b - t_o} \tag{28}$$

which gives

$$\tau = \tau_b \quad \text{for} \quad t = t_o \tag{29a}$$

and

$$\tau = \tau_o \quad \text{for} \quad t = t_b \tag{29b}$$

It is noted from Eq. (28) that

$$\tau_b - \tau_o = \tau_b - t_o \tag{30a}$$

$$\tau = \tau_b + t_o - t \tag{30b}$$

$$d\tau = -dt \tag{30c}$$

$$\frac{d}{d\tau} = \frac{d}{dt} \tag{30d}$$

and

$$\bar{y}(x,t) = y(x, \tau = \tau_b + t_o - t) \tag{30e}$$

Let us assume that the adjoint system shown in Figure 1 gives

$$\bar{y}(x,\hat{t}=t) = \hat{y}(x,\hat{t}=t_b+t_o-t) \tag{31a}$$

$$\bar{y}_t(x,\hat{t}=t) = -\hat{y}_t(x,\hat{t}=t_b+t_o-t) \tag{31b}$$

$$\bar{y}_x(x,\hat{t}=t) = \hat{y}x(x,\hat{t}=t_b+t_o-t) \tag{31c}$$

where $\hat{t}$ is a dummy variable for t.

We may define the adjoint system as the image reflection in the time domain of the original system. Equation (31) yields the following known initial conditions

$$\bar{y}(x,\hat{t}=t_b) = \hat{y}(x,\hat{t}=t_o) \quad \text{(known)} \tag{32a}$$

$$\bar{y}_t(x,\hat{t}=t_b) = -\hat{y}_t(x,\hat{t}=t_o) \quad \text{(known)} \tag{32b}$$

The interpretation of the above equations gives the initial conditions of the original system as the far end conditions for the adjoint system, since the adjoint system is a reflected mirror of the original system in time.

V. INITIAL CONDITIONS FOR THE ADJOINT SYSTEM. We take a symmetry approach for the initial conditions of the adjoint system as

$$\bar{y}(x,t=t_b) = y(x,t-t_o) \quad , \quad \bar{y}_t(x,t-t_b) = -y_t(x,t=t_o) \tag{33}$$

$$\bar{y}(x,t=t_o) = y(x,t-t_b) \quad , \quad \bar{y}_t(x,t-t_o) = -y_t(x,t-t_b) \tag{34}$$

where $y$ and $\bar{y}$ are given in Eq. (10). Thus Eq. (25) becomes

$$D_1 = \int_{x_o}^{x_b} b_1^2 dx\{[w(x,t=t_o)w_t(x,t=t_b) + w(x,t=t_b)w_t(x,t=t_o)]$$

$$- [w(x,t=t_b)w_t(x,t=t_o) + w(x,t=t_o)w_t(x,t=t_b)]\}$$

$$+ \int_{x_o}^{x_b} b_2^2 dx\{u(x,t=t_o)u_t(x,t=t_b) + u(x,t=t_b)u_t(x,t=t_o)]$$

$$- [u(x,t=t_b)u_t(x,t=t_o) + u(x,t=t_o)u_t(x,t=t_b)]\} = 0 \tag{35}$$

Since the integrand of Eq. (35) is zero, the above satisfies Eq. (25).

VI. THE GENERALIZED BOUNDARY CONDITIONS. Let us consider the operator L in Eqs. (19) through (21). It is assumed that elastic springs are installed at the ends such that

$$y_x(x_b,t) = K_b y(x_b,t) \quad , \quad \bar{y}_x(x_b,t) = K_b \bar{y}(x_b,t) \tag{36a}$$

$$y_x(x_o,t) = -K_o y(x_o,t) \quad , \quad \bar{y}_x(x_o,t) = -K_o \bar{y}(x_o,t) \tag{36b}$$

where $K_b$, $K_o$ are diagonal matrices. If Eq. (36) is substituted into Eq. (26), we have

$$D_2 = 0 \tag{37}$$

Since $D_1 = D_2 = 0$, Eq. (24) becomes $D_3$ as given in Eq. (27)

$$D_e = D_3 \tag{38}$$

VII. CONDITIONS FOR THE COUPLING TERMS. The sum of the functionals $J_1$ + I is obtained by adding Eqs. (6c) and (18) as

$$J_1 + I + \int_{x_o}^{x_b} \int_{t_o}^{t_b} (\bar{Q}y + \bar{y}Q)dxdt + T + B + V + \frac{1}{2} D_3 \tag{39}$$

where

$$T = \frac{1}{2} \int_{x_o}^{x_b} [b_1^2(\bar{w}_t w + \bar{w}_t w) + b_2^2(\bar{u}_t u + \bar{u}_t u)]_{t_o}^{t_b} dx \tag{40}$$

$$B = -\frac{1}{2} \int_{t_o}^{t_b} [a_1^2(\bar{w}_x w + \bar{w}_x w) + a_2^2(\bar{u}_x u + \bar{u}_x u)]_{x_o}^{x_b} dt \tag{41}$$

$$V = \frac{1}{2}\int_{t_o}^{t_b} [(\gamma_1+\eta_1)\bar{u}\bar{w} + (\gamma_2\eta_2)\bar{w}\bar{u}]_{x_o}^{x_b} dt \tag{42}$$

and

$$D_3 = \frac{1}{2}\int_{t_o}^{t_b} [(\gamma_1-\eta_1)\bar{u}\bar{w} + (\gamma_2-\eta_2)\bar{w}\bar{u}]_{x_o}^{x_b} dt \tag{43}$$

From the last two equations, one obtains

$$V + \frac{1}{2} D_3 = \int_{t_o}^{t_b} (\gamma_1\bar{u}\bar{w}+\gamma_2\bar{w}\bar{u})]_{x_o}^{x_b} dt \tag{44}$$

We can let Eq. (44) vanish by choosing

$$\gamma_1 = \gamma_2 = 0 \tag{45}$$

which gives

$$V + \frac{1}{2} D_3 = 0 \tag{46}$$

From Eq. (22) one obtains

$$\eta_1 = 1 + \gamma_1 = 1 \tag{46a}$$

$$\eta_2 = 1 + \gamma_2 = -1 \tag{46b}$$

Thus the functional for the original variables and adjoint variations becomes

$$J_1 = -I + \langle\bar{Q},y\rangle + \langle Q,\bar{y}\rangle + T + B \tag{47}$$

The sum of the two functionals $J_2 + I$ is obtained by adding Eqs. (6d) and (18) as

$$J_2 + I = \int_{x_o}^{x_b}\int_{t_o}^{t_b} (\bar{Q}y+\bar{y}Q)dxdt + T + B + V - \frac{1}{2} D_3 \tag{48}$$

where T, B, V, and (1/2)$D_3$ are given in Eqs. (40) through (43). By subtracting (1/2)$D_3$ from V we have

$$V - \frac{1}{2} D_3 = \int_{t_o}^{t_b} (\eta_1\bar{u}\bar{w}+\eta_2\bar{w}\bar{u})]_{x_o}^{x_b} dt \tag{49}$$

In this case we let

$$\eta_1 = \eta_2 = 0 \tag{50}$$

Then from Eq. (22) one obtains

$$\gamma_1 = -1 + \eta_1 = -1 \tag{51}$$

$$\gamma_2 = 1 + \eta_2 = 1 \tag{52}$$

Thus the functional for the adjoint variables and original variations becomes

$$J_2 = -I + \langle \bar{Q}, y \rangle + \langle Q, \bar{y} \rangle + T + W \tag{53}$$

which gives the same form as $J_1$ shown in Eq. (47).

VIII. THE FIRST VARIATION. By taking the variations $\delta\bar{y}$ and $\delta y$ separately, we let

$$\delta J = \delta J(\delta\bar{y}) + \delta J_2(\delta y) = 0 + 0 \tag{54}$$

Then one obtains from Eqs. (40), (41), and (47) that

$$\delta J_1(\delta\bar{y}) = -\delta I(\delta\bar{y}) + \iint Q\delta\bar{y}\, dxdt + \delta T(\delta\bar{y}) + \delta B(\delta\bar{y}) = 0 \tag{55}$$

where

$$\delta T(\delta\bar{y}) = \frac{1}{2}\int_{x_o}^{x_b} [b_1^2(w_t\delta\bar{w}+w\delta\bar{w}_t) + b_2^2(u_t\delta\bar{u}+u\delta\bar{u}_t)]_{t_o}^{t_b} dx \tag{56}$$

$$\delta B(\delta y) = -\frac{1}{2}\int_{t_o}^{t_b} [a_1^2(w_x\delta\bar{w}+w\delta\bar{w}_x) + a_2^2(u_x\delta\bar{u}+u\delta\bar{u}_x)]_{x_o}^{x_b} dt \tag{57}$$

and $-\delta I(\delta\bar{y})$ can be derived from Eq. (18) with $\gamma_1 = \gamma_2 = 0$ and $\eta_1 = \eta_2 = 1$

$$-\delta I(\delta\bar{y}) = -\int_{x_o}^{x_b}\int_{t_o}^{t_b} \{[b_1^2 w_t\delta\bar{w}_t - p_1^2 w\delta\bar{w} - a_1^2 w_x\delta w_x] +$$

$$[b_2^2 u_t\delta\bar{u}_t - p_2^2 u\delta\bar{u} - a_2^2 u_x\delta\bar{u}_x] + [u_x\delta\bar{w} - w_x\delta\bar{u}]\}dtdx \tag{58}$$

The second term on the right side of Eq. (54) is

$$\delta J_2(\delta y) = -\delta I(\delta y) + \iint \bar{Q}\delta y\, dxdt + \delta T(\delta y) + \delta B(\delta y) = 0 \tag{59}$$

where

$$\delta T(\delta y) = \frac{1}{2}\int_{x_o}^{x_b} [b_1^2(\bar{w}\delta w_t+\bar{w}_t\delta w) + b_2^2(\bar{u}\delta u_t+\bar{u}_t\delta u)]_{t_o}^{t_b} dx \tag{60}$$

$$\delta B(\delta y) = -\frac{1}{2}\int_{t_o}^{t_b} [a_1^2(\bar{w}\delta w_x+\bar{w}_x\delta w) + a_2^2(\bar{u}\delta u_x+\bar{u}_x\delta u)]_{x_o}^{x_b} dt \tag{61}$$

and $\delta I(\delta y)$ can be derived from Eq. (18) with $\eta_1 = \eta_2 = 0$ and $-\gamma_1 = \gamma_2 = 1$.

$$-\delta I(\delta y) = -\int_{x_o}^{x_b}\int_{t_o}^{t_b} \{[b_1^2\bar{w}_t\delta w_t - p_1^2\bar{w}\delta w - a_1^2\bar{w}_x\delta w_x] +$$

$$[b_2^2\bar{u}_t\delta u_t - p_2^2\bar{u}\delta u - a_2^2\bar{u}_x\delta u_x] + [-\bar{w}_x\delta u + \bar{u}_x\delta w]\}dtdx \tag{62}$$

The adjoint equation has the same form of the original equation by dropping and adding the bars simultaneously on every variable.

Equations (55) through (58) are the key equations to be used for the finite element method. It is noted that the first variation $\delta J_1(\delta y)$ is the same as the first variation $\delta J_2(\delta y)$ by adding or dropping the bar on top of the variables and their variations. We do not need to solve for the adjoint system in Eqs. (39) through (41) since they give exactly the same solutions as that of the original system.

IX. CONCLUSIONS. The functional in bilinear matrix vector form is symmetrical about the original variables and the adjoint variables. The Euler Lagrange equations for the coupling systems are derived using the fundamental lemma of the calculus of variations. By integrating the bilinear matrix vector expression by parts, one can obtain the bilinear concomitant in terms of initial and boundary terms. The adjoint system can be arranged in a manner that it is a reflected mirror of the original system in time. Thus the initial conditions for the bilinear concomitant become zero. Generalized boundary conditions using many types of "springs" relating the various spatial partial derivatives can be defined to satisfy the boundaries of the concomitant. Algorithms are developed for use in the finite element method by taking the first variations of the functional. There algorithms are simplified because the adjoint system gives exactly the same solutions as that of the original system.

## REFERENCES


1. Stacey, Weston, M., Jr., Variational Methods in Nuclear Reactor Physics, Academic Press, 1974.

2. Shen, C. N. and Wu, Julian, J., "A New Variational Method for Initial Value Problems, Using Piecewise Hermite Polynomial Spline Functions," ARO Report 81-3, Proceedings of the 1981 Army Numerical Analysis and Computers Conference, 1981.

3. Shen, C. N., "Variational Principle for Gun Dynamics With Adjoint Variable Formulation," Proceedings of the Third US Army Symposium on Gun Dynamics, Volume II, May 1982, p. IV-108.

4. Shen, C. N., "On the Extremum of Bilinear Functional for Hyperbolic Type P.D.E.," ARO Report 84-1, Transactions of the First Army Conference on Applied Mathematics and Computing.

## APPENDIX

The wave equation in rods due to T. W. Wright is given as the following system

$$\begin{cases} w_{\xi\xi} + \dfrac{2\lambda}{\lambda+2\mu} u_{\xi} = \dfrac{c^2}{c_1^2} w_{\tau\tau} & \text{(A-1a)} \\ u_{\xi\xi} - \left[\delta \dfrac{\lambda+\mu}{\mu} u + 4 \dfrac{\lambda}{\mu} w_{\xi}\right] = \dfrac{c^2}{c_2^2} u_{\tau\tau} & \text{(A-1b)} \end{cases}$$

which can be transformed to

$$\begin{cases} \left(\dfrac{\lambda+2\mu}{2\lambda} w_{tt} - \dfrac{\lambda+2\mu}{2\lambda} w_{xx}\right) - u_x = 0 & \text{(A-2a)} \\ w_x + \left(\dfrac{\mu}{4\lambda} \dfrac{c^2}{c_2^2} u_{tt} - \dfrac{\mu}{4\lambda} u_{xx} + \dfrac{2(\lambda+\mu)}{\lambda} u\right) = 0 & \text{(A-2b)} \end{cases}$$

With appropriate group of parameters, we have the following form

$$\begin{cases} b_1^2 w_{tt} - a_1^2 w_{xx} + p_1^2 w - u_x = 0 & \text{(A-3a)} \\ w_x + b_2^2 u_{tt} - a_2^2 u_{xx} + p_2^2 u = 0 & \text{(A-3b)} \end{cases}$$

The above system of equations can be expressed by a matrix vector form of equations as

$$\begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{(A-4)}$$

where

$$L_{11} = b_1^2 \frac{\partial^2}{\partial t^2} - a_1^2 \frac{\partial^2}{\partial x^2} + p_1^2 \quad , \quad L_{12} = -\frac{\partial}{\partial x} \qquad \text{(A-5)}$$

$$L_{21} = \frac{\partial}{\partial x} \quad \text{and} \quad L_{22} = b_2^2 \frac{\partial^2}{\partial t^2} - a_2^2 \frac{\partial^2}{\partial x^2} + p_2^2 \qquad \text{(A-6)}$$

which can be written as Eq. (1) in the text.

The notations for the wave equation in rods are:

$w$ = axial displacement

$u$ = radial strain

$\xi = z/a$ = nondimensional axial coordinates

$\tau = ct/a$ = nondimensional time

$a$ = elastic stored energy per unit length

$c_1 = \sqrt{(\lambda+2\mu)/\rho}$ = longitudinal wave speed

$c_2 = \sqrt{\mu/\rho}$ = shear wave speed

$\lambda$ and $\mu$ are Lame' constants

The above system of equations is first derived in different form by Mindlin and Herrmann and can be grouped into a single equation as:

$$\left[\left(\frac{\partial^2}{\partial\xi^2} - \frac{c^2}{c_1^2}\frac{\partial^2}{\partial\tau^2}\right)\left(\frac{\partial^2}{\partial\xi^2} - \frac{c^2}{c_2^2}\frac{\partial^2}{\partial\tau^2}\right) - \delta\frac{\lambda+\mu}{\mu}\frac{c_b^2}{c_1^2}\left(\frac{\partial^2}{\partial\xi^2} - \frac{c^2}{c_b^2}\frac{c^2}{c_b^2}\right)\right](w \text{ or } u) = 0 \quad (A\text{-}7)$$

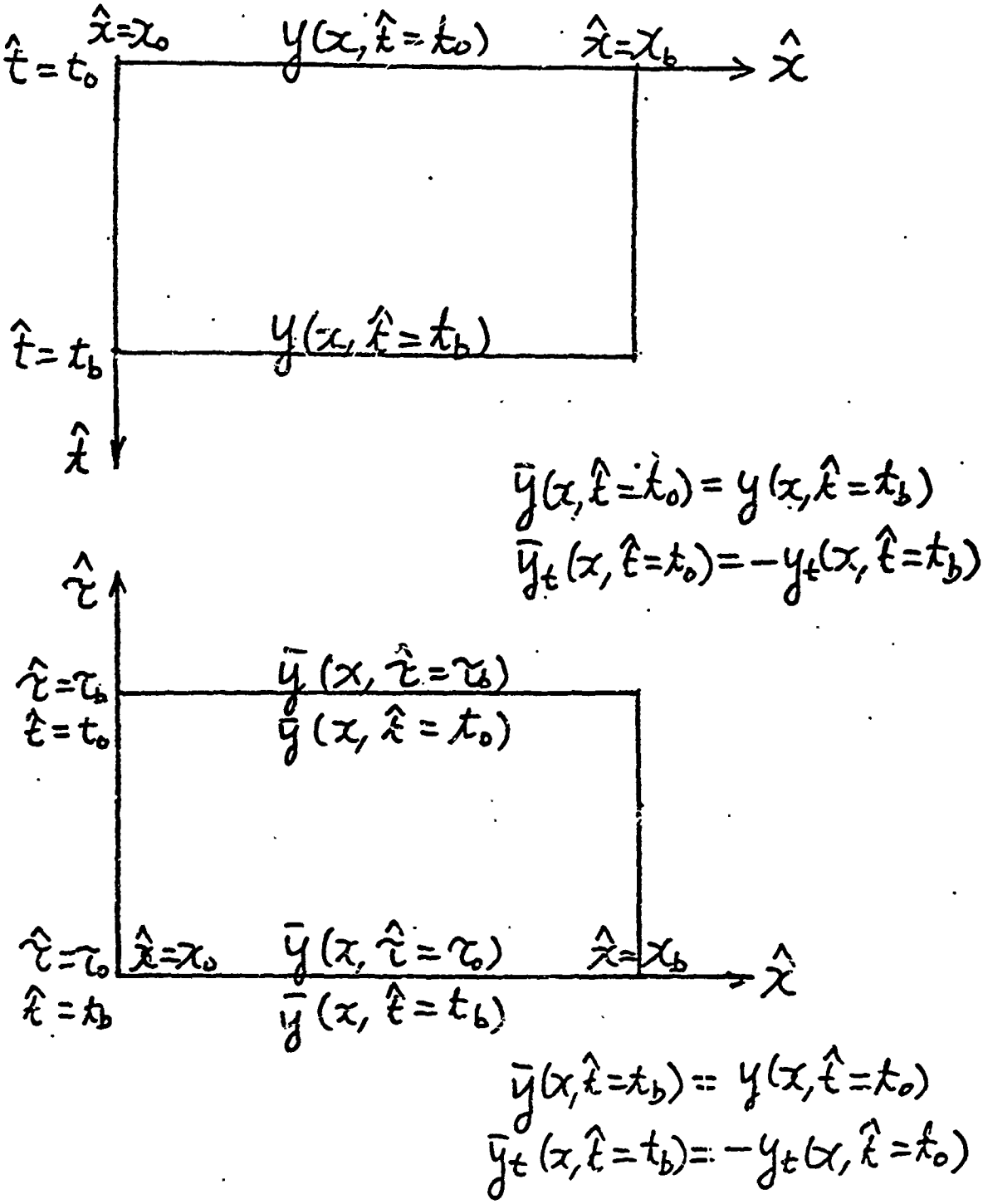



Figure 1. Image Reflection of the Adjoint System.

# REPRESENTATION OF TWO-PHASE FLOWS BY AVERAGING

Aivars Celmiņš
Ground Mobility & Firepower Branch
Vulnerability/Lethality Division
Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005

## ABSTRACT

We consider descriptors of gas-particle aggregates which represent space averaged local properties of the aggregate. We show that such descriptors have undulations due to the finite size of the averaging volume, and derive an estimate for the bounds of the amplitude of the undulations. According to that estimate, one obtains reasonably accurate averages only if the averaging volume contains at least 30-150 particles. In terms of the size of the averaging volume this limitation means that the diameter of the volume should equal at least four to five mean distances between particle centers. A consequence for the modeling of two-phase flows through tubes is that space averaged descriptors cannot resolve the radial structure of such flows unless the mean particle distance is much smaller than one tenth of the tube diameter. This condition excludes, for instance, the use of average descriptors for radial flow resolution of some interior ballistics flows, where the particulate phase consists of propellant grains. We also derive an approximate characterization of the boundary of a particle aggregate for different sizes of the averaging volume.

## 1. INTRODUCTION

Properties of gas-particle mixtures commonly are characterized in terms of descriptors which represent space averages over a finite volume. For instance, one may define the particle temperature at any point of the mixture as the average temperature of all particles within a finite neighborhood of the point, and use a similar definition for average gas properties by averaging the local gas properties over a volume representing the neighborhood. In order to obtain reasonable averages, the averaging volume should be large, for instance, larger than the particles. On the other hand, by averaging over a large volume, one loses information about local flow structures, for instance, the boundaries of particle aggregates are washed out. Therefore, one wants the averaging volume to be small. One may find a reasonable compromise between the contradicting requirements if the requirements are quantitated. Such a quantitation is the purpose of the present paper.

We approach the problem by investigating the undulations of the gas volume fraction $\alpha$ in the averaging volume in terms of the distribution of the particles, and of the size and location of the averaging volume. The investigation produces a bound for the undulations. By specifying a tolerance level for the undulations, one can use this bound to obtain a minimum size of the averaging volume which would insure that undulations of $\alpha$ are below the tolerance level. Undulations of other flow descriptors are shown to be proportional to the undulations of $\alpha$.

PREVIOUS PAGE IS BLANK

The smoothing effect of volume averaging on flow structures is illustrated by a representation of a particle aggregate boundary by the value of $\alpha$. In concept, such a boundary consists of a narrow transition zone between regions with $\alpha=1$ (gas only) and $\alpha=\bar{\alpha}$ (the average gas volume fraction in the mixture region). One finds that the width of the transition zone, which equals about one diameter of the averaging volume is not narrow in comparison to a mean distance between particles. The transition profile can be represented by an approximate formula.

In Section 2, we present and discuss the principal results of the investigation, i.e., the estimated bound of undulations of $\alpha$. Discussions of some properties of the gas volume fraction $\alpha$ are presented in Section 3, which also contains an outline of the derivation of the bound. The representation of a particle aggregate boundary in terms of $\alpha$ is discussed in Section 4. Section 5 contains a concluding discussion of the results.

## 2. PRINCIPAL RESULTS

Let the averaging volume be a sphere with radius R and let the particles be spheres with radius s arranged in a three-dimensional lattice. The gas volume fraction $\alpha$ in the averaging sphere is generally different for different positions of the sphere. However, the limit value of $\alpha$ as R becomes infinite is independent of the location of the center of the averaging sphere, only depending on the lattice and on the particle radius s. Let the limit value be $\bar{\alpha}$. We define a mean distance $L_m$ between particle centers by the equation

$$L_m = 2s(1-\bar{\alpha})^{-1/3} . \tag{2.1}$$

A motivation for this definition is given in Appendix B, where it is also shown that in general, $L_m$ is 10-24 percent larger than the smallest distance between particle centers, i.e., between the lattice points of the particle arrangement.

The difference

$$\Delta\bar{\alpha} = \alpha - \bar{\alpha} , \tag{2.2}$$

between the gas volume fraction in the averaging sphere and the limit value of $\alpha$ depends on the size and position of the sphere as well as on $\bar{\alpha}$ and the particle arrangement. In Section 3, we show by sample calculations that the magnitude of $\Delta\bar{\alpha}$ is bounded by

$$|\Delta\bar{\alpha}| < 0.5(1-\bar{\alpha})\bar{\alpha}^2(L_m/R)^2, \tag{2.3}$$

independently of the position of the averaging sphere. The bound Eq. (2.3) has been tested by calculations for four different lattices (defined in Appendix A) and for $1\leq R/L_m\leq 4$ and $0.5\leq\bar{\alpha}\leq 0.9$. Because the formula has the correct limit $\Delta\bar{\alpha}=0$ for $\bar{\alpha}=1$, it can be used as an estimate for all $\bar{\alpha}\geq 0.5$.

Extrapolations to smaller values of $\bar{\alpha}$ should be done with caution and the same applies to extrapolations to $R/L_m>4$. The domain $R/L_m<1$ is of little practical interest, because there the undulations are too large.

Let $|\Delta\bar{\alpha}|_{tol}$ be an acceptable tolerance level for the undulations. Then one can reformulate Eq. (2.3) as a condition for $R/L_m$ obtaining

$$R/L_m > \bar{\alpha}[0.5(1-\bar{\alpha})/|\Delta\bar{\alpha}|_{tol}]^{1/2}. \tag{2.4}$$

If $R/L_m$ satisfies Eq. (2.4), then the undulations of $\alpha$ are less than $|\Delta\bar{\alpha}|_{tol}$.

If $\bar{\alpha}$ is close to one, i.e., if most of the volume is occupied by gas, then it is reasonable to choose a tolerance level that is proportional to $1-\bar{\alpha}$, instead of a constant level. For instance, the tolerance level could be

$$\left.\begin{array}{ll} |\Delta\bar{\alpha}|_{tol} & \text{for } \bar{\alpha} \le \bar{\alpha}_t, \\ [(1-\bar{\alpha})/(1-\bar{\alpha}_t)]|\Delta\bar{\alpha}|_{tol} & \text{for } \bar{\alpha} \ge \bar{\alpha}_t. \end{array}\right\} \tag{2.5}$$

(We assume that $\bar{\alpha}_t$ and $|\Delta\bar{\alpha}|_{tol}$ are such that $|\Delta\bar{\alpha}|_{tol}/(1-\bar{\alpha}_t)<1$.) The condition Eq. (2.4) with these tolerance levels can be expressed in the form

$$R/L_m > \bar{\alpha}[\max(1-\bar{\alpha}, 1-\bar{\alpha}_t)]^{1/2}[0.5/|\Delta\bar{\alpha}|_{tol}]^{1/2}. \tag{2.6}$$

We restrict for simplicity $\bar{\alpha}_t$ to $\bar{\alpha}_t \ge 0.85$ and observe that for this range the factor of $|\Delta\bar{\alpha}|_{tol}^{-1/2}$ in Eq. (2.6) has a maximum at $\bar{\alpha}=2/3$, a minimum at $\bar{\alpha}=\bar{\alpha}_t$, and is linearly increasing between $\bar{\alpha}=\bar{\alpha}_t$ and $\bar{\alpha}=1$. Hence if $\bar{\alpha}$ is not known, then Eq. (2.6) should be used with $\bar{\alpha}=2/3$, and if, for instance, $\bar{\alpha}$ is known to be larger than $\bar{\alpha}_t$, then Eq. (2.6) should be used with $\bar{\alpha}=1$.

Figure 1 is a graphical display of Eq. (2.6) for $\bar{\alpha}_t=0.9$. It shows, for instance, that for $|\Delta\bar{\alpha}|_{tol}=0.01$ the averaging sphere radius should be larger than $2.7L_m$ if $\bar{\alpha}$ is not known, and larger than $2.2L_m$ if $\bar{\alpha}$ is known to be larger than 0.867. The solid lines in the figure indicate the domain in which sample calculations have been made. Extrapolations are indicated by dashed lines.

The relation Eq. (2.3) also can be expressed in terms of the number N of particles in the averaging volume. We arrive at such an expression by using the approximation (see Appendix B)

$$N \approx (2R/L_m)^3. \tag{2.7}$$

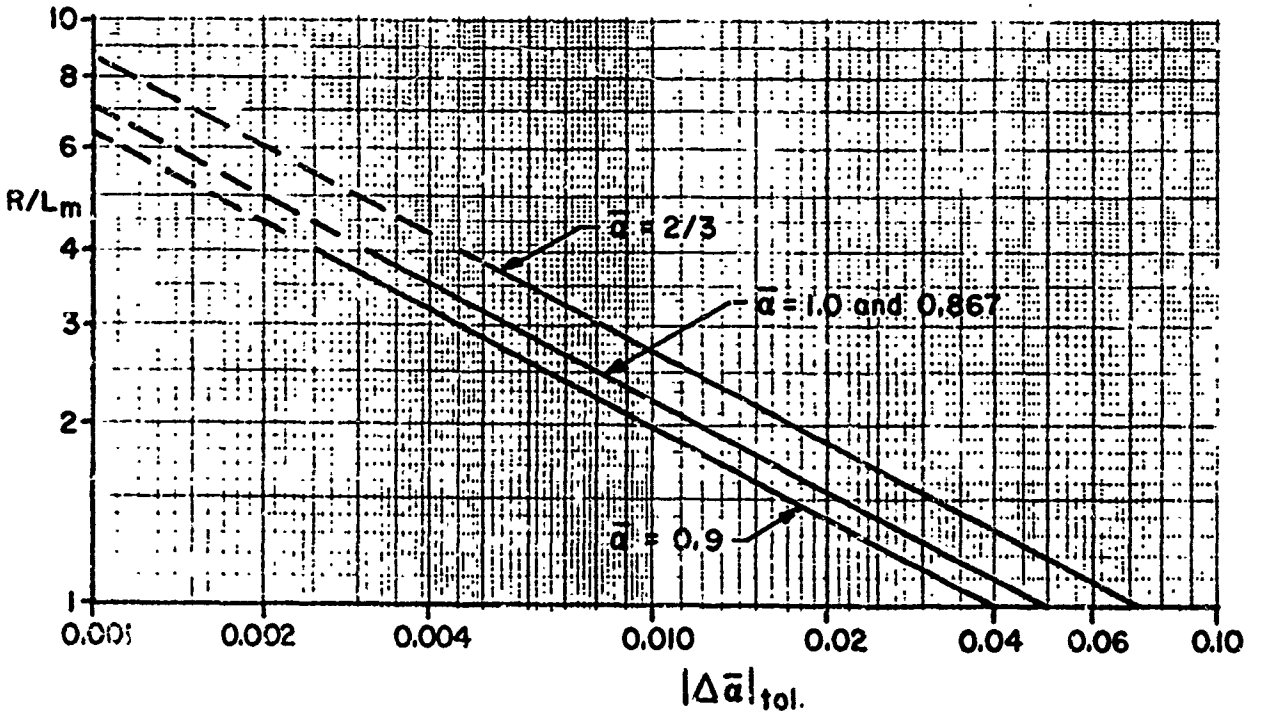



Figure 1. Averaging Sphere Radius for Given Tolerance and $\bar{\alpha}_t = 0.9$

Substituting this expression in Eq. (2.3) we obtain

$$|\Delta\bar{\alpha}| < 2.0(1-\bar{\alpha})\bar{\alpha}^2 N^{-2/3}, \tag{2.8}$$

and the tolerance condition (2.6) becomes

$$N > \bar{\alpha}^3[\max(1-\bar{\alpha}, 1-\bar{\alpha}_t)]^{3/2} [2/|\Delta\bar{\alpha}|_{tol}]^{3/2}. \tag{2.9}$$

The tolerance condition (2.9) is displayed in Figure 2 for $\bar{\alpha}_t$=0.9. It shows that for $|\Delta\bar{\alpha}|_{tol}$ =0.01 the minimum number of particles in the averaging volume is between 65 and 160, depending on $\bar{\alpha}$. The condition has been tested by calculations with N between 3 and 512. Extrapolation beyond the tested range is indicated in Figure 2 by dashed lines.

The undulations of averages of flow descriptors other than $\alpha$, such as gas density, are closely related to the undulations of $\alpha$. We investigate this relation in Appendix C and show by examples that the amplitude $\Delta\phi$ of particle induced undulations of a descriptor $\phi$ is proportional to $\Delta\alpha$. If $\phi$ has a constant gradient and changes by $\delta\phi$ along a distance equal to the diameter of the averaging sphere, then

$$|\Delta\phi| \approx |\delta\phi||\Delta\alpha|/(3\alpha). \tag{2.10}$$

If $\delta\phi$=0 but there are local particle induced inhomogeneities, then the undulations of $\phi$ again are proportional to $\Delta\alpha$ and to the amplitude of the local inhomogeneities. If $d\phi$ is the change of the constant average value of $\phi$ due to such disturbances then

$$|\Delta\phi| \approx |d\phi| \cdot |\Delta\alpha|/(\alpha(1-\alpha)). \tag{2.11}$$

Using the relations (2.10) and (2.11) one can obtain tolerance conditions for $R/L_m$ and N in terms of a tolerance level $|\Delta\phi|_{tol}$ for any flow descriptor $\phi$.

Next, we discuss some consequences of these findings for the representation of two-phase flows through tubes. In order to resolve the radial structure of such flows, the averaging sphere radius should be a small fraction of the tube diameter, say

$$R < cD, \tag{2.12}$$

where c is of the order 0.05 or smaller. (With such a R, one can describe the flow in the core of the tube and up to the distance cD from the wall. If $c \approx 0.05$, then the averaged descriptor could be used to represent gross features of the radial structure.) Assuming e.g., that $|\Delta\bar{\alpha}|_{tol}$=0.01, one

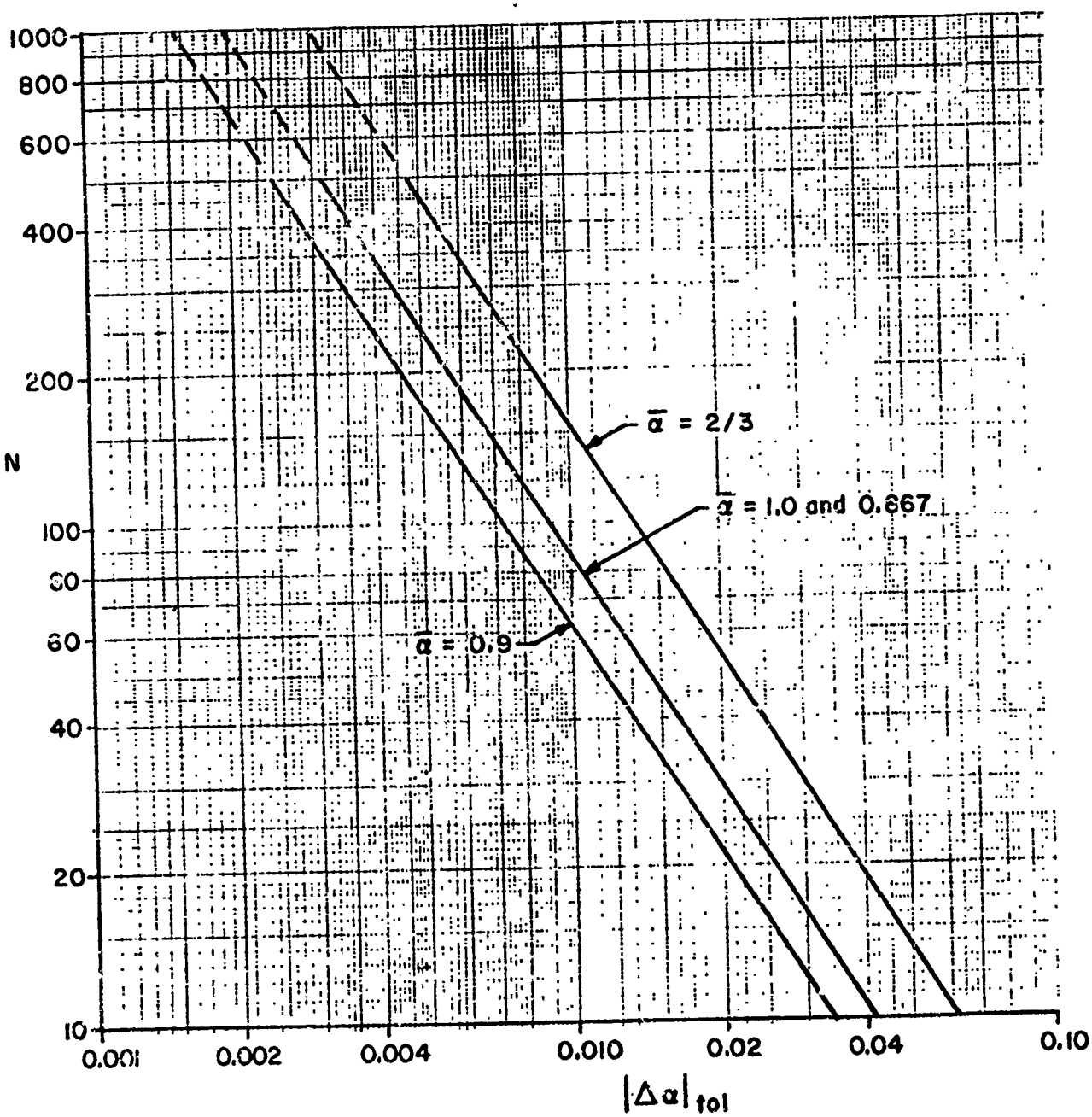



Figure 2. Minimum Number of Particles in an Averaging Volume for Given Tolerance and $\alpha_t = 0.9$

finds from Figure 1, $R>2.7L_m$ and therefore,

$$L_m < Dc/2.7 = D0.02. \tag{2.13}$$

Hence, in order to reasonably represent by averaged descriptors gross features of the radial structure of the flow, the mean distance between the particle centers should be less than 0.02 of the tube diameter. By relaxing the requirements on the tolerance level, one of course can arrive at a less stringent condition. However, Eq. (2.13) likely indicates the correct order of magnitude of an upper bound for $L_m$. A simple consequence for interior ballistics modeling, where the particulate phase consists of propellant grains, is that space averaged descriptors are not adequate for the resolution of radial flow structures, because the condition (2.13) is typically violated in such flows. More discussions of the interior ballistics problem are presented in Section 5.

Volume averaging smoothes and distorts not only the undesirable heterogeneities that are caused by single particles, but also other flow structures, particularly when they have extensions less than a diameter of the averaging volume. An example of a flow structure with a small extension is the boundary of a particle aggregate. Let the average gas volume fraction in the aggregate be $\bar{\alpha}$ and let the aggregate occupy the half-space $z>0$. The conceptual image of the boundary $z=0$ is a narrow transition zone between a region with $\alpha=1$ (gas only) and $\alpha=\bar{\alpha}$ (gas-particle mixture). In fact, the width of the transition zone equals a diameter of the averaging volume and is, therefore, large compared to the mean distance between the particles. If the averaging volume is a sphere with radius R, then the transition profile is approximately given by the following function

$$\tilde{\alpha}(z) = \begin{cases} 1 & \text{if } z \le -R, \\ 1-(1-\bar{\alpha})\left(1+\frac{z}{R}\right)^2\left(2-\frac{z}{R}\right)\frac{1}{4}, & \text{if } -R < z < R, \\ \bar{\alpha} & \text{if } z \ge R. \end{cases} \tag{2.14}$$

This function is an idealization derived under the assumption that undulations of $\alpha$ are zero inside the particle aggregate, but it is not a limit curve for $R\to\infty$. (Such a limit is $\alpha \equiv (1+\bar{\alpha})/2$ for any aggregate occupying a half-space). The real transition profiles are different for different trajectories of the averaging sphere and undulate around a curve approximated by Eq. (2.14). As shown in Section 4, the amplitude of the undulations are bounded by Eq. (2.3) in which $\bar{\alpha}$ is replaced by $\tilde{\alpha}(z)$ from Eq. (2.14).

## 3. UNDULATIONS OF THE GAS VOLUME FRACTION

We consider the gas volume fraction $\alpha$ in an averaging sphere with the radius R and center coordinates $\vec{X}$. We assume that the particles are spheres with radius s and arranged in a three-dimensional lattice with the lattice constant L. Then the particle aggregate is completely described by dimensionless geometrical parameters and two length scales, s and L. Let $\bar{\alpha}$ be

the gas volume fraction of an infinite averaging volume, i.e., of the whole space. Then one can define a mean distance $L_m$ between the particle centers by

$$L_m = 2s(1-\bar{\alpha})^{-1/3} \tag{3.1}$$

(A motivation for this definition is given in Appendix B.) One can show that the ratio $L/L_m$ is for any given lattice independent of s and $\bar{\alpha}$. Therefore, a particle aggregate in lattice form is completely described by dimensionless geometrical parameters, the ratio $L/L_m$, the value of $\bar{\alpha}$ and a single length scale $L_m$.

The gas volume fraction $\alpha$ of the averaging sphere depends on the parameters of the aggregate, and on the location and size of the sphere. We express this dependence by

$$\alpha = f(\vec{X}/L_m, R/L_m, \bar{\alpha}), \tag{3.2}$$

where the function f is completely determined by dimensionless parameters describing the geometry of the lattice. Next, we investigate the dependence of $\alpha$ on the arguments $\vec{X}/L_m$ and $R/L_m$. If R is very small, then $\alpha$ has the value one or zero, depending whether the center position $\vec{X}$ is inside or outside a particle. As R increases, $\alpha$ approaches the value $\bar{\alpha}$, independently of the value of $\vec{X}$, i.e., of the location of the averaging sphere. The transition between these limits is illustrated in Figures 3 and 4. The two curves in Figure 3 correspond to two different positions of the averaging sphere in a square cylinder lattice. (Lattices are defined in Appendix A.) One of the positions is at the point of origin which is a lattice point and occupied by a particle. Therefore, the corresponding curve starts with $\alpha=0$ and $\alpha$ starts to increase only as R becomes larger than s. The other curve is for the averaging sphere's center position $\vec{X}/L_m=(0.4, 0.2, 0.0)$. It starts with the value $\alpha=1$. Both curves approach the limit value $\alpha=\bar{\alpha}$ and undulate about that value as R increases. Figure 4 shows two similar curves for a different value of $\bar{\alpha}$ and in a different lattice. The center positions of the two averaging spheres are the same as in Figure 3. The curves shown in Figures 3 and 4 are typical for all computed examples.* The larger undulations about $\bar{\alpha}$ have a wave length of the order of $L_m$ and an amplitude that decreases proportionally to a negative power of $R/L_m$. The curves belong to a family with three parameters, namely, the three components of the averaging sphere's center coordinate $\vec{X}/L_m$. We are interested in the envelopes of the family of curves. The envelopes were obtained numerically (using a simplex algorithm) for various values of $\bar{\alpha}$ and for the four different lattices described in Appendix A. Some of the results are shown in Figures 5, 6 and 7.

*Additional examples are presented in A. K. R. Celmiņš, "Averaging Effects in Models of Three-Dimensional Two-Phase Flows," BRL Technical Report in Publication.*

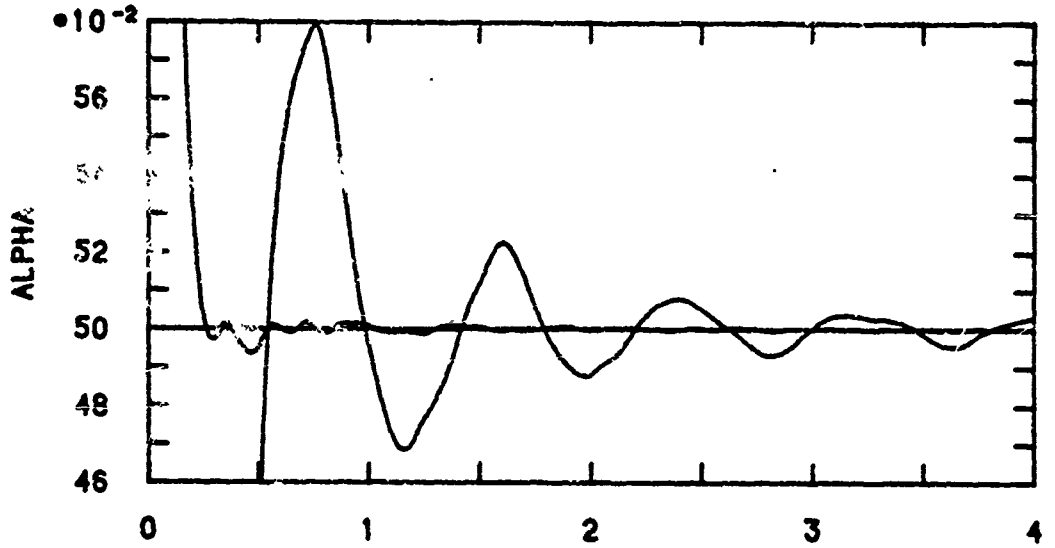



Figure 3. Gas Volume Fraction Dependence on Averaging Sphere Radius Square Cylinder Lattice, $\bar{\alpha} = 0.5$, $s/L_m = 0.3969$

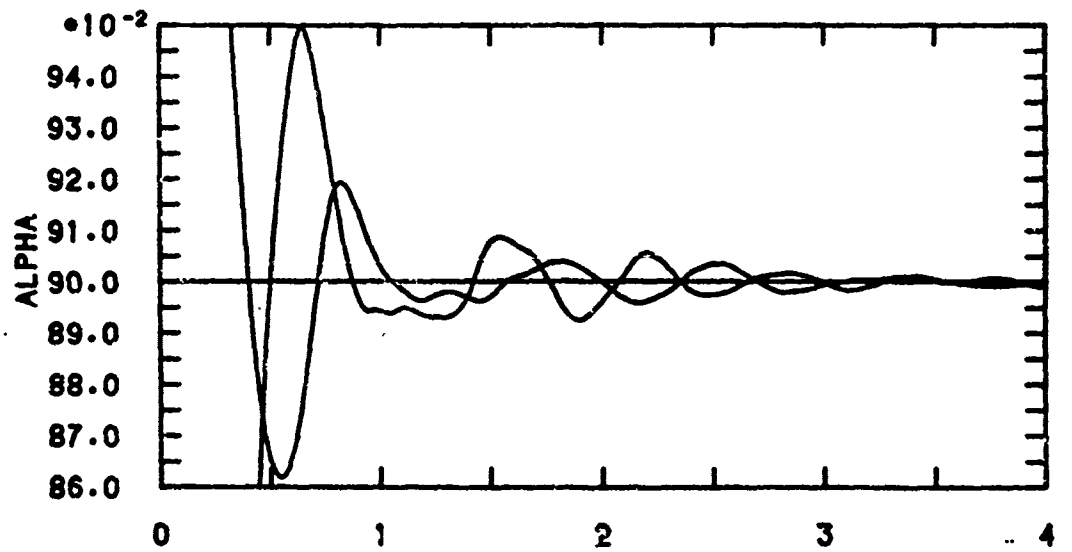



Figure 4. Gas Volume Fraction Dependence on Averaging Sphere Radius Triangular Cylinder Lattice, $\bar{\alpha} = 0.9$, $s/L_m = 0.2321$

Figure 5 shows a regular pattern of the envelopes, but such a regularity was found to be an exception. If one reduces the sizes of the particles in the same lattice, then one obtains the envelopes shown in Figure 6. Figure 7 shows envelopes for a different lattice, but for the same limit value $\bar{\alpha}$ as in Figure 6. The envelopes in the latter two figures are typical for all calculations. They indicate that positive and negative deviations from $\bar{\alpha}$ have in general the same magnitude. Also comparing results for different lattices and the same $\bar{\alpha}$, one finds that none of the lattices produce consistently larger or smaller undulations of $\alpha$. An overview of the trend of the extreme values of the envelopes can be obtained by plotting the extreme values versus the radius of the averaging sphere. Figure 8 shows such a plot in log,log-scale. The figure displays the extreme positive and negative values of the envelopes for four different lattices but the same $\bar{\alpha}$. The different symbols signify different lattices and different signs of the deviations. The line in Figure 8 represents an estimated bound of the deviations. The equation of the line is

$$|\Delta\bar{\alpha}| = 0.5(1-\bar{\alpha})\bar{\alpha}^2 \, (R/L_m)^{-2}. \qquad (3.3)$$

The equation was determined by comparing plots of extreme deviations of $\alpha$ for different values of $\bar{\alpha}$. Next, we discuss the validity of the bound (3.3).

The bound (3.3) is based on sample calculations of envelopes within the range $1.0 \leq R/L_m \leq 4.0$, for $\bar{\alpha}=0.5$, 0.667 and 0.9, and for the four lattices described in Appendix A. Because the lattices have quite different symmetries and because their maximal packing densities range from $\bar{\alpha}_{min}=0.260$ to $\bar{\alpha}_{min}=0.476$, one can assume that the results are valid for all reasonable lattices. It is possible to construct symmetric particle arrangements with undulations in excess of Eq. (3.3), by clustering the particles in a periodic manner. However, one can debate whether such arrangements can be considered as particle aggregates with uniform particle number density. In any case, the estimate by Eq. (3.3) is optimal in the sense that additional computations of examples can only make the bound larger. The limitation of calculations to $\bar{\alpha} \geq 0.5$ was motivated by practical considerations. The minimum $\bar{\alpha}$ is for the square cylinder lattice 0.476. This means that values of $\bar{\alpha}<0.5$ only can be obtained by special packings of the particles, while the results for $\bar{\alpha}>0.5$ have more general validity. The upper limit of $\bar{\alpha}$ is one (gas only) and Eq. (3.3) produces at this limit the correct value $\Delta\bar{\alpha}=0$. Therefore, Eq. (3.3) likely can be used for the whole range $0.5 \leq \bar{\alpha} \leq 1.0$. The range of $R/L_m$ was limited from below to $R/L_m \geq 1$ because for smaller R the undulations become excessively large (see Figures 5, 6 and 7). At $R/L_m=4.0$, the amplitude of the undulations approaches the order of $10^{-3}$, which can be assumed sufficiently small for most applications.

Eq. (3.3) also can be used as a tolerance condition for $R/L_m$. Such an application of the formula is discussed in Section 2.

## 4. GAS VOLUME FRACTION PROFILES

Let the particle aggregate occupy the half-space $z>0$, and let $\bar{\alpha}$ be the gas volume fraction in the half-space. We investigate how $\alpha$ can be used as a descriptor of the extension of the particle aggregate.

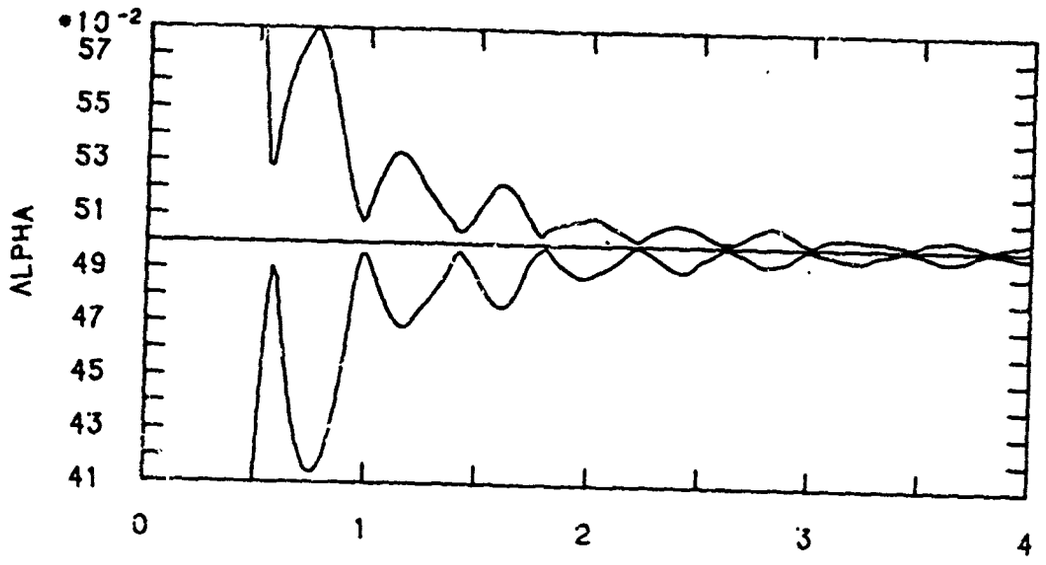



Figure 5. Envelopes of Gas Volume Fraction Curves
Square Cylinder Lattice, $\bar{\alpha} = 0.5$

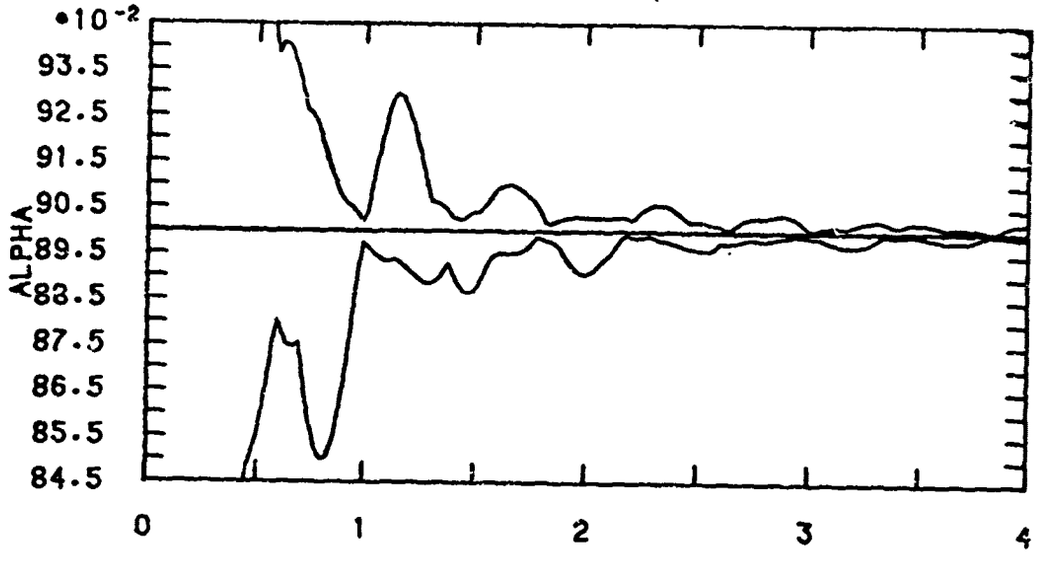



Figure 6. Envelopes of Gas Volume Fraction Curves
Square Cylinder Lattice, $\bar{\alpha} = 0.9$

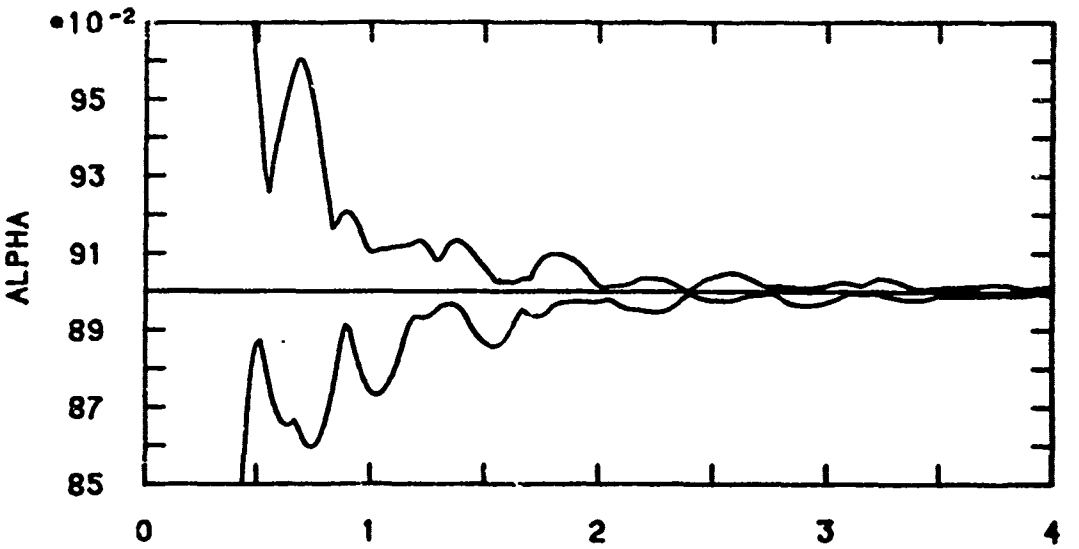



Figure 7. Envelopes of Gas Volume Fraction Functions
Leap-Frog Triangular Lattice, $\alpha = 0.9$

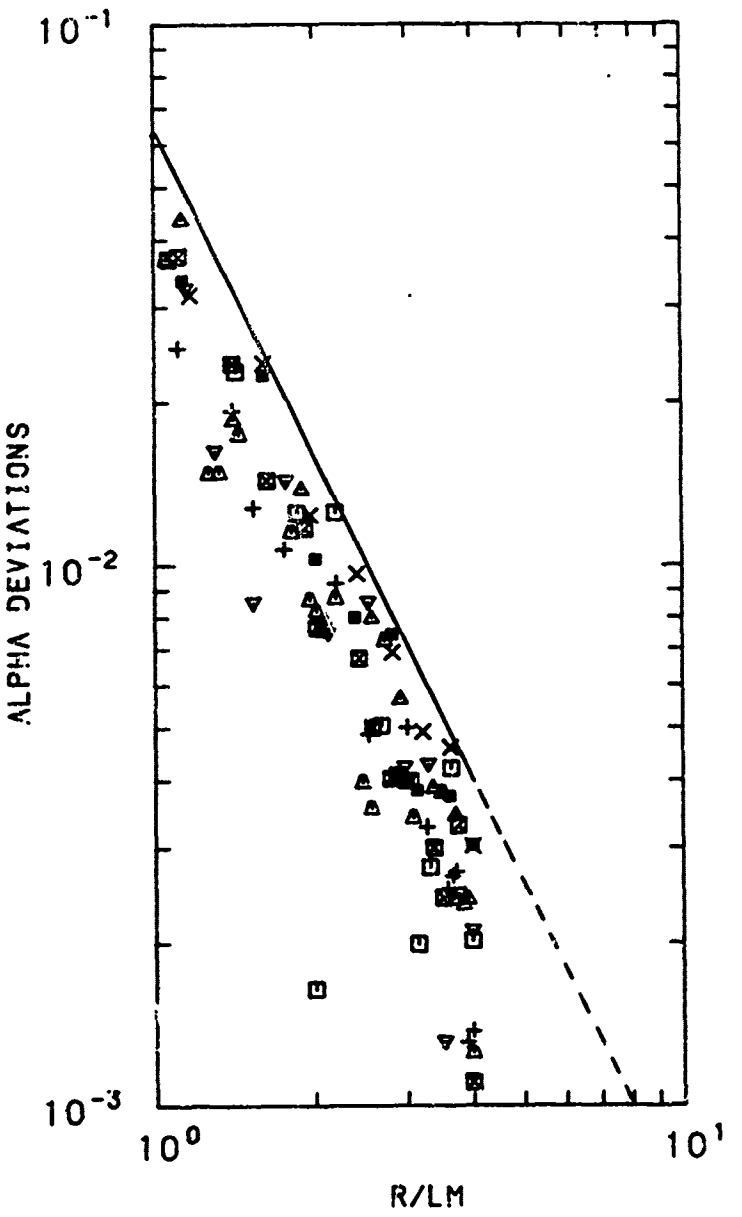



Figure 8. Extreme Deviations of Gas Volume Ratio, Combined Plot for Four Lattices, and Positive and Negative Deviations $\bar{\alpha} = 0.5$

If the averaging sphere is placed outside the aggregate, then one obtains the value $\alpha=1$. For positions of the sphere inside the aggregate, $\alpha$ is approximately equal to $\bar{\alpha}$. The boundary $z=0$ of the aggregate is signified by a transition $\alpha(z)$ from one to $\bar{\alpha}$. Such a transition is illustrated in Figure 9. For the examples shown, the particle aggregate is a lattice with particles placed only at those lattice points which have a non-negative z-coordinate. One can, therefore, define as aggregate boundary the plane $z=-s$ (tangential to the extreme particles). The four calculated transition profiles correspond to two averaging spheres with radiuses $R=L_m$ and $R=2L_m$, respectively, and two trajectories for each sphere. One trajectory is the z-axis and the other trajectory is the line $x/L_m=0.5$, $y/L_m=0.5$.

One observes inside the aggregate the expected periodic undulations about $\bar{\alpha}$. The amplitude of the undulations is bounded by Eq. (2.3), and the wave length is about $1.5L_m$. (For other examples, the wave length was found to be as low as $1.0L_m$.) In the transition zone, the undulations are about a transition curve. This transition curve may be approximately computed by assuming that the gas volume fraction equals $\bar{\alpha}$ in those parts of the averaging sphere which are inside the aggregate. (Beyond the plane $z=-s$ in the present example.) The maximum slope of this approximation is $-(1-\bar{\alpha})(3/2)(L_m/2R)$, and a formula for the curve $\tilde{\alpha}(z)$ is given by Eq. (2.14). One may estimate the undulations in the transition zone by Eq. (2.3) in which $\bar{\alpha}$ is replaced by the appropriate value $\tilde{\alpha}(z)$. Test calculations confirm the validity of this estimate. However, because the aggregate boundary introduces a new structure into the problem, systematic deviations are possible of $\tilde{\alpha}(z)$ from the "ideal" transition curve without undulations. This is illustrated in Figures 10, 11 and 12.

Figure 10 shows envelopes of the transition profiles of Figure 9 with $R/L_m=1.0$. The envelopes were numerically determined by a simplex algorithm whereby the x- and y-coordinates of the averaging sphere were treated as free parameters. The figure also shows the transition curve $\tilde{\alpha}(z)$ with bounds of undulations calculated by Eq. (2.3) as described above. One sees that the difference between both envelopes is for all z-values smaller than the corresponding difference between the estimated bounds. However, at the top of the transition curve, both envelopes are larger than the curve $\tilde{\alpha}(z)$, indicating a positive systematic deviation. A negative systematic deviation can be seen in Figures 11 and 12, which show profiles for the same particle arrangement, but for a different value of $\bar{\alpha}$ and two different values of $R/L_m$.

It appears from the calculated examples that bounds for deviations of a real profile $\alpha(z)$ from $\tilde{\alpha}(z)$ (Eq. (2.14)) can be estimated by Eq. (2.3), evaluated at $\alpha=\tilde{\alpha}(z)$ (instead of $\alpha=\bar{\alpha}$), and multiplied by the factor $1+[\tilde{\alpha}(z)-\bar{\alpha})$, i.e.,

$$|\Delta\alpha| < (1+\tilde{\alpha}(z)-\bar{\alpha})\ 0.5\ (1-\tilde{\alpha}(z))\ \tilde{\alpha}^2(z)\ (L_m/R)^2 . \qquad (4.1)$$

The formula has not been tested systematically. For constant $\alpha \equiv \bar{\alpha}$, it reduces to Eq. (2.3), but it increases the estimated bound within the transition zone by up to a factor $2-\bar{\alpha}$, thus taking care of systematic trends in the transition profiles.

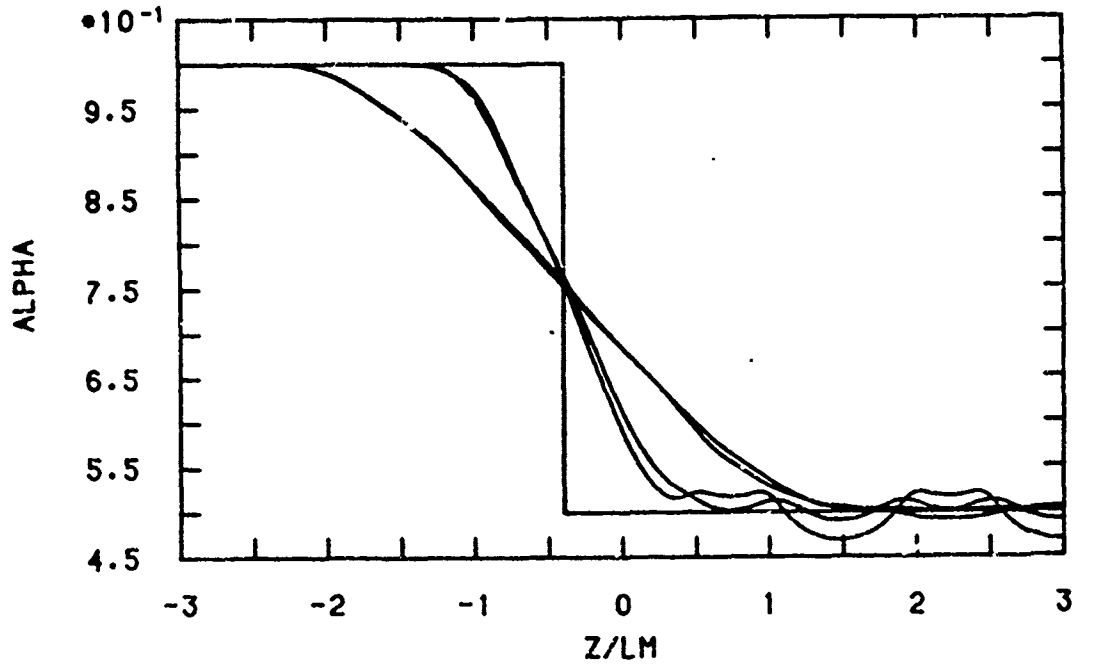



Figure 9. Gas Volume Fraction Profiles at a Particle Aggregate Boundary Leap-Frog Triangular Lattice, $\bar{\alpha} = 0.5$, $s/L_m = 0.3969$ $R/L_m = 1.0$ and 2.0

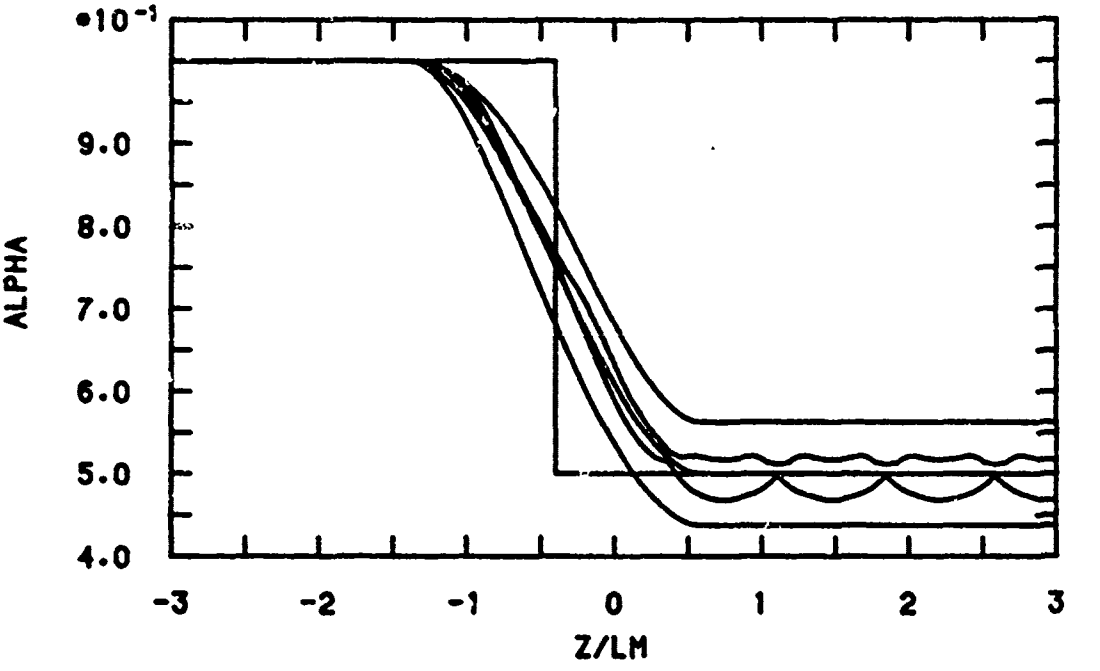



Figure 10. Profile Envelopes and Estimated Bounds of Gas Volume Fraction Leap-Frog Triangular Lattice, $\bar{\alpha} = 0.5$, $R/L_m = 1.0$

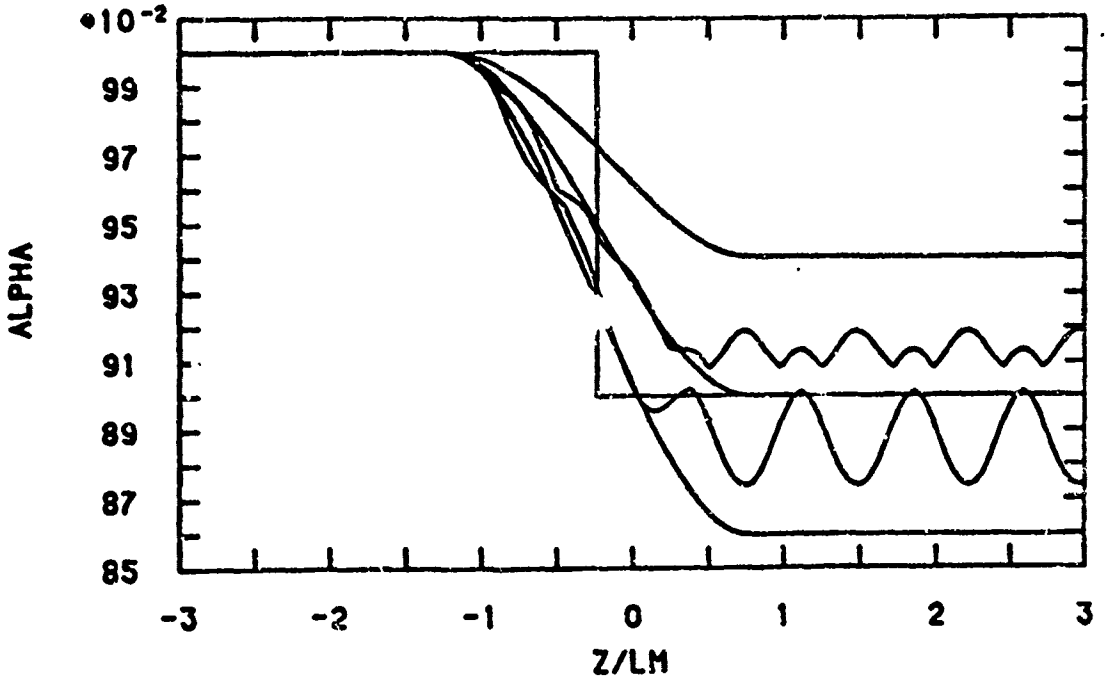



Figure 11. Profile Envelopes and Estimated Bounds of Gas Volume Fraction Leap-Frog Triangular Lattice, $\bar{\alpha}$ = 0.9, $R/L_m$ = 1.0

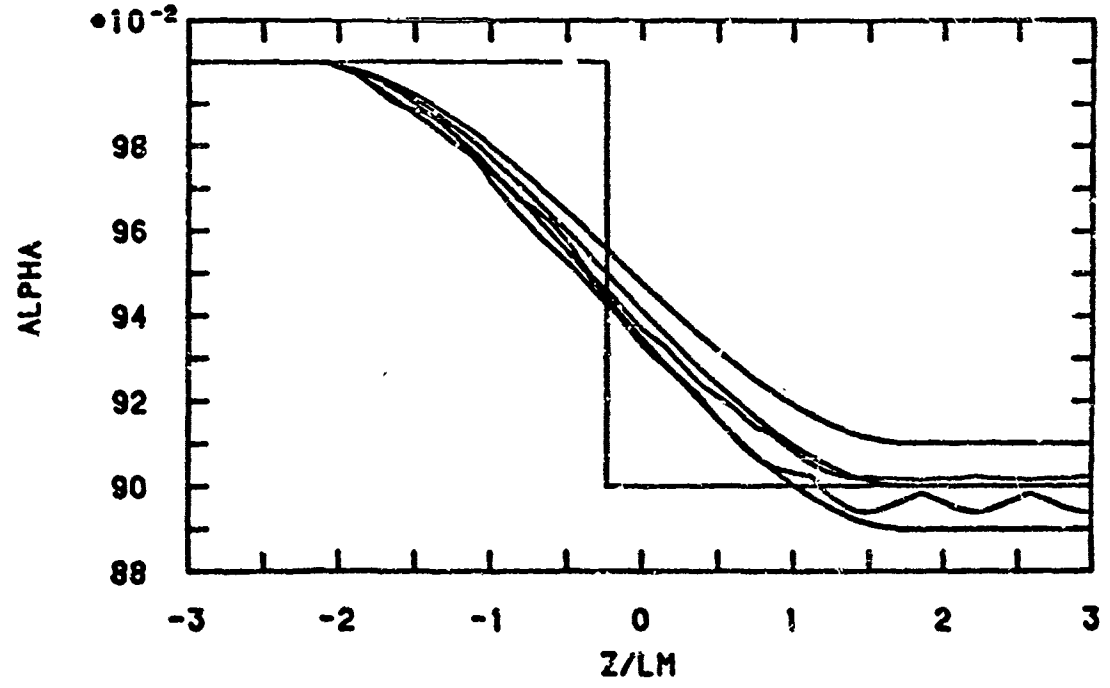



Figure 12. Profile Envelopes and Estimated Bounds of Gas Volume Fraction Leap-Frog Triangular Lattice, $\bar{\alpha}$ = 0.9, $R/L_m$ = 2.0

## 5. DISCUSSIONS OF RESULTS

The sample computations presented in Sections 3 and 4 show that the gas volume fraction $\alpha$ in an averaging sphere with radius R is a function of the sphere's location even in a uniform particle aggregate. In such an aggregate, the function undulates about a limit value $\bar{\alpha} = \lim_{R\to\infty} \alpha(R)$ which is independent of the sphere's location. The wave length of the undulations is up to $1.5L_m$, where $L_m$ is a mean distance between particles. The amplitude of the undulations is bounded by Eq. (2.3). Using Eq. (2.6) or Figure 1, one can determine the minimum size of the averaging sphere's radius R for any particular application. For instance, one finds that for a tolerance level of 0.01 for the undulations, $R/L_m$ must be larger than 2.7, and for a tolerance level of 0.03, $R/L_m$ must be larger than 1.5. We assume that the minimum of the radius is determined to be $R_{min}=rL_m$, and discuss some consequences of such a restriction.

First, we notice that any two-phase flow structures with extensions less than $R_{min}$ will be reduced in amplitude and stretched out to a size of $2R_{min}$ or larger. (See, for instance, the particle aggregate boundary representation by $\alpha$ in Section 4.) Consequently, a complete and accurate representation of the flow field can be done in a mesh with a mesh constant $0.5R_{min}$. Any mesh refinement can be done by interpolation in such a net, and a finer net does not provide a more accurate description of the flow. The same applies to measurements in a two-phase flow field. It suffices to present such measurements at intervals of $0.5R_{min}$. Ideally, local measurements should be made in a finer mesh, and the appropriate average descriptors obtained by averaging over the averaging sphere. This also applies to measurements near boundaries, where local measurements should be averaged and can be presented in a coarse mesh. Boundary values of average descriptors are not necessarily equal to the local boundary values and a mesh refinement in presentation of the results is not needed. (Mesh refinement might be needed for the observations, however.)

Similar consequences can be drawn for the computation of a two-phase flow, for instance, by solving average flow equations. A reasonable computing mesh constant is of the order of $0.5R_{min}$, so that flow structures extending more than, say, $R_{min}$ can be represented. Any refinement of the mesh below $0.5R_{min}$ has the effect of interpolation. If flow structures extending less than $R_{min}$ appear in the solution (using a fine mesh) then they should be interpreted as numerical artifacts or noise, because they cannot be interpreted as average flow properties.

When averaged descriptors are used to represent such transient two-phase flows in which $\alpha$ approaches one in some parts of the flow field, then one should distinguish between the possible causes for $\alpha$ approaching one. If the reason for the disappearance of the particles is a reduction of their sizes (for instance, by combustion), then the condition $R\geq R_{min}=rL_m$ is not violated ($L_m$ does not change.) Therefore, flow regions with $\alpha\approx 1$ can be represented by the same average descriptors as the rest of flow field. If, however, $\alpha$ approaches one because particles diffuse from the mixed phase region, then the

condition $R \geq R_{min} = rL_m$ eventually will be violated ($L_m$ increases due to the diffusion). When that happens, the tolerance level of the undulations is not guaranteed and as a result the representation by averaged descriptors becomes less accurate. A better approach in this case is to model the rarified parts of the aggregate by some other method than averaging, for instance, by individually tracing the diffused particles.

We consider as an example the situation in an interior ballistics two-phase flow where the particulate phase consists of propellant grains. Let the chamber volume of the gun be $W_0$ and the volume of the barrel be $W$. At the time when the projectile exits from the barrel, the volume available for the gas-particle mixture is $W_0+W$. Let the number of propellant grains be $m$. Then at the beginning of the firing cycle the mean distance between particle centers is (see Appendix B)

$$L_{mo} = \left(\frac{6}{\pi}\frac{W_0}{m}\right)^{1/3}, \tag{5.1}$$

and at muzzle time the mean distance is

$$L_m = \left(\frac{6}{\pi}\frac{W_0 + W}{m}\right)^{1/3}. \tag{5.2}$$

Hence,

$$\frac{L_m}{L_{mo}} = \left(1 + \frac{W}{W_0}\right)^{1/3}. \tag{5.3}$$

Let the initial gas volume fraction in the gun be $\alpha_0$, and the initial particle radius be $s_0$. Then $L_{mo}=2s_0\,(1-\alpha_0)^{-1/3}$ and

$$L_m = 2s_0(1 + W/W_0)^{1/3}(1-\alpha_0)^{-1/3}. \tag{5.4}$$

Typical for interior ballistics is $\alpha_0$ between 0.4 and 0.6 and $W/W_0$ is approximately equal to 10. Therefore, the maximum of $L_m/(2s_0)$ is for a typical gun between 2.64 and 3.02. Hence, in order to represent the whole transient firing cycle by averaging descriptors, one has to use an averaging sphere with the radius of about $2.7L_m$, or eight initial particle diameters. For most guns, the diameter of such an averaging sphere is of the same order as the caliber of the tube. This means that the radial flow structure cannot be represented by space averaged descriptors. Such descriptors only can be used by averaging over cross-sectional segments of the tube, in which case one obtains the representation of the core flow.

## Appendix A

## LATTICES

We describe in this appendix the four lattices which were used to define particle positions for calculations of gas volume fractions.

1. Square Cylinder Lattice. We construct the lattice by first arranging the particles in a square mesh with the mesh constant L in the x,y-plane, and then translating the mesh by multiples of L in the z-direction. Each square thereby generates a square cylinder. In this lattice, each particle has six neighbors at the distance L.

2. Triangular Cylinder Lattice. The lattice is constructed by first arranging the particles in the x,y-plane in an equilateral triangle mesh with the mesh constant L, and then translating the mesh in the z-direction by multiples of L. Each triangle thereby generates a triangular cylinder. The number of neighbor particles at distance L is for this lattice eight.

3. Leap-Frog Square Lattice. This lattice is constructed by starting with a square mesh in the x,y-plane, with the lattice constant L and the sides of the squares parallel to the axes. Then, the mesh is translated by multiples of $L/\sqrt{2}$ in the z-direction and by multiples of L/2 in the x- and y-directions. Thus, the pattern is translated in a leap-frog manner from one z-plane to the next. Each particle in this lattice has 12 neighbor particles at distance L.

4. Leap-Frog Triangular Lattice. The lattice is constructed by first arranging the particles in an equilateral triangular mesh in the x,y-plane, with the mesh constant L and one side of the triangle parallel to the x-axis. Then the mesh is translated in the z-direction by multiples of $L\sqrt{2/3}$, and in the y-direction alternatively by $\pm L/\sqrt{3}$. Thus, the triangular mesh is shifted in a leap-frog manner from one z-plane to the next. The number of neighbors at distance L from any particle is 12 in this lattice.

The minimum value of the gas volume fraction (closest packing of spheres) is obtained in the four lattices by setting the particle radius s equal to L/2. The numerical values of $\bar{\alpha}_{min}$ are as follows:

| | | |
|---|---|---|
| Square Cylinder | $\bar{\alpha}_{min} = 1 - \pi/6$ | $= 0.476$ |
| Triangular Cyliner | $\bar{\alpha}_{min} = 1 - \pi/(3\sqrt{3})$ | $= 0.395$ |
| Leap-Frog Lattices | $\bar{\alpha}_{min} = 1 - \pi/(3\sqrt{2})$ | $= 0.260$ |

Both leap-frog lattices are arrangements with closest packing of spheres in three dimensions.

The relation between $\bar{\alpha}$ and s/L is for all four lattices given by

$$\bar{\alpha} = 1 - 8(1-\bar{\alpha}_{min})(s/L)^3, \qquad (A.1)$$

and the relation between L and the mean distance $L_m$ is

$$L/L_m = (1-\bar{\alpha}_{min})^{1/3}. \qquad (A.2)$$

(See Appendix B.) Substituting the values of $\bar{\alpha}_{min}$ in Eq. (A.2) one finds that the value of $L_m/L$ is between 1.24 (square cylinder lattice) and 1.10 (leap-frog lattices).

## Appendix B

### Mean Distance and Number of Particles

Let m be the number of particles in an aggregate, W be the volume occupied by the aggregate, and v be the volume of each particle. Then one can conceptually assign to each particle the fraction W/m of the aggregate volume, and represent the fraction as a virtual sphere with the distance $L_m$. This diameter we define as the mean distance between the particle centers. It is given by the formula

$$\frac{W}{m} = \frac{\pi}{6} L_m^3 . \tag{B.1}$$

The gas volume fraction $\bar{\alpha}$ of the aggregate volume W is related to $L_m$ by

$$1 - \bar{\alpha} = \frac{mv}{W} = \frac{6v}{\pi} L_m^{-3} . \tag{B.2}$$

If the particles are spheres with the radius s, then one obtains from Eq. (B.2)

$$1 - \bar{\alpha} = 8(s/L_m)^3. \tag{B.3}$$

The minimum value of $\bar{\alpha}$ is obtained if s is a maximum. For the four lattices defined in Appendix A, the maximum value of s is L/2. Therefore, one obtains from Eq. (B.3) for the four lattices

$$1 - \bar{\alpha}_{min} = (L/L_m)^3. \tag{B.4}$$

If the averaging volume V is a sphere with radius R and N is the number of particles within the sphere, then one obtains from Eq. (B.1) the approximation

$$N \approx V(R) \cdot \frac{6}{\pi} L_m^{-3} = (2R/L_m)^3. \tag{B.5}$$

The approximation is due to the definition of $L_m$ as a mean distance for the whole particle aggregate, whereas Eq. (B.5) is for the number N of particles in V(R) which is only a fraction of W. If the aggregate occupies the whole space then Eq. (B.5) is exactly valid at the limit $R \to \infty$ :

$$L_m = \lim_{R\to\infty} \left(2R/N(R)^{1/3}\right). \tag{B.6}$$

## Appendix C

### Undulations of Average Descriptors

Let $\tilde{\phi}$ be a local gas property, for instance, density. Then the corresponding average descriptor $\phi$ is defined by

$$\phi = \frac{1}{V_{gas}} \int_{V_{gas}} \tilde{\phi}\, dv, \tag{C.1}$$

where $V_{gas}$ is that part of the averaging volume V, which is occupied by gas. If $\tilde{\phi}$ is constant, then $\phi \equiv \tilde{\phi}$ for any positive value of $V_{gas}=\alpha V$. If $\tilde{\phi}$ is not constant, then the distribution of particles in V does affect the value of the average descriptor $\phi$. We estimate the influence of undulations of the gas volume fraction $\alpha$ on the value of $\phi$ in the case where $\tilde{\phi}$ is a function with a constant gradient.

Let $\tilde{\phi}$ be the function

$$\tilde{\phi}(x,y,z) = \phi_0 + x\phi_x\ , \tag{C.2}$$

with constant $\phi_0$ and $\phi_x$. Let the particle volumes v be small compared to the averaging volume V and let $x_i$ be the x-coordinates of the centers of the N particles that are in V. Then the average descriptor $\phi$ can be approximated by

$$\phi(x) = \frac{1}{V-\Sigma v}\left\{(\phi_0+x\phi_x)V-\sum_{i=1}^{N}(\phi_0+x_i\phi_x)v\right\} = \phi_0 + x\phi_x - \phi_x\xi(1-\alpha)/\alpha, \tag{C.3}$$

where $\xi$ is the average of the deviations $x_i-x$ of the particle positions $x_i$ from the center of x of V.

Next, we relocate the averaging volume to the position $x+\Delta x$. The relocation will in general change the number of particles in V. Let $\Delta M$ be the number of particles that are added (scooped up) by the relocation, and $\Delta m$ be the number of particles that are lost by the relocation. The total number of particles in V is changed by the relocation by

$$\Delta N = \Delta M - \Delta m, \tag{C.4}$$

and the gas volume fraction changes by

$$\Delta\alpha = -\frac{V}{V}\,\Delta N = -(1-\alpha)\Delta N/N. \tag{C.5}$$

The new value of the the average descriptor is, within the same approximation as Eq. (C.3),

$$\phi(x+\Delta x) = \phi_0 + (x+\Delta x)\phi_x - \frac{1-\alpha}{\alpha+\Delta\alpha}\phi_x\left[\xi-\Delta x+\frac{1}{N}\left(\sum^{\Delta M}-\sum^{\Delta m}\right)(x_i-x-\Delta x)\right]. \tag{C.6}$$

We assume that V is a sphere with radius R. Then the average x-coordinate of the added particles may be estimated by $x+2R/3+\Delta x/2$, and the average x-coordinate of the lost particles by $x-2R/3+\Delta x/2$. The number of added and lost particles in a uniformly distributed aggregate (no undulations of $\alpha$) are

$$\Delta M = \Delta m = (1-\alpha)\Delta V/v = N\cdot 3\Delta x/(4R), \tag{C.7}$$

where $\Delta V = V3\Delta x/(4R)$ is the volume newly covered and lost by the relocation.

Next, we assume that the particle distribution is not uniform so that one has an excess of added particles over lost ones. We express this by setting

$$\left.\begin{aligned}\Delta M &= N\ 3\Delta x/(4R) + \Delta N,\\ \text{and}\qquad \Delta m &= N\ 3\Delta x/(4R).\end{aligned}\right\} \tag{C.8}$$

Substituting these values and the average coordinate estimates of added and lost particles in Eq. (C.6) one obtains

$$\begin{aligned}\phi(x+\Delta x) &= \phi_0 + (x+\Delta x)\phi_x - \frac{1-\alpha}{\alpha+\Delta\alpha}\phi_x\left[\xi+\frac{\Delta N}{N}\left(\frac{2R}{3}-\frac{\Delta x}{2}\right)\right] =\\ &= \phi_0 + (x+\Delta x)\phi_x - \frac{1-\alpha}{\alpha+\Delta\alpha}\phi_x\left[\xi-\frac{\Delta\alpha}{1-\alpha}\left(\frac{2R}{3}-\frac{\Delta x}{2}\right)\right].\end{aligned} \tag{C.9}$$

The difference between $\phi(x+\Delta x)$ and $\phi(x)$ is, therefore,

$$\phi(x+\Delta x) - \phi(x) = \Delta x \phi_x - \frac{1}{\alpha+\Delta\alpha}\phi_x 2R \frac{\Delta\alpha}{3}\left[1 - \frac{3}{2}\frac{\Delta x}{2R} + \frac{1-\alpha}{\alpha} 3 \frac{\xi}{2R}\right]. \qquad (C.10)$$

The last term on the right-hand side of Eq. (C.10) is the change of $\phi$ which is caused by an undulation $\Delta\alpha$ over the distance $\Delta x$. From Section 4, we know tnat typical wave lengths of the major undulations are between $L_m$ and $1.5L_m$. In order to obtain the amplitude of a corresponding undulation of $\phi$ one should set $\Delta x$ equal to one-fourth of the wave length. By setting, e.g., $\Delta x = L_m/3$ one obtains the following estimate of the amplitude of the undulation

$$|\Delta\phi| \approx |2R\phi_x| \frac{\Delta\alpha}{\alpha+\Delta\alpha}\frac{1}{3}\left[1 - \frac{L_m}{4R} + \frac{1-\alpha}{\alpha} 3 \frac{\xi}{2R}\right]. \qquad (C.11)$$

In Section 2 we found that a reasonable value of R is such that $L_m/4R<0.25$. Also, the ratio $\xi/R$ is likely much smaller than one. The expression $2R\phi_x$ is the change of $\phi$ along a diameter of the averaging volume, if undulations are not present. Let that change be $\delta\phi$ and let $|\Delta\alpha| << \alpha$. Then Eq. (C.11) can be simplified to

$$|\Delta\phi| \approx |\delta\phi|\ \Delta\alpha/(3\alpha). \qquad (C.12)$$

Undulations of $\alpha$ also can affect such average descriptors which have a zero gradient ($\delta\phi=0$), if the local property $\tilde{\phi}$ is affected by the presence of particles. One example of such a situation is the average temperature of a gas in which one places particles with a different temperature. If the gas is heated by conduction, then a short time later the gas temperature will have changed in a zone around each particle, with a corresponding change of the average temperature. That average value undulates with an amplitude proportional to the amplitude $\Delta\alpha$.

We demonstrate this by considering a simple model in which the gas has a constant local property $\phi_o+\delta_b\phi$ in a boundary region with a volume $\Delta v$ around each particle, and the constant local property $\phi_o$ elsewhere in the flow field. Then the average descriptor is

$$\phi = \frac{1}{V-Nv}\left[\left(V-N(v+\Delta v)\right)\phi_o + N\Delta v(\phi_o+\delta_b\phi)\right] = \phi_o + \delta_b\phi \frac{\Delta v}{v}\frac{1-\alpha}{\alpha}$$

$$= \phi_o + d\phi, \qquad (C.13)$$

where $d\phi$ is the change of the constant value $\phi_o$ due to the different local value in the boundary region. If $\alpha$ changes by $\Delta\alpha$, then the corresponding

change of $\phi$ is from Eq. (C.13)

$$\Delta\phi \approx -\delta_b\phi \frac{\Delta v}{v}\frac{\Delta\alpha}{\alpha^2} = -d\phi \frac{\Delta\alpha}{\alpha(1-\alpha)} . \tag{C.14}$$

It is interesting to notice that in this example the average descriptor of the particles is not affected by undulations of $\alpha$. The particle descriptor is defined by

$$\psi = \frac{1}{V_{par}} \int_{V_{par}} \tilde{\psi}\, dV, \tag{C.15}$$

where $V_{par}$ is the union of all particles that are located in the averaging volume. Let $\psi_o$ be a constant particle property and $\psi_o+\psi_b\delta$ be a property in a boundary region with the volume $\Delta v$ inside each particle. Then the average descriptor is

$$\psi = \frac{1}{Nv}\left[N(v-\Delta v)\psi_o + N\Delta v(\psi_o+\delta_b\psi)\right] = \psi_o + \delta_b\psi\frac{\Delta v}{v}. \tag{C.16}$$

Because $\psi$ is independent of $\alpha$, undulations of $\alpha$ do not influence the value of the average descriptor.

# NUMERICAL INVESTIGATION OF THE STABILITY OF DIFFUSION FLAMES NEAR EXTINCTION AND IGNITION*

Y.S. Choi, C. Laine-Schmidt and G.S.S. Ludford
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca, NY 14853

ABSTRACT. A numerical investigation is made of the near-ignition and near-extinction characteristics of chambered diffusion flames for arbitrary Lewis numbers. In particular, for an S-shaped response curve, both dynamic ignition and extinction are found to occur at the turning points.

1. INTRODUCTION. While the stability of premixed flames has received considerable attention in the mathematical theory of laminar flames that has been developed in the last decade or so, diffusion flames have almost been ignored. Even in the context of plane flames subjected to one-dimensional disturbances there is much to be done, in particular when the response curve of steady states is S-shaped (figure 1).

To be sure, Matalon & Ludford [1] have considered (numerically) the near-ignition stability of chambered diffusion flames, but only for unit Lewis numbers $L_F, L_O$ of fuel and oxidant. For $L_F, L_O$ arbitrary, even the steady states have only recently been determined (Choi [2]). Likewise, Buckmaster, Nachman & Taliaferro [3] have determined (analytically) the near-extinction stability characteristics of the counterflow diffusion flame for $L_F = L_O = 1$. (The problem is identical to that for a chambered flame.)

In view of their effect in premixed flames, it is of some importance to consider Lewis numbers different from 1, and that is the subject of the present paper. We find (numerically) that there is no effect on near-ignition stability of chambered diffusion flames: whatever the values of $L_F, L_O$, neutral stability occurs at the turning point, which is therefore the dynamic ignition point. Dynamic extinction also occurs at the (corresponding) turning point, but for a different reason: for all values of $L_F, L_O$, stability persists at the turning point. This conclusion contradicts Buckmaster, Nachman & Taliaferro and, moreover, reveals the existence of inaccessible steady states that are stable.

2. GOVERNING EQUATIONS FOR NEAR-IGNITION ANALYSIS. As is shown in [2], the equation for the steady state near ignition is

$$\frac{d^2 t_s}{d\eta^2} + Q\eta^{L_O - 2}(1 - \eta)^{L_F} e^{t_s} = 0, \tag{1}$$

$$t_s(0) = t_s(1) = 0, \tag{2}$$

where $t_s$ is the temperature perturbation in the flame zone and Q is a given positive constant. Numerics show that for each Q less than a certain



PREVIOUS PAGE IS BLANK

$Q_0(L_0, L_F)$. there are two solutions; for $Q = Q_0$ there is one; and for $Q > Q_0$ none.

From these solutions, the response

$$R = 1 + \delta_0 \left(\frac{dt_s}{d\eta}\right)\Big|_{\eta=1}$$

can be calculated, where $\delta_0 << 1$ is a (known) small positive constant. It gives rise to the upper and middle branches of the S-shaped response curve sketched in figure 1. The bend is approached as $Q \to Q_0$ and remote parts of the two branches as $Q \to 0$.

Stability analysis involves the solution of

$$\eta^2 \frac{d^2\phi}{d\eta^2} + [\lambda + Q\eta^{L_0}(1-\eta)^{L_F} e^{t_s}]\phi = 0, \qquad (3)$$

$$\phi(0) = \phi(1) = 0.$$

Here $\lambda$ is the eigenvalue: if the spectrum has non-negative real part for a steady state $t_s$, that the state is stable to the class of disturbances considered; otherwise it is unstable.

It can be shown that an unstable eigenvalue, if it exists, is real. Moreover, through asymptotic analysis as $Q \to 0$, the remote upper branch of the S-curve is found to be stable while the remote lower branch is unstable. In the latter case, there is just one unstable eigenvalue.

3. NUMERICAL RESULTS FOR IGNITION. In the work of Matalon & Ludford [1] on ignition for $L_0 = L_F = 1$, the steady state is found by shooting and then the stability problem is tackled by a Galerkin method. A different numerical scheme is adopted here, in which both problems are solved simultaneously and discretization of the differential operator is in terms of Tchebytchev polynomials. In addition, a continuation subroutine is incorporated to facilitate treatment of the bending point and automation of the program. The numercal result obtained by Matalon for unit Lewis number has been recomputed and confirmed.

For general Lewis number, we reach the same conclusion as in the special case $L_0 = L_F = 1$, namely,

i) for all $L_0, L_F$, the neutrally stable point is found to be exactly at the turning point. (i.e. the static ignition point);

ii) when an unstable eigenvalue exists, there is exactly one and it is real.

4. GOVERNING EQUATIONS FOR NEAR-EXTINCTION ANALYSIS. The near-extinction steady states (corresponding to the bottom half of the S-response) have also been described in [1]. The perturbation temperature satisfies the differential equation

$$\frac{d^2 t_s}{d\xi^2} = -KL_0 L_F (k_1\xi + k_2 - t_s)(k_3\xi + k_4 - t_s) e^{t_s} \tag{5}$$

and the boundary conditions

$$t_s = \begin{cases} k_1\xi - k_3 B + O(1) \\ \qquad\qquad\qquad \text{as} \quad \xi \to \mp\infty, \\ k_3\xi + k_1 C + O(1) \end{cases} \tag{6}$$

where $k_1$ is a known positive constant,

$k_3$ is a known negative constant,

$k_2 = PC$, $k_4 = QC$,

P, Q are known constant,

B, C are unknown constants.

Like Q in section 2, the constant K is positive. The numerics determine exactly two solutions for each K greater than a certain $K_0(L_0, L_F)$, exactly one for $K = K_0$ and none for $K < K_0$.

From these solutions the two responses

$$R = -\delta_a C,$$

where $\delta_a \ll 1$ is a (known) small positive constant, can be calculated for each $K > K_0$, thereby generating the middle and lower branches of the S-shaped response curve in figure 1. The bend is approached as $K \to K_0$ and remote parts of the two branches as $K \to \infty$.

The corresponding stability problem is

$$-\left[\frac{d^2\phi_T}{d\xi^2} + \lambda\phi_T\right] = L_0^{-1}\frac{d^2\phi_{Y_0}}{d\xi^2} + \lambda\phi_{Y_0} = L_F^{-1}\frac{d^2\phi_{YF}}{d^2} + \lambda\phi_{YF}.$$

$$= Ke^{t_s}(y_{0S}\phi_{YF} + y_{FS}\phi_{Y0} + y_{FS}y_{0S}\phi_T), \tag{7}$$

$$\phi_T(\pm\infty) = \phi_{Y0}(\pm\infty) = \phi_{YF}(\pm\infty) = 0, \tag{8}$$

where $y_{0s} = L_0(k_1\xi + k_2 - t_s)$, $y_{FS} = L_F(k_3\xi + k_4 - t_s)$

are the mass fractions corresponding to the steady state temperature perturbation $t_s$. Here $\lambda$ is the eigenvalue; if the spectrum has non-negative real part for a steady state $t_s$, that state is stable to the class of disturbance considered; otherwise it is unstable.

It can be shown analytically that the remote lower branch is stable for $L_0 = L_F$. For remote upper branch, the steady-state response depends on whether $k_1 < \frac{1}{2}$ or $k_1 > \frac{1}{2}$, but in either case we are unable to extract any analytical result about the stability.

5. NUMERICAL RESULTS FOR EXTINCTION. Due to the nature of the steady-state problem, it has to be treated separately from the stability problem, so that the numerical procedure is somewhat more complicated than that for ignition. The steady numerics are based on shooting for the correct value of C. Thus, the differential equation (5) is integrated backwards under the boundary condition (6b) for each estimate of C until $\frac{dt_s}{d\xi} \to k_1$ as $\xi \to -\infty$. The constant B is not involved. Once the results for two values of K are obtained, a continuation subroutine can be incorporated as for ignition. For the stability problem, the Tchebytchev polynomials are again used for the discretization of the differential operator.

Figure 2 gives the steady-state response for $L_F = 1$ and various values of $L_0$. The holes in the curves show where the steady state changes from being stable to unstable. In all cases the hole lies on the upper branch, implying that the lower part of the S in figure 1, including a portion of the middle branch, is stable. This result is true for other values of $L_F$ also, and we conclude that

(i) for all $L_0, L_F$, the neutrally stable point is above the turning point;

(ii) when an unstable eigenvalue exists, there is exactly one and it is real.

Note that the dynamic extinction point still lies at the turning point and that there are inaccessible stable states on the middle branch of S.

Conclusion (i) contradicts the recent analytical results [3] of Buckmaster, Nachman & Taliaferro for $L_0 = L_F = 1$, who found that the neutrally stable point lies exactly at the turning point. (They consider the counterflow diffusion flame instead of the chambered diffusion flame, but the stability problems are identical.) We are presently trying to reconcile our work with theirs.

REFERENCES.


[1] M. Matalon & G.S.S. Ludford (1983). On the near-ignition stability of diffusion flames. Int. J. Engng. Sci. 18, 1017.

[2] Y.-S. Choi (1985). Chambered Diffusion Flames for Arbitrary Lewis Numbers. Ph.D. Thesis, Cornell University.

[3] J. Buckmaster, A. Nachman & S. Faliaferro (1983). The fast-time instability of diffusion flames. Physica D9, 408.

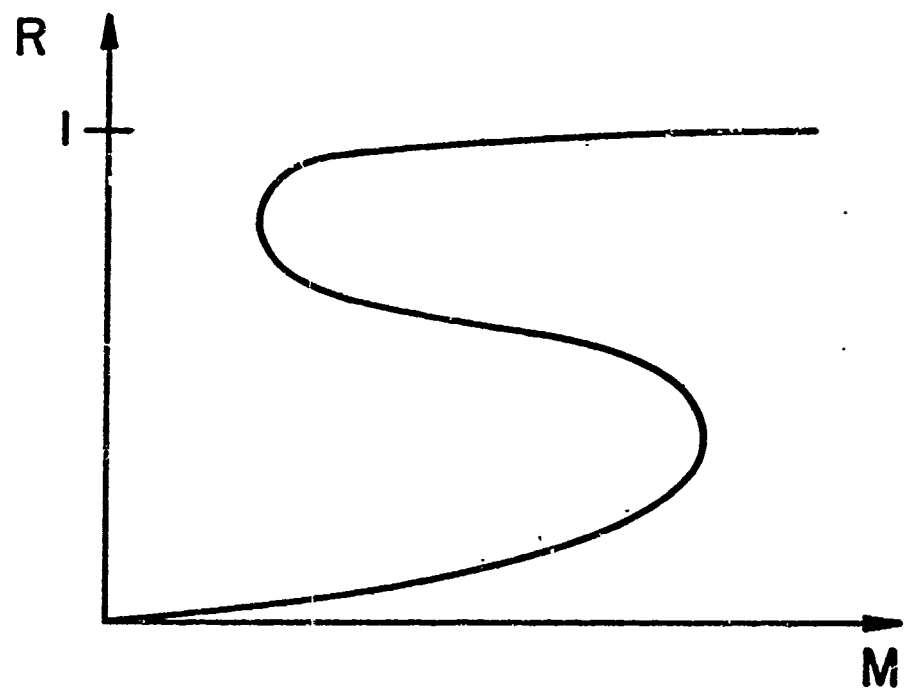



Figure 1. S-shaped response for chambered diffusion flame: R is the fraction of unburnt fuel and M is the injection rate.

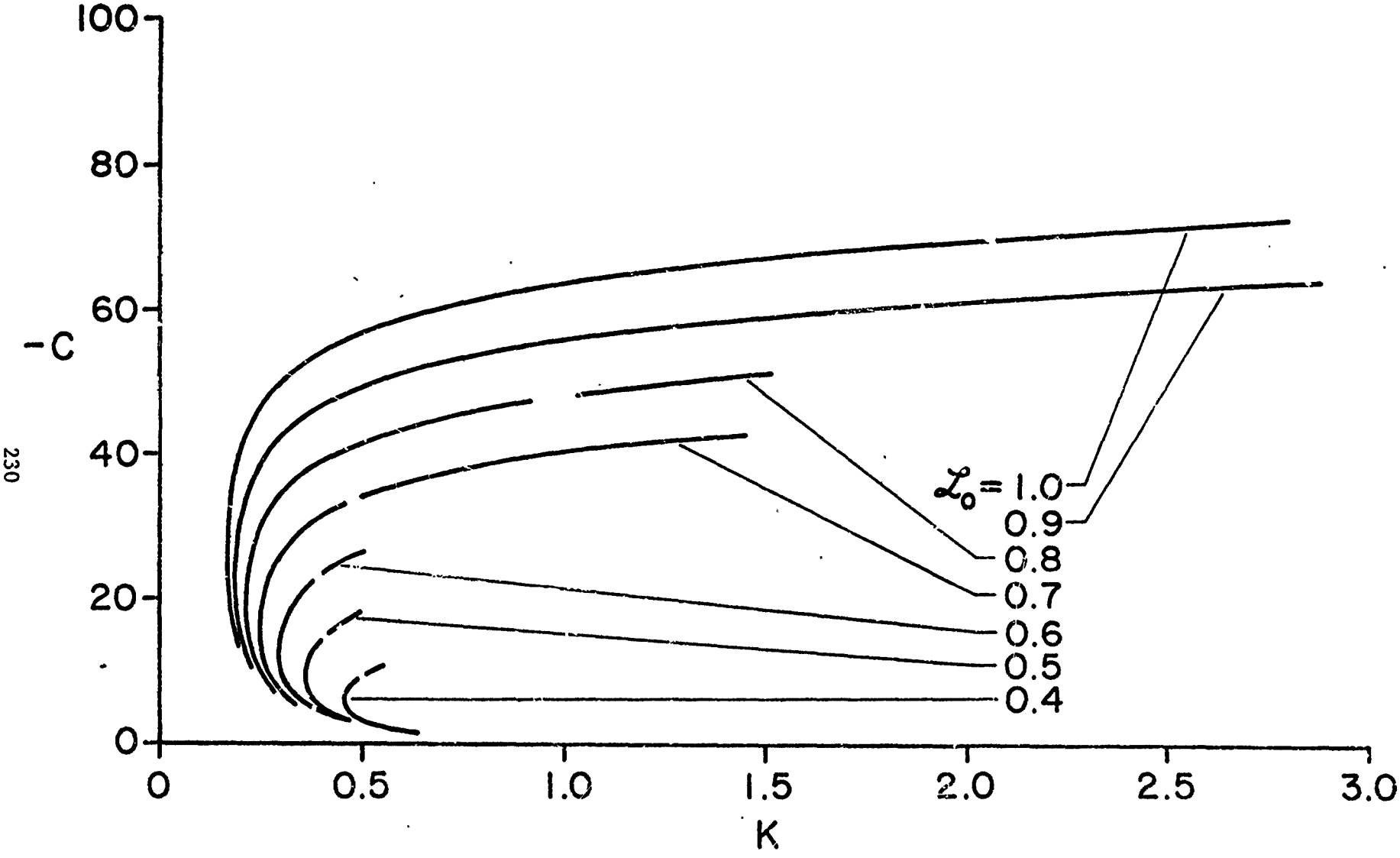



Figure 2. Stability of the steady states near extinction for $L_p = 1$ .

# FLUID MECHANICS OF QUENCHING

Donald A. Drew, Ronald Brent,
Susan Melly, William Schroeder and Stephen Wells
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12181

ABSTRACT

An array of heated rods is lowered vertically in a cold water bath at a constant speed V in order to quench them to obtain desired mechanical properties. Relative to the rods, the water flows in a subchannel, is heated, and boils, while cooling the rods. A model is proposed and studied which considers a one dimensional flow in a subchannel. It is argued that the heat release occurs in a thin region, where water is heated to boiling conditions and boils completely to steam. Above this boiling layer, steam flows rapidly against the friction of the rod bundle. Below the boiling layer, the water flow is approximately hydrostatic. This results in the boiling layer moving at a constant speed proportional to V. The effect of cross flow (leaking into or out of the channel) is also investigated, and the results discussed.

1. INTRODUCTION Metal components can be hardened by heating, followed by rapid cooling. The rapid cooling is often accomplished by immersing the components in a fluid (often water).

In the problem we consider, the component is a metal rod. For manufacturing efficiency, a relatively large array of heated rods is lowered lengthwise into the fluid. The purpose of the model derived and analyzed in this paper is to describe the fluid mechanics in the subchannels of the rod array. In so doing, we gain some insight into the mechanisms involved in the process, and their role in normal and abnormal operation.

An array of rods (Fig. 1 represents a cross section) is lowered vertically into a cold quiescent fluid at a constant speed V. The rods are assumed to be at a constant uniform initial temperature. We model the situation as a number of vertical subchannels which are interconnected through the gaps between the rods. For simplicity, we shall consider one subchannel, and assume that it consists of several rods, one interior flow area, and a number of gaps allowing subchannel fluid and surrounding fluid to mix. See Figure 2. The analysis presented herein is based on the assumption that the cross-flow through the gaps is small, and utilizes a perturbation analysis starting from the flow in the subchannel without cross-flow.

The approach taken is a modification of one used by Lahey and Moody 1977, and Achard, Drew and Lahey 1981. Let us consider one-dimensional flow in a vertical channel of cross-sectional area A. We shall describe the flow relative to a coordinate system fixed in the rods, with z=0 defining the inlet to the subchannel at the bottom of the rods. We shall assume that boiling starts at $z = \lambda_1(t)$, and that by $z = \lambda_2(t)$, the liquid has boiled away. Thus, the region $0 < z < \lambda_1(t)$ contains liquid only, the region $\lambda_2(t) < z < L$ contains steam only, while the region $\lambda_1(t) < z < \lambda_2(t)$ contains a mixture of steam and water. The boiling boundary $z = \lambda_1(t)$ occurs when the incoming fluid reaches the boiling temperature $T_b$. The dryout point $z = \lambda_2(t)$ occurs when the two-phase mixture has absorbed enough heat to boil all the liquid.

Boiling is a very efficient heat transfer process (Rohsenow and Hartnett 1973). In this problem, if liquid contacts the heat transfer surface (i.e. the rod) and the surface is at a temperature above the boiling temperature, the liquid will boil until the surface temperature is equal to the boiling temperature. In contrast, in regions ($z > \lambda_2(t)$) where only steam contacts the rods, the heat transfer is very low. Thus, the rods lose large amounts of heat in a short time in the region $\lambda_1(t) < z < \lambda_2(t)$. This region is therefore relatively thin. We shall assume that this region is negligibly thin. We then have $\lambda_1(t) = \lambda_2(t) = \lambda(t)$. We shall refer to $z = \lambda(t)$ as the position of boiling heat transfer.

We shall further assume that the convective heat transfer to the liquid dominant in the region $0 < z < \lambda(t)$ (where $T < T_b$) is also efficient, so that we can assume that the rods and the liquid are all at the ambient liquid temperature $T_0$ there.

2. GOVERNING EQUATIONS Let us now discuss the equations of governing the motion in the various regions. First, we consider the subcooled region $0 < z < \lambda(t)$. The equation of conservation of mass is

$$\frac{\partial j}{\partial z} = \frac{h}{A} w(z,t), \tag{1}$$

where j is the velocity of the liquid, h is the gap, and w is the cross-flow velocity (positive into the channel).

The equation of conservation of momentum in the z-direction is

$$\frac{\partial j}{\partial t} + \frac{\partial j^2}{\partial z} = - \frac{1}{\rho} \frac{\partial p}{\partial z} - g - \frac{P}{\rho_\ell A} \tau + \frac{hw}{A} v_c , \tag{2}$$

where p is the pressure, g is the gravitational acceleration, P is the frictional perimeter of the channel, i.e. the part of the channel boundary which contacts the rods. Also, $\tau$ is the frictional force per unit area at this boundary. Finally, the last term represents the rate of addition of z-momentum to the channel fluid due to the cross-flow. The cross flow carries with it velocity $v_c$. If $w<0$, we shall assume that $v_c = V$. This assumes that the fluid coming into the channel is the quiescent fluid outside the channel. If $w<0$, we shall assume $v_c = j$.

The frictional force is modelled by

$$\tau = f\rho\, j^2 , \tag{3}$$

where f is called the Fanning friction factor. It is usually assigned a value of $f = 0.02$, although its value depends on the geometry and the flow regime (laminar or turbulent).

As discussed, the energy equations result in

$$T_r(z,t) = T(z,t) = T. \tag{4}$$

for the subcooled region.

For the post dry-out region $\lambda(t) < z < L$, where L is the length of the

rods, we assume that the mass and momentum balances are similar to those for the subcooled region. Thus, we have

$$\frac{\partial \rho_s}{\partial t} + \frac{\partial \rho_s j_s}{\partial z} = \frac{h}{A} w \rho_s \tag{5}$$

The momentum equation is

$$\rho_s \left( \frac{\partial j_s}{\partial t} + j_s \frac{\partial j_s}{\partial z} \right) = - \frac{\partial p}{\partial z} - \rho_s g - \frac{Pf}{A} \rho_s j_s^2 + \frac{hw}{A} \rho_s v_c \tag{6}$$

The assumption of negligibly small heat transfer between the steam and the rods leads in

$$T_s = T_b \tag{7}$$

$$T_r = T_\infty \tag{8}$$

An equation of state is needed. We shall take

$$p = \rho_s R T_s \tag{9}$$

The boiling occurs in a relatively thin region around $z = \lambda(t)$. It is important to conserve mass and energy across this thin region. Consider a moving control volume of thickness $\Delta z$ which straddles $z = \lambda(t)$. Conservation of energy in the fluid stream requires

$$\rho_\ell c_\ell T_0 A \left( j - \frac{d\lambda}{dt} \right)\Bigg|_{z = \lambda - \frac{\Delta z}{2}} - \rho_s c_s A T_b \left( j_s - \frac{d\lambda}{dt} \right)\Bigg|_{z = \lambda + \frac{\Delta z}{2}} + \int_{\lambda - \frac{\Delta z}{2}}^{\lambda + \frac{\Delta z}{2}} Pq(z',t)dz' - \int_{\lambda - \frac{\Delta z}{2}}^{\lambda + \frac{\Delta z}{2}} h_{fg} \Gamma A \, dz' = 0 \ . \tag{10}$$

Here $c_\ell$, $c_s$, and $c_r$ are the specific heats of the liquid, steam and the rods, A is the cross sectional area of the rods; q is the rate of heat flow from the bars to the fluid per unit area, and $\Gamma$ is the rate of change of liquid to steam per unit volume.

Conservation of energy for the rods gives

$$\rho_r c_r A_r T_r \left(-\frac{d\lambda}{dt}\right)\Bigg|_{z=\lambda-\frac{\Delta z}{2}} - \rho_r c_r A_r T_r \left(-\frac{d\lambda}{dt}\right)\Bigg|_{z=\lambda+\frac{\Delta z}{2}} \tag{11}$$

$$= \int_{\lambda-\frac{\Delta z}{2}}^{\lambda+\frac{\Delta z}{2}} Pq\, dz' \quad .$$

Conservation of total mass in the control volume is given by

$$\rho_\ell A\left(j_\ell - \frac{d\lambda}{dt}\right)\Bigg|_{\lambda-\frac{\Delta z}{2}} = \rho_s A\left(j_s - \frac{d\lambda}{dt}\right)\Bigg|_{\lambda+\frac{\Delta z}{2}} \tag{12}$$

Lastly, conservation of mass of liquid in the control volume gives

$$\rho_\ell A\left(j_\ell - \frac{d\lambda}{dt}\right)\Bigg|_{\lambda-\frac{\Delta z}{2}} = \int_{\lambda-\frac{\Delta z}{2}}^{\lambda+\frac{\Delta z}{2}} \Gamma A\, dz' \tag{13}$$

Substitution of 11-13 into (10), assuming that $T_r(\lambda + \frac{\Delta z}{2}) = T_0$, $T_r(\lambda - \frac{\Delta z}{2}) = T_\infty$ and letting $\Delta z \to 0$ yields

$$\frac{d\lambda}{dt} = j(\lambda(t),t)\, Q \quad , \tag{14}$$

where

$$Q = \frac{\rho_\ell A(c_s T_b + h_{fg} - c_\ell T_0)}{[\rho_r c_r A_r (T_\infty - T_0) + \rho_\ell A(c_s T_b + h_{fg} - c_\ell T_0)]} \tag{15}$$

The momentum jump condition across the boiling heat transfer region is

$$p(\lambda^+,t) - p(\lambda^-,t) + \rho_s^+ j_s^+ \left(j_s^+ - \frac{d\lambda}{dt}\right) - \rho_\ell j_\ell^- \left(j_\ell^- - \frac{d\lambda}{dt}\right) = 0 \tag{16}$$

The cross-flow velocity at any level z is related to the pressure drop across the gap connecting the channel with its surroundings. Thus, we have

$$p^{(0)} - p^{(s)} = \Delta p = K\rho^* |w| w \quad , \tag{17}$$

where w is the cross-flow velocity in a single gap, $p^{(0)}$ is the pressure outside the channel, $p^{(s)}$ is the stagnation pressure inside the channel, defined by $p^{(s)} = p + \rho j^2/2$, $\rho^*$ is the density of the fluid flowing through the gap, and K is an orifice parameter.

For an ideal orifice, the parameter K is given by the Venturi relation

$$K = \left(1 - \frac{A_2^2}{A_1^2}\right) \quad , \tag{18}$$

where $A_2$ is the gap area, and $A_1$ is the unrestricted area. For a square array of rods[2] $A_2^2/A_1^2 = [h/(h+2r_0)]^2$ where h is the gap width and $r_0$ is the radius of the rods.

Finally, the pressure in the fluid outside the channel is assumed to be hydrostatic, so that

$$p^{(0)}(z,t) = \rho_\ell g(Vt-z) + p_a \quad . \tag{19}$$

A pressure boundary condition is needed at the top of the array. We shall assume that Bernoulli's equation holds in the exiting stream. Thus,

$$p(L,t) + \tfrac{1}{2}\rho_s j_s^2 \Big|_{z=L} = p_a \tag{20}$$

We shall make two further approximations. They are

$$\rho_s/\rho_\ell << 1 \tag{21}$$

and

$$|p-p_a|/p_a << 1 \quad . \tag{22}$$

The first assumption (21) results in several further approximations. Consider the momentum equation in the post dry-out region (6). The pressure must drop from roughly hydrostatic at $z=\lambda$ to atmospheric at z=L. This implies that $j_s$ must be appreciable, so that $\rho_s j_s^2 \sim \rho_\ell g L$. Further, examining (12) shows that

$$j_s >> j_\ell \tag{23}$$

his implies that

$$\rho_\ell j_\ell^2 << \rho_\ell g L . \tag{24}$$

hus, we can take the pressure in the subcooled region to be approximately ydrostatic:

$$p = \rho_\ell g(Vt-z) + p_a \tag{25}$$

or $0 < z < \lambda(t)$.

This further implies that

$$w \equiv 0 \tag{26}$$

or $0 < z < \lambda(t)$. Therefore, from eq. (1),

$$j_\ell = j_\ell(t) \tag{27}$$

hich, from eq. (14) implies

$$\frac{d\lambda}{dt} = j_\ell(t)\ Q \tag{28}$$

Assumption (22), with the equation of state (9) and eq. (7) gives

$$\rho_s = \rho_{s_0} = \text{const.} \tag{29}$$

f the steam density is constant, then eq. (5) gives

$$\frac{\partial j_s}{\partial z} = \frac{h}{A} w . \tag{30}$$

Lastly, the perturbed pressure $p' = p - p_a$, in the post dry-out region is iven approximately by

$$\rho_{s_0} j_s \frac{\partial j_s}{\partial z} = - \frac{\partial p'}{\partial z} - F\ \rho_{s_0} j_s^2 . \tag{31}$$

here $F = Pf/A$. Note that use of eq. (31) means that the initial condition on s is not needed. Indeed, the approximation implied by $j_s >> j_\ell$ results in he singular perturbation of eq. (6), with eq. (31) giving the outer solution. resumably, an initial condition not satisfying eq. (31) will rapidly equilirate to eq. (31).

Finally, eq. (12) combined with eq. (21) and eq. (28) gives

$$\rho_{s_0} j_s^+ = \rho_\ell\left(j_\ell(t) - \frac{d\lambda}{dt}\right) = \rho_\ell\left(\frac{1-Q}{Q}\right)\frac{d\lambda}{dt} , \tag{32}$$

and eq. (16) gives

$$p(\lambda^+,t) = p_a + \rho_\ell g(Vt-\lambda) - \frac{\rho_\ell^2}{\rho_{s_0}}\left(\frac{\varrho-1}{\varrho}\right)^2\left(\frac{d\lambda}{dt}\right)^2 \tag{33}$$

3. PERTURBATION SOLUTION. We shall assume that the cross-flow velocity w is small compared to $j_s$. We write $w = \delta w_1$, where $\delta << 1$. Then we expand the the relevant functions $j_s$, $p'$ as $(\ )_0 + \delta(\ )_1$. From eqs. (30) and (32), we have

$$j_{s_0} = \frac{\rho_\ell}{\rho_{s_0}}\left(\frac{1-\varrho}{\varrho}\right)\frac{d\lambda_0}{dt} \quad . \tag{34}$$

Eqs. (31) and (33) then gives

$$p_0(L,t) = -[1+F(L-\lambda_0)]\frac{\rho_\ell^2}{\rho_{s_0}}\left(\frac{\varrho-1}{\varrho}\right)^2\left(\frac{d\lambda_0}{dt}\right)^2 + \rho_\ell g(Vt-\lambda_0) \tag{35}$$

Finally, eq. (20) gives an equation for $\lambda(t)$. It is

$$\frac{d\lambda_0}{dt} = \left(\frac{\rho_{s_0} g}{\rho_\ell}\right)^{1/2}\left(\frac{\varrho}{1-\varrho}\right)\left(\frac{Vt-\lambda_0}{\frac{1}{2}+F(L-\lambda_0)}\right)^{1/2} \tag{36}$$

Nondimensionalization with $\hat{\lambda}_0 = \lambda_0/L$, $\hat{t} = t/(L/V)$ results in

$$\frac{d\hat{\lambda}_0}{d\hat{t}} = \left(\frac{\rho_{s_0} gL}{\rho_\ell V^2}\right)^{1/2}\left(\frac{\varrho}{1-\varrho}\right)\left(\frac{\hat{t}-\hat{\lambda}_0}{\frac{1}{2}+FL(1-\hat{\lambda}_0)}\right)^{1/2} \quad . \tag{37}$$

Let us now calculate the $O(\delta)$ correction to the flow due to cross-flow. First, we substitute the $O(\delta^0)$ pressures in eq. (17), we find

$$\delta w_1 = j_{s_0}\left(\frac{F}{K}\right)^{1/2}[(L-z) + r(Vt-z)] \tag{38a}$$

for $\lambda_0 < z < Vt$, where $r = \dfrac{\rho_\ell g}{\rho_{s_0} F j_{s_0}^2}$ and

$$\delta w_1 = j_{s_0}\left(\frac{F}{K}\right)^{1/2}(G-z)^{1/2} \tag{38b}$$

for $Vt < z < L$. Integrating eq. (5) for the two subregions gives

$$\delta j_{s_1} = C(t) + \beta \frac{1}{1+r} j_{s_0} L^{-3/2}[(L-z) + r\,(Vt-z)]^{3/2} \tag{39a}$$

for $\lambda_0 < z < Vt$, where $\beta = \frac{2}{3}\frac{hL}{A}\left(\frac{\overline{FL}}{K}\right)^{\frac{1}{2}}$ and

$$\delta j_{s_1} = D(t) + (\rho_a/\rho_\ell)\beta\, j_{s_0} L^{-3/2}(L-z)^{3/2}$$

for $Vt < z < L$. Here $\rho_a$ is the density of the atmosphere around the exposed top of the bundle. If we assume $\rho_a << \rho_\ell$, we have

$$\delta j_{s_1} \cong D(t), \tag{39b}$$

The velocity will be continuous at $z = Vt$ if

$$D(t) = C(t) + \beta\, j_{s_0} L^{-3/2}(L-Vt)^{3/2}/(1+r) \tag{40}$$

The pressure is given by

$$\delta p_1 = -\rho_{s_0} j_{s_0}\left[2F\int_{\lambda_0}^{z} \delta j_1\, dz' + \delta j_{s_1}(z,t) - \delta j_{s_1}(\lambda_0,t)\right] + \delta p_1(\lambda_0^+,t) \tag{41a}$$

for $\lambda_0 < z < Vt$, and

$$\delta p_1 = -\rho_{s_0} j_{s_0}\left[2F\int_{\lambda_0}^{z} \delta j_1\, dz' + \delta j_{s_1}(z,t) - \delta j_{s_1}(L,t)\right] + \delta p_1(L,t) \tag{41b}$$

for $Vt < z < L$.

With

$$\delta p_1(L,t) = -\rho_{s_0} j_{s_0} \delta j_{s_1}(L,t)$$

$$= -\rho_{s_0} j_{s_0} D(t) \tag{42a}$$

and

$$\delta p_1(\lambda_0^+,t) = -\frac{\partial p_0}{\partial z}(\lambda_0^+,t)\delta\lambda_1(t) + \delta p_1(\lambda_0^-,t)$$

$$+ \frac{\partial p_0}{\partial z}(\lambda_0^-,t)\delta\lambda_1(t) - 2\rho_{s_0} j_{s_0}\delta j_{s_1}(\lambda_0^+,t)$$

$$= (F\rho_{s_0} j_{s_0}^2 - \rho_\ell g)\delta\lambda_1(t)$$

$$-2\rho_{s_0} j_{s_0}\left[C(t) - \beta\frac{1}{1+r} j_{s_0} L^{-3/2}(L+rVt - (1+r)\lambda_0)^{3/2}\right] \quad (42b)$$

substituted in eqs. (41), we have

$$\delta p_1(z,t) = -\rho_{s_0} j_{s_0}\left\{2F\left[C(t)(z-\lambda_0) - \frac{2}{5}\beta j_{s_0} L^{-3/2}/(1+r)^2\right.\right.$$

$$\left[(L-z) + r(Vt-z)\right]^{5/2}]$$

$$+ \beta j_{s_0}\frac{1}{1+r} L^{-3/2}\left(\left[(L-z) + (Vt-z)\right]^{3/2}\right.$$

$$\left.\left.+ \left[(L-\lambda_c) + r(Vt-\lambda_0)\right]^{3/2}\right)\right]\Big\}$$

$$+ (F\rho_{s_0} j_{s_0}^2 - \rho_\ell g)\delta\lambda_1(t)$$

$$- 2\rho_{s_0} j_{s_0} C(t) \quad (43a)$$

for $\lambda_0 < z < Vt$, and

$$\delta p_1(z,t) = -2\rho_{s_0} j_{s_0} D(t)[F(z-L) + 1] \quad (43b)$$

for $Vt < z < L$. Continuity of the pressure at $z = Vt$ gives

$$C = \frac{1}{2F(L-\lambda_0)} \left[-\beta j_{s_0} \left[2F\, L^{-3/2}\, (L-Vt)^{5/2} \frac{1}{5} \frac{3+5r}{(1+r)^2}\right.\right.$$

$$+ L^{-3/2} \left\{(L-Vt)^{3/2} - [L-\lambda_0 + r(Vt-\lambda_0)]^{3/2}\right\} \frac{1}{1+r}$$

$$\left.\left. - Fj_{s_0}(1-r)\delta\lambda_1\right] \right. \tag{44}$$

From eq. (14), we have

$$\frac{d\delta\lambda_1}{dt} = Q\, \delta j_{\ell 1} \tag{45}$$

and, from eq. (32) we have

$$\frac{d\delta\lambda_1}{dt} = \frac{\rho_{s_0}}{\rho_\ell} \frac{Q}{1-Q}\, \delta j_{s_1} \tag{46}$$

Defining $\delta\hat{\lambda}_1 = \delta\lambda_1/L$ and $\hat{j}_{s_0} = j_{s_0}/V$, we have

$$\frac{d\delta\hat{\lambda}_1}{d\hat{t}} = \frac{\rho_{s_0}}{\rho_\ell} \frac{Q}{1-Q} \frac{1}{FL(1-\hat{\lambda}_0)} \left\{-\beta\hat{j}_{s_0} \left[\frac{2}{5} FL(1-\hat{t})^{5/2} \frac{3+5r}{(1+r)^2}\right.\right.$$

$$\left. + (1-\hat{t})^{3/2} \frac{1}{1+r}\right\}$$

$$+ \frac{\hat{j}_{s_0}(1-r)\delta\hat{\lambda}_1}{2(1-\hat{\lambda}_0)}$$

$$\left. + \frac{1}{1+r}\left(1 - \frac{1}{2FL(1-\hat{\lambda}_0)}\right)\beta\hat{j}_{s_0}(1-\hat{\lambda}_0 + r(\hat{t}-\hat{\lambda}_0))^{3/2}\right. \tag{47}$$

The initial condition on $\hat{\lambda}_1$ is

$$\delta\hat{\lambda}_1(0) = 0 \ . \tag{48}$$

4. DISCUSSION. Figures 3 through 6 show several cases of $\lambda$ versus t, for various values of h, $r_0$ and L. The quantity of practical importance is the part of the rods exposed to water on the outside ($z < Vt$) and to steam on the inside ($z > \lambda(t)$). Generally, the maximum fraction of the rod exposed to this extreme situation occurs somewhere in the middle of the process. In Figures 4 and 5, however, the extreme occurs in the middle of the process. Figures 7 through 10 show the maximum fraction exposed versus the relative gap size $h/r_0$ for various values of $r_0$ and L. Note that increasing the spacing between the rods decreases the maximum fraction of the bar exposed.

Let us discuss the model and the process. The model attempts to capture the essential physics leading to extreme behavior in the quenching fluid. The two essential effects resisting the hydrostatic filling of a fluid subchannel are friction and "rocketing". Friction is largest in the steam. "Rocketing" is a large momentum change near the relatively thin boiling region. A Bernoulli effect at the channel outlet assists the filling. Indeed, in this model the effect of crossflow from the outside into the subchannel is to increase the Bernoulli effect at the top, and hence to raise the level of the liquid.

The model suggests that decreasing resistance to steam flow in the array will raise the water level, and hence decrease the maximum fraction exposed to extreme conditions. This can be done by increasing the spacing between the bars. This also will increase crossflow, and effect which also seems to and the minimization of maximum fraction exposed. These predictions concur with several observations about the process.

1. The process works for one rod. For no friction, the model predicts $\lambda = Vt$.

2. Lengthening the rods and decreasing their diameter (and, presumably, packing them more closely in the bundle) increases friction, and increases the maximum fraction expcsed. The process is known to work for shorter, thicker rods, and has some problems with longer, thinner rods.

Acknowledgement

The work reported herein was started as a project in the course Advanced Mathematical Modeling at RPI. The students (RB, SM, WS and SW) worked out the framework for the analysis. Dr. John Vasilakis of Benet Laboratories, Watervliet Arsenal, assisted in monitoring the project. Partial support of the U. S. Army Research Office is greatfully acknowledged.

## REFERENCES


[1] Achard, J. L., Drew, D. A. and Lahey, R. T. 1981. The Effect of Gravity and Friction on the Stability of Boiling Flow in a Channel, Chem. Eng. Comm. 8.

[2] Lahey, R. T. and Moody, F. 1977. The Thermal-Hydraulics of a Boiling Water Nuclear Reactor, ANS Monograph.

[3] Rohsenow, W. and Hartnett, J. P. 1973. Handbook of Heat Transfer, McGraw Hill.

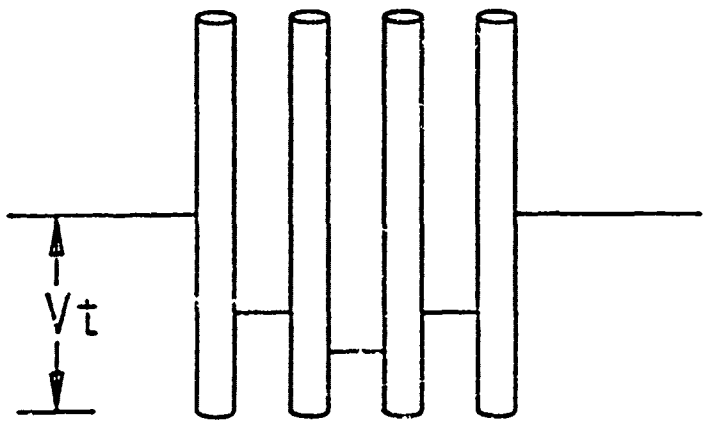



Figure 1. Flow Geometry

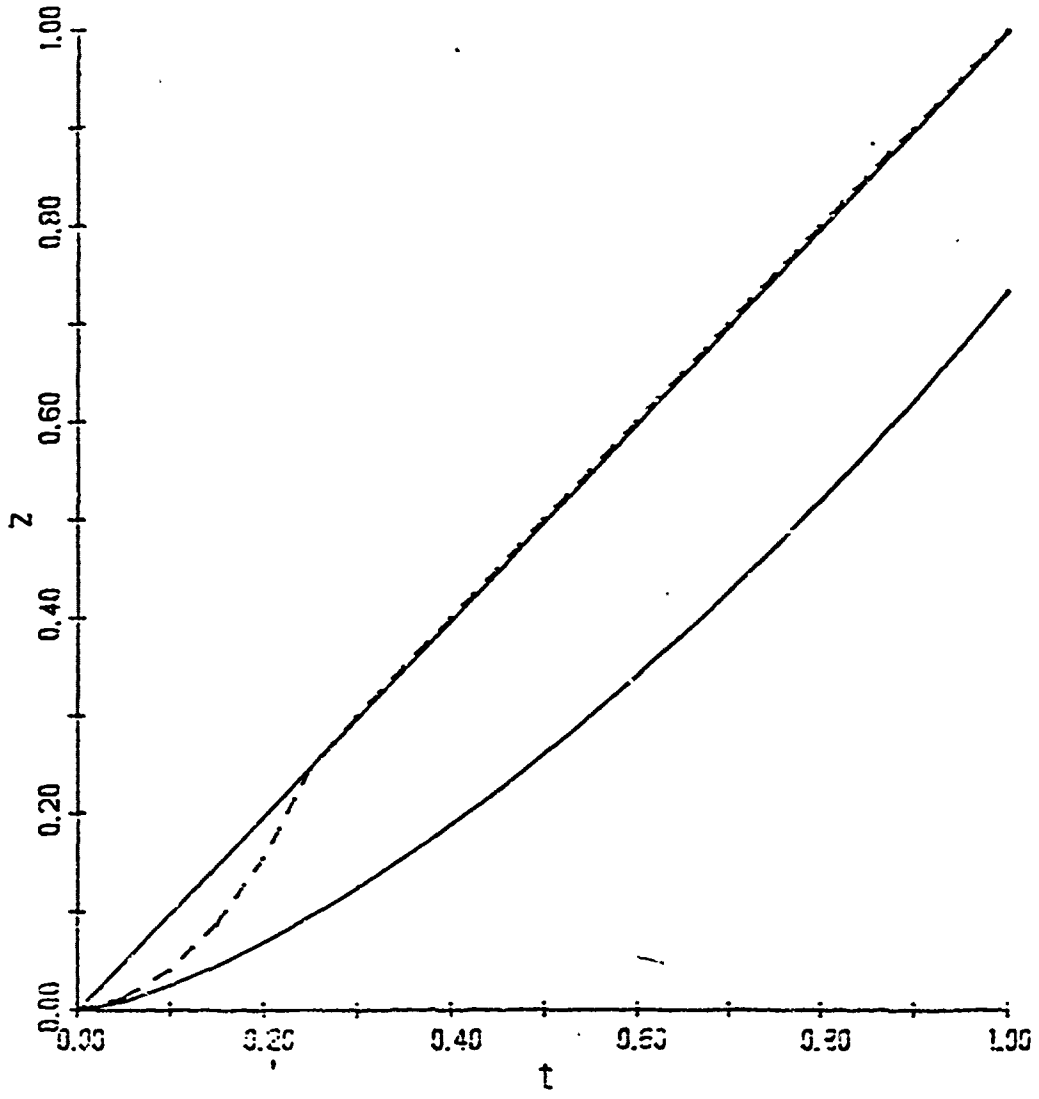



Figure 3. The height of liquid in the channel $z = \lambda(t)$ (—) without crossflow, and (---) with crossflow. The outside level $z = Vt$ is also shown. Here $r = 0.0127$m, $L = 0.457$m, and $h/r = 0.015$.

PREVIOUS PAGE IS BLANK

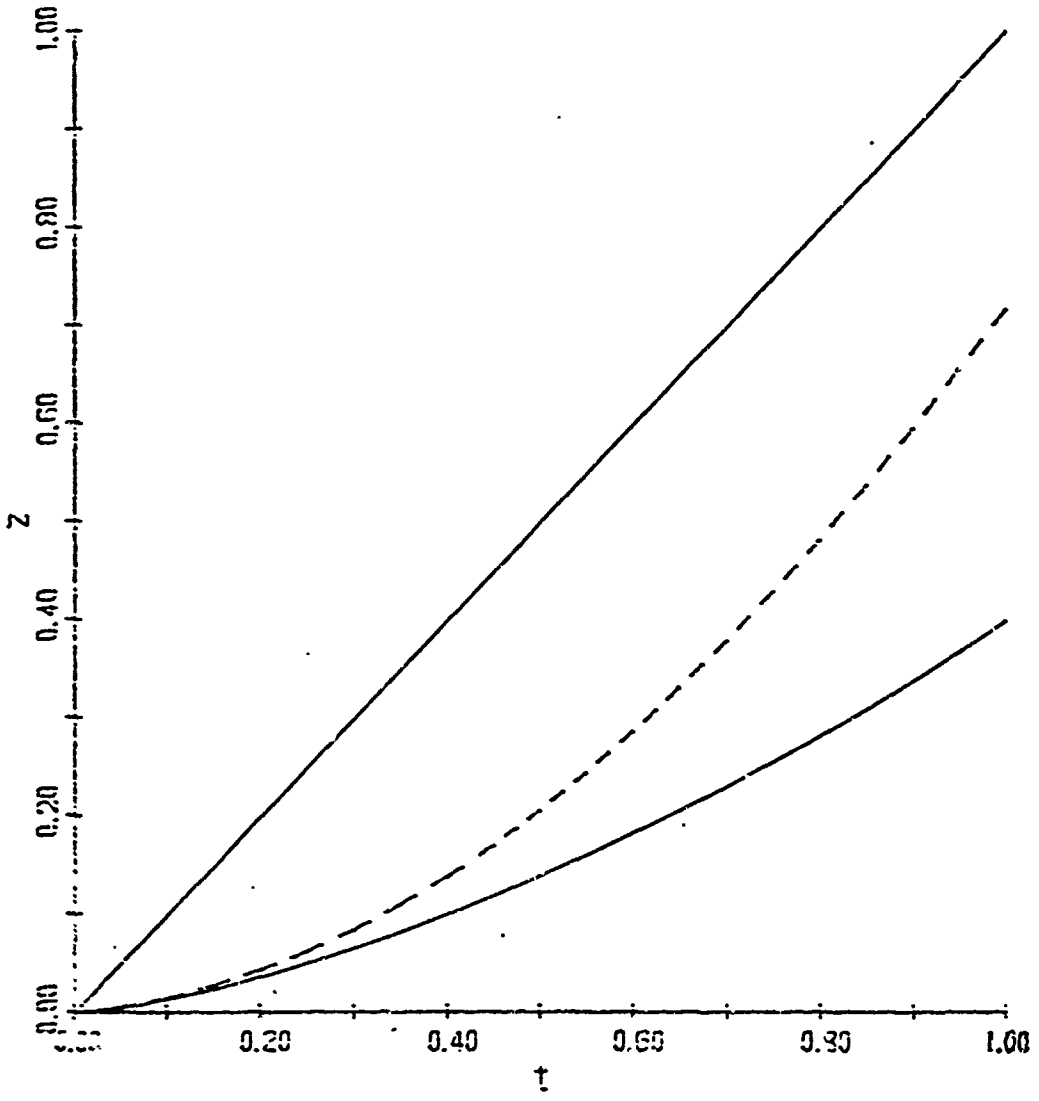



Figure 4. Same as Fig. 3 except h/r = 0.005

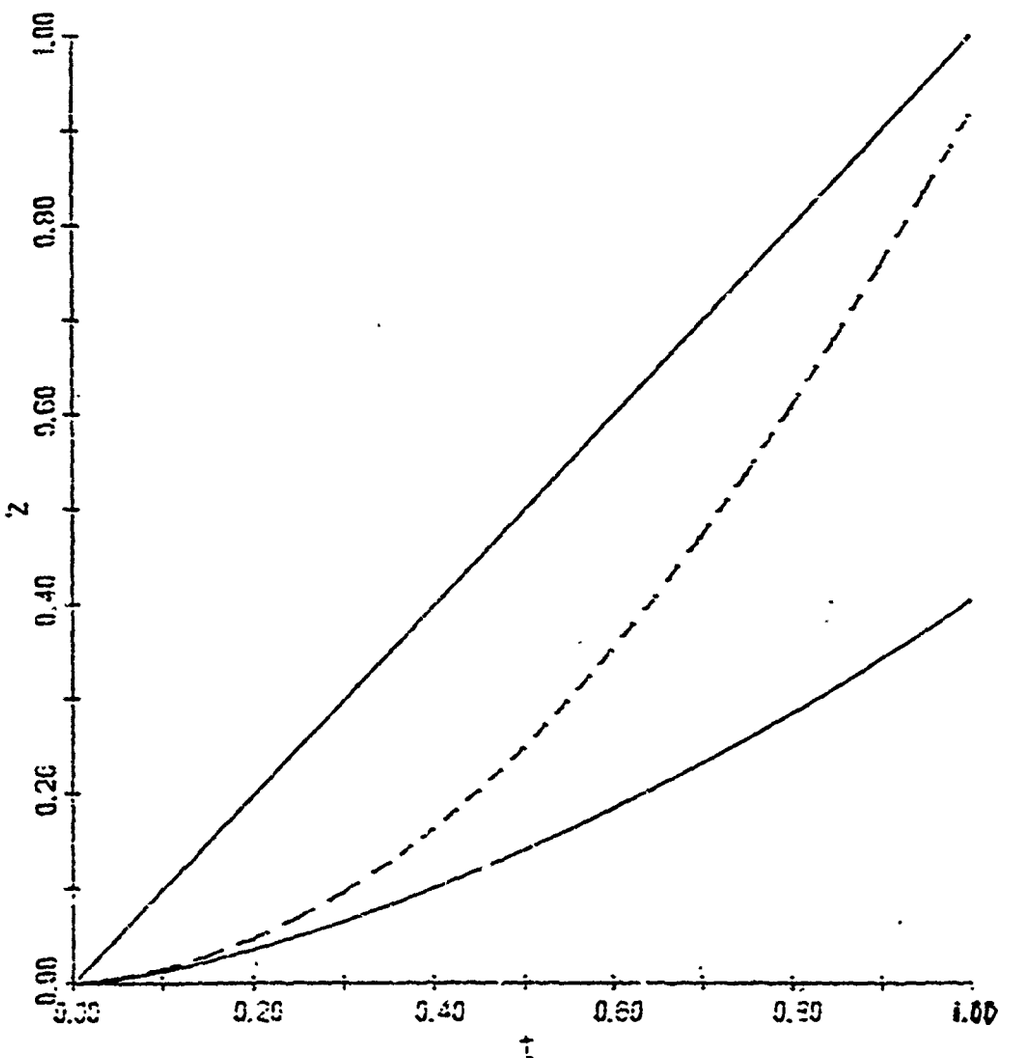
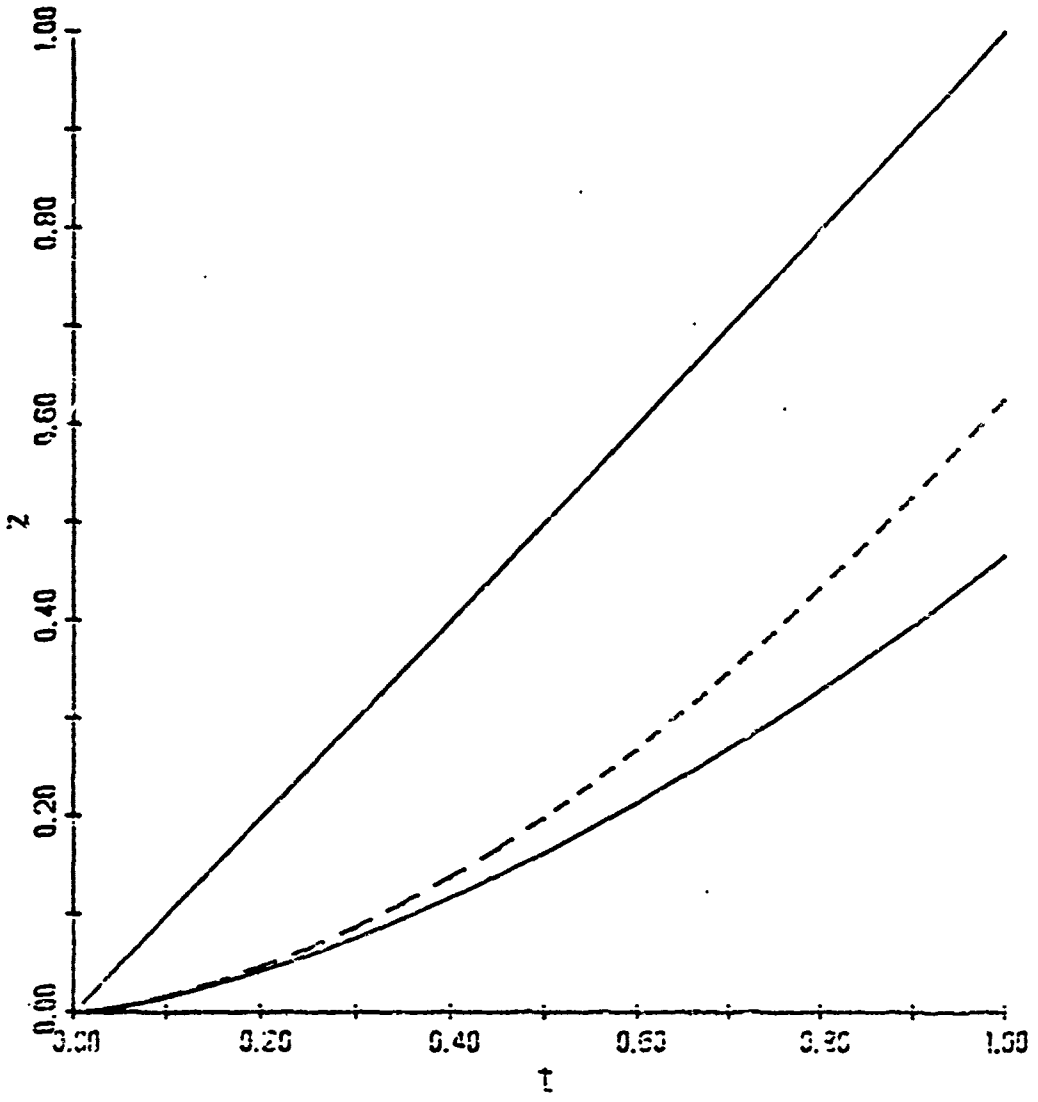



Figure 5. Same as Fig. 3 except r = 0.019m, L = 0.609m, and h/r = 0.005




gure 6. Same as Fig. 3 except r = 0.0127m, L = 0.609m, and h/r = 0.005

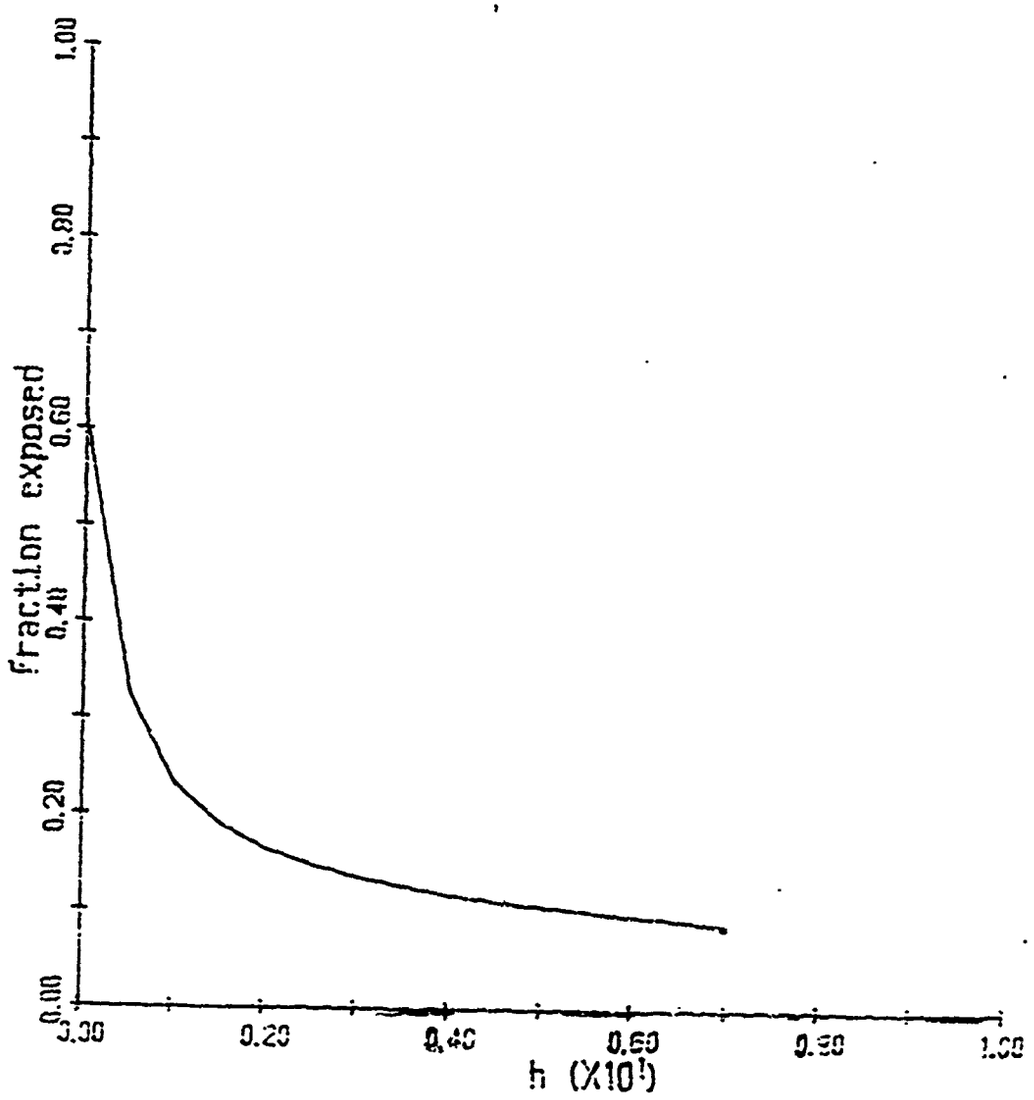



Figure 7. Maximum fraction of the rod exposed as a function of h/r, for r = 0.0127m and L = 0.457m

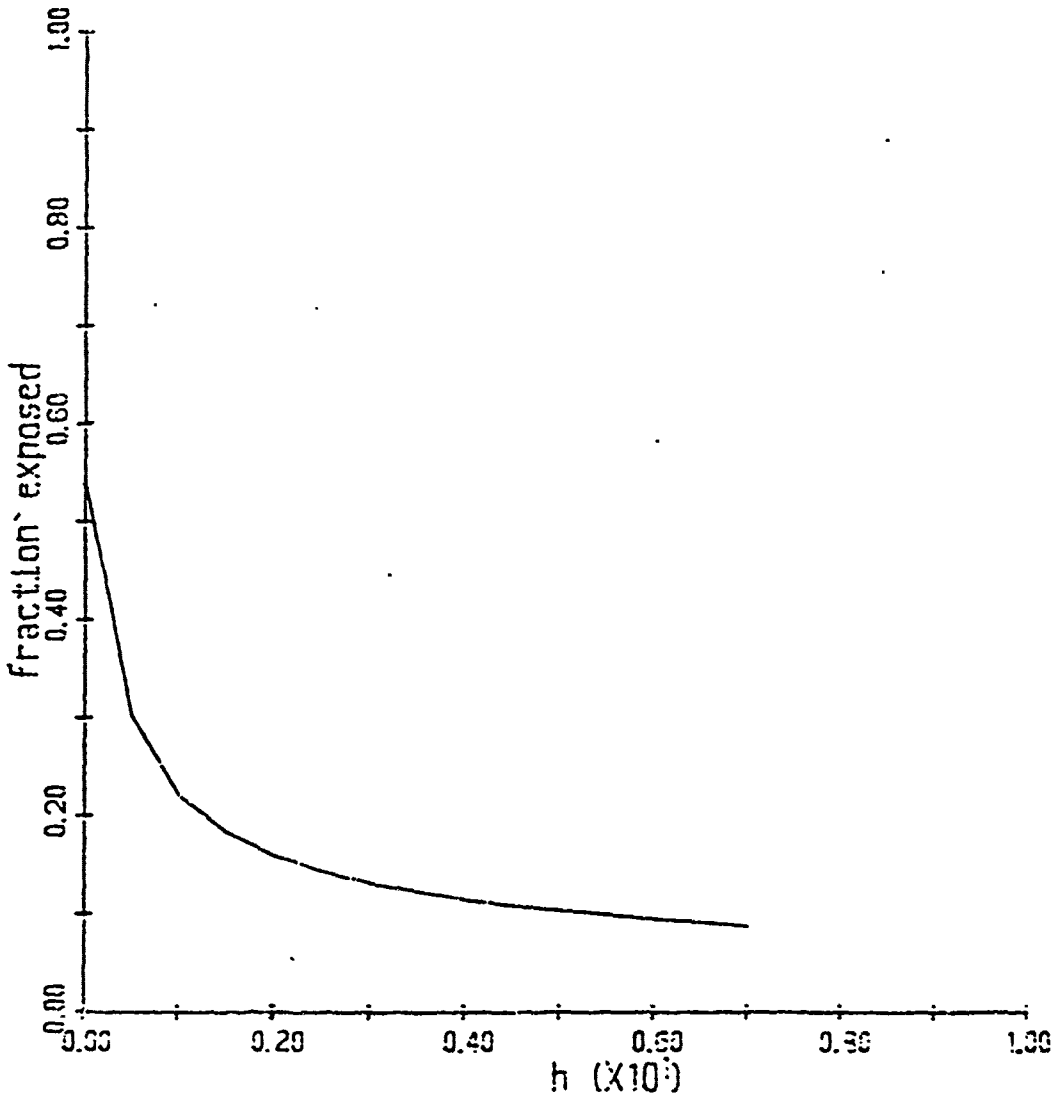



Figure 8. Same as Fig. 7 except r = 0.019m and L and L = 0.609m

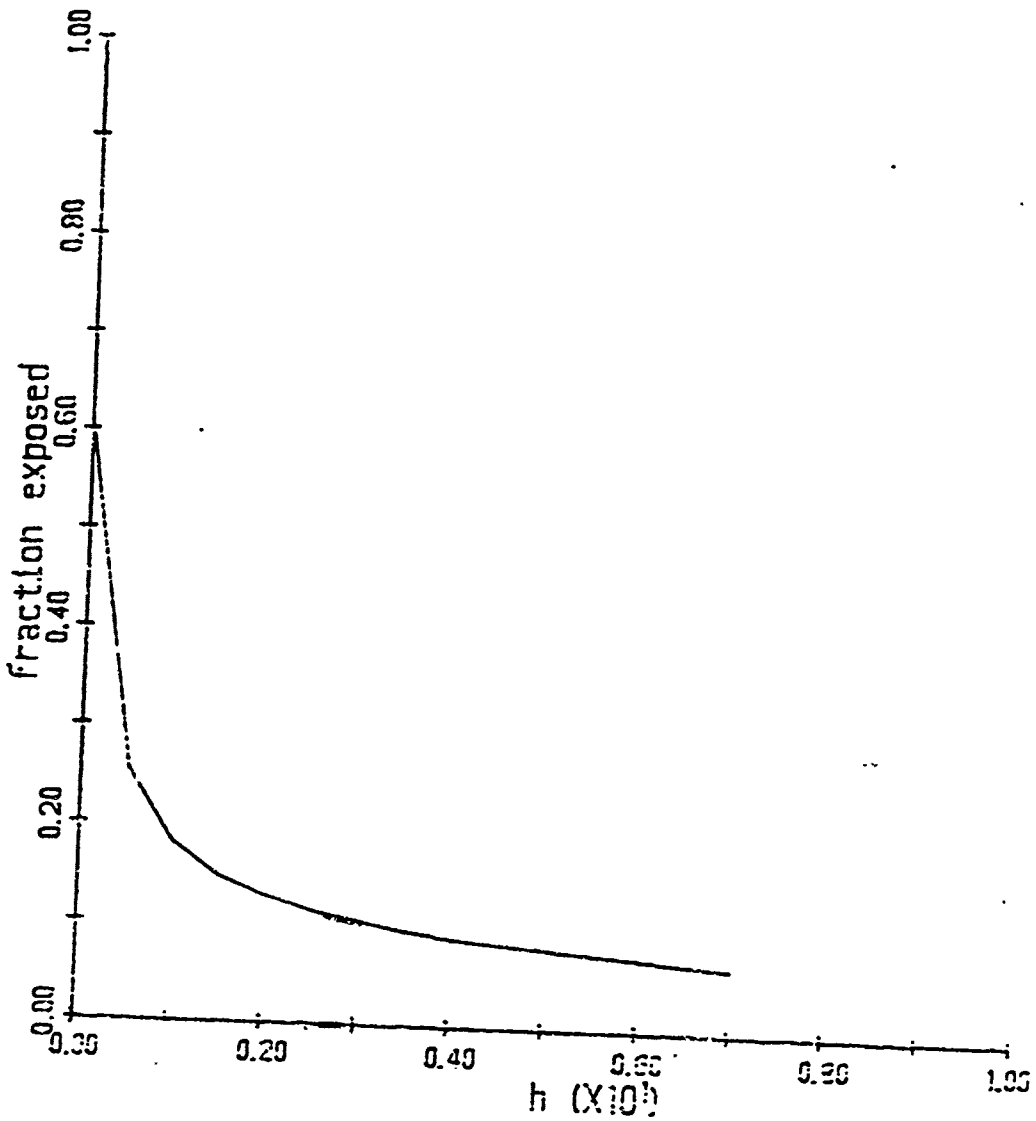



Figure 9. Same as Fig. 7 except r = 0.0127m and L = 0.609m

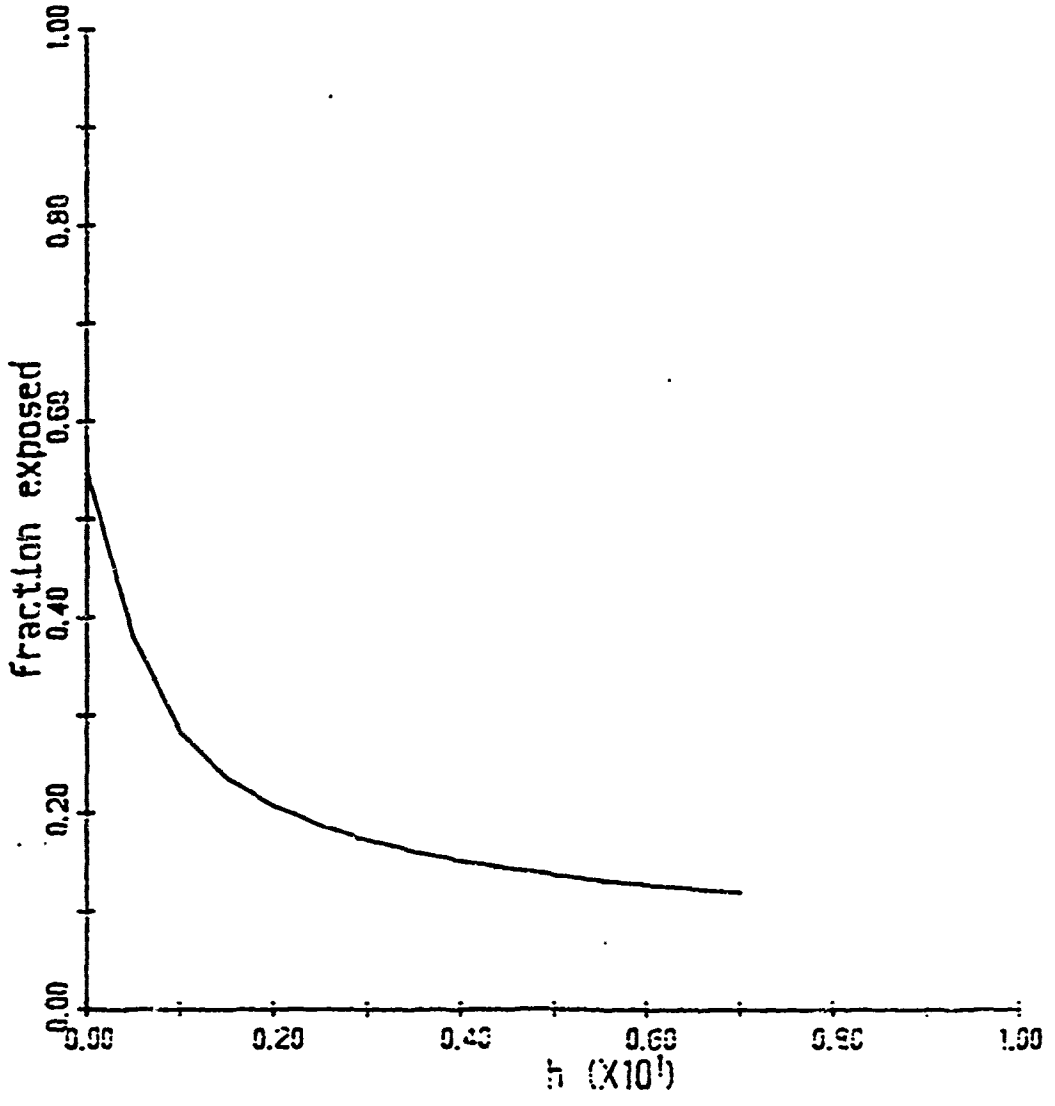



Figure 10. Same as Fig. 7 except r = 0.019m and L = 0.457m

# BEYOND STEFAN PROBLEMS: THE STRUCTURE OF SHEARED SOLIDIFICATION FRONTS*

F. S. HALL & G. S. S. LUDFORD
Theoretical and Applied Mechanics
Cornell University
Ithaca, NY 14853

ABSTRACT. Determination of the motion of a solidification front into a molten metal, a so-called Stefan problem, does not depend on detail of the solidification process inside the front. Such details are needed, however, when the liquid is being stirred by a magnetic field, since they determine the structure of the coincident Hartman (viscous) layer, and hence the secondary motion in the liquid. Within this layer the viscosity is variable, due to the spacial variation in the concentration of solid particles. In the absence of a theory in the nucleation literature, a growth law for the concentration of solid particles is proposed for the small undercooling that must occur in such a layer. The (asymptotic) analysis is carried out for the slow solidification of a cylinder of molten metal that is being stirred by a rotating, uniform magnetic field, but the results are of general validity.

I. INTRODUCTION. The propagation of a solidification front into a pure molten metal at rest is a consequence of the solid sending out shoots into the liquid. As a shoot grows it sends out its own shoots, the process repeating itself to form a tree-like structure known as a dendrite. The growth of a dendrite is limited only by its neighbors and the availability of liquid between its shoots. A rotating magnetic field causes the liquid metal to flow along the solidification front so as to break off the dendritic shoots as soon as they form, i.e. while they are still small. These minute solid particles migrate through the shear layer towards the liquid core, providing nuclei for the solidification process. The layer is therefore viewed as a gradation of neutrally buoyant particles, their concentration tending to zero on the liquid side and to one on the solid side.

In the absence of a theory determining the rate at which the dendritic fragments grow, we shall propose a growth law. The object of this paper is then to show how the structure of the viscous layer is modified when it is also a solidification layer, and to determine the disturbance to the main core flow that is caused by this viscous layer.

*Supported by the U.S. Army Research Office.

II. MAGNETOHYDRODYNAMIC STIRRING. The structure of the viscous layer is governed by the continuity and Navier-Stokes equations, the latter being modified by the Lorentz body force. In dimensionless forms these equations are

$$\nabla\cdot\mathbf{v}=0, \quad N^{-1}D\mathbf{v}/Dt = -\nabla p + \mathbf{j}\times\mathbf{B} + M^{-2}\nabla\cdot\{\eta(\nabla\mathbf{v}+\nabla\mathbf{v}^{T})\}, \quad (1)$$

where N and M are the interaction and Hartmann numbers, respectively. We consider the case of slow stirring (i.e. $N\to\infty$), so that the inertia terms are neglected. The inverse of the Hartmann number, $M^{-1}$, is a measure of the "thickness" of the viscous layer. Allowance has been made in equations (1) for a variable viscosity (dimensionless).

In order to determine how the velocity field varies as we traverse this viscous-solidification layer, we need to know how the viscosity varies through the layer. Certainly, the value of the viscosity at any point within the layer is determined by the concentration of solid (or liquid) at that point. With this in mind, we now consider the solidification process.

III. SOLIDIFICATION PROCESS. The model that we shall adopt is as follows. The dendritic shoots that form at the solid are supposed to be broken off by the scouring action of the shear flow while they are still small. The solidification layer therefore consists of particles whose concentration tends to one on the solid side, and to zero on the liquid side of the layer.

In the absence of such particles, i.e. for so-called homogeneous nucleation, the solidification rate depends on the amount of the under cooling from Te, the temperature at which solid and liquid are in equilibrium. For each temperature less than Te, a potential nucleus must have a critical size in order to grow, and it is a matter of determining how many of the particles, i.e. clusters of atoms that are continually forming and dissolving, are large enough. The relation between the temperature and the rate of increase in concentration of solid turns out to be

$$\frac{d}{dt}(1-c) = \kappa c \exp\left[\frac{-\alpha}{(1-T)^{2}T}\right], \quad (2)$$

where c is the liquid concentration, the unit of temperature is Te, and $\kappa$ is a (very large) constant. The value of $\alpha$, also a constant, is 1.14 for aluminum. Near the equilibrium temperature, i.e., for small values of 1-T, the right side is negligibly small, corresponding to the requirement of a large critical size; as 1-T increases the exponential slowly increases until at some value, $11\times 10^{-1}$ for aluminum (corresponding to $104^{\circ}$K undercooling from Te = $933^{\circ}$K), the coefficient of c suddenly becomes appreciable and solidification occurs in a fraction of a second. One concludes that there is a definite temperature at which nuclei of critical size are present in sufficient numbers to cause

almost instantaneous solidification. This temperature will be referred to as the solidification temperature, Ts.

When extraneous particles are present to act as nuclei, less undercooling is needed for solidification. The number of particles is increased, so that there are now more of critical size at any termperature. There is no theory of the nucleation rate in such circumstances, so we propose to use the same formula (2) with a smaller value of $\alpha$. (More precisely, we may argue that the effective change in free energies from solid to liquid are smaller because the energies of the extraneous particles are not to be counted; since the two energies, volume and surface, are reduced proportionately, the formula follows from the same analysis as for homogeneous nucleation.)

In applying the law (2) to a cylindrically converging solidification front, we will go to a frame moving with the front. Then, if the speed of the front is small enough for the process to be considered quasi-steardy relative to the front, we may write

$$V_0 \frac{dc}{dr} = -\kappa c \exp\left[\frac{-\alpha}{(1-T)^2}\right], \tag{3}$$

where $V_0$ is the (very small) dimensionless speed, and the time derivative has been replaced by $V_0 d/dr$. To complete the problem we must add the heat equation

$$V_0 \frac{dT}{dr} - Pe^{-1} \frac{1}{r} \frac{d}{dr}\left(r \frac{dT}{dr}\right) = -KV_0 \frac{dc}{dr}, \tag{4}$$

where K and Pe are constants (Pe is the Peclet number, which is supposed to have the same constant value in the solid as it has in the liquid; this assumption, equality of the two thermal diffusivities, is easily removed).

IV. THE SOLIDIFICATION LAYER. If, as we shall suppose, the solidification front has a dimensionless thickness comparable to $M^{-1}$, its structure must be investigated in order to determine how the effective viscosity $\eta$ changes from 1 to $\infty$ in going from the liquid to the solid side of layer. Several theories (Mancini, 1984) for the effective viscosity have been proposed, each yielding a different dependence on concentration. Since we do not have enough information about the dendritic particles to select anyone of these theories, we shall take the simplest relation.

$$\eta = 1/c^2 \tag{5}$$

as as illustration. Once c, and hence $\eta$, is found, we can determine the structure of the viscous layer (i.e., the velocity field.)

Whenever there is a layer, there must be a small parameter $\varepsilon$, and the question is to identify $\varepsilon$. The amount of undercooling is certainly a candidate, but the amount that occurs in ordinary solidification processes is too small, leading to thicknesses much less than $M^{-1}$. However, when dendritic shoots are prevented from forming, and only small particles are in contact with the liquid, the cooling may be much larger (though still small). Such is the case here, and we shall proceed on the assumption that the undercooling is sufficient to make the solidification layer of thickness $M^{-1}$. (This assumption should be verified experimentally).

To determine the amount of undercooling

$$\varepsilon=1-T_s \tag{6}$$

necessary to make the two layers of the same thickness, we set

$$\alpha=\tilde{\alpha}\varepsilon^{\ell},T=T_s+\varepsilon^{m}T_1/2\tilde{\alpha}+\ldots, \tag{7}$$

where $\tilde{\alpha},\ell,m$ are positive constants and $T_1$ represents the temperature perturbation (about $T_s$) in the layer. The exponent in equation (3) is then

$$\frac{-\alpha}{(1-T)^2T}=-\alpha\frac{\tilde{\varepsilon}^{\ell-2}}{1-\varepsilon}-\varepsilon^{\ell+m-3}T_1+\ldots \tag{8}$$

(Expansion of the exponent is similar to that in the asymptotic approach to combustion theory). The first term is cancelled by the pre-exponential factors, i.e.

$$\tilde{\alpha}\varepsilon^{\ell-2}=\ell n(K/V_o) \tag{9}$$

to leading order, implying $\ell<2$; the second is expected to provide $O(1)$ variations in the layer, i.e.,

$$\ell+m-3=0. \tag{10}$$

requiring $m<1$. It follows that the layer variable for solidification must be

$$\tilde{R}=2\tilde{\alpha}\tilde{\beta}\varepsilon^{n-m}(1-r) \tag{11}$$

if a small temperature gradient $\beta\varepsilon^{n}(n>0)$ outside the layer is to be matched.

Now the solidification layer is equal thickness to the viscous layer. This amounts to identifying the solidification variable $\tilde{R}$ with the appropriate variable R for the viscous layer. But the viscous layer is

the region where the viscous term in equations (1) is as important as the other terms (in the limit as $M\to\infty$); so that

$$R=M(1-r). \tag{12}$$

Setting

$$\tilde{R}=R \tag{13}$$

now shows that

$$\varepsilon^{m-n}=2\tilde{\alpha}\tilde{\beta}M^{-1} \tag{14}$$

where $\varepsilon^{m-n}$ measures the thickness of solidification layer, and $M^{-1}$ measures the thickness of viscous layer.

Equations (9), (10), and (14) show that

$$\varepsilon^{1-n}=2\tilde{\beta}M^{-1}\ell n(K/V_0). \tag{15}$$

This relation is of some importance because it does not depend on $\alpha$, a quantity not amenable to experimental determination. More information about $\varepsilon$ is obtained by writing the basic equation (3) and (4) to leading order in the layer variables.

$$\frac{dc}{dR}=ce^{-T},\quad \frac{d^2T}{dR^2}=-\frac{dc}{dR}. \tag{16}$$

Here we have set

$$\frac{\tilde{\alpha}\varepsilon^{\ell-2}}{(1-\varepsilon)}-\ell n\ (K/V_0)=(m-n)\ \ell n\varepsilon-\ell n(2\tilde{\alpha}\tilde{\beta}),\tilde{\beta}\varepsilon^n=KV_0Pe \tag{17}$$

In equation (16), $T_1$ has been replaced by T, and c now stands for its leading term in an expansion of the type (7b). Equation (17a) is just a more accurate version of the requirement (9), but equation (17b) gives an additional relation for $\varepsilon$. Combining equations (15) and (17b) determines $\varepsilon$ in terms of known constants.

V. SOLUTION OF LAYER PROBLEM. There is no temperature gradient in the melt to order $\varepsilon^n$, so that the layer equations (16) have to be solved under the boundary conditions

$$c\to 1,\ \frac{dT}{dR}\to 0\quad \text{as } R\to+\infty \tag{18}$$

on the liquid side. The requirement

$$\frac{dT}{dR}\to 1 \text{ as } c\to 0 \tag{19}$$

is then automatically satisfied, as is seen from the integral

$$S \equiv dT/dR = 1-c \tag{20}$$

of equation (16b). This expresses nothing more than the jump condition in the Stefan problem.

If we substitute equation (20) into the remaining layer equation (16a), we obtain

$$\frac{d^2T}{dR^2} = \left(\frac{dT}{dR}-1\right) e^{-T} . \tag{21}$$

The solution, in parametric form, is

$$T= -\ell n\ [-\ell n(1-S)-S],\ R= \int_{\frac{1}{2}}^{S} \frac{dS}{(1-S)[S+\ell n(1-S)]}, \tag{22}$$

The second of the two equations is of greater importance since, according to definition (20), S is the solid concentration in the layer, and hence, directly related to the viscosity $\eta$ by equation (5), i.e.

$$\eta=(1-S)^{-2}. \tag{23}$$

This formula for $\eta$ , when substituted into the Navier-Stokes equations, allows us to determine the structure of the viscous layer.

VI. DISCUSSION. We have made a number of assumptions in the preceding analysis which should be explored here. The viscous layer and the solidification layer were required to have the same thickness. The assumption enabled the structure of the viscous layer to be determined without reference to the details of the solidification process.

If no assumption is made about these two thicknesses, details of the solidification process must be understood in order to determine how the effective viscosity varies from that of the pure liquid to infinity (the value in the pure solid). A complete treatment would require theories of

i) the size and shape of the fragements broken of by the shear flow. This is a very complicated problem, which involves relating the breaking strength of the dendritic shoots to the distributed forces applied to them by the liquid. To our knowledge, no investigations have been done in this area.

ii) the rate at which these dendritic fragments grow. Even the formation of nuclei (i.e. clusters of electrically neutral atoms) in the absence of extraneous particles must be reinvestigated. In the classical neucleation theory, electrically neutral liquid atoms come together to from solid particles, but this must be modified in the present case. Because of the applied rotating magnetic field, currents are generated; and, since a current is nothing more than a flow of electrons, this implies that the atoms in the solidification layer are

no longer neutral. The theory would then have to be extended so as to take account of the extraneous particles.

iii) the effective viscosity of the resulting suspension. There are almost as many laws relating viscosity to concentration as there are investigators. Each gives different results depending, for example, on the size and shape of the particles. We chose a simple law relating viscosity to concentration for discussion purposes only. However, we do not expect the qualitative nature of the results to depend on the particular viscosity-concentration relation used. Even if a theory is formulated to describe the size and shape of the particles, a more sophisticated and accurate relation between viscosity and concentration can be constructed.

We sidestepped the above three items because our main purpose was to determine the qualitative structure of the viscous layer, and hence the disturbance that it causes in the core flow. A fuller treatment of the subject is given by Hall, Ludford & Walker (1984).

## REFERENCES


Hall, F.S., Ludford, G.S.S. & Walker, J.S. (1984). Hartmann layers in slowly solidifying liquids. Progress in Astronautics and Aeronautics. In press.

Mancini, F. (1984). The Viscosity of Suspensions. M.S. Thesis, Cornell University.

# SHOCK-INDUCED THERMAL RUNAWAY

T. L. Jackson and A. K. Kapila
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

ABSTRACT. Ignition of an initially cold combustible gas is studied when chemical reaction is switched on due to the passage of a piston-supported shock of sufficient strength. The time and space history of the shocked gas up to the instant of thermal runaway is described, by using a combination of asymptotics and numerics.

1. INTRODUCTION. Consider a combustible material confined to the half space $x > 0$ and capable of undergoing an exothermic chemical reaction of the Arrhenius type. Suppose that the temperature of the material is so low that it is practically inert, and will not burn if left alone. Let combustion be initiated by applying an ignition stimulus at the boundary $x = 0$. It is of considerable interest to obtain a mathematical description of the events following the application of the stimulus. In particular, if a combustion wave propagating through the material is eventually established, it is important to understand the evolutionary process that gives rise to it.

Recently, Kapila [1] has treated the case where burning is initiated by applying a heat flux at the boundary. Building on the pioneering work of Liñan and Williams [2] and Kassoy [3], the analysis employs a combination of large activation energy asymptotics and numerics, and is able to describe rather completely the entire course of events culminating in the development of a well-defined deflagration wave. A major drawback of the study, however, is that it completely ignores the motion and deformation of the material, and is therefore of limited value for gaseous combustibles. Application of heat causes thermal expansion of the gas, thus producing a strong coupling between gasdynamic and chemical aspects of the problem. In one particular context, such a coupling has been studied, independently, by Blythe [4] and Clarke [5]. They consider a spatially homogeneous, weakly reactive atmosphere into which small-amplitude but rapidly-varying gasdynamic disturbances are introduced. Their work examines the effect of chemical heat release on shock formation and the influence of pressure disturbances on thermal runaway.

The present analysis is very much in the spirit of [4-5] but is concerned with a problem more like the gaseous counterpart of that treated in [1]. The ignition stimulus is modelled by an impulsively started piston moving into the gas at a constant speed. This creates a shock wave which runs ahead of the piston, and will travel at a constant speed if the gas were inert. It is assumed here, however, that the shock is just strong enough to raise the temperature of the gas behind it to the ignition temperature, so



PREVIOUS PAGE IS BLANK

that the chemical reaction is switched on due to the passage of the shock. The purpose of the analysis is to describe the state of the gas between the piston and the shock, and to determine the extent to which the shock is accelerated by chemical heat release, up to the instant of thermal runaway. Post-runaway events are currently under study.

2. GOVERNING EQUATIONS AND ASSUMPTIONS. The equations of reactive gasdynamics for plane, one-dimensional, unsteady motion are [6]

$$\rho_t + u\rho_x + \rho u_x = 0 \ , \tag{1}$$

$$\rho(u_t + uu_x) + \frac{1}{\gamma} p_x = 0 \ , \tag{2}$$

$$y_t + uy_x = -w \ , \tag{3}$$

$$p = \rho T \ , \tag{4}$$

$$\rho(T_t + uT_x) - \frac{\gamma - 1}{\gamma} (p_t + up_x) = \beta\rho w \ , \tag{5}$$

where

$$w = \frac{\varepsilon y}{\beta} \exp[\frac{\theta}{T_o} - \frac{\theta}{T}] \ , \tag{6}$$

and

$$\varepsilon = T_o^2/\theta \ . \tag{7}$$

Here, $y$ is the mass fraction of the reactant in the combustible gas, $T$ the temperature, $\rho$ the density, $p$ the pressure and $u$ the velocity. The variables $y$, $T$, $\rho$ and $p$ have been made dimensionless by referring them to the corresponding quantities in the cold, stationary gas ahead of the shock. The scale for velocity is taken to be the frozen sound speed in the cold gas, while time is referred to the homogeneous induction time of the shocked gas at the instant the shock is generated. The initial (dimensionless) shock speed is denoted by $M_o$, such that $M_o - 1 = O(1)$, and $\gamma$ is the specific-heats ratio. Transport effects are ignored, and a one-step first-order Arrhenius reaction postulated, with $\theta$ the dimensionless activation energy and $\beta$ the heat-release parameter. The analysis will proceed on the assumption that

$$\theta >> 1 \quad , \quad \text{i.e.,} \quad \varepsilon << 0 \ . \tag{8}$$

3. PERTURBATION ANALYSIS. In the absence of chemistry ($w \equiv 0$), the state of the gas is given by

$$p = \rho = T = y = 1 \quad , \quad u = 0 \quad \text{for} \quad x > M_o t \ , \tag{9a}$$

$$p = p_o, \ \rho = \rho_o, \ T = T_o, \ y = 1, \ u = u_o \quad \text{for} \quad u_o t < x < M_o t \ , \tag{9b}$$

where $p_o$, $\rho_o$, $T_o$, and $u_o$ are related to $M_o$ through the Rankine-Hugoniot conditions

$$p_o = \frac{2\gamma M_o^2 + 1 - \gamma}{\gamma + 1} \quad , \quad \rho_o = \frac{(\gamma + 1) M_o^2}{(\gamma - 1) M_o^2 + 2} \ ,$$

$$T_o = \frac{(2\gamma M_o^2 + 1 - \gamma)\{(\gamma - 1) M_o^2 + 2\}}{(\gamma + 1)^2 M_o^2} \quad , \quad u_o = \frac{2(M_o^2 - 1)}{(\gamma + 1) M_o} \ . \tag{10}$$

Observe, in particular, that $u_o$ is also the speed of the piston.

The presence of $T_o$ in the exponent of (6) indicates that in the limit $\theta \to \infty$ being considered here, $T_o$ is the switch-on temperature for the chemical reaction. In the initial stages of evolution, therefore, the state of the shocked gas will deviate from (9b) by an $O(\varepsilon)$ amount, as will the speed of the shock. If the perturbed speed of the shock is taken to be

$$M = M_o + \varepsilon M_1 \ , \tag{11}$$

the conditions (10) show that the state of the gas immediately behind the shock will become, to $O(\varepsilon)$,

$$p_s \sim p_o \left\{ 1 + \frac{4\varepsilon\gamma M_o}{2\gamma M_o^2 + 1 - \gamma} m_1 \right\} \ , \tag{12a}$$

$$\rho_s \sim \rho_o \left\{ 1 + \frac{4\varepsilon}{M_o\{(\gamma - 1)M_o^2 + 2\}} m_1 \right\} \ , \tag{12b}$$

$$T_s \sim T_o \left\{ 1 + \frac{4\varepsilon}{M_o} \left( \frac{\gamma M_o^2}{2\gamma M_o^2 + 1 - \gamma} - \frac{1}{(\gamma - 1)M_o^2 + 2} \right) m_1 \right\} \ , \tag{12c}$$

$$u_s \sim u_o \left\{ 1 + \frac{\varepsilon(M_o^2 + 1)}{M_o(M_o^2 - 1)} m_1 \right\} \ . \tag{12d}$$

Let the state of the gas between the piston and the shock be expanded as

$$p \sim p_o + \varepsilon p_1 + \ldots, \quad \rho \sim \rho_o + \varepsilon \rho_1 + \ldots,$$

$$T \sim T_o + \varepsilon T_1 + \ldots,$$

$$u \sim u_o + \varepsilon u_1 + \ldots,$$

$$y \sim 1 + \varepsilon y_1 + \ldots \quad . \tag{13}$$

All perturbation quantities will depend upon $x$ and $t$, except the shock-speed perturbation $m_1$, which will depend upon $t$ alone. Substitution of (13) into (1)-(6) yields the following equations for the disturbances:

$$\rho_{1_t} + u_o \rho_{1_x} + \rho_o u_{1_x} = 0 , \tag{14}$$

$$\rho_o(u_{1_t} + u_o u_{1_x}) + \frac{1}{\gamma} p_{1_x} = 0 , \tag{15}$$

$$p_1 = \rho_o T_1 + T_o \rho_1 , \tag{16}$$

$$\rho_o(T_{1_t} + u_o T_{1_x}) - \frac{\gamma - 1}{\gamma} (p_{1_t} + u_o p_{1_x}) = \rho_o e^{T_1} , \tag{17}$$

$$y_{1_t} + u_o y_{1_x} = - \frac{1}{\beta} e^{T_1} \quad . \tag{18}$$

The reactant equation (18) is uncoupled from the set (14)-(18) which, except for the nonlinear source term, governs linearized acoustics in a steadily moving medium. It is convenient to introduce a new spatial variable $\xi$ via the transformation

$$x = u_o t + a_o \xi ,$$

where

$$a_o = \sqrt{T_o} \tag{19}$$

is the initial acoustic speed (dimensionless) in the shocked gas. In the new coordinates, eqns. (14)-(18) transform into

$$\rho_{1_t} + \frac{\rho_o}{a_o} u_{1_\xi} = 0, \quad \rho_o u_{1_t} + \frac{1}{\gamma a_o} p_{1_\xi} = 0, \quad p_1 - \rho_o T_1 - T_o \rho_1 = 0 ,$$

$$\rho_o T_{1_t} - \frac{\gamma - 1}{\gamma} p_{1_t} = \rho_o e^{T_1} , \tag{20}$$

and

$$y_{1_t} = -\frac{1}{\beta} e^{T_1} \quad . \tag{21}$$

Also, the piston is brought to rest at the location $\xi = 0$, while the shock trajectory is given by $\xi = \int_0^t V dt$, where

$$V \equiv \frac{M - u_o}{a_o} = V_o + \varepsilon V_1 \; ; \quad V_o = \frac{M_o - u_o}{a_o} \; , \quad V_1 = \frac{m_1}{a_o} \quad . \tag{22}$$

On using the relevant expressions from (10), $V_o$ can be written as

$$V_o = \left\{ \frac{(\gamma - 1)M_o^2 + 2}{2\,M^2 - \gamma + 1} \right\}^{\frac{1}{2}} \quad . \tag{23}$$

Equations (20) are to be solved subject to the requirements that at the piston, $u_1$ vanishes and at the shock, the conditions (12a-d) must be satisfied. Since analysis is only being carried to $O(\varepsilon)$, the shock conditions can be applied at the undisturbed shock path $\xi = V_o t$. Henceforth we shall ignore eqn. (21) because once $T_1$ is known, $y_1$ can be determined by a simple integration of (21) under the condition that $y_1$ vanishes initially. Also, in the set (20), $p_1$, $T_1$ and $u_1$ will be treated as the fundamental variables, because $\rho_1$ is defined by the third equation of (20).

It is a simple matter to distill from (20) the following single partial differential equation for $T_1$ :

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial \xi^2}\right) T_{1_t} = \left(\gamma \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial \xi^2}\right) e^{T_1} \quad . \tag{24}$$

This equation clearly reflects the coupling between acoustic motion and thermal explosion, and signals the presence of the isothermal sound speed in addition to the usual frozen sound speed. It is not, however, particularly conducive to the construction of a solution. For the latter purpose it is best to return to the set (20) and rewrite it in the characteristic form

$$\frac{\partial}{\partial r}\left(\tilde{p}_1 + \gamma\, a_o\, \rho_o\, \tilde{u}_1\right) = \frac{\gamma \rho_o}{2} e^{\tilde{T}_1} \quad , \tag{25a}$$

$$\frac{\partial}{\partial s}\left(\tilde{p}_1 - \gamma\, a_o\, \rho_o\, \tilde{u}_1\right) = \frac{\gamma \rho_o}{2} e^{\tilde{T}_1} \quad , \tag{25b}$$

$$\frac{\partial}{\partial t}\left(\rho_o\, T_1 - \frac{\gamma - 1}{\gamma} p_1\right) = \rho_o\, e^{T_1} \quad , \tag{25c}$$

where the characteristic coordinates $r$ , $s$ are defined by

$$t + \xi = r \quad , \quad t - \xi = s \tag{26}$$

and the notation

$$\phi(\xi, t) = \tilde{\phi}(r, s) \tag{27}$$

is employed.

Fig. 1 shows the geometry of the situation, including a typical point P between the piston and the shock and the three characteristic lines passing through it. It is a simple matter to carry out numerical integration of equations (25) along the characteristics.

## 4. NUMERICAL RESULTS

The numerical calculations were performed at $\gamma = 1.4$ and $M_o = 2.646$ . Then, the remaining parameters characterizing the shock at $t = 0$ are found to be

$$p_o = 8.000, \quad \rho_o = 3.500, \quad T_o = 2.286, \quad u_o = 1.890,$$

$$a_o = 1.512, \quad V_o = 0.500, \quad W_o = 0.571 \quad .$$

A small-time solution, developed analytically, was employed to begin integration at $t = 0.001$ . Integration was terminated at $t = 0.875$ when the development of large temporal gradients in the solution signalled the imminence of thermal runaway. Figures 2-4 exhibit the graphs of $T_1$, $p_1$ and $u_1$ against $\xi$ at various values of $t$ while Fig. 5 shows the variation with time of the shock-speed perturbation $m_1$ .

As one would expect, the greatest temperature rise is at the piston face, where the residence time for the shocked gas is the longest; at any given time, temperature decreases monotonically away from the piston. It is instructive to compare this induction process with the constant-volume, spatially homogeneous chemical heating which would occur if the piston were held stationary and the entire bulk of the gas were brought instantaneously to the ignition temperature $T_o$ at $t = 0$ . The corresponding solution, obtained from equations (20) in the absence of spatial gradients and under null initial conditions, is

$$T_1 = -\ln(1 - \gamma t) \ ,$$

for which time-to-thermal runaway is $1/\gamma = 0.714$ , as against the value $0.875$ obtained for the shock-induced case.

At thermal runaway and beyond, unboundedness of the solution in the vicinity of the piston renders the expansions (13) invalid. Further development of the explosive process is under study.

An expanded version of this paper will appear in the SIAM Journal on Applied Mathematics.

## ACKNOWLEDGMENTS

Professor David Kassoy is acknowledged for pointing out to us that a similar problem has also been considered by Clarke and Cant [7], who presented it at the Ninth International Colloquium on Dynamics of Explosions and Reactive Systems, Poitiers, July 3-8, 1983.

## FIGURE CAPTIONS

Fig. 1. The flow geometry.

Fig. 2. Evolution of temperature perturbation $T_1$ .

Fig. 3. Evolution of pressure perturbation $p_1$ .

Fig. 4. Evolution of velocity perturbation $u_1$ .

Fig. 5. History of shock-speed perturbation $m_1$ .

## REFERENCES

1. Kapila, A. K. (1981). "Evolution of deflagration in a cold combustible subject to a uniform energy flux," International Journal of Engineering Science, 19, 495-509.

2. Liñan, A., and Williams, F. A. (1971). "Theory of ignition of a reactive solid by a constant energy flux," Combustion Science and Technology, 3, 91-98.

3. Kassoy, D. R. (1976). "Extremely rapid transient phenomena in combustion, ignition and explosion," SIAM-AMS Proceedings, 10, 61-72. (Asymptotic Methods and Singular Perturbations, ed. R. E. O'Malley, Jr..)

4. Blythe, P. A. (1978). "Wave propagation and ignition in a combustible mixture," Seventeenth International Symposium on Combustion, 909-916, Pittsburgh: The Combustion Institute.

5. Clarke, J. F. (1978). "Small-amplitude gasdynamic disturbances in an exploding atmosphere," Journal of Fluid Mechanics, 89, 343-379.

6. Williams, F. A. (1965). Combustion Theory, Addison Wesley.

7. Clarke, J. F., and Cant, R. S. (1984) "Nonsteady gasdynamic effects in the induction domain behind a strong shock wave." Progress in Aeronautics and Astronautics, to appear.

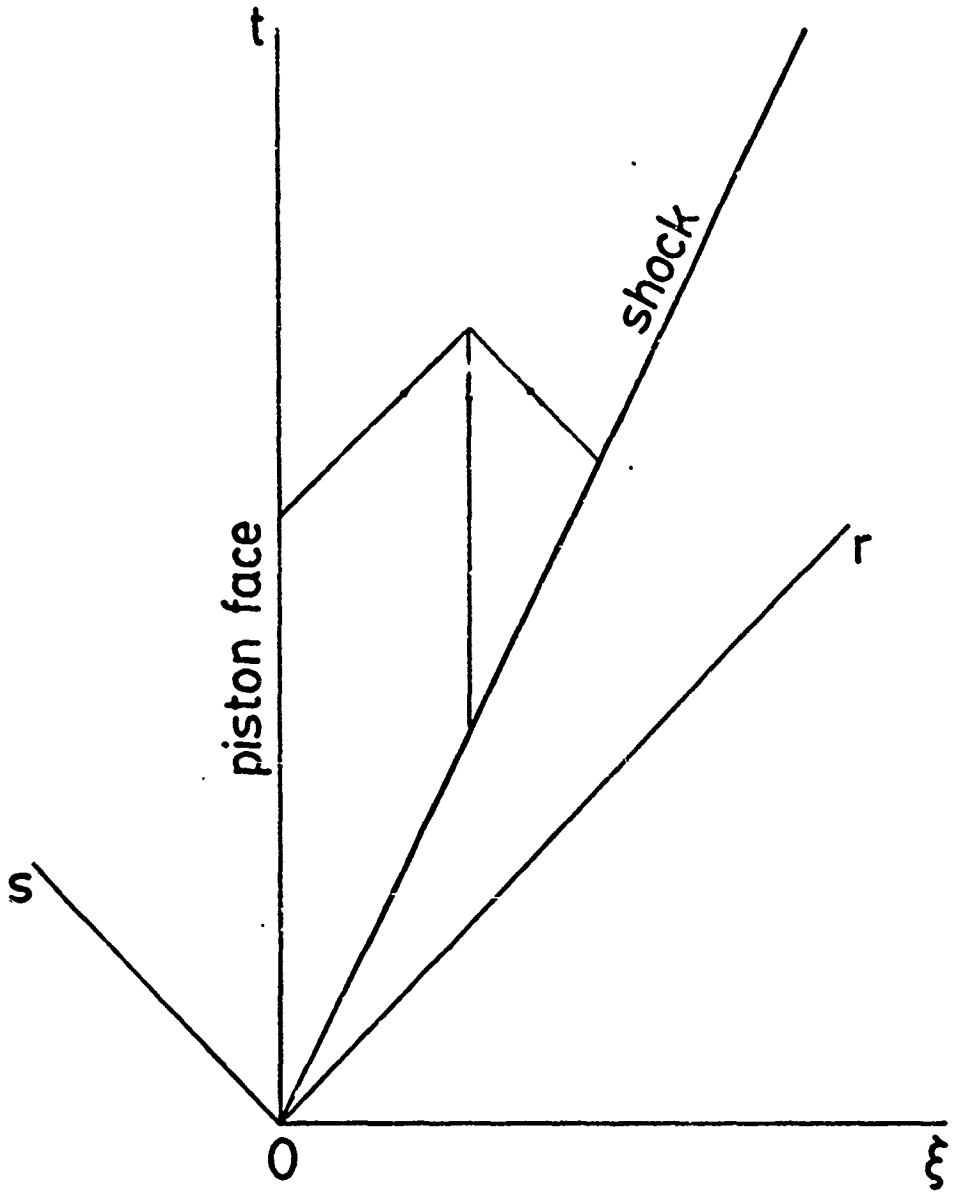



Figure 1

**The flow geometry.**

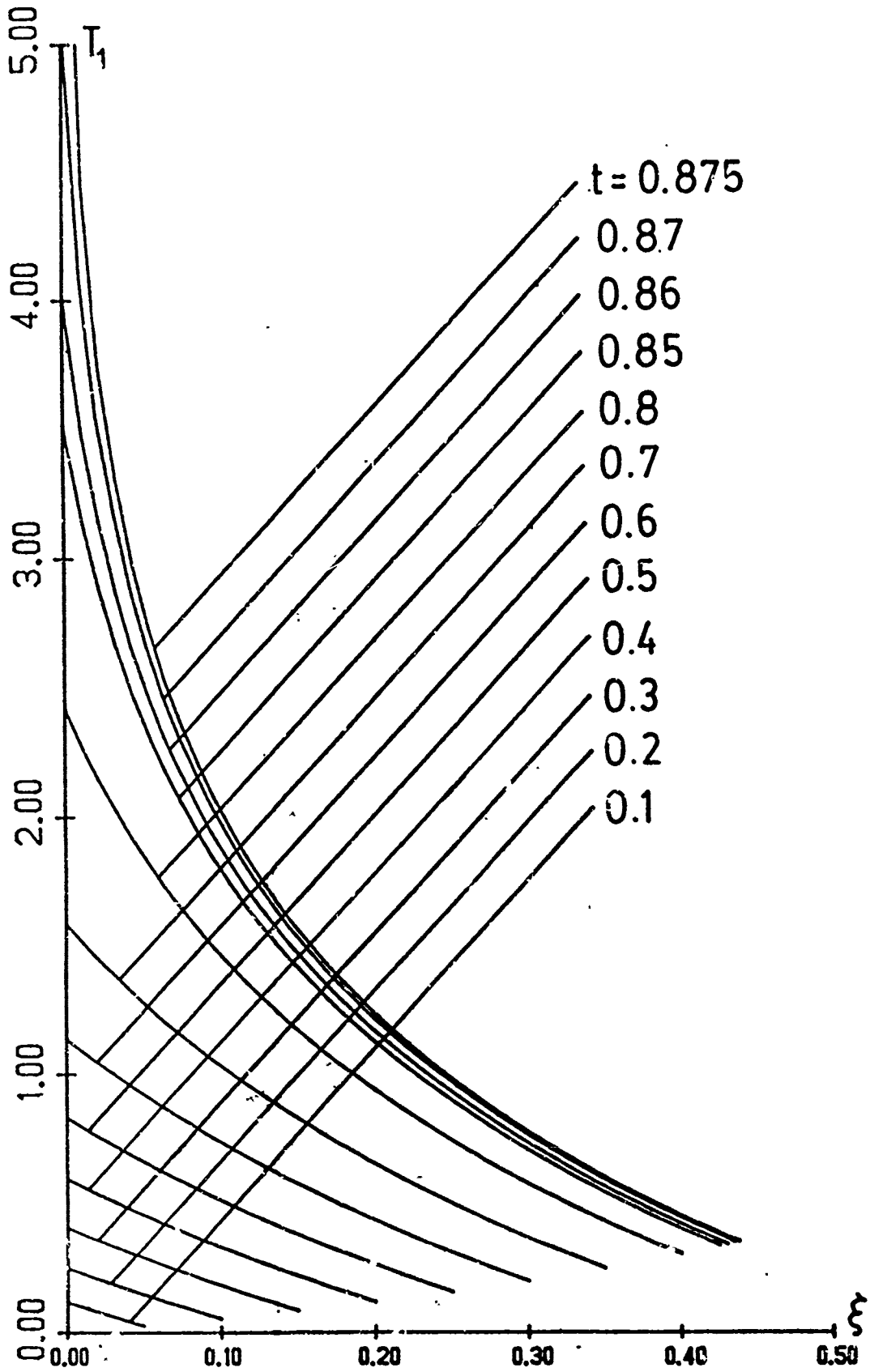
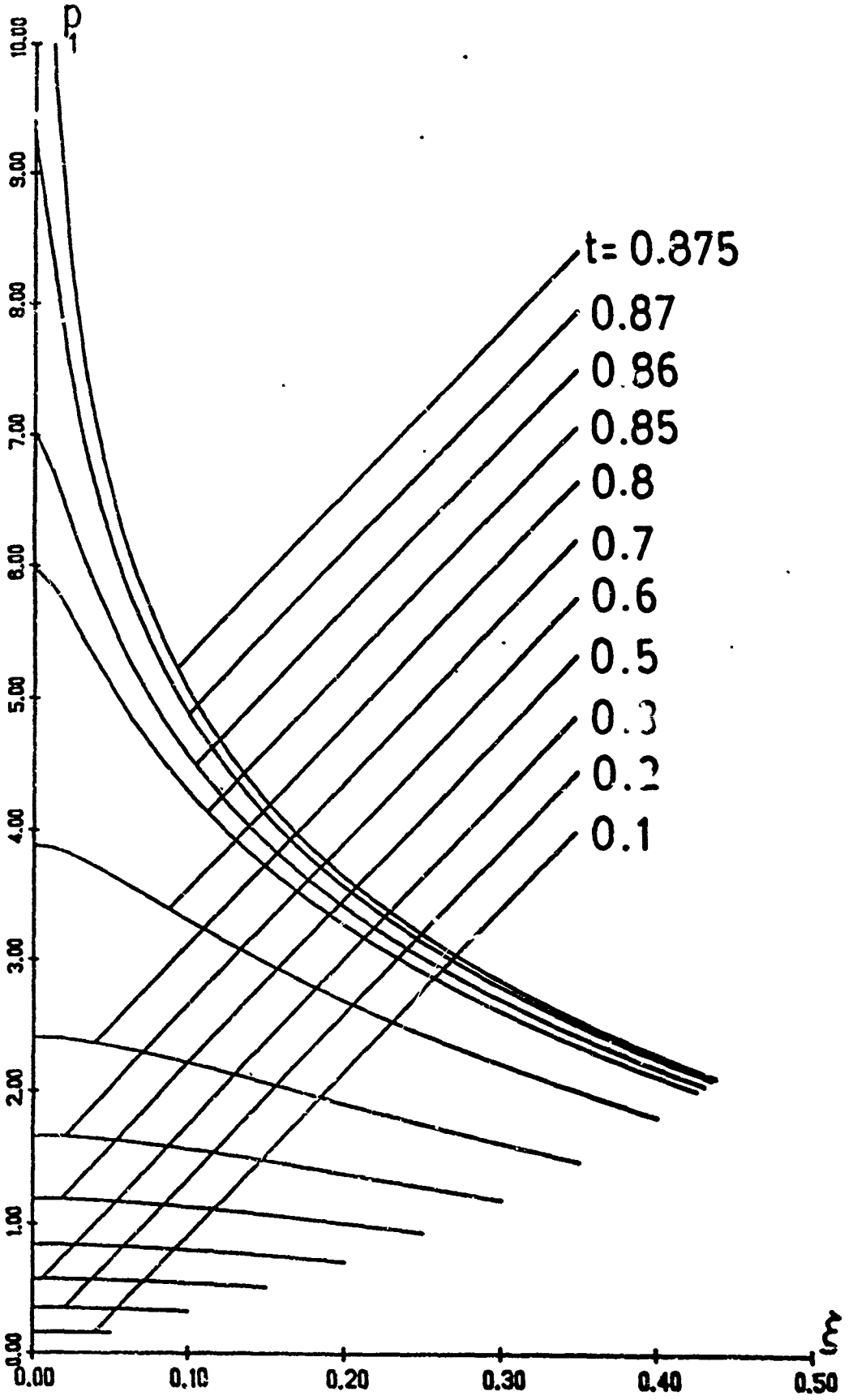



Figure 2

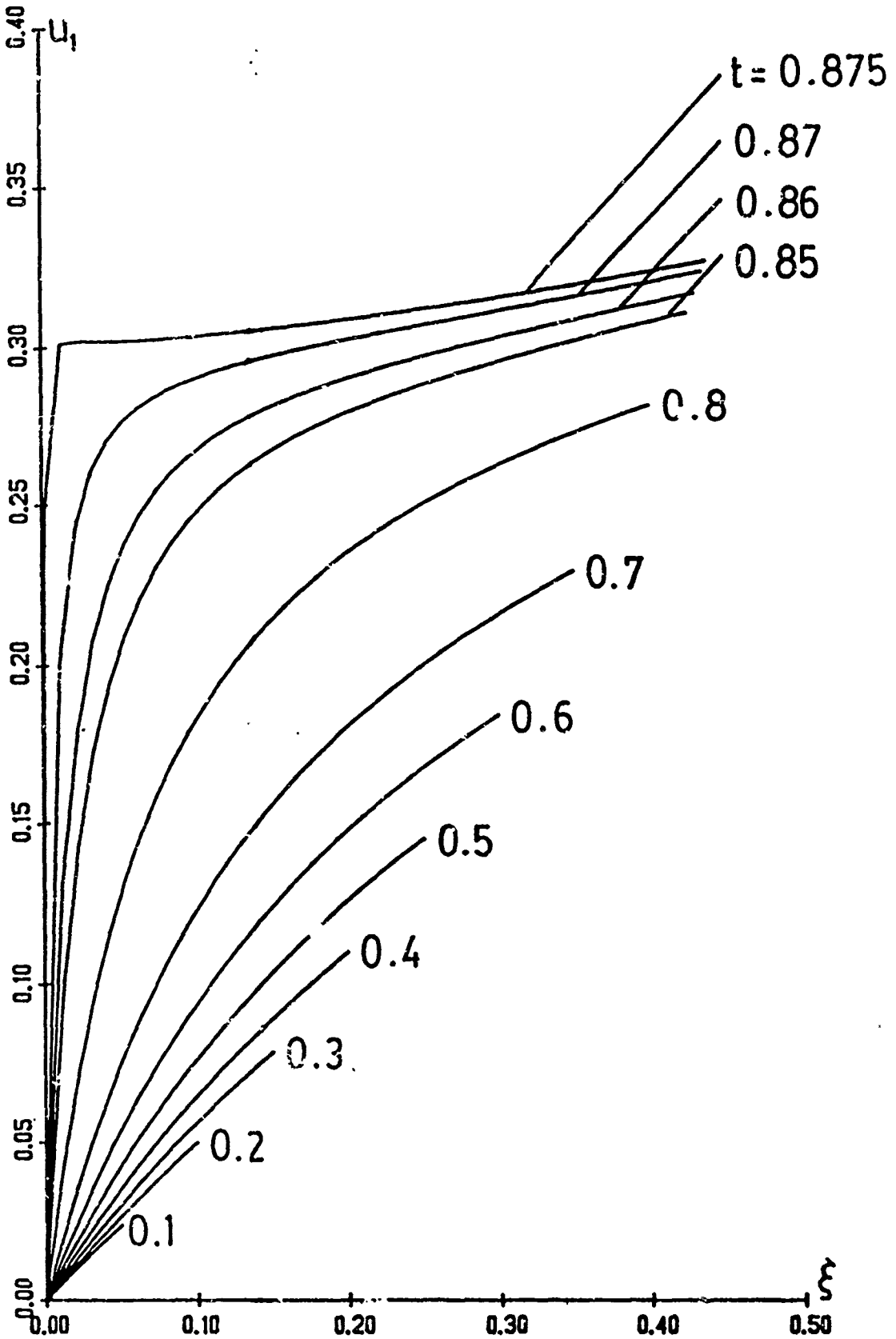
Evolution of temperature perturbation $T_1$ .




Evolution of pressure perturbation $p_1$ .



Figure 3




Evolution of velocity perturbation $u_1$.

Figure 4

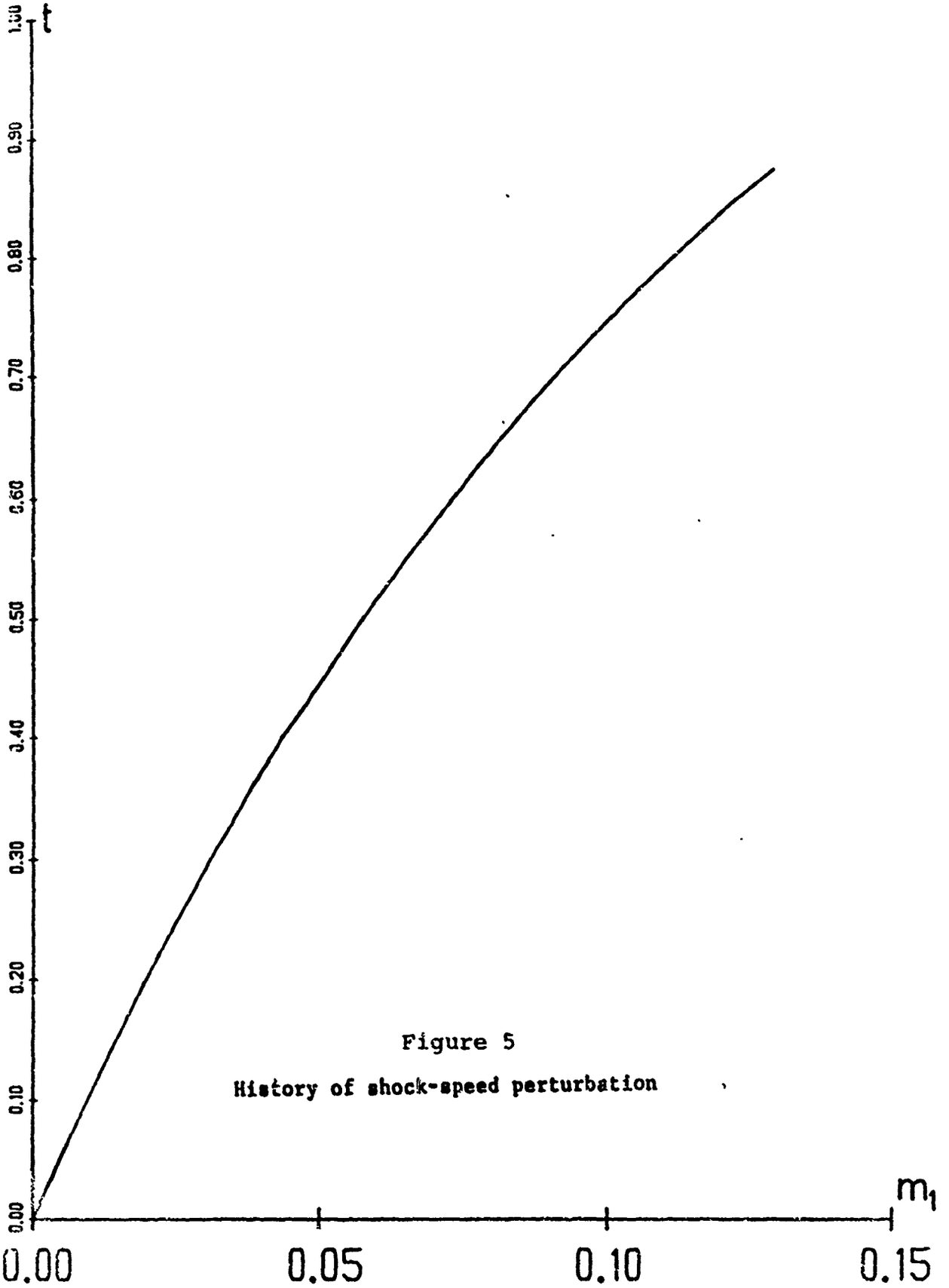



Figure 5

History of shock-speed perturbation

# THE STEFAN PROBLEM OF DETONATION THEORY*

A.A. Oyediran and G.S.S. Ludford
Department of Theoretical and Applied Mechanics
Cornell University, Ithaca, NY 14853

ABSTRACT. A certain model of one-dimensional detonation waves leads to a Stefan problem: the unknown f satisfies Burgers equations on the two sides of a moving discontinuity at which it is given ($f_*$, say) and the jump in its derivative (corresponding to the exothermic reaction) is prescribed. An alternative formulation of the problem can be obtained by means of the Hopf-Cole transformation, which replaced the Burgers equations by diffusion - type equations.

The problem possesses a steady solution, the discontinuity moving with constant speed and f depending only on distance from it. This solution is stable for a range of the parameter f, and unstable otherwise, as was shown at the First Army Conference, when preliminary results on the subsequent evolution of the instability were presented. The instability has now been reexamined, using three computation schemes on each of the two formulations of the problem, resulting in the more definite conclusions presented here.

I. INTRODUCTION. Work on detonation waves by Stewart and Ludford [1] (presented at the First Army Conference) showed that galloping detonations evolve with time. We have now re-examined this instability with a more refined numerical technique. In addition, using asymptotic analysis for small time step we show how the velocity of the wave is determined as it evolves. This velocity depends, in general, only on the gradients and curvatures on the two sides of the flame sheet; but, when the gradients are equal and opposite, the third derivatives are involved.

Our model of a one-dimensional detonation wave [1] leads to

$$f_T + \frac{\gamma+1}{2}[k(T)-f]f_\eta = \frac{\gamma}{2} f_{\eta\eta}, \tag{1}$$

subject to

$$\begin{aligned} &f(0,T) = f_* \text{ (given)}, \quad [f_\eta] \equiv f_\eta(0^+,T)-f_\eta(0^-,T) = -(\gamma+1)\alpha^2/2\gamma \\ &f(-\infty,T) = 0, \quad f(+\infty,T) = f_+ \text{ (given)}, \end{aligned} \tag{2}$$

with a given initial condition $f(\eta,0) = G(\eta)$.

The unknown f satisfies a Burgers equation on the two sides of the moving discontinuity (which has been reduced to rest in the formulation above) and represents the disturbance of a quiescent state ahead of the wave ($\eta \to -\infty$). The velocity of the wave is given by k(T). All the reaction takes place at the moving discontinuity, where $f = f_* > 0$ always. The

*Supported by the U.S. Army Research Office

derivative of f takes a jump from one side to the other of the discontinuity (flame). The constants $\alpha$ and $\gamma$ (ratio of specific heats) are assignable, as is $f_+(>0)$.

In the classical Stefan problem of ice-solidification, the jump in slope at the moving discontinuity is proportional to k(T) and the nonlinear term is missing from equation (1). The first of these ensures that the velocity is always determined by the gradients and curvatures at the discontinuity, while the second makes no difference.

II. RESULTS FROM STABILITY ANALYSIS. The problem as it stands is over determined. If k(T) is supposed given: there are three conditions at $\eta = 0$, namely

$$f(0^{\pm},T) = f_* \ , \quad [f_\eta] = -(\tfrac{\gamma+1}{2\gamma})\alpha^2. \tag{3}$$

The extra condition determines the speed k(T) of the flame. Prescribing $f_+$, $\alpha$ and restricting $f_*$ to a certain interval, namely $f_* \in (\alpha^2/f_+, (f_+^2+\alpha^2)/f_+)$, yields a steady solution with

$$k = k_S = \frac{f_+^2 + \alpha^2}{2f_+} \geq \alpha. \tag{4}$$

The steady solution may be written

$$f_s = \begin{cases} 2k_s f_* e^{-\xi_-}/[2k_s - f_* + f_+ e^{\xi_-}] \\ [f_-(f_+ - f_*) + f_+(f_* - f_-)e^{\xi_+}]/[(f_+ - f_*) + (f_* - f_-)e^{\xi_+}] \end{cases} \quad \text{for } \eta \lesseqgtr 0. \tag{5}$$

provided $k_S > \alpha$; here

$$\xi_- = (\gamma+1)k_s\eta/\gamma, \quad \xi_+ = (\gamma+1)(f_+ - f_-)\eta/2\gamma, \quad f_- = \alpha^2/f_+ \tag{6}$$

Examining the stability of the steady-state solution by setting

$$k = k_S + e^{\lambda T}, \quad f = f_S(\eta) + F(\eta)e^{\lambda T} \quad (\varepsilon << 1) \tag{7}$$

and substituting into the problem (1) and (2), we find that the eigenvalue $\lambda$ satisfy the dispersion relation

$$(f_b - f_*)f_*(2k_S - f_*) = (f_a - f_*)(f_* - f_-)(f_+ - f_*), \tag{8}$$

where

$$f_a = k_S + \sqrt{k_S^2+\Lambda},\ f_b = k_S - \sqrt{k_S^2-\alpha^2+\Lambda},\ \Lambda = 8\gamma\lambda/(\gamma+1)^2. \tag{9}$$

This result is due to Stewart [2]. Examination of this dispersion relation reveals that $\lambda$ is always real and that its sign is as shown below.

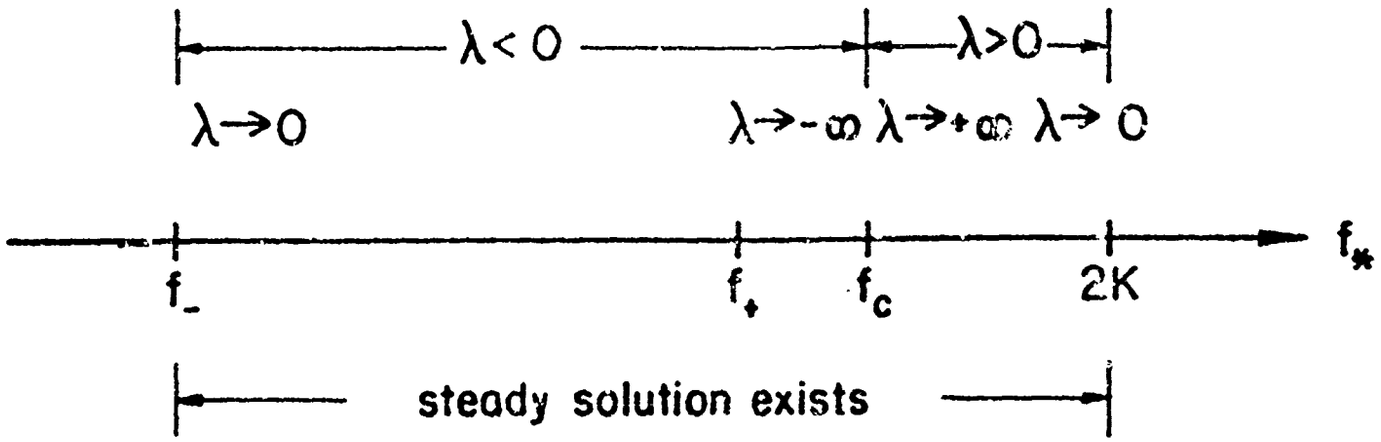



Figure 1. Showing the signs of the eigenvalue. $\lambda < 0$ corresponds to stable while $\lambda > 0$ is unstable.

III. NUMERICAL RESULTS. Stewart and Ludford [1] examined this problem at the First Army Conference; in particular, some numerical calculations were made with $f_* = 1.50$, $f_+ = 0.9128$ and $\alpha = \sqrt{2/(\gamma+1)} = .9128$ (for $\gamma = 7/5$) using a perturbed steady-state profile as initial conditions. One result is shown in Figure 2, and another (exhibiting peaks ) in Figure 5.

The problem was re-examined, using a more accurate integration scheme. The numerical procedure is schematically as follows:

$$f_T = S(f_{\eta\eta}, f_\eta, f, k), \tag{10}$$

$$f(0^{\pm},T) = f_*,\quad f(\infty,T) = f_+,\quad f(-\infty,T) = 0. \tag{11}$$

together with initial conditions. In Stewart and Ludford [1] explicit schemes used centered fifferences in evaluating S; here a machine package was used to do this more accurately. We search for K(T) until

$$R = [f_\eta] + 0.7142, \tag{12}$$

was less than $10^{-7}$. (Here 0.7142 is the value of $(\gamma+1)\alpha^2/2\gamma$ for the $\alpha,\gamma$ given above.) The integration goes forward in time until no value of k could be found to make R small enough. By using standard IMSL routines, the repsonse in Figure 3 was obtained. At the point of breakdown, a plot of R against k immediately before breakdown is shown in Figure 4.

IV. ANALYTICAL DETERMINATION OF k(T). Finally, we present some analytical results concerning the determination of $k$ at any particular time step, which we may take to be $T = 0$. Setting

$$f = \frac{\alpha}{\sqrt{2}} F, \; T = \frac{4\gamma}{(\gamma+1)^2\alpha^2} t, \; \eta = \frac{\sqrt{2}\gamma}{(\gamma+1)} x, \; k = \frac{\alpha}{\sqrt{2}} K \tag{13}$$

shows that the problem may be written

$$\begin{aligned} &F_t + (K-F)F_x = F_{xx}, \\ &F(0^{\pm},t) = F_* \; (\equiv\sqrt{2}f_*/\alpha), \; F_x(0^+,t)-F_x(0^-,t) = -1, \\ &F(-\infty,t) = 0, \quad F(+\infty,t) = F_+ (\equiv\sqrt{2}f_+/\alpha). \end{aligned} \tag{14}$$

We may suppose that the initial data $F_0(x)$ satisfies conditions at $x = 0$, i.e.

$$F_0(0^{\pm}) = F_*, \quad F_0'(0^+)-F_0'(0^-) = -1. \tag{15}$$

Away from the origin, we may use the outer expansion

$$F = F_0(x) + tF_1(x) + \ldots \text{ for } t \ll 1. \tag{16}$$

Close to the origin, the inner expansion takes the form

$$F = F_* + t^{\frac{1}{2}}f_1(\xi) + tf_2(\xi) + t^{3/2}f_3(\xi) + \ldots \text{ with } \xi = x/t^{\frac{1}{2}}. \tag{17}$$

By matching and satisfying all the conditions at the origin, we find

$$f_1(\xi) = G_{\pm}\xi \quad \text{with} \quad G_{\pm} = F_0'(0^{\pm}). \tag{18}$$

Similarly we have $f_2$ determined as

$$f_2(\xi) = -(K_0-f_*) + \tfrac{1}{2}C_{\pm}(\xi^2+2) + B_{\pm}\left[\int_{|\xi|}^{\infty} e^{-\xi^2/4} \frac{d\xi}{(\xi^2+2)^2}\right](\xi^2+2) \tag{19}$$

$$\text{with} \quad C_{\pm} = F_0''(0^{\pm}),$$

while requiring that $f_2(0^{\mp}) = 0$, $[f_2'] = 0$ leads to a set of three algebraic equations

$$-K_0^* G_{\pm} + C_{\pm} + B_{\pm}\sqrt{\pi}/4 = 0, \quad B_+ + B_- = 0. \tag{20}$$

where $K_0^* = K_0 - F_*$.

The unknowns in the set (20) are $B_\pm$ and $K_0^*$, while the determinant of the system is $G_+ + G_-$. If $G_+ + G_-$ does not vanish, we have a unique solution

$$K_0^* = \frac{C_+}{G_+} = \frac{C_-}{G_-} = V \text{ (say)}, \quad B_\pm = 0. \tag{21}$$

(The equality $C_+/G_+$ and $C_-/G_-$ may be assumed during the evolution since otherwise the differential equation (14a) is violated.) Thus the velocity of the wave is determined. On the other hand, if $G_+ + G_-$ does vanish, the solution may be written

$$B_\pm = 4(K_0^* - V)G_\pm/\sqrt{\pi} \quad \text{with} \quad K_0^* \text{ undetermined.} \tag{22}$$

To find $K_0^*$, we proceed to the next step in the calculations, i.e. the determinant of $f_3(\xi)$. Three algebraic equations, corresponding to the system (20) are obtained, whose determinant is again $G_+ + G_-$; now, however, there is a consistency requirement

$$K_0^{*2} - 3VK_0^* + 2(T_- - T_+) = 0 \quad \text{with} \quad T_\pm = F_0'''(0^\pm) \tag{23}$$

when the determinant vanishes. If there is a value of $K_0$, it must satisfy this quadratic equation; note that there are real roots if and only if

$$T_- - T_+ \leq 9V^2/8. \tag{24}$$

In the calculations of Figure 3, it was observed that the numerical integration failed around $T = 80 \times 10^{-4}$. The values of the slopes $G_\pm$ at the origin were found to have a sum $G_+ + G_-$ close to zero. The quadratic (23) presumably did not have real roots, although third derivatives were not obtained accurately enough to exhibit a violation of the inequality (24).

We have not been able to reproduce Stewart and Ludford's spikes (Figure 5) in our computations so far but, if they exist, the above analysis provides an explanation. Instead of a cusp, there is actually a jump in K from its limiting value as $G_+ + G_- \to 0$ to a root of the quadratic (23). Figure 6 sketches such jumps, which could appear to be cusps if the resolution in time were not good enough.

The theory is currently being tested numerically, by applying initial conditions for which $G_+ + G_- = 0$ and $G_+/C_+ = G_-/C_-$, with third derivatives satisfying and violating the conditions (24). We are also trying to find a (singular) solution to take over when the condition is violated, i.e. when there is no finite velocity with which the discontinuity can move away.

REFERENCES.

[1] D.S. Stewart and G.S.S. Ludford. Near Chapman Jouget detonations. Transactions of the 1st Army Conference on Applied Mathematics and Computing, pp. 801-811, ARO-Report 84-1, 1984.

[2] D.S. Stewart. Private Communication.

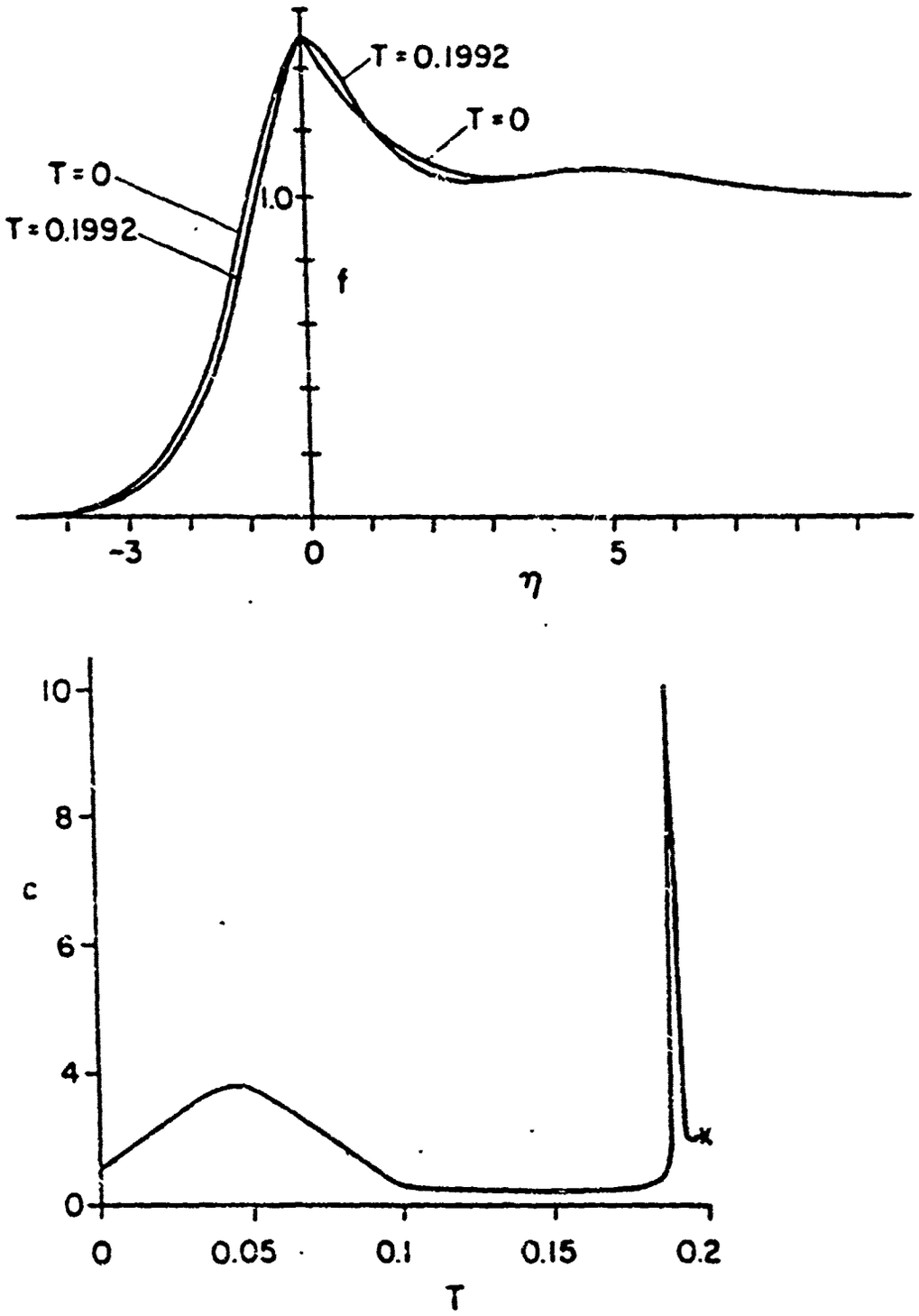



Figure 2. Response of a neutrally stable CJ-detonation to a rarefaction in the burnt region. Breakdown at T = 0.1992 is not shown by f-profile.

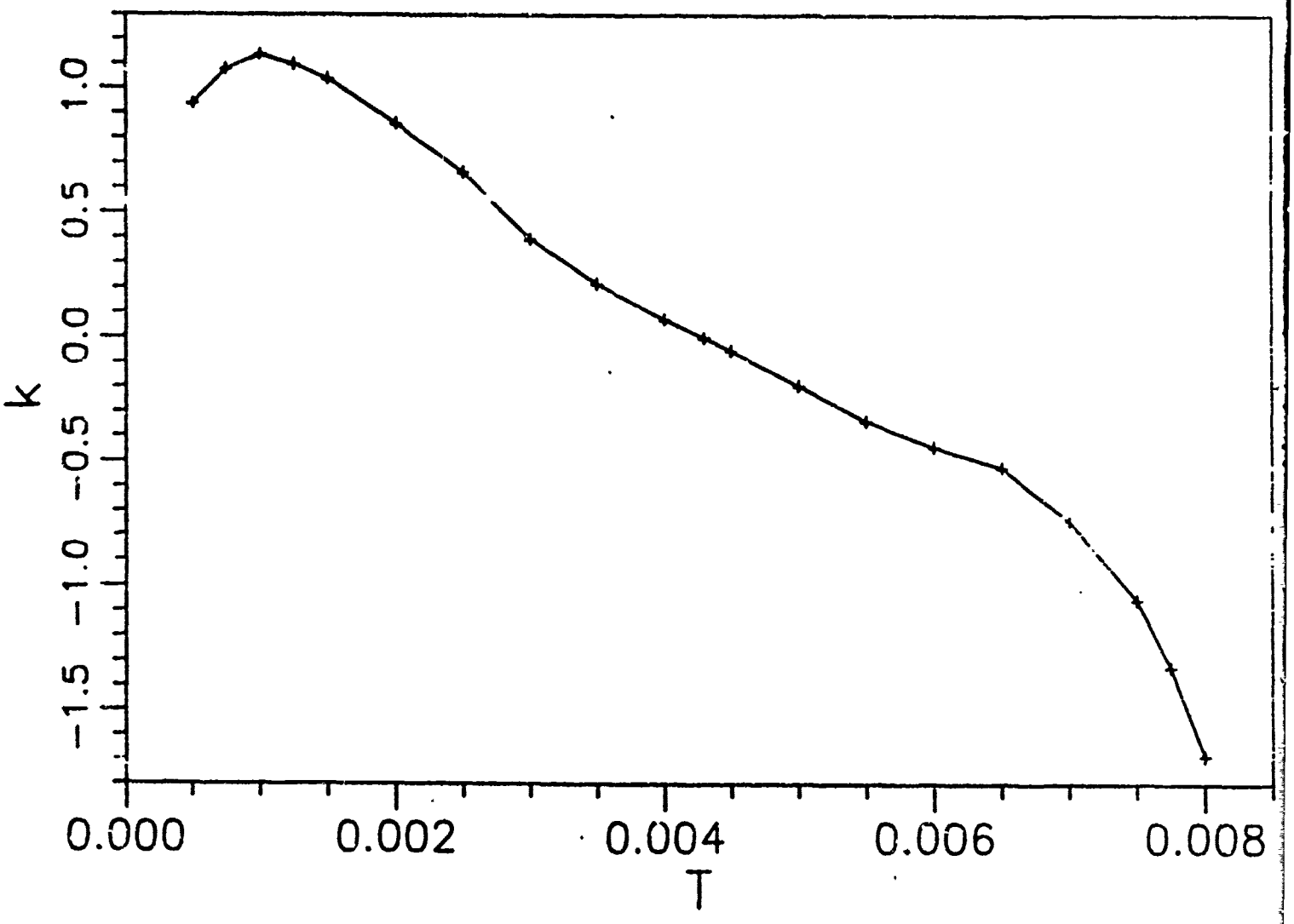



Figure 3. Response of a neutrally stable C-J deontation to a slight rarefaction in the burnt region of Figure 2. Integration fails at $T = 80 \times 10^{-4}$ ($F_* = 1.50$).

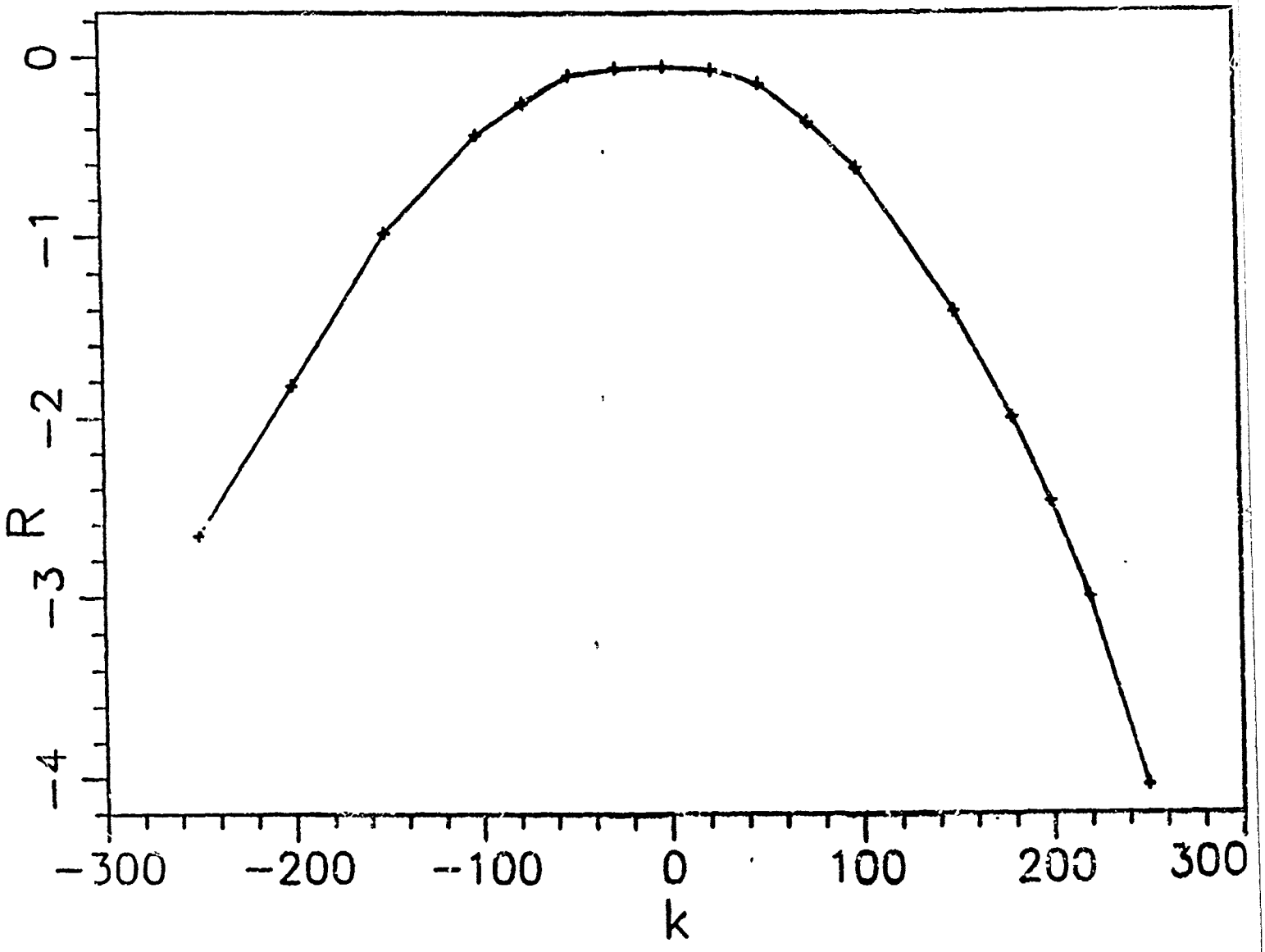
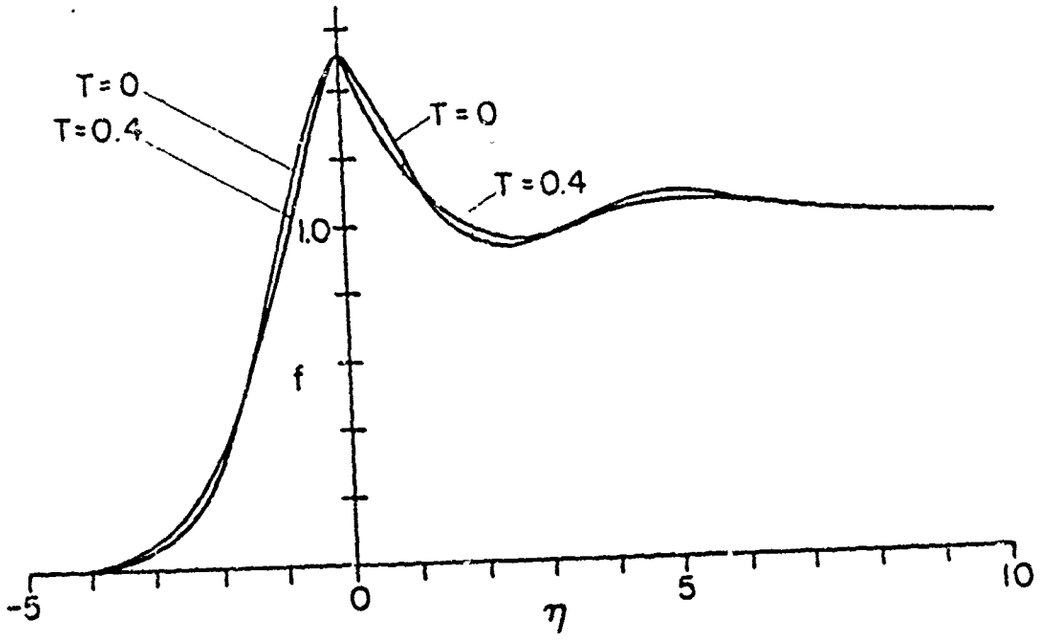



Figure 4. Immediately after breakdown in Figure 3 ($T = 80 \times 10^{-4}$), a plot of R v. K was carried out at $T = 80.5 \times 10^{-4}$ as shown above.



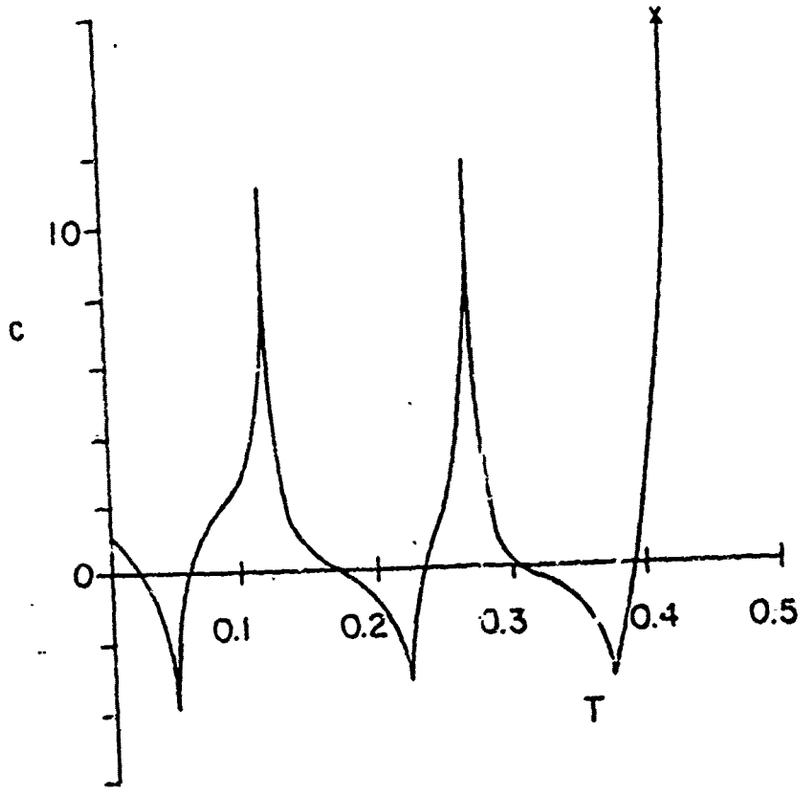




Fig. 5, Response of CJ-detonation in Figure 2 to a slightly stronger rarefaction. At breakdown the f-profile is indistinguishable from that shown for T = 0.4 .

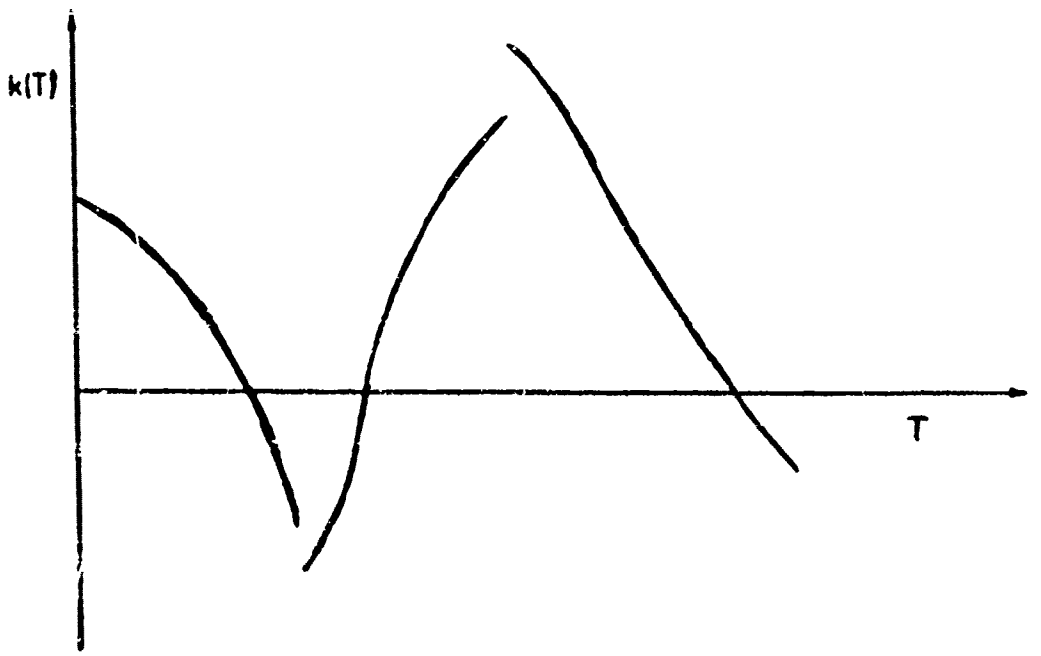



Figure 6. Satisfying and violating the conditions (24). We are also trying to find a (singular) solution to take over when the condition is violated, i.e. when there is no finite velocity with which the discontinuity can move away.

# FINITE INCREMENT FORMULATION OF THE PRANDTL-REUSS CONSTITUTIVE EQUATIONS

Russell L. Mallett

Department of Mechanical Engineering,
Aeronautical Engineering & Mechanics
Rensselaer Polytechnic Institute
Troy, New York 12180-3590

ABSTRACT. The early finite increment formulations of elastic-plastic constitutive laws for use in large deformation finite-element codes placed severe limits on step size in order to avoid significant error accumulation. In this paper finite increment forms of the constitutive equations for Prandtl-Reuss materials are investigated, two primary sources of finite increment error are identified, and procedures for elimination of these errors are developed. By adopting a geometric description in nine-dimensional stress space, the relevant stress and strain increments are shown to all lie in a two-dimensional subspace. This allows the nature of the errors and the description of procedures which overcome them to be easily and precisely visualized with the aid of planar vector diagrams.

I. INTRODUCTION. In order to perform deformation and stress analyses for a material being finitely deformed, a spacially discretized model of the continuum usually must be introduced and this is commonly achieved by the finite-element method. A technique must then be developed to solve the nonlinear governing equations. The character of the solution technique depends in part on the nature of the material's constitutive law. An incremental procedure is appropriate in the case of plasticity because of its path dependent nature. This requires that the constitutive equations be cast in finite increment form but, when this is done in the simplest and most direct fashion, two problems develop. First, stress points which should remain on a convex yield surface are forced to move finite distances in a tangential direction during each incremental step and thus they drift outward from the yield surface with each step. Second, stress points tend to drift off of the strain hardening curve for essentially the same reason. In both cases the finite increment errors accumulate monotonically since they are essentially in the same direction at each step. As will be shown, suitable modification of the incremental constitutive equations can completely eliminate such errors.

After defining the physical variables of interest and briefly reviewing the relevant continuum equations, a geometric characterization of the deviatoric Prandtl-Reuss constitutive equations in nine-dimensional Kirchhoff stress space will be developed. Then it will be shown that the relevant stress, stress rate and deformation rate tensors all lie in a two-dimensional subspace of stress space. This naturally leads to a simple description in terms of a planar

PREVIOUS PAGE IS BLANK

vector diagram. Finite-increment forms of the Prandtl-Reuss equations will then be analyzed with the aid of the corresponding planar vector diagram.

For the case of a nonstrain-hardening material, a geometric description will be given of finite increment procedures which insure that stress points remain on the fixed yield surface. The preferred procedure will then be generalized to the case of strain hardening where the added complication of drift from the strain hardening curve arises. A geometric description of a procedure for precise tracking of the strain hardening curve will be given.

II. CONTINUUM EQUILIBRIUM AND CONSTITUTIVE EQUATIONS. In recent years it has become possible to apply the finite-element method to the difficult problem of analyzing metal forming processes involving large deformation. Hill [1] developed the governing continuum equations (in variational form) for rate independent materials at finite strain and Nibbitt, Marcal and Rice [2] developed a corresponding Lagrangian finite-element procedure. McMeeking and Rice [3] proposed an alternative finite-element procedure emphasizing the current rather than the initial configuration because it is more appropriate for analyzing plastic material behavior. The variational basis of their method will be given following a brief review of relevant definitions and notation.

Let $x_i$ denote the three rectangular Cartesian coordinates of a material particle in the current (deformed) configuration and let $X_i$ denote its coordinates in the reference (undeformed) configuration. The particle's trajectory is then given by

$$x_i = x_i(X_1, X_2, X_3, t), \quad i = 1, 2, 3 \tag{1}$$

where t is the elapsed time. Taking first derivatives gives the material velocity vector $\underset{\sim}{V}$ and the deformation gradient tensor $\underset{\sim}{F}$

$$\underset{\sim}{V} = \left\{\frac{\partial x_i}{\partial t}\right\}, \quad \underset{\sim}{F} = \left(\frac{\partial x_i}{\partial X_j}\right). \tag{2}$$

A second differentiation gives the velocity gradient tensor $\underset{\sim}{L}$ whose symmetric and skew-symmetric parts are the deformation rate $\underset{\sim}{D}$ and the material spin $\underset{\sim}{W}$, respectively

$$\underset{\sim}{L} = \left(\frac{\partial V_i}{\partial x_j}\right) = \dot{\underset{\sim}{F}}\underset{\sim}{F}^{-1} = \underset{\sim}{D} + \underset{\sim}{W} \tag{3}$$

The material mass density ratio is given by

$$\frac{\rho}{\rho_o} = \frac{1}{\det \underset{\sim}{F}} \tag{4}$$

When finite deformations are considered, care must be taken to distinguish the various possible stress tensor definitions. When the current configuration is emphasized, the Kirchhoff stress $\underset{\sim}{\tau}$ and the true or Cauchy stress $\underset{\sim}{\sigma}$ are particularly useful. They are related by

$$\underset{\sim}{\sigma} = \frac{\rho}{\rho_o} \underset{\sim}{\tau} \tag{5}$$

It is also necessary to distinguish various rates of stress. Constitutive laws, for example, require the use of an objective stress rate but it is the material rate that must be integrated to determine stress evolution. The material rate $\dot{\underset{\sim}{\tau}}$ and the objective Jaumann rate $\overset{*}{\underset{\sim}{\tau}}$ of Kirchhoff stress are related by the equation

$$\dot{\underset{\sim}{\tau}} = \overset{*}{\underset{\sim}{\tau}} + \underset{\sim}{W}\underset{\sim}{\tau} - \underset{\sim}{\tau}\underset{\sim}{W} \tag{6}$$

which can be interpreted as a decomposition of $\dot{\underset{\sim}{\tau}}$ separating the effects of material deformation and rotation.

The (corrected) rate form of the principle of virtual work developed by McMeeking and Rice is

$$\int \frac{\rho}{\rho_o} \left[ \overset{*}{\tau}_{ij} \delta D_{ij} - \frac{1}{2} \tau_{ij} \delta (2D_{ik}D_{jk} - V_{k,i}V_{k,j}) \right] dv = .... \tag{7}$$

where integration is over the volume of the current configuration and the missing terms on the right-hand side account for the effects of prescribed surface tractions and body forces. It is the velocity field that is subject to variation and the required auxiliary equations are the constitutive law in the form

$$\overset{*}{\tau}_{ij} = \mathcal{L}_{ijk\ell} D_{k\ell} \tag{8}$$

and the mass conservation law

$$\dot{\rho} = - \rho D_{kk} . \tag{9}$$

After spacial (finite-element) discretization, Eqs.(7) and (8) are applied incrementally with the determined velocity field used to update the configuration and Eqs.(6) and (9) used to update the material density and stress fields at the end of each incremental loading step.

In order to cast the constitutive equations for a time independent elastic-plastic material in the form of Eq.(8), it is necessary to obtain a rate form of the elasticity equations and combine them with the appropriate plasticity equations. This is most easily accomplished by making the kinematic assumption that the deformation rate can be linearly decomposed into elastic and plastic parts

$$\underset{\sim}{D} = \underset{\sim}{D}^{e} + \underset{\sim}{D}^{p} . \tag{10}$$

The desired result is obtained by substitution from the elasticity relation

$$\underset{\sim}{D}^{e} = C \overset{\nabla}{\underset{\sim}{\tau}}^{*} \tag{11}$$

and the plasticity flow rule

$$\underset{\sim}{D}^{p} = \frac{1}{h} \left( \frac{\partial f}{\partial \underset{\sim}{\tau}} : \overset{\nabla}{\underset{\sim}{\tau}}^{*} \right) \frac{\partial f}{\partial \underset{\sim}{\tau}} , \tag{12}$$

where h denotes the tangent modulus of the strain hardening law and f denotes the yield function in the yield criterion

$$f(\tau) = Y , \tag{13}$$

and by inverting the resulting equation to solve for $\overset{\nabla}{\underset{\sim}{\tau}}^{*}$.

For the case of isotropic elasticity, a Mises yield criterion and isotropic strain hardening, the above procedure produces the equations for a Prandtl-Reuss material characterized by two elastic constants, such as the shear modulus G and the bulk modulus K, and a strain hardening curve

$$Y = Y(\bar{\varepsilon}^{p}) \tag{14}$$

where Y is the yield stress and $\bar{\varepsilon}^{p}$ is the equivalent plastic strain determined by integration of

$$\dot{\bar{\varepsilon}}^{p} = \left( \frac{2}{3} D^{p}_{ij} D^{p}_{ij} \right)^{1/2} . \tag{15}$$

Separating the result into hydrostatic and deviatoric parts gives

$$\overset{\nabla}{\tau}_{ii} = K D_{ii} \tag{16}$$

$$\overset{*}{\tau}'_{ij} = 2G\left(D'_{ij} - \frac{\beta}{1+h/3G}\,\frac{\frac{3}{2}\,\tau'_{ij}\tau'_{k\ell}}{\bar{\tau}^2}\,D'_{k\ell}\right) \tag{17}$$

where primes denote deviators, $\bar{\tau}$ the equivalent Kirchhoff stress

$$\bar{\tau} = \left(\frac{3}{2}\,\tau'_{ij}\tau'_{ij}\right)^{1/2}, \tag{18}$$

h the tangent modulus

$$h = \frac{dY}{d\bar{\varepsilon}^p} \tag{19}$$

and where $\beta$ is unity during plastic flow ($\bar{\tau} = Y$) and zero otherwise. Since the situation is trivial when plastic flow is not occurring, we shall only consider the case of plastic flow in the remainder of the paper and thus $\beta$ will be set equal to unity.

III. GEOMETRIC REPRESENTATION OF STRESS TENSORS. It is easily verified that a stress tensor $\underset{\sim}{\sigma}$ may be regarded as a vector in a nine-dimensional vector space. In particular, its Cartesian components $\sigma_{ij}$ can be interpreted as nine rectangular Cartesian coordinates of a point in a nine-dimensional Euclidean space called stress space. Since $\underset{\sim}{\sigma}$ is a symmetric tensor, it actually lies in a six-dimensional subspace of stress space called reduced stress space. The usual decomposition of a stress tensor into its deviatoric and hydrostatic parts,

$$\sigma_{ij} = \sigma'_{ij} + \frac{1}{3}\,\delta_{ij}\sigma_{kk} \tag{20}$$

becomes a vector decomposition into two orthogonal components, the orthogonality resulting from the fact that

$$\sigma'_{ij} \cdot \frac{1}{3}\,\delta_{ij}\sigma_{kk} = 0\,. \tag{21}$$

Thus reduced stress space can be decomposed into two orthogonal subspaces, a one-dimensional subspace of hydrostatic stresses and a five-dimensional subspace of deviatoric stresses.

A direct consequence of the vector representation is the fact that stresses can be visualized with ordinary vector diagrams whenever they lie in a subspace of dimension three or less. The orthogonal decomposition just mentioned, for example, can be represented by a planar figure where $\underset{\sim}{\sigma}_{dev}$ and $\underset{\sim}{\sigma}_{hyd}$ are perpendicular and have lengths of $(\sigma'_{ij}\sigma'_{ij})^{1/2}$ and $\sigma_{kk}/\sqrt{3}$, respectively (Figure 1).

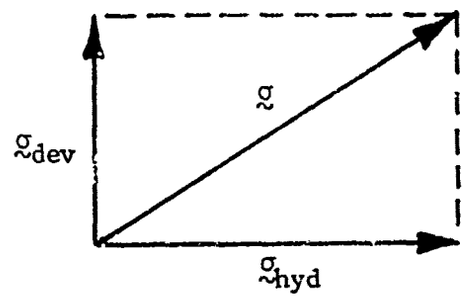



Figure 1 Decomposition of Stress

Another easily visualized situation occurs when the principal directions of the stress tensors under consideration coincide and remain fixed (in physical space) so that the stress tensors all lie in a three-dimensional subspace of stress space called principal stress space. Decomposition of this subspace into its hydrostatic and deviatoric parts leads to the well known $\Pi$ plane. Unfortunately, this representation is often misused by employing it when the principal directions do not in fact remain fixed.

The deviatoric Prandtl-Reuss equations can be written in symbolic tensor form as

$$\overset{*}{\underset{\sim}{T}}' = 2G(\underset{\sim}{D}' - \underset{\sim}{D}^P) . \tag{22}$$

If we permit stress rates to be represented as vectors in stress space (since they are in essence stress increments), then Eq.(22) can be interpreted as a decomposition of the stress rate vector $\overset{*}{\underset{\sim}{T}}'$ into two components, $2G\underset{\sim}{D}'$ and $2G\underset{\sim}{D}^P$, in Kirchhoff stress space. Since $2G\underset{\sim}{D}^P$ is proportional to $\underset{\sim}{T}'$, and thus the corresponding vectors are parallel, it is clear that all of the vectors of interest are contained in the two-dimensional subspace spanned by the stress vector $\underset{\sim}{T}'$ and the stress rate vector $2G\underset{\sim}{D}'$ and can be represented in a planar diagram.

If a unit vector $\underset{\sim}{n}$ in the direction of $\underset{\sim}{T}'$ is introduced, a more explicit form of Eq.(22) is obtained, namely:

$$\overset{*}{\underset{\sim}{T}}' = 2G\left[\underset{\sim}{D}' - \frac{1}{1+h/3G}\,\underset{\sim}{n}(\underset{\sim}{n}\cdot\underset{\sim}{D}')\right] = P(h,\underset{\sim}{n})\cdot\underset{\sim}{D}' \tag{23}$$

where $P(h,\underset{\sim}{n})$ designates a linear vector operator. Note that P becomes a projection operator in the case of perfect plasticity ($h = 0$) since it projects $\underset{\sim}{D}'$ onto the hyperplane normal to $\underset{\sim}{n}$. Noting that the Mises yield surface

$$T'_{ij}T'_{ij} = \frac{2}{3}Y^2 \tag{24}$$

becomes a circle in the two-dimensional subspace, Figure 2 is obtained. Note that $\underset{\sim}{n}$ can be interpreted as the yield surface unit normal vector.

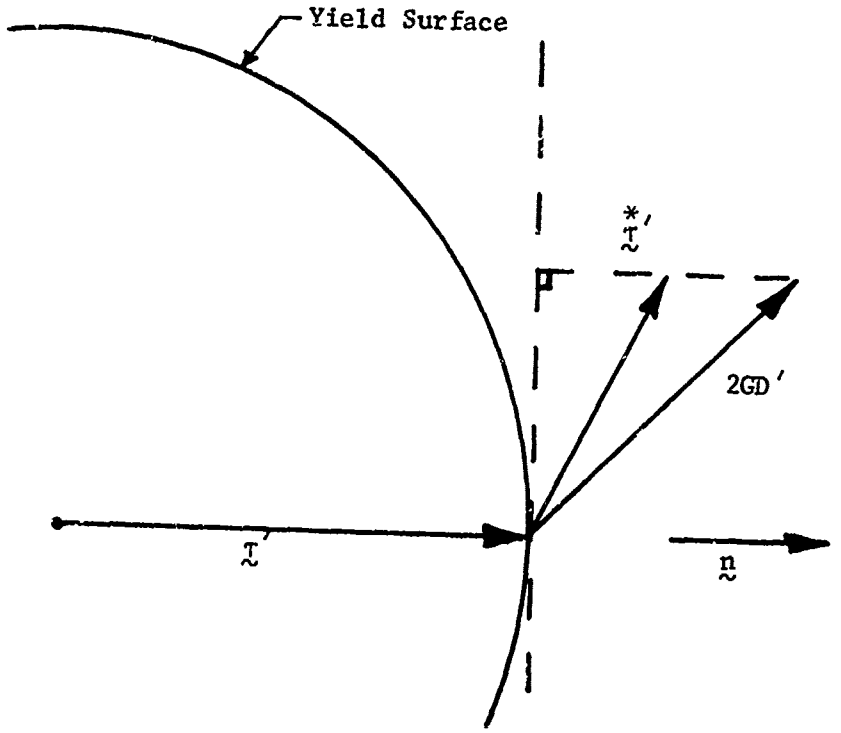



Figure 2 Geometric Representation of the Deviatoric Prandtl-Reuss Equations

IV. FINITE INCREMENTS FOR CASE OF PERFECT PLASTICITY. Consider first the case of perfectly plastic materials (no strain hardening). If we introduce the notation

$$\Delta\underset{\sim}{\tau}' = \overset{*}{\underset{\sim}{\tau}}'\Delta t, \quad \Delta\underset{\sim}{\varepsilon}' = \underset{\sim}{D}'\Delta t \tag{25}$$

for finite increments of stress and strain, and set $h = 0$, then a finite increment form of Eq.(23) is

$$\Delta\underset{\sim}{\tau}' = 2G(\underset{\sim}{I} - \underset{\sim}{n}\underset{\sim}{n}) \cdot \Delta\underset{\sim}{\varepsilon}' \tag{26}$$

where $\underset{\sim}{I}$ is the identity operator, and $\underset{\sim}{n}\underset{\sim}{n}$ a dyad in nine-dimensional stress space. This gives the vector diagram shown in Figure 3.

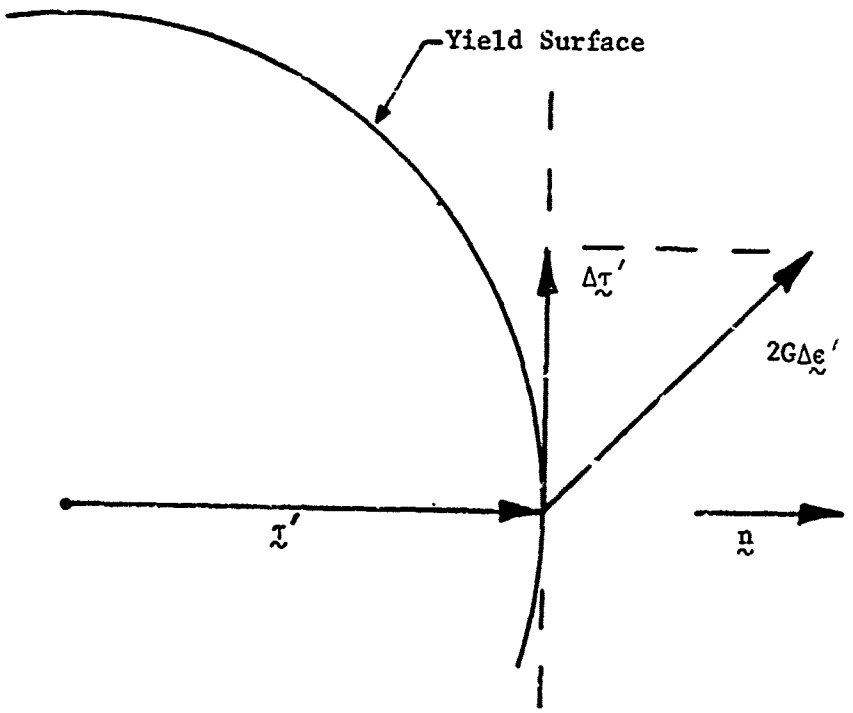



Figure 3 Finite Increment Prandtl-Reuss Diagram for Perfect Plasticity Case (h = 0)

As indicated previously this result is unacceptable because the stress vector $\underset{\sim}{\tau}' + \Delta\underset{\sim}{\tau}'$ does not remain on the yield surface as it should One somewhat crude but often used procedure to eliminate this defect, called radial return, is to simply shift the tip of the $\Delta\tau'$ vector toward the origin until it meets the Mises yield circle. Another is to let the tip of the $\Delta\underset{\sim}{\tau}'$ vector coincide with the point of the yield circle determined by the line from the origin to the tip of the $2G\Delta\underset{\sim}{\varepsilon}'$ vector. A much more rational procedure is obtained by noting that improved accuracy is always obtained by using midstep values of the coefficients in an equation relating increments. In the current situation that suggests that we replace the initial yield surface normal $\underset{\sim}{n}$ by some midstep normal $\hat{\underset{\sim}{n}}$, so that

$$\Delta\underset{\sim}{\tau}' = 2G(I - \hat{\underset{\sim}{n}}\hat{\underset{\sim}{n}}) \cdot \Delta\underset{\sim}{\varepsilon}' . \tag{27}$$

In particular, if we determine the locus of the tip of the $\Delta\underset{\sim}{\tau}'$ vector for all possible vectors $\hat{\underset{\sim}{n}}$ and determine where that locus intersects the yield circle, the best $\hat{\underset{\sim}{n}}$ will be determined. Since Eq.(27) forces $\Delta\underset{\sim}{\tau}'$ to be perpendicular to $\hat{\underset{\sim}{n}}$, the desired locus is a circle with the vector $2G\Delta\underset{\sim}{\varepsilon}'$ as a diameter as shown in Figure 4. Simple geometry

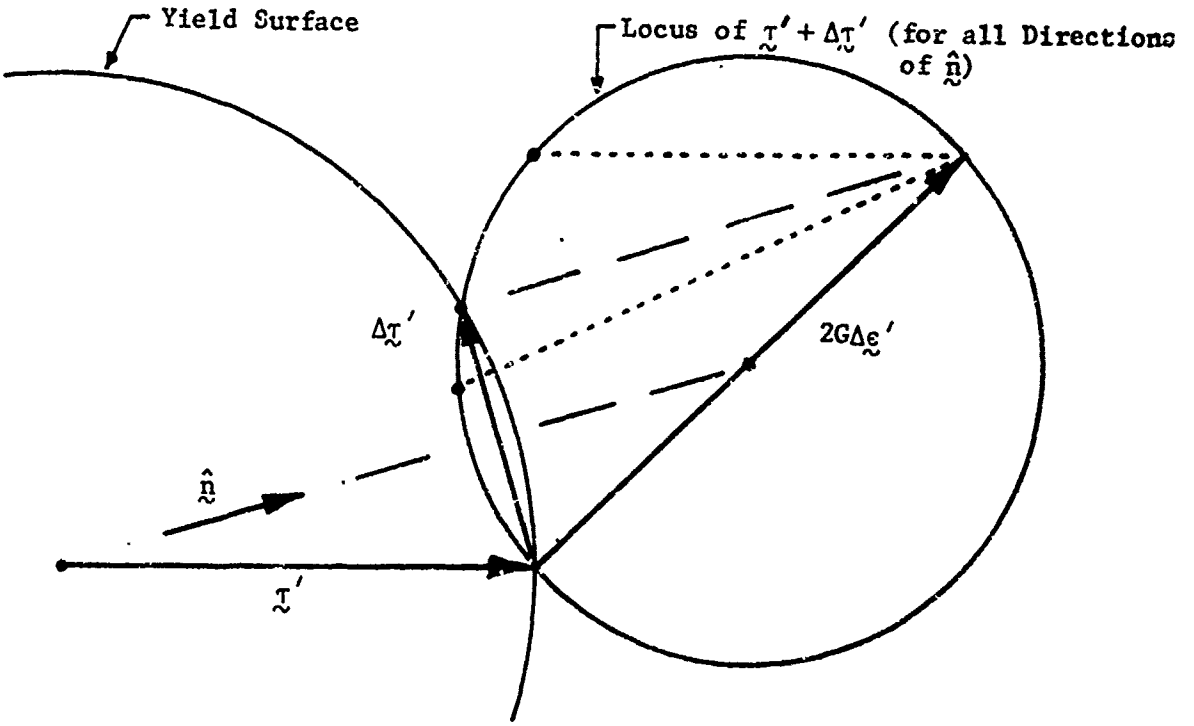



Figure 4 Modified Finite Increment Prandtl-Reuss Diagram for Perfect Plasticity Case (h = 0)

then shows that the $\hat{\underset{\sim}{n}}$ vector which keeps the stress vector $\underset{\sim}{\tau}' + \Delta\underset{\sim}{\tau}'$ on the yield circle is directed along the line connecting the origin and the center of the circular locus, i.e.

$$\hat{\underset{\sim}{n}} = \frac{\underset{\sim}{\tau}' + G\Delta\underset{\sim}{\varepsilon}'}{|\tau' + G\Delta\underset{\sim}{\varepsilon}'|} \tag{28}$$

This result was originally given by Rice and Tracey [4] who defined $\hat{\underset{\sim}{n}}$ as a unit vector in the direction of the average of $\underset{\sim}{\tau}'$ and $\underset{\sim}{\tau}' + 2G\Delta\underset{\sim}{\varepsilon}'$ and then demonstrated that $\hat{\underset{\sim}{n}}$ has the "remarkable feature" that the stress tensor $\underset{\sim}{\tau} + \Delta\underset{\sim}{\tau}'$ obtained by using $\hat{\underset{\sim}{n}}$ in place of $\underset{\sim}{n}$ satisfies the yield criterion precisely. Since they represented stress states as vectors in the $\Pi$-plane, their result should have been restricted to cases where principal directions remain fixed in the material. However, we have shown here that, if stress states are properly represented as vectors, the result can be derived and is not subject to restrictions.

The dotted lines in Figure 4 indicate the results that would be given by Eq.(27) if the yield surface normal at the beginning or end of the step were used for $\hat{\underset{\sim}{n}}$. Note that if radial return was used in these cases, to force the stress point to lie on the yield circle, the resulting

stress points would bracket the optimum result and thus it could be claimed that using the initial normal leads to an overestimation of $\Delta\underset{\sim}{\tau}'$ while the end of step normal leads to an underestimate.

V. FINITE INCREMENTS FOR CASE OF STRAIN HARDENING PLASTICITY. Now we turn to the case where the material exhibits strain hardening ($h \neq 0$). The finite increment form of Eq.(23) can be written as

$$\Delta\underset{\sim}{\tau}' = 2G(\Delta\underset{\sim}{\varepsilon}' - \Delta\underset{\sim}{\varepsilon}^p) \tag{29}$$

where

$$\Delta\underset{\sim}{\varepsilon}^p = \frac{1}{1+h/3G}\,\hat{\underset{\sim}{n}}(\hat{\underset{\sim}{n}} \cdot \Delta\underset{\sim}{\varepsilon}') \tag{30}$$

If we assume that $\hat{\underset{\sim}{n}}$ is selected according to Eq.(28), then the tip of the $\Delta\underset{\sim}{\tau}'$ vector in Figure 4 will be shifted along the dashed line a distance determined by the factor $(1+h/3G)^{-1}$. The result is indicated in Figure 5. However, if we use the value of h corresponding to the beginning of the incremental step, the point on the strain-hardening curve will move along a tangent to that curve and thus will drift above it as indicated in Figure 6.

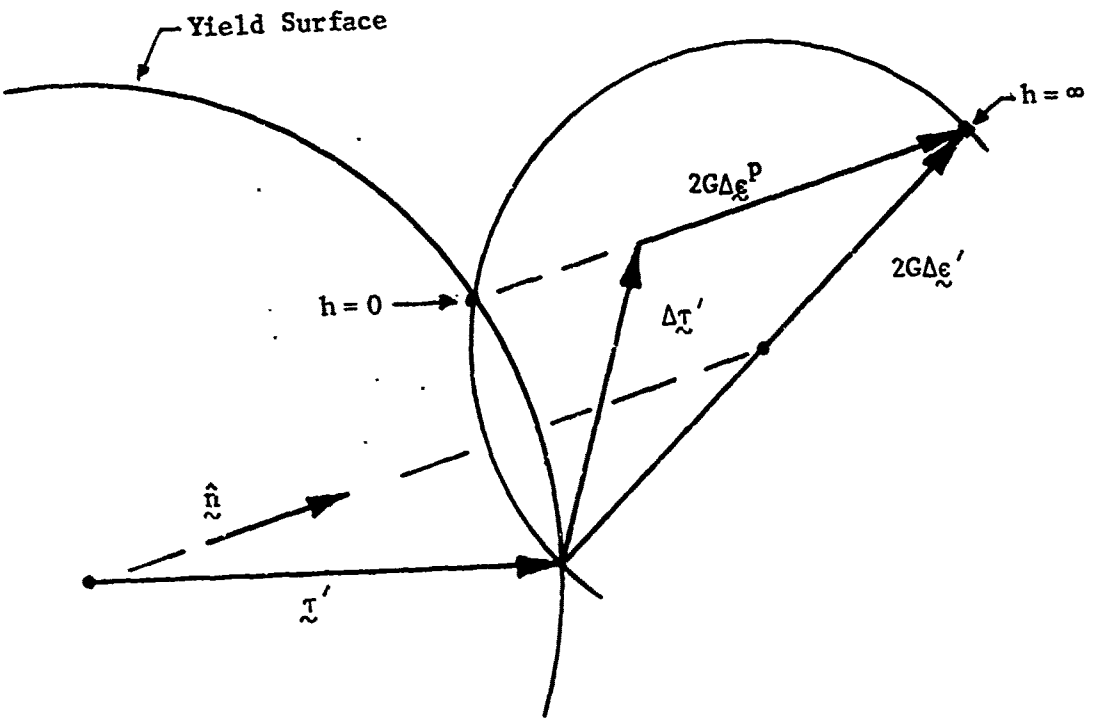



Figure 5 Finite-Increment Prandtl-Reuss Diagram for Strain Hardening Case ($h \neq 0$)

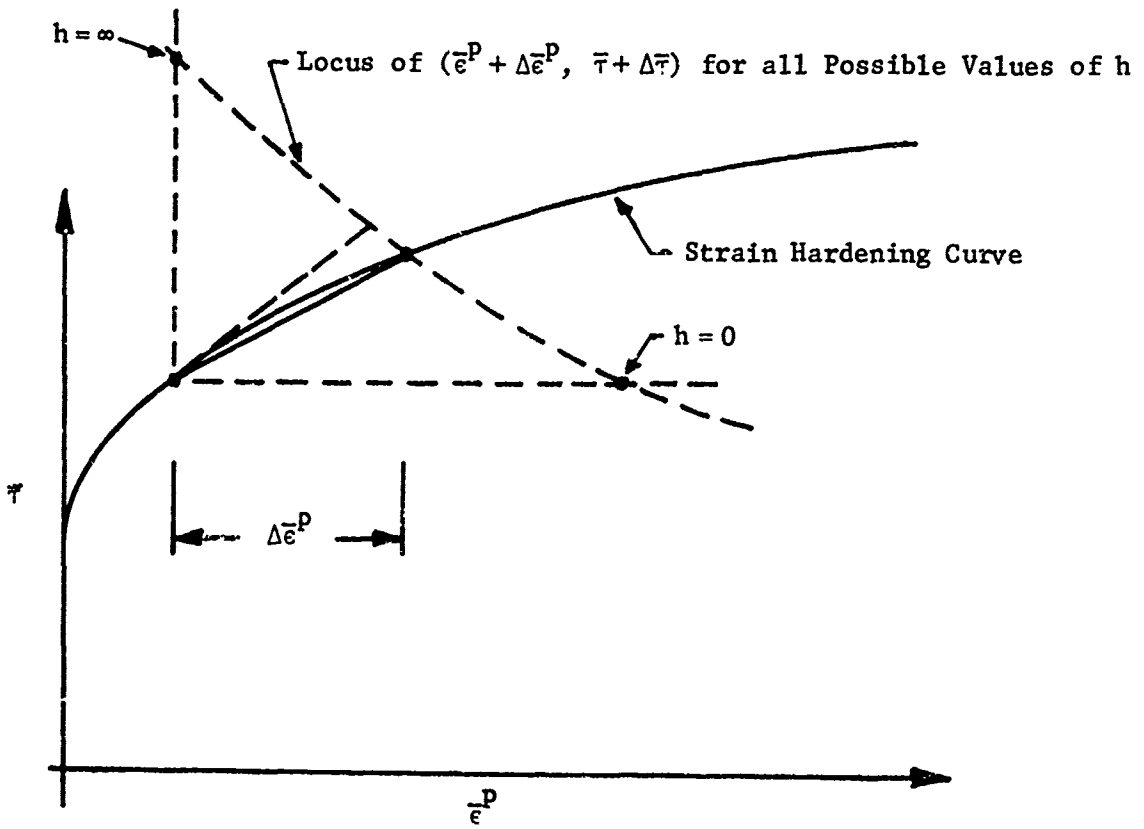



Figure 6 Diagram for Determining Midstep Tangent Modulus

Once again it is advisable to try to find an appropriate midstep value, this time for the tangent modulus h so that the stress point remains on the strain hardening curve. This can be accomplished by finding the locus of the points $(\bar{\epsilon}^p + \Delta\epsilon^p,\ \bar{\tau} + \Delta\bar{\tau})$ corresponding to all positive values of h and determine where this locus intersects the strain hardening curve.

An equation for the locus is easily determined by resolving the vector $\underset{\sim}{\tau}' + \Delta\underset{\sim}{\tau}'$ into components parallel and perpendicular to $\hat{\underset{\sim}{n}}$ and then expressing its squared length as the sum of the squares of its components. Referring to Figure 5 gives the following result:

$$|\underset{\sim}{\tau}' + \Delta\underset{\sim}{\tau}'|^2 = \left(\hat{\underset{\sim}{n}} \cdot (\underset{\sim}{\tau}' + 2G\Delta\underset{\sim}{\epsilon}') - |2G\Delta\underset{\sim}{\epsilon}^p|\right)^2 + |G\Delta\underset{\sim}{\epsilon}'|^2 - (\hat{\underset{\sim}{n}} \cdot G\Delta\underset{\sim}{\epsilon}')^2. \qquad (31)$$

Noting that the last two terms can be alternately expressed as $|\underset{\sim}{\tau}'|^2 - (\underset{\sim}{n} \cdot \underset{\sim}{\tau}')^2$ and that

$$\bar{\tau} + \Delta\bar{\tau} = \sqrt{3/2}\ |\underset{\sim}{\tau}' + \Delta\underset{\sim}{\tau}'| \qquad (32)$$

$$\Delta\bar{\varepsilon}^p = \sqrt{2/3}\ |\Delta\underset{\sim}{\varepsilon}^p|\ , \tag{33}$$

the equation of the locus becomes

$$(\bar{\tau} + \Delta\bar{\tau})^2 = (P - 3G\Delta\bar{\varepsilon}^p)^2 + \bar{\tau}^2 - Q^2 \tag{34}$$

where

$$P = \sqrt{3/2}\ \hat{\underset{\sim}{n}} \cdot (\underset{\sim}{\tau}' + 2G\Delta\underset{\sim}{\varepsilon}') \tag{35}$$

$$Q = \sqrt{3/2}\ \hat{\underset{\sim}{n}} \cdot \underset{\sim}{\tau}' . \tag{36}$$

This equation relating $\Delta\bar{\tau}$ and $\Delta\bar{\varepsilon}^p$ is clearly a hyperbola. Its upper branch intersects the strain hardening curve, as indicated in Figure 6, and thus determines a chord whose slope is the desired midstep value of the tangent modulus h. In fact, if the strain hardening curve is specified in tabular fashion, so that it is in effect given as a piecewise linear function, then simple closed form expressions for $\Delta\bar{\varepsilon}^p$, $\Delta\bar{\tau}$ and $h = \Delta\bar{\tau}/\Delta\bar{\varepsilon}^p$ can be obtained.

The accuracy of this procedure has been verified by comparing its results with the results obtained by numerical integration for cases where $\underset{\sim}{D}'$ remains constant over the time interval $\Delta t$. It was found to be highly accurate even for large increment cases where $|2G\Delta\underset{\sim}{\varepsilon}'| = |\underset{\sim}{\tau}'|$.

In situations where a stress point lies on the yield surface but does not move very much over a period of time the above procedure is very stable. When less accurate procedures are used for the strain-hardening case there is a tendency for such a stress point to vascillate between plastic loading and elastic unloading in successive steps. This can cause considerable difficulty because the incremental stiffness matrix must be recomputed whenever a stress point changes from loading to unloading or vice versa and this change can cause other stress points to change which in turn can cause the first stress point to change again. In fact there is the possibility that this sequence of recomputations would never end, although we have never encountered such a situation in practice.

VI. MIDSTEP YIELDING. Since the onset of yielding represents a discontinuity in material behavior, it would seem advantageous to adjust the size of each incremental step so that such discontinuities always occur at the end of a step. However, this would result in a large number of very small steps. Marcal and King [5] suggested that such a limitation on step size could be avoided by allowing midstep yielding as indicated in Figure 7.

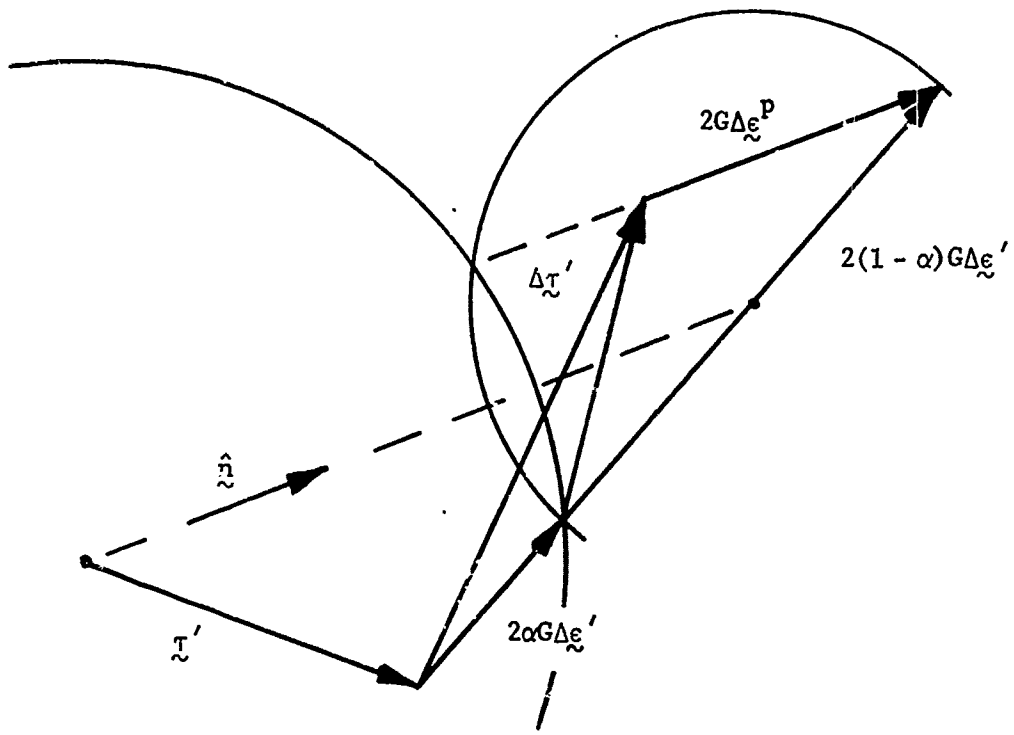



Figure 7 Finite Increment Prandtl-Reuss Diagram for Midstep Yielding

If $\alpha$ denotes the fraction of the step that is required for a stress point to reach the yield surface, then $\Delta\underset{\sim}{\tau}'$ can be separated into two parts, namely a purely elastic part, $2\alpha G\Delta\underset{\sim}{\varepsilon}'$, followed by an elastic-plastic part determined by the procedure in the previous section. The fraction $\alpha$ is determined by the requirement that

$$\sqrt{3/2}\ |\underset{\sim}{\tau}' + 2\alpha G\Delta\underset{\sim}{\varepsilon}'| = Y, \tag{37}$$

which leads to a quadratic equation for $\alpha$, and $\hat{\underset{\sim}{n}}$ is given by

$$\hat{n} = \frac{\underset{\sim}{\tau}' + (1+\alpha)G\Delta\underset{\sim}{\varepsilon}'}{|\underset{\sim}{\tau}' + (1+\alpha)G\Delta\underset{\sim}{\varepsilon}'|} \ . \tag{38}$$

The $\Delta\bar{\tau}$ versus $\Delta\bar{\varepsilon}^p$ locus is still given by Eq.(34) if the expression for Q is changed to

$$Q = \sqrt{3/2}\ \hat{n} \cdot (\underset{\sim}{\tau}' + 2\alpha G\Delta\underset{\sim}{\varepsilon}') \ . \tag{39}$$

VII. TOTAL STRESS INCREMENT. The total stress increment

$$\Delta \underset{\sim}{\tau} = \dot{\underset{\sim}{\tau}} \Delta t \tag{40}$$

must be determined in order to update stress at the end of each incremental step. So far we have only determined $\Delta \underset{\sim}{\tau}'$, the deviatoric part of the stress increment, corresponding to the Jaumann stress rate $\overset{*}{\underset{\sim}{\tau}}$. Two additional stress increments are required, one corresponding to the deviatoric part of $\overset{*}{\underset{\sim}{\tau}}$ and another corresponding to the tensor $\underset{\sim}{W}\underset{\sim}{\tau} - \underset{\sim}{\tau}\underset{\sim}{W}$ appearing in Eq.(6) which accounts for rotational effects. Since $\underset{\sim}{W}\underset{\sim}{\tau} - \underset{\sim}{\tau}\underset{\sim}{W}$ is deviatoric, the first of these two stress increments is simply the increment in $\underset{\sim}{\tau}_{hyd}$ which, according to Eq.(16), can be written in the stress vector form

$$\Delta \underset{\sim}{\tau}_{hyd} = \underset{\sim}{e} \Delta \tau_{ii} = \underset{\sim}{e} K D_{ii} \Delta t \tag{41}$$

where $\underset{\sim}{e}$ is the vector in stress space corresponding to the tensor $\delta_{ij}$.

If we let $\overset{o}{\underset{\sim}{\tau}}$ designate the stress vector corresponding to the tensor $\underset{\sim}{W}\underset{\sim}{\tau} - \underset{\sim}{\tau}\underset{\sim}{W}$, its increment becomes

$$\Delta_R \underset{\sim}{\tau} = \overset{o}{\underset{\sim}{\tau}} \Delta t \, . \tag{42}$$

The finite increment form of Eq.(6) then becomes

$$\Delta \underset{\sim}{\tau} = \Delta \underset{\sim}{\tau}_{yhd} + \Delta \underset{\sim}{\tau}' + \Delta_R \underset{\sim}{\tau} \tag{43}$$

and it remains to find an expression for $\Delta_R \underset{\sim}{\tau}$.

It is readily shown that $\overset{o}{\underset{\sim}{\tau}}$ is deviatoric and that $\underset{\sim}{\tau}_{hyd}$, $\underset{\sim}{\tau}_{dev}$ $(= \underset{\sim}{\tau}')$ and $\overset{o}{\underset{\sim}{\tau}}$ are directed in mutually orthogonal directions in stress space. In the absence of deformation rates $\underset{\sim}{\tau}_{hyd}$ remains fixed and $\underset{\sim}{\tau}'$ changes direction but remains fixed in length (and thus remains on a hypersphere in the five-dimensional stress deviator subspace). If the initial value of $\overset{o}{\underset{\sim}{\tau}}$ is used in Eq.(42), then $\Delta_R \underset{\sim}{\tau}$ is orthogonal to $\underset{\sim}{\tau}'$ and thus $\underset{\sim}{\tau}'$ grows in length during the incremental step and gives another example of a finite increment error which is biased so that it accumulates monotonically with each step. The simplest remedy is to apply the radial return technique, i.e., reduce the length of $\underset{\sim}{\tau}' + \Delta_R \underset{\sim}{\tau}$ without changing its direction, at the end of each step.

The more rational approach of employing a midstep value of $\overset{o}{\underset{\sim}{\tau}}$ in Eq.(42) is complicated by the fact that $\underset{\sim}{\tau}'$ does not remain confined to a two-dimensional subspace of stress space. Fortunately an exact solution for the increment itself can be obtained for the case of constant spin. The rotation tensor $\underset{\sim}{R}(t)$, associated with a spin tensor $\underset{\sim}{W}(t)$ is the solution of the initial value problem

$$\dot{\underset{\sim}{R}} = \underset{\sim}{W}\underset{\sim}{R}, \quad \underset{\sim}{R}(0) = \underset{\sim}{I}. \tag{44}$$

and for constant $\underset{\sim}{W}$ that solution, with the help of the Cayley-Hamilton theorem, becomes

$$R = e^{\underset{\sim}{W}t} = \underset{\sim}{I} + c_1\underset{\sim}{W} + c_2\underset{\sim}{W}^2 \tag{45}$$

where, at $t = \Delta t$,

$$c_1 = \frac{\sin(\omega\Delta t)}{\omega}, \quad c_2 = \frac{1 - \cos(\omega\Delta t)}{\omega^2} \tag{46}$$

$$\omega^2 = -\frac{1}{2}\,\mathrm{tr}(\underset{\sim}{W}^2). \tag{47}$$

The exact rotational stress increment is given by

$$\underset{\sim}{\tau} + \Delta_R\underset{\sim}{\tau} = \underset{\sim}{R}\,\underset{\sim}{\tau}\,R^T. \tag{48}$$

Substitution of the expression for $\underset{\sim}{R}$ leads to

$$\Delta_R\underset{\sim}{\tau} = c_1\underset{\sim}{\beta}_1 + c_2\underset{\sim}{\beta}_2 - c_1c_2\underset{\sim}{\beta}_3 + c_2^2\underset{\sim}{\beta}_4 \tag{49}$$

where the four deviatoric tensors $\underset{\sim}{\beta}_i$ are defined as

$$\underset{\sim}{\beta}_1 = \underset{\sim}{W}\underset{\sim}{\tau} - \underset{\sim}{\tau}\underset{\sim}{W} \qquad \underset{\sim}{\beta}_2 = \underset{\sim}{W}^2\underset{\sim}{\tau} + \underset{\sim}{\tau}\underset{\sim}{W}^2 - 2\underset{\sim}{W}\underset{\sim}{\tau}\underset{\sim}{W}$$

$$\underset{\sim}{\beta}_3 = \underset{\sim}{W}(\underset{\sim}{W}\underset{\sim}{\tau} - \underset{\sim}{\tau}\underset{\sim}{W})\underset{\sim}{W} \qquad \underset{\sim}{\beta}_4 = \underset{\sim}{W}^2\underset{\sim}{\tau}\underset{\sim}{W}^2 + \omega^2\underset{\sim}{W}\underset{\sim}{\tau}\underset{\sim}{W} \tag{50}$$

Note that the four terms in Eq.(49) are of successively higher order in $\Delta t$.

In the general case, where $\underset{\sim}{\tau}'$ and the four $\underset{\sim}{\beta}_i$ are linearly independent, the locus of the stress point determined by $\underset{\sim}{\tau}' + \Delta_R\underset{\sim}{\tau}$ (for all values of $\Delta t$), which is a path on a hypersphere of radius $|\underset{\sim}{\tau}'|$, enters all five dimensions of the stress deviator subspace. On the other hand, in the special case of plane strain it turns out that $\underset{\sim}{\beta}_3 = \omega^2\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_4 = -1/2\,\omega^2\underset{\sim}{\beta}_2$ so that $\underset{\sim}{\tau}' + \Delta_R\underset{\sim}{\tau}$ is confined to the subspace spanned by $\underset{\sim}{\tau}'$, $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ and a three-dimensional visualization is possible. In this case the locus of the stress point lies on the intersection of a plane and a sphere, i.e., on a circle, as indicated in Figure 8.

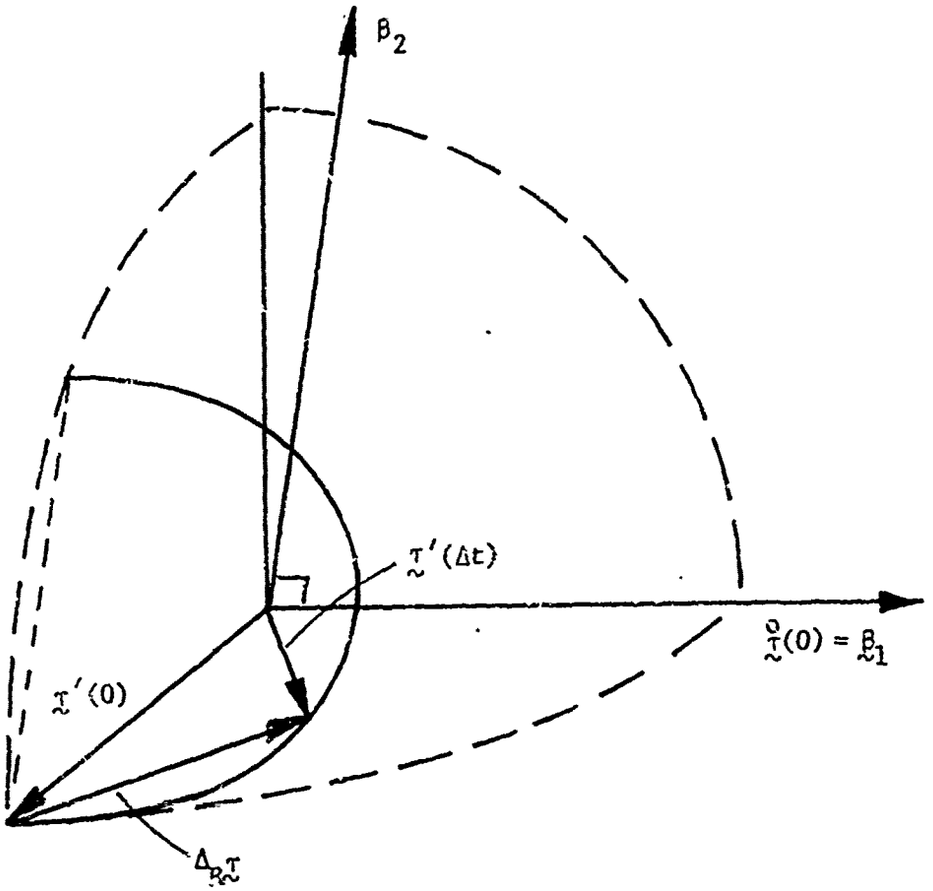



Figure 8 Rotational Stress Increment for Plane Strain with Constant Spin

Finally, it should be pointed out that the trigonometric functions in Eq.(46) can be replaced by rational approximations, for computational simplicity, without reintroducing the biased finite increment error. This is accomplished by making the substitution

$$\sin(\omega\Delta t) = \frac{2f}{1+f^2}, \quad 1-\cos(\omega\Delta t) = \frac{2f^2}{1+f^2} \tag{51}$$

where

$$f = \tan\frac{\omega\Delta t}{2} \tag{52}$$

and then introducing <u>any</u> rational approximation for f. The simplest approximation, $f \approx \omega\Delta t/2$, for example, causes an angle of $\omega\Delta t = 0.1\,\pi$ radians $= 18^\circ$ to be approximated by $17.85^\circ$. It can be shown that this particular approximation is identical to the one suggested by Hughes and Winget [6].

VIII. <u>SUMMARY</u>. A geometric characterization of stress has been used to describe certain finite-increment errors which arise when incremental expressions that are only first order in $\Delta t$ are employed. These errors are biased in one direction so that they accumulate

monotonically in successive incremental steps. An easily visualized characterization was described for a procedure developed by Rice and Tracey [4] to eliminate such an error, namely drift from the yield surface in the case of perfect plasticity. Their technique was generalized to include the case of strain hardening where the additional problem of drift from the strain hardening curve had to be eliminated also. The technique of Marcal and King [5] to allow midstep yielding was then incorporated into the procedure. Finally, a source of biased finite-increment error in treating the effect of material spin was characterized and a procedure for eliminating it was given. Although a geometric visualization was not possible in the general case, a three-dimensional illustration was given for the plane strain case.

## REFERENCES


1. R. Hill, "Some Basic Principles in the Mechanics of Solids Without a Natural Time," J. Mech. Phys. Solids, 7, 209, 1959.

2. H.D. Hibbitt, P.V. Marcal and J.R. Rice, "A Finite Element Formulation for Problems of Large Strain and Large Displacement," Int. J. Solids Struct., 6, 1069, 1970.

3. R.M. McMeeking and J.R. Rice, "Finite Element Formulations for Problems of Large Elastic-Plastic Deformation," Int. J. Solids Struct., 11, 601, 1975.

4. J.R. Rice and D.M. Tracey, "Computational Fracture Mechanics," in Numerical and Computer Methods in Structural Mechanics, ed. S.J. Fenves et al., Academic Press, 1973, pp. 585-623.

5. P.V. Marcal and I.P. King, "Elastic-Plastic Analysis of Two-Dimensional Stress Systems by the Finite Element Method," Int. J. Mech. Sci., 9, 143, 1967.

6. T.J.R. Hughes and J. Winget, "Finite Rotation Effects in Numerical Integration of Rate Constitutive Equations Arising in Large-deformation Analysis, Int. J. Num. Meth. Engng., 15, 1862, 1980.

# CONSTITUTIVE FEATURES OF SOLIDS AT SHOCK-WAVE LOADING RATES

Dennis E. Grady†
Thermomechanical and Physical Division
Sandia National Laboratories
Albuquerque, NM 87185

ABSTRACT. Solids subjected to high-velocity impact or explosive loading exhibit unusual transient and post-shock properties during the extremely brief period associated with the shock-wave risetime and release. These features can include a unique solid-state shock viscosity behavior, anomalous transient shock-hardening effects, heterogeneous shear effects during the shock risetime, and shock-induced solid state and metallurgical transformations. Improved methods in time-resolved instrumentation have been critical in the emerging understanding of these constitutive features. An increasing sophistication in physical and computational modeling is required to incorporate these effects in applied problems of dynamic solid mechanics.

I. SHOCK-WAVE CONCEPTS. An explosion, radiation deposition, or high-velocity impact can lead to a brief but intense pressure loading of a solid body through the propagation of a compression shock wave. Pressures achieved by conventional methods range from a few to several hundred GPa. Fundamental properties of shock waves are most readily appreciated through consideration of a normal one-dimensional shock such as might be produced by the planar impact of flat plates [1-3]. The shock transition in a single phase material occurs through the passage of a pressure wave with a risetime so brief that it is usually regarded as a discontinuity. The shock state is characterized by changes in the pressure, $P$, material velocity, $u_p$, specific volume, $V$, and specific internal energy, $E$. The kinematics of the wave are determined by a unique shock velocity, $U_s$. Fundamental laws governing conservation of mass, momentum, and energy lead to the Hugoniot conditions relating variables through the shock transition,

$$V/V_o = (U_s - u_p)/U_s, \tag{1}$$

† This work performed at Sandia National Laboratories supported by the U. S. Department of Energy under contract number DE-AC04-76-DP00789.

PREVIOUS PAGE IS BLANK

$$P - P_o = U_s u_p / V_o, \tag{2}$$

$$E - E_o = \frac{1}{2}(P + P_o)(V - V_o). \tag{3}$$

A thermodynamic description of the shock transition for a particular material is completed with a representation for the Hugoniot, which is an experimentally determined pressure-volume relation for the locus of shock pressure-shock volume states achieved through a sequence of increasing amplitude shock waves.

Under sufficiently high shock pressures in a solid, material response has usually been regarded as fluid in the sense that deviatoric stress and strain values are neglected in relation to hydrodynamic values [1-3]. In shock loading to pressures of about 20 GPa or lower in metals and perhaps twice that in refractory materials, this assumption is becoming recognized as a poor approximation in that effects of material strength can profoundly influence physical processes [4-6]. In addition, the fact that a shock wave has a brief but finite risetime cannot always be ignored in consideration of the mechanisms for the shock-induced processes [7,8].

## II. MICROSTRUCTURAL FEATURES IN THE SHOCK TRANSITION.

Microstructural features in the deforming solid are becoming increasingly recognized as important to the stress wave and flow process. Constitutive modeling efforts are tending toward better understanding and explicit treatment of the material microstructure in addressing elastic responses, strength, the yield process, flow and phase transformation. Two microstructural aspects appear important. First is the pre-existing microstructure including grain structure, porosity, and internal stress fields. The second is a transient deformation microstructure induced during the shock process. The former microstructure dominates wave propagation for stress amplitudes near the strength limit of the material but becomes of decreasing importance for shocks significantly stronger than the material strength. In contrast, an induced deformation microstructure achieves increasing importance in the extreme high-rate flow associated with a strong shock wave. These microstructural effects appear to play a significant role in both stress-wave propagation and the material processes occurring under stress-wave loading.

The pre-existing microstructure of a solid governs the details of nonlinear stress waves which load the material beyond the level of plastic yield or fracture. Microstructural features profoundly influence elastic wave propagation through frequency dependence and dispersion. In large amplitude wave propagation such elastic response leads to dispersed or ramped waves and governs the loading rate at which yield or other critical stress levels are achieved. Grain structure, through size and anisotropy, affects onset of plastic flow or brittle failure. Grain size influences the yield stress level and, under impact loading rates, can lead to rate-dependent

yield phenomena. Grain anisotropy can cause broadening of the yield process and leads to recoverable elastic response in multiple-wave shock loading experiments [9]. Porosity is accompanied by strong local stress fluctuations, providing sites for premature yielding, fracture and phase transformation. Large local deformation leads to transient local hot spots which can accelerate thermally-activated rate processes under shock loading. Collapse of pore volume at shock deformation rates involves microinertial effects, which can lead to rate-dependent response of the bulk material [10,11]. Dynamic fracture or spall is governed by the nucleation rate and growth of microstructural defects and flaws inherent in the material. Under impulse loading, time-dependent damage growth and rate-dependent spall strengths are observed [12,13]. Tensile fracture damage can occur through growth and coalescence of a population of microcracks in brittle solids or through void nucleation at atomic or microstructural defects followed by catastrophic cavitation in ductile solids [14,15]. The intensity of damage in terms of the density of microcracks or voids per unit volume appears to be controlled by the loading rate through a balance of the rates of nucleation and growth, and microstructural energy effects [16].

III. HETEROGENEOUS SHEAR. As the amplitude of the shock wave becomes increasingly larger than the strength of the material, existing microstructure properties become less important. Complex wave structure due to yielding or phase transformation tends to become overdriven and the shock wave degenerates toward an extremely rapid rising pressure pulse, perhaps a few nanoseconds in duration and a few tens of micrometers in extent. The width of the wave is governed by an effective viscous response of all of the dissipative processes occurring within the shock. Although gross microstructure such as porosity continues to influence the shock process through void volume crushup and intense shock heating, more subtle microstructure, such as grain size and orientation, and existing dislocation structure, defects, or impurities, seems to be of lesser importance.

Modeling the shock pressure as a brief, intense, homogeneous deformation from the initial to the shock state with the attendant homogeneous temperature and entropy rise appears to be too simplistic, however. This approach is incapable of explaining a rich body of shock-wave phenomena, including electric and magnetic effects, partial melting and thermally-activated solid state transformations, shock-induced chemical changes, and a host of metallurgical effects. A large body of post-shock metallurgical investigation exists in the literature (see, for instance, Mikkola [17], Murr [18], Grady *et al.* [8]). Although fraught with interpretational difficulties due to uncertainties in the shock unloading path and post-shock metallurgical changes, metallographic optical and transmission electron microscope studies seem to indicate that the shock deformation is an extremely heterogenous and turbulent process. A highly heterogeneous deformation process is further strengthened by recent advances in methods for measuring time-resolved stress waves in solids which reveal features in the shock deformation wave structure that are not easily reconciled with present theories of homogeneous shock deformation and high rate flow [19,20].

Recent theoretical efforts which attempt to account for the heterogeneous deformation process during shock loading reveal that significantly higher than average temperatures may persist briefly in deformation zones within the shock wave [7,21,22], and calculations indicate local temperature rises of a few hundred to a few thousand degrees Kelvin, depending on the magnitude of dissipation, the thermal conductivity, and the mass fraction of intensely deformed material. Such temperatures are sufficient to complete phase transformation by thermal activation within the shock risetime and localized melting within the shock can occur [23]. Dimensions of shear zones and temperature gradients expected to occur after passage of the shock wave suggest extremely rapid cooling rates (on the order of $10^{11}$ K/s), capable of quenching transformed material and submicrostructure within the high-pressure shear-banded material.

IV. SHOCK-WAVE VISCOSITY. An important aspect of the microstructure and material property changes which occur during passage of a shock wave is the time duration within which they must occur. Improved methods for measuring time-resolved profiles show that shock compression is not discontinuous but occurs within one to a few hundred ns over the stress range of about 1 to 10 GPa with the risetime decreasing rapidly with increasing stress amplitude. Irreversible deformation processes occur within the plastic portion of the wave. The plastic wave can change with time during the early evolution of the profile but achieves a steady shape after a short propagation distance due to a balance between the nonlinearity of the material compression behavior and rate-dependent dissipative processes which tend to disperse the wave.

In the hydrodynamic approximation of shock compression, a viscous relation has been found useful in characterizing material behavior. In more general elastic-plastic response, a more complicated viscous behavior is expected. It has been useful, however, to classify the dissipation over a steady-wave shock compression process by an effective viscosity. The viscosity coefficient is quantified experimentally as the ratio of the maximum viscous stress, which is proportional to the maximum difference between the Rayleigh line and the Hugoniot, and the strain rate from the maximum slope of the wave profile.

Recently, steady-wave profile data on a number of metals and nonmetals have been examined for risetime behavior [24]. A plot of steady-wave stress jump against strain rate for materials which include copper, aluminum, beryllium, iron, quartz, and magnesium oxide indicate unexpected consistencies. Strain rate increases as the fourth power of the stress jump for all material examined. This implies that the shock viscosity decreases as the square root of strain rate exhibiting a non-Newtonian behavior.

Swegle [24] has incorporated a square root viscous relation within a general Maxwell-like plasticity model and has readily reproduced the work hardening and steady-wave response observed experimentally. These calculations were performed without including artificial viscosity.

Factors governing shock viscosity and risetime of the plastic wave have not yet been determined. Viscous flow should be associated with the microscopic process of dislocation multiplication and motion, vacancy production, precipitate alteration, *etc.* There are tentative indications, however, that shock wave risetimes and viscosity are governed by more fundamental, mechanism independent, energy principles. If so, the microscopic shock process would be an effect rather than a cause, occurring in the most energetically favorable way, consistent with the time constraints. This idea is speculative, but it is clear that a better understanding is necessary here before a comprehensive theory of the shock deformation process will emerge.

V. ANOMALOUS SHOCK HARDENING. Recent measurements of plastic wave profiles in metals such as aluminum [4], copper [25], and beryllium [26], indicate strength properties at the Hugoniot state and viscous effects within the shock front which are unique in behavior and not readily explained [4,8,25,26]. Attempts to rationalize metallographic studies of shocked samples, which indicate strong heterogeneities in the microscale deformation with the very high rate of flow determined from the measured wave profiles, indicate that adiabatic shear deformation and thermal trapping may play an important role here also.

A unique shock wave experiment in which a second unloading or reshock wave is passed through the metal within microseconds after the initial deformation shock wave reveals further elastic-plastic response with significantly enhanced material strength. It is difficult to explain this effect without a thermal mechanism. Usual concepts of plastic flow suggest that the Hugoniot stress state should reside on the yield surface. The data show only a small residual state of shear stress relative to the strength at Hugoniot states greater than about 5 GPa. A transition from normal elastic-plastic response to the observed anomalous behavior appears to occur at about this stress level. A reduced state of shear stress on the Hugoniot and enhanced strength would be expected if the flow stress were small and if some rapid strength recovery mechanism were operating during the microsecond or less before the strength is tested with a release or reloading wave. Heterogeneous deformation and thermal trapping during the high-rate deformation process would be expected to cause reduced flow stress, and microscale thermal quenching after passage of the shock wave could provide the recovery process. Such an explanation has yet to be verified, although, model calculations indicate local shear temperatures consistent with the interpretation [7].

VI. SHOCK-INDUCED PHASE TRANSFORMATION. Processes of coherent phase transformation occurring under shock-wave compression provide a striking example of phenomena affected by microstructures. Coherent transformation processes include recrystallization and twinning, coherent precipitation, and displacive, martensitic or semi-reconstructive transformations in solids, although all of these have not yet been observed under shock loading. These processes are commonly reversible or exhibit little hysteresis in the transition between states which occur through the motion of a coherent interface. Further, nonhydrostatic stresses

markedly influence the conditions of phase coexistence both in the bulk and on the microscale where structural defects provide sites of second-phase nucleation. Paterson [27] has reviewed the theoretical development of nonhydrostatic thermodynamics applicable to coherent transformation prior to 1973, and several authors, including Kamb [28], Fletcher [29], and Robin [30], have noted the difficulty in establishing generalized thermodynamic potentials independent of the specific process. They note that coexistence conditions depend on the coherent interface orientation with respect to crystal axes and interface accommodated stress discontinuities in those components not required for stress equilibrium.

The I-II transformation in the naturally occurring mineral calcite is a coherent displacive transformation which has been investigated extensively under stress-wave loading [31,32]. The transformation initiates and proceeds within the elastic range of the material and, in polycrystalline specimens, is sensitive to both the shear stress state and the microstructure of the body. Phase change through the stress wave involves a transformation shape change as well as volume change, and the process leads to highly nonlinear stress-wave response, including wave splitting and rarefaction shocks. Accurate characterization of the stress wave response requires the inclusion of microstructural parameters to account for local stress heterogeneities which affect the range and shear sensitivity of the coherent transformation.

Perhaps the most significant shock-wave phenomena for which a plausible understanding has emerged within the context of a model of heterogeneous deformation and accompanying adiabatic shear and temperature trapping is the 4-to-6 fold coordination quartz-to-stishovite reconstructive phase transformation which occurs in crystalline $SiO_2$ [33,34]. Equivalent shock-induced phase transformation effects have been observed in a number of silicate minerals as well as other materials [35], however, the behavior of quartz is representative and has historic interest. This thermally-activated transformation requires minutes to hours to complete under static high pressure but is completed within a few ns under shock compression.

Shock-wave studies also note an anomalous metastable Hugoniot [35,36] response through the quartz-stishovite mixed phase region, which relates a fixed mass fraction of transformed material at a particularly Hugoniot pressure. Shock pressures in excess of 20 GPa over the initiation transformation pressure are necessary for complete transformation to the stishovite phase. In addition to the unusual Hugoniot behavior, release wave studies from shock pressures indicate fluid or fluid-like response at the Hugoniot state [34]. The reverse transformation during shock unloading reveals large hysteresis in the complete shock transformation cycle. Shock recovery experiments uncover complex deformation fabric through optical and TEM microscopy. Traces of stishovite and significant quantities of high-density glass are seen in the recovered samples [37,38].

These and other curious shock effects associated with the phase change in quartz as well as other materials are readily understood within the context of a

heterogeneous deformation and accompanying adiabatic shear and thermal trapping effect. Calculations show that temperatures associated with localized adiabatic shearing are adequate to accommodate reconstructive transformation through thermally-activated rate processes within the risetime of the shock wave, and thermal quenching after passage of the shock wave accounts for the mass fraction transformed on the metastable Hugoniot [21,23]. Also, local temperatures and thermal conduction rates are consistent with the persistence of laminar melt domains at the shock state and, therefore, the fluid-like release wave behavior. High-density glass in recovered shocked minerals, as well as minerals recovered from impact meteor craters, have been explained within the context of a thermal heterogeneous shock deformation process [39].

VII. SUMMARY. The present report reviews studies focused on understanding and modeling large-amplitude, nonlinear stress-wave propagation in solids. Recently developed time-resolved measuring techniques are providing constraining data in terms of the structure and evolution of stress and particle velocity profiles. The data indicates that microstructural effects are fundamental to the stress-wave propagation phenomena. Constitutive modeling of the dynamic deformation process, with explicit treatment of both the existing and evolving microstructure, is needed to calculate complex stress-wave propagation. More specifically, only through microstructural considerations will important shock-wave effects involving unique physical, chemical, and metallurgical processes be understood and exploited.

REFERENCES


1. W. Band and G. E. Duvall, Amer. J. Phys. 29, 780 (1961).
2. G. E. Duvall and G. R. Fowles, in *High Pressure Physics and Chemistry,* Volume 2, edited by R. S. Bradley, pp. 209-287, Academic Press (1963).
3. I. C. Skidmore, Appl. Mat. Res. 4, 131 (1965).
4. J. R. Asay and L. C. Chhabildas, in *Shock Waves and High-Strain Rate Phenomena in Metals,* edited by M. A. Meyers and L. E. Murr, pp. 417-432, Plenum Press (1981).
5. D. C. Wallace, Phys. Rev. B. 22, 1487 (1980).
6. G. R. Fowles, J. Geophys. Res. 72, 5729 (1967).
7. D. E. Grady and J. R. Asay, J. Appl. Phys. 53, 7350 (1982).
8. D. E. Grady, J. R. Asay, R. W. Rohde, and J. L. Wise, in *Material Behavior under High Stress and Ultrahigh Loading Rates,* edited by J. Mescall and V. Weiss, Plenum Press, 81-100 (1983).
9. J. R. Asay and J. Lipkin, J. Appl. Phys. 49, 4242 (1978).

10. B. M. Butcher, *Shock Waves and the Mechanical Properties of Solids,* edited by J. J. Burke and V. Weiss, Syracuse University Press, p. 277 (1971).

11. M. M. Carrol and A. C. Holt, J. Appl. Phys. 43, 1626 (1972).

12. D. A. Shockey, D. R. Curran, L. Seaman, J. T. Rosenberg, and C. F. Peterson, Int. J. Rock Mech. Min. Sci. 11, 303 (1974).

13. D. E. Grady and M. E. Kipp, Int. J. Rock Mech. Min. Sci. 17, 147 (1980).

14. L. Seaman, D. R. Curran, and D. A. Shockey, J. Appl. Phys. 47, 4814 (1976).

15. A. L. Stevens, L. Davison, and W. E. Warren, J. Appl. Phys. 43, 4922 (1972).

16. D. E. Grady, J. Appl. Phys. 53, 322 (1982).

17. D. E. Mikkola and R. N. Wright, *Shock Waves in Condensed Matter - 1981 (AIP),* edited by W. J. Nellis, L. Seaman, and R. A. Graham (1981).

18. L. E. Murr, in *Shock Waves and High-Strain-Rate Phenomena in Metals,* edited by M. A. Meyers and L. E. Murr, Plenum Press, p. 607 (1981).

19. J. R. Asay and L. C. Chhabildas, *Shock Waves and High-Strain-Rate Phenomena in Metals,* edited by M. A. Meyers and L. E. Murr, Plenum Press, p. 417 (1981).

20. L. C. Chhabildas, J. L. Wise, and J. R. Asay, *Shock Waves in Condensed Matter - 1981 (AIP),* edited by W. J. Nellis, L. Seaman, and R. A. Graham (1981).

21. D. E. Grady, J. Geophys. Res. 85, 913 (1980).

22. Y. Horie, Phys. Rev. B. 21, 5549 (1980).

23. D. E. Grady, *High Pressure Research: Applications in Geophysics,* edited by M. Manghnani and S. Akimoto, Academic Press, p. 389 (1976).

24. J. W. Swegle and D. E. Grady, in preparation.

25. L. C. Chhabildas and J. R. Asay, in *High Pressure in Research and Industry,* edited by C. M. Backman, T. Johannison, and L. Tegner, pp. 183-189 (1982).

26. L. C. Chhabildas, J. L. Wise, and J. R. Asay, *Shock Waves in Condensed Matter - 1981 (AIP),* edited by W. J. Nellis, L. Seaman, and R. A. Graham, 422-426 (1981).

27. M. S. Paterson, Rev. Geophys. 11, 355 (1973).

28. W. B. Kamb, J. Geophys. Res. 66, 259 (1962).

29. R. C. Fletcher, J. Geophys. Res. 78, 7661 (1973).

30. P. Y. F. Robin, Amer. Mineral 59, 1286 (1974).

31. D. E. Grady, R. E. Hollenbach, and K. W. Schuler, J. Geophys. Res. 83, 2839 (1978).

32. D. E. Grady, J. Geophys. Res. 84, 7549 (1979).

33. T. J. Ahrens and J. T. Rosenberg, in *Shock Metamorphism of Natural Minerals*, edited by B. French and N. Short (1968).

34. D. E. Grady, W. J. Murri, and G. R. Fowles, J. Geophys. Res. 79, 332 (1974).

35. G. E. Duvall and R. A. Graham, Rev. Modern Phys. 49, 523 (1977).

36. R. G. McQueen, S. P. Marsh and J. N. Fritz, J. Geophys. Res. 72, 4999 (1967).

37. P. S. DeCarli and D. J. Milton, Science 147, 144 (1965).

38. D. Stoffler, Fortschr. Mineral 49, 50 (1972).

39. J. Arndt, W. Hummel, and I. Gonzalez-Cabeza, Phys. Chem. Minerals 8 230 (1982).

PREVIOUS PAGE IS BLANK

# EXAMPLES AND SIGNIFICANCE OF CHANGE OF TYPE IN VISCOELASTICITY

Daniel D. Joseph[1], Michael Renardy[2] and Jean-Claude Saut[3]

ABSTRACT. The equations governing the flow of viscoelastic fluids are classified according to the symbol of their differential operators. Conditions for a change of type in steady two-dimensional flows are derived for a three-constant Oldroyd model. We find a change of type in the vorticity equation when a critical condition involving speeds and stresses is satisfied. We also sketch how change of type can be discussed for more general models.

I. INTRODUCTION. An important dimensionless quantity characterizing the flow of viscoelastic fluids is the Weissenberg or Deborah number. The exact definition of this quantity varies with the constitutive model and the flow under consideration, but, roughly speaking, it measures the ratio of elastic to viscous forces, or, alternatively, of a time characteristic of the fluid to a time characteristic of the flow.

Numerical calculations of steady flows in viscoelastic fluids typically fail if this Weissenberg number is high or even moderate. It is not well understood why and the reason is probably not always the same. Experimentally, qualitative changes in the flow behaviour are often observed at high Weissenberg numbers.

In a recent paper [6], we advance the idea that some of these effects are related to a change of type in the governing equations. We discuss change of type in detail for a three-constant Oldroyd model, but also sketch an analysis for more general models. This study extends earlier work of Rutkevich [10], Ultman and Denn [11], and Luskin [7]. When discussing change of type we have to distinguish between two cases:

1. There is a change of type for the equations governing steady flow as well as for the time-dependent equations. This leads to Hadamard instability and ill-posedness of the initial value problem. This kind of situation is familiar from the theory of phase transitions.
2. There is a change of type in the steady equations, but not in the unsteady equations. This happe when the speed of the fluid exceeds a wave propagation speed as in a sonic transition in gas dynamics. There is no Hadamard instability associated with this.

(1) Dept. of Aerospace Engineering, University of Minnesota, 110 Union St. S.E., Minneapolis, MN 55455.
(2) Dept. of Mathematics and Mathematics Research Center, University of Wisconsin, 610 Walnut St., Madison, WI 53705.
(3) Dept. of Mathematics, Universite de Paris-Sud, F-91405 Orsay, France.

(1) The United States Army under Contract No. DAAG-29-82-0051 and the the Fluid Mechanics Branch of the National Science Foundation.
(2) The United States Army under Contract No. DAAG29-80-C-0041. This material is based upon work supported in part by the National Science Foundation under Grant Nos. MCS-8210950 and MCS-8215064.

Several papers in the literature have attempted to link experimental observations to change of type. Hunter and Slemrod [4], and, on the basis of a different model, Becker and his coworkers [2] have tried to explain melt fracture by a change of type leading to Hadamard instability (see [1] for a detailed and critical discussion of Becker's theory). Ultman and Denn [11] refer to an observation of James [5] on heat transfer in flows past a cylinder. It appears that there is a discontinuity in slope when heat transfer coefficient is plotted against the speed of the fluid. Ultman and Denn suggest that a sonic transition occurs at the speed where the slope is discontinuous. Recently, Yoo, Ahrens and Joseph [12] have discussed experiments by Metzner, Uebler and Fong [8] on tube entry flows from a conical region. At high Weissenberg number, the flow partitions into an interior cone, where the streamlines are approximately straight towards the sink, and an outer region of recirculation. The boundary between these regions seems to be rather sharp, and there is an apparent discontinuity in the vorticity (see Fig. 11 in [8]). Yoo, Ahrens and Joseph relate this observation to our analysis of Oldroyd models. All these studies are rather tentative, and at present not enough is known either experimentally or theoretically to make strong claims.

In section 2, we give basic definitions relating to change of type in first order systems of partial differential equations. These are applied in section 3 to the study of two-dimensional steady flows for a class of three-constant Oldroyd models [9]. A criterion for criticality is given, and the vorticity is identified as the variable associated with the change of type. In section 4 we demonstrate how similar ideas can be extended to general fluids with fading memory. However, it is in general not possible to decouple the characteristic equation and isolate a vorticity equation as in the case of the three-constant Oldroyd model.

2. BASIC DEFINITIONS. The equations for viscoelastic flow discussed below have the form of quasilinear first order systems. In this section, we give some definitions relating to characteristics and change of type in such systems (see e.g. [3]). We are concerned with equations of the form

$$\sum_{\ell=0}^{n} \underline{\underline{A}}_\ell(\underline{x},\underline{u}) \frac{\partial \underline{u}}{\partial x_\ell} = \underline{f}(\underline{x},\underline{u}) \tag{2.1}$$

where $\underline{u}$ is a k-vector and the $\underline{\underline{A}}_\ell$ are $k \times k$-matrices. The term "quasi-linear" means that $\underline{\underline{A}}_\ell$ and $\underline{f}$ may depend on $\underline{x}$ and $\underline{u}$, but not on derivatives of $\underline{u}$, i.e. the highest order derivatives occur in the equations in a linear way. For every choice of $\underline{x}$ and $\underline{u}$, we define characteristic surfaces as follows: A surface given by an equation $\phi(t,x_1,\ldots,x_n) = 0$ is characteristic if

$$\det\Big(\sum_{\ell=0}^{n} \underline{\underline{A}}_\ell \frac{\partial \phi}{\partial x_\ell}\Big) = 0 \quad . \tag{2.2}$$

The system is called elliptic if there are no real characteristic surfaces. Hyperbolic systems are characterized as the opposite extreme, namely, there is a maximal number of real characteristics. More precisely, a system is called hyperbolic, if one of the matrices $\underline{\underline{A}} = \underline{\underline{A}}_\mu$ is non-singular and, for every choice of real parameters $(\lambda_\ell,\ \ell = 0,1,\ldots,n;\ \ell \neq \mu)$, the roots $\alpha$ of the eigenvalue problem

(2.3) $$\det\Big(\alpha \underline{\underline{A}} - \sum_{\substack{\ell=0 \\ \ell\neq\mu}}^{n} \lambda_\ell \underline{\underline{A}}_\ell\Big) = 0$$

are real and semisimple. The equations of viscoelastiticity are neither elliptic nor hyperbolic. However, we will encounter situations where the number of real characteristic surfaces changes. In this case, we say there is a change of type.

The phenomenon of Hadamard instability is closely related to this. It is evident that, if (2.3) has complex roots, then Im $(\alpha)$ can be made arbitrarily large by making the $\lambda_\ell$ large. If we choose $\mu = 0$ and interpret the first coordinate $x_0 = t$ as time, then this means that the linearization of (2.1) will have rapidly growing solutions when the initial data are very oscillatory. This kind of catastrophic instability is referred to as "Hadamard instability".

3. CHANGE OF TYPE IN TWO-DIMENSIONAL STEADY FLOWS OF THREE-CONSTANT OLDROYD FLUID. We consider differential models with a constitutive law of the form

(3.1) $$\lambda \frac{D\underline{\underline{\tau}}}{Dt} + \underline{\underline{\tau}} = 2\eta \underline{\underline{D}}$$

where $D/Dt$ denotes a frame invariant time derivative expressed as

(3.2) $$\frac{D\underline{\underline{\tau}}}{Dt} = \frac{\partial \underline{\underline{\tau}}}{\partial t} + (\underline{u}\cdot\nabla)\underline{\underline{\tau}} + \underline{\underline{\tau}}\,\underline{\underline{\Omega}} - \underline{\underline{\Omega}}\,\underline{\underline{\tau}} - a(\underline{\underline{\tau}}\,\underline{\underline{D}} + \underline{\underline{D}}\,\underline{\underline{\tau}}) \ .$$

Here we have split the velocity gradient $\nabla \underline{u}$ with components $(\nabla \underline{u})_{ij} = \partial u_i/\partial x_j$ into its symmetric part $\underline{\underline{D}} = 1/2\,(\nabla \underline{u} + (\nabla \underline{u})^T)$ and its antisymmetric part $\underline{\underline{\Omega}} = 1/2\,(\nabla \underline{u} - (\nabla \underline{u})^T)$. The special cases $a = 1$, $a = -1$ and $a = 0$ are known as the upper convected, lower convected and corotational Maxwell model, respectively.

In steady two-dimensional flows, we denote velocity components by $u$ and $v$, and the extra stress tensor is written in the form

(3.3) $$\underline{\underline{\tau}} = \begin{pmatrix} \sigma & \tau \\ \tau & \gamma \end{pmatrix} \ .$$

The constitutive law (3.1), together with the equation of motion and the incompressibility condition leads to the following quasilinear first order system

$$u\sigma_x + v\sigma_y + \tau(v_x - u_y) - a[2\sigma u_x + \tau(u_y + v_x)] - 2\frac{\eta}{\lambda} u_x = -\frac{\sigma}{\lambda}$$

$$u\tau_x + v\tau_y + \frac{1}{2}(\sigma-\gamma)(u_y - v_x) - \frac{a}{2}(\sigma+\gamma)(u_y + v_x) - \frac{\eta}{\lambda}(u_y + v_x) = -\frac{\tau}{\lambda}$$

$$u\gamma_x + v\gamma_y + \tau(u_y - v_x) - a[2\gamma v_y + \tau(u_y + v_x)] - 2\frac{\eta}{\lambda} v_y = -\frac{\gamma}{\lambda} \qquad (3.4)$$

$$\rho(uv_x + vu_y) + p_x - \sigma_x - \tau_y = 0$$

$$\rho(uv_x + vv_y) + p_y - \tau_x - \gamma_y = 0$$

$$u_x + v_y = 0 \ .$$

We can apply the definitions of section 2 to this system. This leads to the following equation for the slope $\alpha = dy/dx$ of characteristic lines.

$$(3.5)\quad (1+\alpha^2)(-\alpha u+v)^2\{\rho(-\alpha u+v)^2 + (\tfrac{\gamma-\sigma}{2})(\alpha^2-1) + 2\tau\alpha - (\alpha^2+1)(\tfrac{\eta}{\lambda} + a(\tfrac{\gamma+\sigma}{2}))\} = 0 \ .$$

We see that the stream lines are double characteristics, and that at two characteristic values are always complex. The interesting factor is the last one. The roots of this factor change from complex to real when the sign of

$$(3.6)\quad [\rho u^2 + \tfrac{\gamma}{2}(1-a) - \tfrac{\sigma}{2}(1+a) - \tfrac{\eta}{\gamma}][(1+a)\tfrac{\gamma}{2} + (a-1)\tfrac{\sigma}{2} + \tfrac{\eta}{\lambda} - \rho v^2] + (\rho uv-\tau)^2$$

changes from negative to positive.

The reason why (3.5) decouples into quadratic factors becomes evident in a streamfunction-vorticity formulation. When the equations are rewritten in this way, one can see that the roots $\alpha = \pm i$ are associated with the equation expressing the vorticity as the Laplacian of the stream function. The third factor is associated with an equation which involves a linear combination of second derivatives of the vorticity and only contains lower order terms otherwise. It is therefore the vorticity which is associated with the change of type. It is interesting in this context that the experiments of Metzner, Uebler and Fong [8] can be interpreted as suggesting a discontinuity in the vorticity.

One can also derive a time dependent vorticity equation, which leads to a criterion for Hadamard instability. Hadamard instability occurs if one of the following conditions is violated

$$(3.7)\quad \lambda^2\tau^2 - [\eta - \lambda(\tfrac{\gamma}{2}(1-a) - \tfrac{\sigma}{2}(1+a))][\eta - \lambda(\tfrac{\sigma}{2}(1-a) - \tfrac{\gamma}{2}(1+a))] < 0$$

$$(3.8)\quad \lambda[\tfrac{\gamma}{2}(1-a) - \tfrac{\sigma}{2}(1+a)] - \eta < 0 \ .$$

Note that (3.7) agrees exactly with (3.6) for zero speeds. Changes of type in steady flow which do not involve Hadamrd instability must therefore require a non-zero speed of the fluid. In fact, the criterion is that the speed of the fluid is faster than a viscoelastic wave speed. In particular, if the stresses vanish, a change of type occurs when the fluid speed exceeds the wave speed of linear viscoelasticity. Since this requires a finite (but not large) Reynolds number, such changes of type are more likely to be found in dilute polymer solutions rather than in melts.

In discussing the criteria (3.6) or (3.7), (3.8), it must be kept in mind that the values of the extra stresses are not arbitrary. The constitutive law (3.1) can be regarded as an evolution problem for the stress with given deformation. However, in the discussion of materials with fading memory, we are not interested in arbitrary solutions of this evolution problem, but only in those that behave reasonably as time tends to $-\infty$. This imposes restrictions on the values of the extra stresses, which can be shown to preclude Hadamard instability if $a = \pm 1$.

For a discussion of particular flow geometries we refer to [6] and [12].

4. CHANGE OF TYPE IN FLUIDS WITH FADING MEMORY. The extra stress $\underline{\underline{\tau}}$ in a simple fluid is given by an isotropic functional of the history of the relative Cauchy strain $\underline{\underline{G}}(s) = \underline{\underline{F}}_t^T(t-s)\underline{\underline{F}}_t(t-s) - \underline{\underline{1}}$, i.e.

$$\underline{\underline{\tau}} = \underline{\underline{F}}[\underline{\underline{G}}(s)]_{s=0}^{\infty} \quad . \tag{4.1}$$

By taking the material derivative of (4.1), we obtain

$$\frac{d\underline{\underline{\tau}}}{dt} = \underline{\underline{F}}_1[\underline{\underline{G}} \mid \frac{d\underline{\underline{G}}}{dt}] \quad . \tag{4.2}$$

Following Coleman and Noll, we assume that the Fréchet derivative $\underline{\underline{F}}_1$ of the functional $\underline{\underline{F}}$ can be represented in the form

$$\underline{\underline{F}}_1[\underline{\underline{G}} \mid \frac{d\underline{\underline{G}}}{dt}] = \int_0^{\infty} \underline{\underline{K}}(s,\underline{\underline{G}}) \frac{d\underline{\underline{G}}(s)}{dt}\, ds \quad . \tag{4.3}$$

Here $\underline{\underline{K}}(s,\underline{\underline{G}})$ is a fourth order tensor depending on $s$ and the values $\{\underline{\underline{G}}(\sigma), 0 \leq \sigma < \infty\}$. For the following, we assume that $K$ and its first derivative with respect to $s$ are integrable.

The material derivative of $\underline{\underline{G}}$ is given by

$$\frac{d\underline{\underline{G}}}{dt} = -\underline{\underline{L}}^T\underline{\underline{G}} - \underline{\underline{G}}\,\underline{\underline{L}} - \frac{d\underline{\underline{G}}}{ds} \tag{4.4}$$

where $\underline{\underline{L}} = \nabla \underline{u}$ is the present value of the velocity gradient. Hence we find

$$\begin{aligned} \int_0^{\infty} K_{ijk\ell}(s,\underline{\underline{G}}) \frac{dG_{k\ell}}{dt}(s)ds &= -\int_0^{\infty}(K_{ijk\ell} + K_{ij\ell k})G_{p\ell}(s)ds \cdot L_{pk}(t) \\ &\quad - \int_0^{\infty} K_{ijk\ell}(s,\underline{\underline{G}}) \frac{d\underline{\underline{G}}(s)}{ds}\, ds \quad . \end{aligned} \tag{4.5}$$

The last term can be integrated by parts and treated as a perturbation of lower differential order. With

$$M_{ijkp} = -\int_0^{\infty}(K_{ijk\ell} + K_{ij\ell k})G_{p\ell}(s)ds \tag{4.6}$$

we can therefore write the equations of viscoelastic fluid motion in the form

$$\begin{aligned} \frac{d\tau_{ij}}{dt} &= M_{ijkp}\frac{\partial u_p}{\partial x_k} + N_{ij} \\ \rho\frac{du_i}{dt} &= -\frac{\partial p}{\partial x_i} + \frac{\partial \tau_{ij}}{\partial x_j} + f_i \\ \frac{\partial u_i}{\partial x_i} &= 0 \quad . \end{aligned} \tag{4.7}$$

This again has the form of a quasilinear first order system, and the definitions of characteristics and change of type apply. In general, however, it is not possible to decouple this system as in section 3 and isolate a vorticity equation. In two-dimensional steady flow, we would still find the stream lines as double characteristics, but the remaining characteristic values would be determined by a fourth order equation, which cannot easily be

factored. In [6], we identify a class of constitutive models which has certain structural similarities with the Oldroyd models above and permits the derivation of a vorticity equation.

REFERENCES


[1] M. Ahrens, D. D. Joseph, M. Renardy and Y. Renardy, Remarks on the stabilities of viscometric flow, to appear, Rheol. Acta.

[2] U. Akbay, E. Becker, S. Krozer and S. Sponagel, Instability of slow viscometric flow, Mech. Res. Comm. 7 (1980), 199-204.

[3] I. M. Gelfand, Some problems in the theory of quasilinear equations, Amer. Math. Soc. Translations 29 (1963), 295-380.

[4] J. K. Hunter and M. Slemrod, Unstable viscoelastic fluid flow exhibiting hysteritic phase changes, Phys. Fluids 26 (1983), 2345-2351.

[5] D. F. James, Laminar flow of dilute polymer solutions around circular cylinders, Ph.D. Thesis, California Inst. of Technology, Pasadena 1967.

[6] D. D. Joseph, M. Renardy and J. C. Saut, Hperbolicity and change of type in the flow of viscoelastic fluids, to appear, Arch. Rat. Mech. Anal.

[7] M. Luskin, On the classification of some model equations for viscoelasticity, to appear, Rheol. Acta.

[8] A. B. Metzner, E. A. Uebler and C. F. Chang Man Fong, Converging flows of viscoelastic materials, AIChE J. 15 (1969), 750-758.

[9] J. G. Oldroyd, Non-Newtonian effects in steady motion of some idealized elasticoviscous liquids, Proc. Roy. Soc. London A 245 (1958), 278-297.

[10] I. M. Rutkevich, The propagation of small perturbations in a viscoelastic fluid, J. Appl. Math. Mech. 34 (1970), 35-50.

[11] J. S. Ultman and M. M. Denn, Anomalous heat transfer and a wave phenomenon in dilute polymer solutions, Trans. Soc. Rheology 14 (1970), 307-317.

[12] J. Y. Yoo, M. Ahrens and D. D. Joseph, Hyperbolicity and change of type in sink flow, to appear.

# A MORE ACCURATE SOLUTION TO THE ELASTIC-PLASTIC PROBLEM OF PRESSURIZED THICK-WALLED CYLINDERS

Peter C. T. Chen
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. A new method has been developed for solving the partially plastic problems of thick-walled cylinders made of strain-hardening or ideally-plastic materials subjected to any combination of internal pressure, external pressure, and end loads. The incremental strains are chosen as the basic unknowns in the finite-difference formulation. The incremental sizes of the applied loading are determined automatically and no iteration is needed. Complete solutions for the stresses, strains, and displacement have been obtained and all numerical results are very accurate. This approach is also efficient and simple, yet quite general, when compared with many solutions in the literature.

I. INTRODUCTION. The partially plastic problem of pressurized thick-walled cylinder is of practical importance to pressure vessels and the autofrettage process of gun barrels. Many solutions for this problem have been reported [1-7]. For thick tubes under very high pressure operation, the elastic-plastic material model should be represented by the von Mises' yield criterion, Prandtl-Reuss' incremental stress-strain laws, the strain-hardening, and compressibility [8]. However, a closed-form solution exists only in the plane-strain case neglecting strain-hardening and compressibility.

For the generalized plane-strain problems considered here, numerical solutions were reported by the finite-difference method [4,7] and finite-element method [5]. The incremental displacements were used as the basic unknowns and a displacement function was assumed in the finite-element method [5]. The incremental stresses and strains were used in [4] as the basic unknowns, but only the incremental strains were used in [7]. The spatial discretization used in [4,7] was based on the forward difference scheme and a fixed sequence of incremental loading was used.

In the present paper, a new method is developed and more accurate numerical results are obtained. The incremental strains are chosen as the basic unknowns in the finite-difference formulation. Both strain-hardening and ideally-plastic materials can be considered. The spatial discretization are based on the central difference scheme and the incremental sizes of the applied loading are determined automatically in the program. The incremental results are calculated directly and no iteration is needed. The convergence of the approach will be discussed and more accurate results will be reported.

II. BASIC EQUATIONS. Assuming small strain and no body forces in the axisymmetric state of generalized plane-strain, the radial and tangential stresses, $\sigma_r$ and $\sigma_\theta$, must satisfy the equilibrium equation,

$$r(\partial\sigma_r/\partial r) = \sigma_\theta - \sigma_r \ ; \tag{1}$$

and the corresponding strains, $\varepsilon_r$ and $\varepsilon_\theta$, are given in terms of the radial displacement, u, by

$$\varepsilon_r = \partial u/\partial r \quad , \quad \varepsilon_\theta = u/r \tag{2}$$

It follows that the strains must satisfy the equation of compatibility

$$r(\partial\varepsilon_\theta/\partial r) = \varepsilon_r - \varepsilon_\theta \tag{3}$$

If the material is assumed to be elastic-plastic, obeying the Mises' yield criterion, the Prandtl-Reuss flow theory, and the isotropic hardening law, the stress-strain relations are [1]:

$$d\varepsilon_i' = d\sigma_i'/2G + (3/2)\sigma_i' d\sigma/(\sigma H) \tag{4}$$

$$d\sigma > 0 \quad \text{for } i = r,\theta,z$$

$$d\varepsilon_m = E^{-1}(1-2\nu)d\sigma_m \tag{5}$$

where E, $\nu$ are Young's modulus, Poisson's ratio, respectively,

$$2G = E/(1+\nu),$$

$$\varepsilon_m = (\varepsilon_r+\varepsilon_\theta+\varepsilon_z)/3 \quad , \quad \varepsilon_i' = \varepsilon_i - \varepsilon_m$$

$$\sigma_m = (\sigma_r+\sigma_\theta+\sigma_z)/3 \quad , \quad \sigma_i' = \sigma_i - \sigma_m$$

$$\sigma = (1/\sqrt{2})[(\sigma_r-\sigma_\theta)^2 + (\sigma_\theta-\sigma_z)^2 + (\sigma_z-\sigma_r)^2]^{1/2} > \sigma_o \tag{6}$$

and $\sigma_o$ is the yield stress in simple tension or compression. For a strain-hardening material, H' is the slope of the effective stress/plastic strain curve

$$\sigma = H(\int d\varepsilon^p) \tag{7}$$

For an ideally-plastic material (H' = 0), the quantity $(3/2)d\sigma/(\sigma H')$ is replaced by $d\lambda$, a positive factor of proportionality. When $\sigma < \sigma_o$ or $d\sigma > 0$, the state of stress is elastic and the second term in Eq. (4) disappears. Following Yamada et al [9], Eqs. (4) and (5) can be rewritten in an incremental form

$$d\sigma_i = d_{ij}d\varepsilon_j \quad \text{for } i,j = r,\theta,z \tag{8}$$

and

$$d_{ij} = [\nu/(1-2\nu) + \delta_{ij} - \sigma_i'\sigma_j'/S]$$

where

$$S = \frac{2}{3}\left(1 + \frac{1}{3}H'/G\right)\sigma^2 \quad , \quad H'/E = \omega/(1-\omega) \tag{9}$$

$\omega E$ is the slope of the effective stress-strain curve, and $\delta_{ij}$ is the Kronecker delta.

This form was used in the finite-element formulation for solving elastic-plastic thick-walled tube problems [5]. In the following section, the incremental stress-strain matrix will be used in the finite-difference formulation.

III. FINITE-DIFFERENCE FORMULATION. Consider a thick-walled cylinder of inner radius a and external radius b. The tube is subjected to inner pressure p, external pressure q, and end force f. The elastic solution for this problem is well-known and the pressure p*, q*, or f* required to cause initial yielding can be determined by using the Mises' yield criterion. For loading beyond the elastic limit, an incremental approach of the finite-difference formulation is used. The cross-section of the tube is divided into n rings with $r_1=a, r_2, \ldots, r_k=\rho, \ldots, r_{n+1}=b$ where $\rho$ is the radius of the elastic-plastic interface. At the beginning of each increment of loading, the distribution of displacements, strains, and stresses are assumed to be known and we want to determine $\Delta u$, $\Delta\varepsilon_r$, $\Delta\varepsilon_\theta$, $\Delta\varepsilon_z$, $\Delta\sigma_r$, $\Delta\sigma_\theta$, $\Delta\sigma_z$ at all grid points. Since the incremental stresses are related to the incremental strains by the incremental form (Eq. (8)) and $\Delta u = r\Delta\varepsilon_\theta$, there exists only three unknowns at each station that have to be determined for each increment of loading. Accounting for the fact that the axial strain $\varepsilon_z$ is independent of r, the unknown variables in the present formulation are $(\Delta\varepsilon_\theta)_i$, $(\Delta\varepsilon_r)_i$, for $i = 1,2,\ldots,n,n+1$, and $\Delta\varepsilon_z$.

The equation of equilibrium (1) and the equation of compatibility (3) are valid for both the elastic and the plastic regions of a thick-walled tube. A first-order-correct finite-difference analog of these two equations at $i = 1,\ldots,n$ has been given in [4,7]. Other finite-difference forms can be written. In the present paper, the difference equations given below are second-order-correct. The equation of compatibility (3) and equation of equilibrium (1) are replaced, respectively, by

$$c_{1i}(\Delta\varepsilon_\theta)_i + c_{2i}(\Delta\varepsilon_r)_i + c_{3i}(\Delta\varepsilon_\theta)_{i+1} + c_{4i}(\Delta\varepsilon_r)_{i+1} = c_{5i} \tag{10}$$

and

$$c_{1i}(\Delta\sigma_r)_i + c_{2i}(\Delta\sigma_\theta)_i + c_{4i}(\Delta\sigma_r)_{i+1} + c_{4i}(\Delta\sigma_\theta)_{i+1} = c_{6i} \tag{11}$$

where

$$c_{1i} = -\frac{3}{2} + \frac{1}{2}\gamma_i \ , \ c_{2i} = \frac{1}{2} - \frac{1}{2}\gamma_i$$

$$c_{3i} = \frac{3}{2} - \frac{1}{2}\gamma_i \ , \ c_{4i} = -\frac{1}{2} + \frac{1}{2}\gamma_i$$

$$c_{5i} = -c_{1i}(\varepsilon_\theta)_i - c_{2i}(\varepsilon_r)_i - c_{3i}(\varepsilon_\theta)_{i+1} - c_{4i}(\varepsilon_r)_{i+1}$$

$$c_{6i} = - c_{1i}(\sigma_r)_i - c_{2i}(\sigma_\theta)_i - c_{8i}(\sigma_r)_{i+1} - c_{4i}(\sigma_\theta)_{i+1}$$

and

$$\gamma_i = r_{i+1}/r_i \tag{12}$$

With the aid of the incremental stress-strain relations (8), Eq. (11) can be written as

$$c_{7i}(\Delta\varepsilon_\theta)_i + c_{8i}(\Delta\varepsilon_r)_i + c_{9i}(\Delta\varepsilon_\theta)_{i+1} + c_{10i}(\Delta\varepsilon_r)_{i+1} + c_{11i}(\Delta\varepsilon_z) = c_{6i} \tag{13}$$

where

$$c_{7i} = c_{1i}(d_{12})_i + c_{2i}(d_{22})_i \quad , \quad c_{8i} = c_{1i}(d_{11})_i + c_{2i}(d_{21})_i$$

$$c_{9i} = c_{3i}(d_{12})_{i+1} + c_{4i}(d_{22})_{i+1} \quad , \quad c_{10i} = c_{3i}(d_{11})_{i+1} + c_{4i}(d_{22})_{i+1}$$

$$c_{11i} = c_{1i}(d_{13})_i + c_{2i}(d_{23})_i + c_{3i}(d_{13})_{i+1} + c_{4i}(d_{23})_{i+1} \tag{14}$$

The boundary conditions for the problem are

$$\Delta\sigma_r(a,t) = -\Delta p \quad , \quad \Delta\sigma_r(b,t) = -\Delta q$$

$$\pi \sum_{i=1}^{n} [r_i(\Delta\sigma_z)_i + r_{i+1}(\Delta\sigma_z)_{i+1}](r_{i+1}-r_i) = \mu\pi a^2\Delta p + \Delta f \tag{15}$$

where $\mu$ is 0 for open-end tubes, and 1 for closed-end tubes. Using the incremental relations (8), we rewrite Eq. (15) as

$$(d_{12})_1(\Delta\varepsilon_\theta)_1 + (d_{11})_1 (\Delta\varepsilon_r)_1 + (d_{13})_1\Delta\varepsilon_z = -\Delta p \tag{16}$$

$$(d_{12})_{n+1}(\Delta\varepsilon_\theta)_{n+1} + (d_{11})_{n+1}(\Delta\varepsilon_r)_{n+1} + (d_{13})_{n+1}\Delta\varepsilon_z = -\Delta q \tag{17}$$

and

$$\sum_{i=1}^{n} [c_{12i}(\Delta\varepsilon_\theta)_i + c_{13i}(\Delta\varepsilon_r)_i + c_{14i}(\Delta\varepsilon_\theta)_{i+1} + c_{15i}(\Delta\varepsilon_r)_{i+1} + c_{16i}(\Delta\varepsilon_z)] = \mu a^2\Delta p + \Delta f/\pi \tag{18}$$

where

$$c_{12i} = (r_{i+1}-r_i)r_i(d_{32})_i \quad , \quad c_{13i} = (r_{i+1}-r_i)r_i(d_{31})_i$$

$$c_{14i} = (r_{i+1}-r_i)r_{i+1}(d_{32})_{i+1} \quad , \quad c_{15i} = (r_{i+1}-r_i)r_{i+1}(d_{31})_{i+1}$$

$$c_{16i} = (r_{i+1}-r_i)[r_i(d_{33})_i + r_{i+1}(d_{33})_{i+1}] \tag{19}$$

Now we can form a system of $2n+3$ equations for solving $2n+3$ unknowns, $(\Delta\varepsilon_\theta)_i$, $(\Delta\varepsilon_r)_i$, at $i = 1,2,\ldots,n,n+1$ and $\Delta\varepsilon_z$. Equations (16), (17), and (18) are taken as the first and last two equations, respectively, and the other $2n$ equations are set up at $i = 1,2,\ldots,n$ using Eqs. (10) and (13). The final system is an unsymmetric matrix of arrow type with the nonzero terms appearing in the last row and column and others clustering about the main diagonal, two

below and two above. In the computer program which was developed, the dimensionless quantities $r/a$, $E\varepsilon_r/\sigma_o$, $E\varepsilon_\theta/\sigma_o$, $E\varepsilon_z/\sigma_o$, $\sigma_r/\sigma_o$, $\sigma_\theta/\sigma_o$, $\sigma_z/\sigma_o$, $p/\sigma_o$, $q/\sigma_o$, $f/(\pi a^2\sigma_o)$ were used in the formulation and the Gaussin elimination method was used to solve these equations. All calculations were carried out on IBM 4341 with double precision to reduce round-off errors.

IV. OPTIMAL INCREMENTAL LOADING. Given any combination of incremental-loading ($\Delta p$, $\Delta q$, or $\Delta f$), we can now determine all incremental results (displacements, strains, and stresses) directly. No iteration is needed, while in [4], many iterations in each step were required because a value for $\Delta\varepsilon_z$ was assumed. The sizes of incremental-loading should be chosen properly in order to obtain accurate results at a reasonable cost. When the total applied pressure p is given, it is natural to divide the loading path in m equal fixed increments such as $\Delta p = (p-p^*)/m$. Larger values of m give more accurate results. A sequence of decreasing load-increments is a better choice than that of equal increments. In order to increase the efficiency without effecting the accuracy, an adaptive algorithm has been implemented on the basis of a scaled incremental-loading approach [5].

In each step, a dummy load-increment such as $\Delta p$ is applied and the incremental results $\Delta\sigma_i$ for $i = r,\theta,z$ at all grids are determined. For all grid points at which $\sigma = ||\sigma_i|| < \sigma$, we compute the scaler $\alpha$'s by the formula

$$\alpha = \frac{1}{2}\{\Gamma + [\Gamma^2 + 4||\Delta\sigma_i||^2(\sigma_o^2 - ||\sigma_i||^2)]^{1/2}\}||\Delta\sigma_i||^2 \tag{20}$$

where

$$\Gamma = ||\sigma_i||^2 + ||\Delta\sigma_i||^2 - ||\sigma_i||^2 \tag{21}$$

and $||\sigma_i||$, $||\Delta\sigma_i||$, $||\sigma_i + \Delta\sigma_i||$ are computed by

$$||\sigma_i||^2 = -[(\sigma_r-\sigma_\theta)^2 + (\sigma_\theta-\sigma_z)^2 + (\sigma_z-\sigma_r)^2] \tag{22}$$

Let $\lambda$ be the minimum of the $\alpha$'s. Then $\lambda$ is the load-increment factor just sufficient to yield one additional point. A sequence of $\lambda(j)$ can be determined for all steps $j = 1,2,\ldots m$ and the updated results are

$$p(j) = p(j-1) + \lambda(j)\Delta p(j)$$

$$\sigma_i(j) = \sigma_i(j-1) + \lambda(j)\Delta\sigma_i(j) \quad , \quad \text{etc.} \tag{23}$$

This sequence of incremental loading is optimal for the present problem because all the coefficients c's in Eqs. (12), (14), and (19) are functions of the previous stresses and strains.

V. CONVERGENCE STUDY. In order to demonstrate the accuracy of the approach, four convergence studies are made. Consider a thick-walled tube of wall ratio $b/a = 2$ and subjected to internal pressure only. The cross-section of the tube is divided into n rings of equal thicknesses, i.e., $h = (b-a)/n$. The first problem is a closed-end tube loaded in the elastic range with $G = 10^5/3$ psi, $\nu = 0.3$, $p = 5$ psi. The numerical results with n =

10, 20, 50, 100 are shown in Table 1 together with the Lame' solution for the hoop stresses and strains at the boundaries a and b. The numerical results are correct up to four digits with n = 100.

TABLE 1. ELASTIC SOLUTION FOR A CLOSED-END TUBE

(b/a = 2, G = $10^5/3$ psi, $\nu$ = 0.3, p = 5 psi)

| | n | $\sigma_\theta/a$ | $\sigma_\theta/b$ | $E\varepsilon_\theta/a$ | $E\varepsilon_\theta/b$ |
|---|---|---|---|---|---|
| Exact | - | 8.3333 | 3.3333 | 9.3333 | 2.8333 |
| FDM | 10 | 8.3000 | 3.3000 | 9.3000 | 2.8000 |
| | 20 | 8.3256 | 3.3250 | 9.3250 | 2.8250 |
| | 50 | 8.3320 | 3.3320 | 9.3320 | 2.8320 |
| | 100 | 8.3330 | 3.3330 | 9.3330 | 2.8330 |

The second problem is the initial yielding solution for a plane-strain tube with $E/\sigma_o$ = 200, $\nu$ = 0.3, $\varepsilon_z$ = 0. The numerical results with n = 10, 20, 50, 100, 200, are shown in Table 2 together with the exact solution for the dimensionless $p = p/60$, $\sigma_\theta = \sigma_\theta/\sigma_o$, $\varepsilon_\theta = (E/\sigma_o)(\varepsilon_\theta)$ at r = a and b.

TABLE 2. INITIAL YIELDING SOLUTION FOR A PLANE-STRAIN TUBE

(b/a = 2, $E/\sigma_o$ = 200, $\nu$ = 0.3)

| | n | $\bar{p}$ | $\bar{\sigma}_\theta/a$ | $\bar{\sigma}_\theta/b$ | $\bar{\varepsilon}_\theta/a$ | $\bar{\varepsilon}_\theta/b$ |
|---|---|---|---|---|---|---|
| Exact | - | .43229 | .72049 | .28820 | .82424 | .26226 |
| FDM | 10 | .00110 | -.00107 | -.00216 | -.00054 | -.00065 |
| | 20 | .00028 | -.00027 | -.00054 | -.00014 | -.00016 |
| | 50 | .00005 | -.00004 | -.00009 | -.00002 | -.00003 |
| | 100 | .00001 | -.00001 | -.00003 | -.00001 | -.00001 |
| | 200 | .00001 | .00000 | -.00001 | .00000 | .00000 |

The third problem is the elastic-perfectly plastic solution for a plane-strain tube with b/a = 2, $E/\sigma_o$ = 200, $\nu$ = 0.3, $\omega$ = 0, $\varepsilon_z$ = 0. The numerical results with n = 10, 20, 50, 100, 200 are shown in Table 3 for the pressure and displacement at the bore corresponding to 50 percent and 100 percent overstrain, i.e., $\rho/a$ = 1.5 and 2.0.

TABLE 3. ELASTIC-PERFECTLY PLASTIC SOLUTION FOR A PLANE-STRAIN TUBE

($b/a = 2$, $E/\sigma_o = 200$, $\nu = 0.3$)

| | 50% Overstrain | | 100% Overstrain | |
|---|---|---|---|---|
| n | $\bar{p}$ | $\bar{\varepsilon}_\theta\|_a$ | $\bar{p}$ | $\bar{\varepsilon}_\theta\|_a$ |
| 10 | .71863 | 1.97725 | .79849 | 3.68937 |
| 20 | .71825 | 1.96969 | .79790 | 3.66831 |
| 50 | .71808 | 1.96780 | .79760 | 3.66280 |
| 100 | .71803 | 1.96759 | .79751 | 3.66211 |
| 200 | .71801 | 1.96757 | .79747 | 3.66199 |

The numerical results converge and are accurate up to four digits with n = 100. There is no closed-form solution available for comparison. The famous paper by Hodge and White [2] has been used quite often as a basis for assessing the accuracy of other approximate methods. However, the numerical integration is cumbersome. The present formulation is much simpler and seems more accurate. In order to further demonstrate the accuracy of the present approach, a convergence study for an incompressible, ideally-plastic thick tube in plane-strain condition has been made and compared with exact solution [2]. The numerical results for a nearly incompressible material ($\nu$ = 0.49999) are shown in Table 4 together with the exact solution ($\nu = 1/2$) for the internal pressure and the displacement at the bore corresponding to 50 percent and 100 percent overstrain.

TABLE 4. INCOMPRESSIBLE, IDEALLY-PLASTIC SOLUTION FOR A PLANE-STRAIN TUBE

($b/a = 2$, $E/\sigma_\theta = 200$, $\nu = 0.49999$)

| | 50% Overstrain | | 100% Overstrain | |
|---|---|---|---|---|
| n | $\bar{p}$ | $\bar{\varepsilon}_\theta\|_a$ | $\bar{p}$ | $\bar{\varepsilon}_\theta\|_a$ |
| 10 | .72069 | 1.95714 | .80015 | 3.48806 |
| 20 | .72077 | 1.95072 | .80034 | 3.47012 |
| 50 | .72078 | 1.94891 | .80038 | 3.46509 |
| 100 | .72078 | 1.94865 | .80038 | 3.46437 |
| Exact | .72078 | 1.94856 | .80038 | 3.46410 |

We may thus conclude that exact solutions can be obtained by this numerical approach.

VI. ADDITIONAL RESULTS. After establishing the convergence and accuracy of this new approach, the numerical results for more general problems have been obtained. Some of the additional results are documented here for future comparison by others. All numerical results presented here are for a thick-walled tube with wall ratio $b/a = 2$, $E/\sigma_o = 200$, $\nu = 0.3$, $n = 100$, $\omega = E_t/E = 0.1$. The numerical results are accurate up to four or five digits. Table 5

shows the results of the dimensionless quantities $p/\sigma_o$, $\sigma_\theta/\sigma_o$, $\sigma_z/\sigma_o$, $(E/\sigma_o)\varepsilon_r$, $(E/\sigma_o)\varepsilon_r$, $(E/\sigma_o)\varepsilon_\theta$, $(E/\sigma_o)\varepsilon_z$ at the inside or elastic-plastic boundary $\rho$ for $\rho/a = 1.0, 1.1, 1.2, \ldots, 2.0$ in a plane-strain tube with strain-hardening parameter $\omega = 0.1$. Tables 6 and 7 show the similar results of the dimensionless stresses, strains, displacement at the bore or elastic-plastic boundary for various stages of elastic-plastic loadings in an open-end or closed-end tube, respectively.

TABLE 5. ELASTIC-PLASTIC SOLUTION FOR A PLANE-STRAIN TUBE

($b/a = 2$, $E/\sigma_o = 200$, $\nu = 0.3$, $\varepsilon_t/E_t = 0.1$, $n = 100$, $\varepsilon_z = 0$)

| $\rho/a$ | $p/\sigma_o$ | Inside $\sigma_\theta/\sigma_o$ | $\sigma_\theta/\sigma_o\|_\rho$ | Inside $\sigma_z/\sigma_o$ | $\sigma_z/\sigma_o\|_\rho$ | $\frac{E}{\sigma_o}\varepsilon_r\|_a$ | $\frac{E}{\sigma_o}\frac{U_a}{a}$ |
|---|---|---|---|---|---|---|---|
| 1.0 | .43229 | .72049 | .72049 | .08646 | .08646 | -.67438 | .82424 |
| 1.1 | .51296 | .66972 | .75016 | .05586 | .10453 | -.91636 | 1.00141 |
| 1.2 | .58362 | .63049 | .78249 | .02537 | .12427 | -1.17381 | 1.20371 |
| 1.3 | .64556 | .60225 | .81739 | -.00409 | .14566 | -1.44917 | 1.43022 |
| 1.4 | .69982 | .58427 | .85479 | -.03148 | .16866 | -1.73882 | 1.68000 |
| 1.5 | .74725 | .57572 | .89457 | -.05589 | .19323 | -2.04304 | 1.95207 |
| 1.6 | .78851 | .57579 | .93667 | -.07658 | .21932 | -2.36107 | 2.24535 |
| 1.7 | .82416 | .58376 | .98092 | -.09302 | .24687 | -2.69205 | 2.55869 |
| 1.8 | .85466 | .59898 | 1.02718 | -.10490 | .27581 | -3.03504 | 2.89081 |
| 1.9 | .88041 | .62086 | 1.07530 | -.11205 | .30606 | -3.38898 | 3.24034 |
| 2.0 | .90177 | .64885 | 1.12509 | -.11447 | .33753 | -3.75274 | 3.60578 |

TABLE 6. ELASTIC-PLASTIC SOLUTION FOR AN OPEN-END TUBE

(b/a = 2, $E/\sigma_o$ = 200, $\nu$ = 0.3, $E_t/E$ = 0.1, n = 100, $\varepsilon_z$ = 0)

| $\rho/a$ | $p/\sigma_o$ | $\frac{\sigma_\theta}{\sigma_o}\|_\rho$ | $\frac{\sigma_\theta}{\sigma_o}\|_\rho$ | $\frac{\sigma_z}{\sigma_o}\|_a$ | $\frac{\sigma_z}{\sigma_o}\|_\rho$ | $\frac{E}{\sigma_o}\varepsilon_r\|_a$ | $\frac{E}{\sigma_o}\frac{U_a}{a}$ | $\frac{E}{\sigma_o}\varepsilon_z$ |
|---|---|---|---|---|---|---|---|---|
| 1.0 | .42857 | .71429 | .71429 | 0.0 | 0.0 | -.64286 | .84286 | -.08571 |
| 1.1 | .50697 | .66772 | .74089 | -.02818 | .00090 | -.86384 | 1.01920 | -.10234 |
| 1.2 | .57515 | .63143 | .76928 | -.05498 | .00334 | -1.09744 | 1.21679 | -.11883 |
| 1.3 | .63434 | .60489 | .79916 | -.08033 | .00689 | -1.34207 | 1.43368 | -.13552 |
| 1.4 | .68546 | .58744 | .83021 | -.10392 | .01106 | -1.59572 | 1.66769 | -.15275 |
| 1.5 | .72929 | .57829 | .86195 | -.12542 | .01529 | -1.85284 | 1.91616 | -.17089 |
| 1.6 | .76641 | .57659 | .89385 | -.14463 | .01895 | -2.11934 | 2.17590 | -.19034 |
| 1.7 | .79730 | .58141 | .92520 | -.16147 | .02135 | -2.38245 | 2.44299 | -.21149 |
| 1.8 | .82236 | .59175 | .95517 | -.17604 | .02181 | -2.64070 | 2.71271 | -.23466 |
| 1.9 | .84191 | .60652 | .98280 | -.18862 | .01965 | -2.88902 | 2.97950 | -.26008 |
| 2.0 | .85630 | .62457 | 1.00708 | -.19963 | .01432 | -3.12192 | 3.23718 | -.28781 |

TABLE 7. ELASTIC-PLASTIC SOLUTION FOR A CLOSED-END TUBE

($b/a = 2$, $E/\sigma_o = 200$, $\nu = 0.3$, $E_t/E = 0.1$, $n = 100$)

| $\rho/a$ | $p/\sigma_o$ | $\frac{\sigma_\theta}{\sigma_o}\|_\rho$ | $\frac{\sigma_\theta}{\sigma_o}\|_\rho$ | $\frac{\sigma_z}{\sigma_o}\|_a$ | $\frac{\sigma_z}{\sigma_o}\|_\rho$ | $\frac{E}{\sigma_o}\varepsilon_r\|_a$ | $\frac{E}{\sigma_o}\frac{U_a}{a}$ | $\frac{E}{\sigma_o}\varepsilon_z$ |
|---|---|---|---|---|---|---|---|---|
| 1.0 | .43301 | .72169 | .72169 | .14434 | .14434 | -.69282 | .80829 | .05774 |
| 1.1 | .51408 | .66902 | .75199 | .11139 | .17341 | -.94569 | .98359 | .06863 |
| 1.2 | .58514 | .62897 | .78518 | .07731 | .20331 | -1.21580 | 1.18564 | .07861 |
| 1.3 | .64758 | .60074 | .82124 | .04394 | .23449 | -1.50300 | 1.41370 | .08814 |
| 1.4 | .70247 | .58332 | .86014 | .01276 | .26717 | -1807142 | 1.66712 | .09746 |
| 1.5 | .75069 | .57573 | .90186 | -.01509 | .30145 | -2.12799 | 1.14532 | .10665 |
| 1.6 | .79290 | .57710 | .94636 | -.03887 | .33724 | -2.46525 | 2.24773 | .11566 |
| 1.7 | .82963 | .58672 | .99359 | -.05811 | .37437 | -2.81846 | 2.57373 | .12432 |
| 1.8 | .86323 | .60610 | 1.04858 | -.07381 | .41641 | -3.22466 | 2.95874 | .13314 |
| 1.9 | .88833 | .62834 | 1.09583 | -.08243 | .45136 | -3.56994 | 3.29352 | .13946 |
| 2.0 | .91091 | .65936 | 1.15052 | -.08775 | .49023 | -3.96597 | 3.68518 | .14507 |

## REFERENCES


1. Hill, R., Mathematical Theory of Plasticity, Oxford University Press, 1950.

2. Prager, W. and Hodge, P. G., Theory of Perfectly Plastic Solids, John Wiley & Sons Publication, Inc., 1951, Chapter 4.

3. Hodge, P. G. and White, G. N., "A Quantitative Comparison of Flow and Deformation Theories of Plasticity," J. Appl. Mech., Vol. 17, 1950, pp. 180-184.

4. Chu, S. C., "A More Rational Approach to the Problem of an Elastoplastic Thick-Walled Cylinder," J. of the Franklin Institute, Vol. 294, 1972, pp. 57-65.

5. Chen, P. C. T., "The Finite Element Analysis of Elastic-Plastic Thick-Walled Tubes," Proc. of Army Symposium on Solids Mechanics, 1972, The Role of Mechanics in Design Ballistic Problems, pp. 243-253.

6. Elder, A. S., Tomkins, R. C., and Mann, T. L., "Generalized Plane-Strain in an Elastic, Perfectly Plastic Cylinder, with Reference to the Hydraulic Autofrettage Process," Tran. 21st Conference of Army Mathematicians, 1975, pp. 623-659.

7. Chen, P. C. T., "Generalized Plane-Strain Problems in an Elastic-Plastic Thick-Walled Cylinder," Trans. 26th Conference of Army Mathematicians, 1980, pp. 265-275.

8. Davidson, T. E. and Kendall, D. P., "The Design of Pressure Vessels for Very High Pressure Operation," Watervliet Arsenal Report WVT-6917. Also in Mechanical Behavior of Materials Under Pressure, (H.L.D. Pugh, Ed.), Elsevier Co., 1970, Chapter 2.

9. Yamada, Y., Yoshimura, N., and Sakurni, T., "Plastic Stress-Strain Matrix and Its Application for the Solution of Elastic-Plastic Problems by the Finite Element Method," Int. J. Mech. Sci., Vol. 10, 1968, pp. 343-354.

# A REFINED SHEAR DEFORMATION THEORY FOR LAMINATED ANISOTROPIC PLATES

J. N. Reddy
Department of Engineering Science and Mechanics
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061 USA

ABSTRACT. An improved plate theory that accounts for the transverse shear deformation is presented. The theory is based on an assumed displacement field in which the inplane displacements are expanded in terms of the thickness coordinate up to the cubic term and the transverse deflection is assumed to be independent of the thickness coordinate. The governing differential equations of the theory are derived from the virtual work principle. The theory eliminates the need for shear correction factors because the transverse shear stresses are represented parabolically. A comparison of the present numerical results for bending with the exact solutions of the three-dimensional elasticity theory shows that the present theory is more accurate than the first order shear deformation theory.

I. INTRODUCTION. The classical laminate theory (CLT), which is an extension of the classical plate theory (CPT) to laminated plates, is inadequate for laminated plates made of advanced filamentary composite materials. This is because the effective elastic modulus to the effective shear modulus ratios are very large for such laminates. An adequate description of the transverse shear stresses, especially near the edges, can be achieved with the use of a shear deformation theory.

The first, stress-based, shear deformation plate theory is due to Reissner [1-3]. The theory is based on a linear distribution of the inplane normal and shear stresses through the thickness,

$$\sigma_1 = \frac{M_1}{(h^2/6)}\frac{z}{(h/2)}, \quad \sigma_2 = \frac{M_2}{(h^2/6)}, \quad \sigma_6 = \frac{M_6}{(h^2/6)}\frac{z}{(h/2)} \tag{2}$$

where $(\sigma_1,\sigma_2)$ and $\sigma_6$ are the normal and shear stresses, $(M_1,M_2)$ and $M_6$ are the associated bending moments (which are functions of the inplane coordinates x and y), z the thickness coordinate and h is the total thickness of the plate. The distribution of the transverse normal and shear stresses ($\sigma_3$, $\sigma_4$ and $\sigma_5$) is determined from the equilibrium equations of the three-dimensional elasticity theory. The differential equations and the boundary conditions of the theory were obtained using Castigliano's theorem of least work.

The origin of displacement-based theories is apparently attributed to Basset [4], who begins his analysis with the assumption that the displacement components in a shell can be expanded in a series of powers of the thickness coordinate $\zeta$. For example, the displacement component $u_1$ is written in the form

PREVIOUS PAGE IS BLANK

$$u_1(\xi_1,\xi_2,\zeta) = u_1^o(\xi_1,\xi_2) + \sum_{n=1} \zeta^n u_1^{(n)}(\xi_1,\xi_2) \qquad (2a)$$

where $\xi_1$ and $\xi_2$ are the curvilinear coordinates in the middle surface of the shell, and $u_1^{(n)}$ have the meaning

$$u_1^{(n)}(\xi_1,\xi_2) = \frac{d^n u_1}{d\zeta^n}\Big|_{\zeta=0} \quad , \quad n = 0,1,2,\ldots \qquad (2b)$$

Basset's work did not receive as much attention as it deserves. In 1949 NACA technical note, Hildebrand, Reissner and Thomas [5] presented (also see Hencky [6]) a displacement-based shear deformation theory for shells (which obviously can be specialized to flat plates). They assumed the following displacement field,

$$\begin{aligned} u_1(\xi_1,\xi_2,\zeta) &= u(\xi_1,\xi_2) = \zeta\ \phi_x(\xi_1,\xi_2) \\ u_2(\xi_1,\xi_2,\zeta) &= v(\xi_1,\xi_2) + \zeta\ \phi_y(\xi_1,\xi_2) \\ u_3(\xi_1,\xi_2,\zeta) &= w(\xi_1,\xi_2) \end{aligned} \qquad (3)$$

The differential equations of the theory are then derived using the principle of minimum total potential energy. This gives five differential equations in the five displacement functions, u, v, w, $\phi_x$, and $\phi_y$.

The shear deformation theory based on the displacement field in Eq. (3) for plates is often referred to as the Mindlin's plate theory. Mindlin [7] presented a complete theory based on the displacement field (3) taken from Hencky [6]. The literature review points out that the basic idea came from Basset [4], Hildebrand, Reissner and Thomas [5], and Hencky [6]. Therefore, by referring the displacement-based shear deformation theory as Mindlin's theory we are not giving due credit to the others. We shall refer to the shear deformation theory based on the displacement field (3) as the first-order shear deformation theory.

Higher-order, displacement-based, shear deformation theories have been investigated by Nelson and Lorch [8], Librescu [9] and Lo, Christensen and Wu [10]. These higher-order theories are cumbersome and computationally demanding because with each additional power of the thickness coordinate an additional dependent unknown is introduced, per displacement component, into the theory. Levinson [11] and Murthy [12] presented third-order theories that assume transverse inextensibility. The nine displacement functions were reduced to five by requiring that the transverse shear stresses vanish on the bounding planes of the plate. However, both authors used the equilibrium equations of the first-order theory in their analysis. Stating in other words, the higher-order terms of the displacement field are accounted only in the calculation of the strains but not in the governing differential equations or in the boundary

conditions. Recently, the present author (see [13,14]) corrected these theories by deriving the governing differential equations from variational principles. In this paper we review the results from [13,14] and present the variational formulation of the boundary-value problem associated with the theory. The existence and uniqueness questions will be investigated elsewhere.

II. THE REFINED THEORY. Experimental evidence shows that the transverse shear stresses vary parabolically through the thickness, especially near the edges of a laminate. A parabolic distribution of the shear stresses is provided by a displacement field in which the inplane displacements vary as cubic functions of the thickness coordinate. We begin with the displacement field of the form (3), except that $u_1$ and $u_2$ are assumed to be cubic functions of the thickness coordinate:

$$\begin{aligned} u_1(x,y,z) &= u(x,y) + z\psi_x(x,y) + z^2\xi_x(x,y) + z^3\zeta_x(x,y) \\ u_2(x,y,z) &= v(x,y) + z\psi_y(x,y) + z^2\xi_y(x,y) + z^3\zeta_y(x,y) \\ u_3(x,y,z) &= w(x,y) \end{aligned} \tag{4}$$

where the xy-plane is taken to coincide with the midplane Ω of the laminate and z is the thickness coordinate (i.e., transverse to the laminate). Using the boundary conditions

$$\sigma_4 \equiv \sigma_{yz} = 0 \quad , \quad \sigma_5 = \sigma_{xz} = 0 \text{ at } z = \pm \frac{h}{2} \tag{5}$$

we can eliminate $\xi_x$, $\xi_y$, $\zeta_x$ and $\zeta_y$ (see Reddy [13]) and obtain

$$\begin{aligned} u_1 &= u + z[\psi_x - \frac{4}{3}(\frac{z}{h})^2(\psi_x + \frac{\partial w}{\partial x})] \\ u_2 &= v + z[\psi_y - \frac{4}{3}(\frac{z}{h})^2(\psi_y + \frac{\partial w}{\partial y})] \\ u_3 &= w \end{aligned} \tag{6}$$

It is not difficult to see that (u,v,w) are the displacements of the point (x,y,0) and $\psi_x$ and $\psi_y$ are the rotations of transverse normals at point (x,y,0) about the y and x axes respectively.

The governing differential equations of the theory can be obtained by employing the principle of vitual displacements (see Reddy [15]):

$$\delta \int_{\text{Vol.}} \frac{1}{2} \sigma_{ij}\varepsilon_{ij} dV + \delta \int_S t_i u_i dS = 0 \tag{7}$$

where $\sigma_{ij}$ and $\varepsilon_{ij}$ re the stress and strain components, respectively, and $t_i$ are the specified boundary tractions on the surface of the laminate.

We obtain the following five differential equations:

$$\frac{\partial N_1}{\partial x} + \frac{\partial N_6}{\partial y} = 0 \quad , \quad \frac{\partial N_6}{\partial x} + \frac{\partial N_2}{\partial y} = 0$$

$$\frac{\partial Q_1}{\partial x} + \frac{\partial Q_2}{\partial y} - \frac{4}{h^2}\left(\frac{\partial R_1}{\partial x} + \frac{\partial R_2}{\partial y}\right) + \frac{4}{3h^2}\left(\frac{\partial^2 P_1}{\partial x^2} + 2\frac{\partial^2 P_6}{\partial x \partial y} + \frac{\partial^2 P_2}{\partial y^2}\right) = f$$

$$\frac{\partial M_1}{\partial x} + \frac{\partial M_6}{\partial y} - Q_1 + \frac{4}{h^2} R_1 - \frac{4}{3h^2}\left(\frac{\partial P_1}{\partial x} + \frac{\partial P_6}{\partial y}\right) = 0$$

$$\frac{\partial M_6}{\partial x} + \frac{\partial M_2}{\partial y} - Q_2 + \frac{4}{h^2} R_2 - \frac{4}{3h^2}\left(\frac{\partial P_6}{\partial x} + \frac{\partial P_2}{\partial y}\right) = 0 \qquad (8)$$

where f is the distributed transverse force on the laminate, and the stress resultants $N_i$, $M_i$, $P_i$, $Q_i$ and $R_i$ are defined by

$$(N_i, M_i, P_i) = \int_{-h/2}^{h/2} \sigma_i (1, z, z^3)\, dz \quad (i = 1, 2, 6)$$

$$(Q_2, R_2) = \int_{-h/2}^{h/2} \sigma_4 (1, z^2)\, dz$$

$$(Q_1, R_1) = \int_{-h/2}^{h/2} \sigma_5 (1, z^2)\, dz \qquad (9)$$

The stress resultants are related to the strains by the laminate constitutive relations

$$\{N\} = [A]\{\varepsilon^0\} + [B]\{\kappa^0\} + [C]\{\kappa^2\}$$
$$\{M\} = [B]\{\varepsilon^0\} + [D]\{\kappa^0\} + [F]\{\kappa^2\}$$
$$\{P\} = [E]\{\varepsilon^0\} + [F]\{\kappa^0\} + [H]\{\kappa^2\}$$

$$\begin{Bmatrix} Q_2 \\ Q_1 \\ R_2 \\ R_1 \end{Bmatrix} = \begin{bmatrix} A_{44} & A_{45} & D_{44} & D_{45} \\ & A_{55} & D_{45} & D_{55} \\ & & F_{44} & F_{45} \\ & & & F_{55} \end{bmatrix} \begin{Bmatrix} \varepsilon_4^0 \\ \varepsilon_5^0 \\ \kappa_4^2 \\ \kappa_5^2 \end{Bmatrix} \qquad (10)$$

where $A_{ij}$, $B_{ij}$, etc., are the plate stiffnesses, defined by

$$(A_{ij},B_{ij},D_{ij},E_{ij},F_{ij},H_{ij}) = \int_{-h/2}^{h/2} Q_{ij}(1,z,z^2,z^3,z^4,z^6)\,dz \qquad (i,j=1,2,6)$$

$$(A_{ij},D_{ij},F_{ij}) = \int_{-h/2}^{h/2} Q_{ij}(1,z^2,z^4)\,dz \quad (i,j=4,5) \tag{11}$$

and $\varepsilon_i^o$, $\kappa_i^o$ and $\kappa_i^2$ (i=1,2,4,5,6) are the strain components

$$\varepsilon_1^o = \frac{\partial u_o}{\partial x} \;,\; \kappa_1^o = \frac{\partial \phi_x}{\partial x} \;,\; \kappa_1^2 = -\frac{4}{3h^2}\left(\frac{\partial \phi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}\right)$$

$$\varepsilon_2^o = \frac{\partial u_o}{\partial y} \;,\; \kappa_2^o = \frac{\partial \phi_y}{\partial y} \;,\; \kappa_2^2 = -\frac{4}{3h^2}\left(\frac{\partial \phi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2}\right)$$

$$\varepsilon_4^o = \phi_y + \frac{\partial w}{\partial y} \;,\; \kappa_4^2 = -\frac{4}{h^2}\left(\phi_y + \frac{\partial w}{\partial y}\right)$$

$$\varepsilon_5^o = \phi_x + \frac{\partial w}{\partial x} \;,\; \kappa_5^2 = -\frac{4}{h^2}\left(\phi_x + \frac{\partial w}{\partial x}\right)$$

$$\varepsilon_6^o = \frac{\partial u_o}{\partial y} + \frac{\partial v_o}{\partial x} \;,\; \kappa_6^o = \frac{\partial \phi_x}{\partial y} + \frac{\partial \phi_y}{\partial x} \;,\; \kappa_6^2 = -\frac{4}{3h^2}\left(\frac{\partial \phi_x}{\partial y} + \frac{\partial \phi_y}{\partial x} + 2\,\frac{\partial^2 w}{\partial x \partial y}\right) \tag{12}$$

III. EXISTENCE AND UNIQUENESS OF SOLUTIONS. Here we consider the Dirichlet problem for the equations in (8). For symmetric cross-ply laminates, the first two equations become uncoupled from the last three equations of Eq. (8). For zero inplane forces we have $u = v = 0$. Thus, we are required to find $w$, $\phi_x$ and $\phi_y$ under given boundary conditions (say, clamped plate) and distributed load $f$. In this part of the discussion we shall use the standard notation of functional analysis (see Rektorys [16]).

The variational problem associated with the last three equations in (8) with the zero essential boundary conditions involves finding $\Lambda = (w,\phi_x,\phi_y) \in \underset{\sim}{H}$ such that

$$B(\bar{\Lambda},\Lambda) = \ell(\bar{\Lambda}) \text{ for all } \bar{\Lambda} \in \underset{\sim}{H} \tag{13}$$

where $\underset{\sim}{H}$ denotes the product space

$$H = H_o^2(\Omega) \times H_o^1(\Omega) \times H_o^1(\Omega) \tag{14}$$

Here $\Omega$ denotes the midplane of the plate (assumed to be a Lipschitzian) with a smooth boundary $\Gamma$, and $H_o^1(\Omega)$ and $H_o^2(\Omega)$ are the Hilbert spaces of order 1 and 2, respectively, with compact support,

$$H_o^m(\Omega) = \{u: \; D^{\underset{\sim}{\alpha}} u \in L_2(\Omega), \; |\underset{\sim}{\alpha}| \leq m, \; D^{\underset{\sim}{\beta}} u = 0 \text{ on } \Gamma,$$

$$|\underset{\sim}{\beta}| \leq m - 1\} \tag{15}$$

Thus, if w is in $H_o^2(\Omega)$ then w and its derivatives of order 1 and 2 are square integrable in $\Omega$ and w and its first derivatives vanish on $\Gamma$. The bilinear form $B(\cdot,\cdot)$ and linear form $\ell(\cdot)$ are defined by

$$B(\bar{\Lambda},\Lambda) = \int_\Omega \{\frac{\partial\bar{\psi}_x}{\partial x}[D_{11}\frac{\partial\psi_x}{\partial x} + D_{12}\frac{\partial\psi_y}{\partial y} + (-\frac{4}{3h^2})F_{11}(\frac{\partial\psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2})$$

$$+ (-\frac{4}{3h^2})F_{12}(\frac{\partial\psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})] + \frac{\partial\bar{\psi}_y}{\partial y}[D_{12}\frac{\partial\psi_x}{\partial x} + D_{22}\frac{\partial\psi_y}{\partial y}$$

$$+ (-\frac{4}{3h^2})F_{12}(\frac{\partial\psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}) + (-\frac{4}{3h^2})F_{22}(\frac{\partial\psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})] + (\frac{\partial\bar{\psi}_x}{\partial y}$$

$$+ \frac{\partial\bar{\psi}_y}{\partial x})[D_{66}(\frac{\partial\psi_x}{\partial y} + \frac{\partial\psi_y}{\partial x}) + (-\frac{4}{3h^2})F_{66}(\frac{\partial\psi_x}{\partial y} + \frac{\partial\psi_y}{\partial x} + 2\frac{\partial^2 w}{\partial x\partial y})]$$

$$+ (-\frac{4}{3h^2})(\frac{\partial\bar{\psi}_x}{\partial x} + \frac{\partial^2\bar{w}}{\partial x^2})[F_{11}\frac{\partial\psi_x}{\partial x} + F_{12}\frac{\partial\psi_y}{\partial y} + (-\frac{4}{3h^2})H_{11}(\frac{\partial\psi_x}{\partial x}$$

$$+ \frac{\partial^2 w}{\partial x^2}) + (-\frac{4}{3h^2})H_{12}(\frac{\partial\psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})] + (-\frac{4}{3h^2})(\frac{\partial\bar{\psi}_y}{\partial y} + \frac{\partial^2\bar{w}}{\partial y^2})[F_{12}\frac{\partial\psi_x}{\partial x}$$

$$+ F_{22}\frac{\partial\psi_y}{\partial y} + (-\frac{4}{3h^2})H_{12}(\frac{\partial\psi_x}{\partial x} + \frac{\partial^2 w}{\partial x^2}) + (-\frac{4}{3h^2})H_{22}(\frac{\partial\psi_y}{\partial y} + \frac{\partial^2 w}{\partial y^2})]$$

$$+ (-\frac{4}{3h^2})(\frac{\partial\bar{\psi}_x}{\partial y} + \frac{\partial\bar{\psi}_y}{\partial x} + 2\frac{\partial^2\bar{w}}{\partial x\partial y})[F_{66}(\frac{\partial\psi_x}{\partial y} + \frac{\partial\psi_y}{\partial x}) + (-\frac{4}{3h^2})H_{66}(\frac{\partial\psi_x}{\partial y}$$

$$+ \frac{\partial\psi_y}{\partial x} + 2\frac{\partial^2 w}{\partial x\partial y})] + (A_{55} - \frac{4}{h^2}D_{55})(\bar{\psi}_x + \frac{\partial\bar{w}}{\partial x})(\psi_x + \frac{\partial w}{\partial x})$$

$$+ (A_{44} - \frac{4}{h^2} D_{44})(\bar{\psi}_y + \frac{\partial \bar{w}}{\partial y})(\psi_y + \frac{\partial w}{\partial y}) + (-\frac{4}{h^2})(D_{44} - \frac{4}{h^2} F_{44})(\bar{\psi}_x$$

$$+ \frac{\partial \bar{w}}{\partial x})(\psi_x + \frac{\partial w}{\partial x}) + (-\frac{4}{h^2})(D_{55} - \frac{4}{h^2} F_{55})(\bar{\psi}_y + \frac{\partial \bar{w}}{\partial y})(\psi_y + \frac{\partial w}{\partial y})\}dxdy$$

$$\ell(\bar{\Lambda}) = \int_\Omega f\bar{w}dxdy \tag{16}$$

The existence and uniqueness of the solution to Eq. (13) follows from the Lax-Milgram theorem (See [16]) if $B(\bar{\Lambda},\Lambda)$ is continuous and H-elliptic, i.e., there exist constants $\mu_o$ and $\mu_1$ such that

$$|B(\bar{\Lambda},\Lambda)| \leq \mu_o \|\bar{\Lambda}\|_H \|\Lambda\|_H$$

$$B(\Lambda,\Lambda) \geq \mu_1 \|\Lambda\|_H^2 \tag{17}$$

and $\ell(\bar{\Lambda})$ is continuous on H. Here $\|\cdot\|_H$ denotes the product norm in H,

$$\|(w,\psi_x,\psi_y)\|_H^2 = \|w\|^2_{H^2(\Omega)} + \|\psi_x\|^2_{H^1(\Omega)} + \|\psi_y\|^2_{H^1(\Omega)} \tag{18}$$

The continuity of $B(\cdot,\cdot)$ and $\ell(\cdot)$ can be easily established using the Cauchy-Schwarz inequality (the constant $\mu_o$ depends on the maximum values of $D_{ij}$, $F_{ij}$ and $H_{ij}$). However, the proof of the H-ellipticity of $B(\cdot,\cdot)$ is considerably involved and will be addressed in a separate publication.

IV. EXACT SOLUTIONS AND NUMERICAL RESULTS. Exact solutions to the boundary value problem described by the first three equations of Eq. (8) with the following simply supported boundary conditions of a rectangular plate can be obtained (see [14]):

$$w(x,0) = w(x,b) = w(0,y) = w(a,y) = 0$$

$$\psi_x(x,0) = \psi_x(x,b) = \psi_y(0,y) = \psi_y(a,y) = 0 \qquad \text{(essential)}$$

$$M_1(0,y) = M_1(a,y) = M_2(x,0) = M_2(x,b) = 0$$

$$P_1(0,y) = P_2(a,y) = P_2(x,0) = P_2(x,b) = 0 \qquad \text{(natural)} \tag{20}$$

where the origin of the coordinate system is taken at the lower left corner of the plate. The (0°/90°/0°) cross-ply laminate with the following lamina properties and under sinusoidal loading is considered ($G_{13} = G_{12}$ and $\nu_{13} = \nu_{14}$):

$$E_1/E_2 = 25,\ G_{12}/E_2 = 0.5,\ G_{23}/E_2 = 0.2,\ \nu_{12} = 0.25$$

Figure 1 contains plots of the nondimensional center deflection,

$$\bar{w} = w(\frac{a}{2}, \frac{b}{2}, 0) \frac{E_2 h^3}{f_o a^4} \times 10^4$$

versus side to thickness ratio of square laminates. The deflections predicted by the higher-order theory are in good agreement with the 3-D elasticity theory (see [17]). The effect of not including the transverse shear strains on the deflections is significant for side to thickness ratios smaller than 20, as can be seen from the difference in the solutions predicted by the classical laminate theory (CLT) and the first and higher-order shear deformation theories (FSDT) and HSDT). The present higher order-theory gives more accurate deflections than the first order theory.

As explained in the introduction, the higher-order theory represents the parabolic distribution of the transverse shear stresses whereas the first-order theory represents only constant distribution through thickness. The distributions of $\sigma_{xz}(=\sigma_5)$ through the thickness (computed from the constitutive equations) as given by the two shear deformation plate theories and the 3-D elasticity theory are compared in Fig. 2. Both plate theories suffer from the discontinuities at the lamina interfaces because the continuity of stresses across interlamina is not imposed. Relatively, the higher-order theory gives more accurate solution than the first order theory. The average value of $\sigma_{xz}$ at the lamina interfaces given by the higher-order theory is closer to the elasticity solution than the first-order theory solution.

Lastly, Fig. 3 contains plots of $\sigma_{yz}(=\sigma_4)$ versus side to thickness ratio. The higher-order theory again shows an improvement over the first order theory.

V. SUMMARY AND CONCLUSIONS. A third-order shear deformation theory based on assumed displacement field is presented for laminated plates. The theory accounts for parabolic distribution of the transverse shear stresses and hence does not require the use of shear correction coefficients. In general, the theory is more accurate than the first-order theory while both theories contain the same displacement functions.

Existence and uniqueness of solutions to the variational problem of the theory are not available at this writing. The proof of the ellipticity in a proper vector space is considerably involved and requires attention.

VI. REFERENCES

[1] Reissner, E., "On the theory of bending of elstic plates," J. Math. Phy., Vol. 23, pp. 184-191, 1944.

[2] Reissner, E., "The effect of transverse shear deformation on the bending of elastic plates," J. Appl. Mech., Vol. 17, No. 1, pp. A-69 to A-77, 1945.

[3] Reissner, E., "On bending of elastic plates," Q. Appl. Math., Vol. 5, pp. 55-68, 1947.

[4] Basset, A. B., "On the extension and flexure of cylindrical and spherical thin elastic shells," Phil. Trans. Royal Soc., (London), Ser. A, Vol. 181, No. 6, pp. 433-480, 1890.

[5] Hildebrand, F. B., Reissner, E. and Thomas, G. B., "Note on the foundations of the theory of small displacements of orthotropic shells," NACA Technical Note No. 1833, March 1949.

[6] Hencky, H., "Über die berucksichtigung der schubverzerrung in ebenen platten," Ing. Archiv., Vol. 16, 1947.

[7] Mindlin, R. D., "Influence of rotatory inertia and shear on flexural motions of isotropic, elastic plates," J. Appl. Mech., Vol. 18, p. A31, 1951.

[8] Nelson, R. B. and Lorch, D. R., "A refined theory for laminated orthotropic plates," J. Appl. Mech., Vol. 41, pp. 177-183, 1974.

[9] Librescu, L., Elastostatics and Kinetics of Anisotropic and Heterogeneous Shell-Type Structures, Noordhoff, The Netherlands, 1975.

[10] Lo, K. H., Christensen, R. M. and Wu, E. M., "A high-order theory of plate deformation," Parts 1 and 2, J. Appl. Mech., pp. 663-676, 1977.

[11] Levinson, M., "An accurate simple theory of the staics and dynamics of elastic plates," Mech. Res. Commun., Vol. 7, p. 343, 1980.

[12] Murthy, M. V. V., "An improved transverse shear deformation theory for laminated anisotropic plates," NASA Technical Paper 1903, November 1981.

[13] Reddy, J. N., "A refined nonlinear theory of plates with transverse shear deformation," Int. J. Solids Struct., to appear, 1984.

[14] Reddy, J. N., "A simple higher-order theory for lamianted composite plates," J. Appl. Mech., to appear, 1984.

[15] Reddy, J. N., Energy and Variational Methods in Applied Mechanics, John Wiley, New York, 1984.

[16] Rektorys, Variational Methods in Mathematics, Science and Engineering, D. Reidel Publi ing Co., Boston, 1975.

[17] Pagano, N. J. and Hatfield, S. J., "Elastic behavior of multilayered bidirectional composites," AIAA J. Vol. 10, pp. 931-933, 1972.

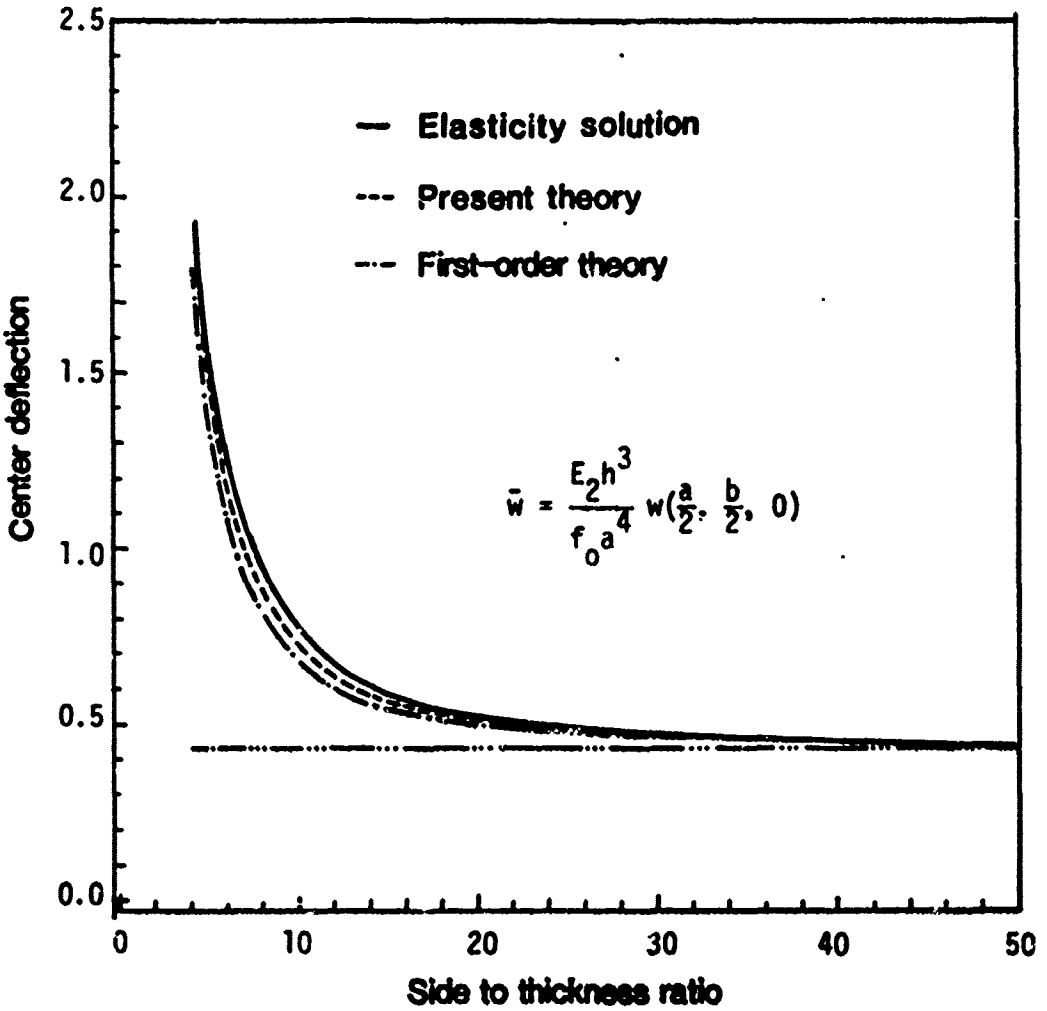



Figure 1 The effect of transverse shear deformation on the deflections of a square laminate (0°/90°/0°) under sinusoidal transverse load

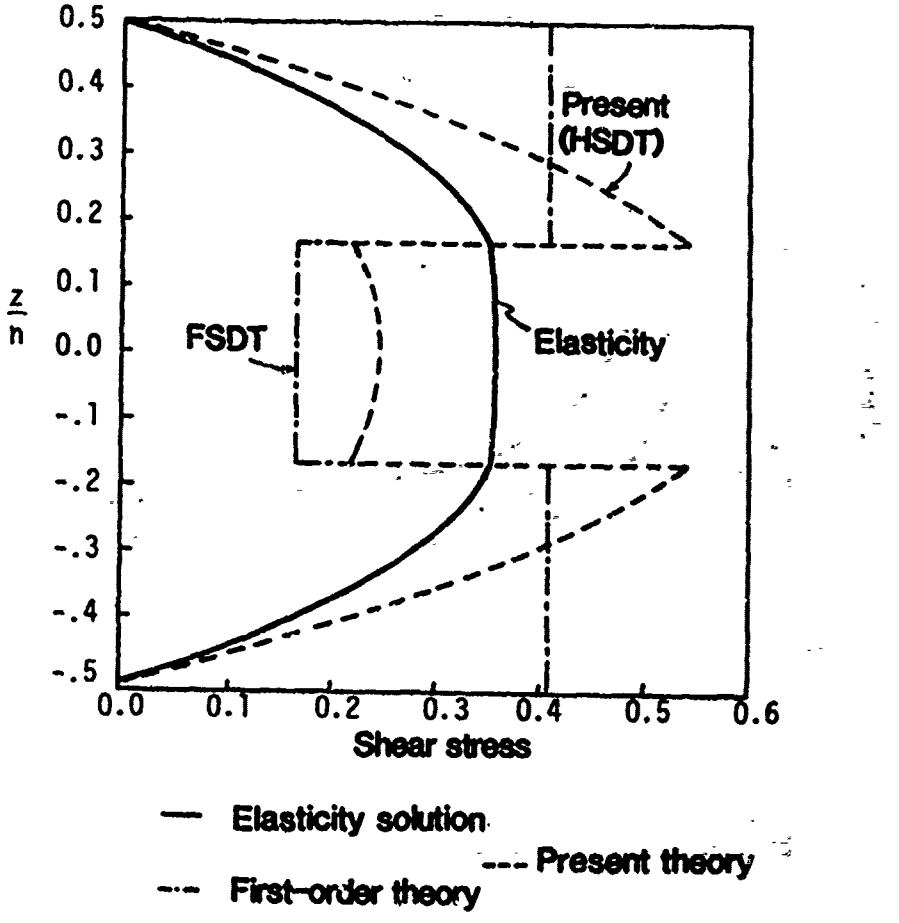



Figure 2 The variation of the transverse shear stress $\bar{\sigma}_{xz}$ through the thickness of a square laminate (0°/90°/0°) under sinusoidal load (a/h=10)

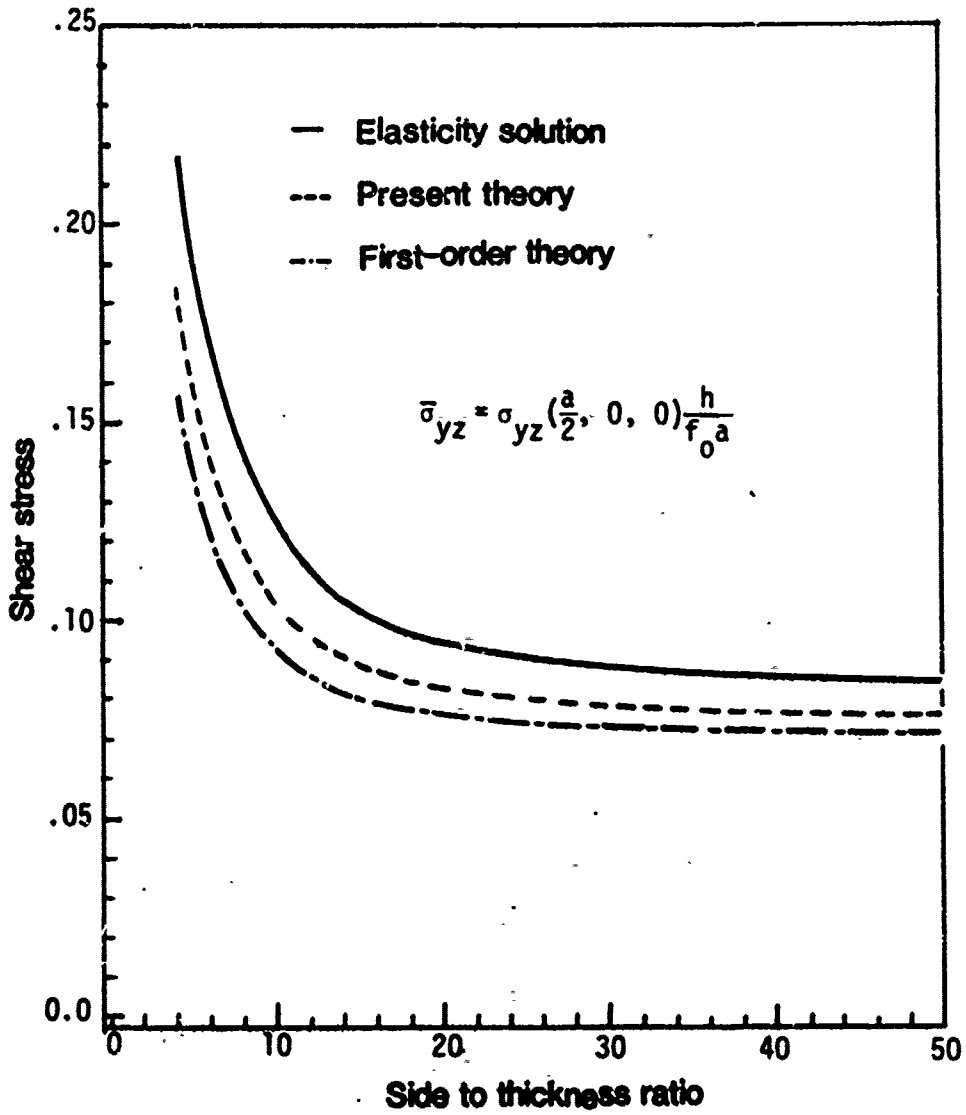



Figure 3 The effect of the transverse shear deformation on the shear stress $\bar{\sigma}_{yz}$ of a square laminate (0°/90°/0°) under sinusoidal transverse load

# ANISOPARAMETRIC FINITE ELEMENT METHODOLOGY FOR PENALTY CONSTRAINT FORMULATIONS IN SOLID MECHANICS

Alexander Tessler

Mechanics of Materials Branch
Army Materials and Mechanics Research Center
Watertown, Massachusetts 02172

ABSTRACT. A discussion of what is called the anisoparametric finite element method, suited for penalty constraint formulations, is presented in the context of displacement flexure elements. Two auxiliary criteria for the kinematic field are elaborated, on which basis the desired interpolation assumptions are established. Much attention is devoted to the lowest order triangular plate element. Several numerical experiments are conducted to ascertain the element performance.

1. INTRODUCTION. Numerous problems in solid mechanics are treated by finite element models based upon (exterior) penalty constraint variational methods (see e.g. [1]). Refined flexure theories that extend classical assumptions to account for transverse shear deformation and rotary inertia effects are representative of these methods (e.g., [2-5]). In these theories, the transverse shear deformation energy serves as a penalty constraint. Its purpose is to fulfill the classical assumption of negligible transverse shear as well as to accommodate the non-classical (shear-weak) deformation regime.

There has been an extensive effort, and for the most part unsuccessful, to adopt the well-established isoparametric methodology to penalty constraint formulations. The essential feature of the isoparametric approach is that interpolation assumptions are independent, and the same parametric shape functions describe spatial variations of field variables and the element geometry ( super- and sub-parametric schemes employ respectively higher and lower order interpolations for the element geometry, see [6]).

One major drawback resulting from the isoparametric approach is manifested by overly stiff behavior in the classical regime. This phenomenon is termed 'shear locking'. Improvements have been achieved by the use of selective/reduced integration procedures, which effectively alleviate locking [7-10]. However, such remedies are often accompanied by spurious mechanisms that render these elements unreliable for general applications.

In this communication, a consistent finite element methodology, termed anisoparametric, is discussed in connection with linear displacement models for penalty constraint flexure elements. The distinguishing feature of this methodology is

manifested by the kinematic interpolations that obey, in addition to the standard convergence criteria (e.g., see [6]), the criteria impelled by the penalty constraints. The kinematic variables, in order to meet these auxiliary requirements, employ distinct degree parametric polynomials (hence, the term anisoparametric). These interpolations make possible a convenient explicit degree-of-freedom reduction procedure performed a priori to the element matrix integrations. Consequently, simple, isoparametric-like nodal patterns and compatible kinematic fields can be obtained. Several such elements have recently been proposed for the analysis of Timoshenko beams [11], axisymmetric shells [12] and plates [13-15]. These elements do not suffer from 'locking' and overall are superior in performance to their reduced integration, isoparametric counterparts.

The latest plate elements [14,15] employ so-called element-appropriate shear corrections, which accelerate the solution convergence. This aspect is addressed here briefly. Much of the discussion is focused on the lowest order triangular element, whose complete derivation may be found in [15]. Several numerical examples are carried out with this element to demonstrate its performance characteristics.

2. 'ISOPARAMETRIC SHEAR LOCKING'. To ascertain the cause of difficulties associated with a typical penalty constraint formulation, we examine a plate element 'displacement' approach according to the refined theories of Reissner [3] or Mindlin [4]. Using either theory, the strain energy functional for a linearly elastic, isotropic plate element of thickness h, area A, Young's modulus E and Poisson's ratio $\upsilon$ takes the form (refer to Fig.1 for the plate notation) :

$$U = \tfrac{1}{2}D\{ \iint [\, \theta_{y,x}^2 + 2\upsilon\theta_{y,x}\theta_{x,y} + \theta_{x,y}^2 + 1/2(1-\upsilon)(\theta_{y,y} + \theta_{x,x})^2 ]\, dxdy$$
$$+ \alpha \iint 1/A[\, \gamma_{xz}^2 + \gamma_{yz}^2 ]\, dxdy\} \qquad (2.1)$$

where the first and second integrals represent nondimensional bending and transverse shear strain energies, respectively; $D=Eh^3/12(1-\upsilon^2)$ is the bending rigidity; $\alpha = 6k(1-\upsilon)A/h^2$ is the penalty parameter; k is the shear correction factor; and

$$\gamma_{xz} = w_{,x} + \theta_y \quad \text{and} \quad \gamma_{yz} = w_{,y} + \theta_x \qquad (2.2)$$

are the transverse shear (penalty) strains; and a comma is used to denote partial differentiation.

The second integral in (2.1) may be interpreted as a penalty constraint functional. Its purpose is twofold: (1) to enforce the classical limit of vanishing transverse shear as the

plate approaches its thin limit (i.e., as h-->0 the Kirchhoff constraints $\gamma_{xz}, \gamma_{yz}$ -->0 are enforced); (2) to accommodate the non-classical (thick) regime accounting for the transverse shear effect.

It is emphasized that the liniting classical regime is the critical one for the element behavior. As the penalty parameter tends to infinity as h-->0, it is desirable to ensure that the vanishing shear strains do not originate any spurious constraining. This is because the spurious constraining, often referred to as 'locking', constitutes a major stumbling block leading to excessively stiff and essentially useless solutions.

Practically all isoparametric elements (i.e., elements employing the same parametric functions for the deflection ,w, and normal rotation , $\theta_x$ and $\theta_y$ , variables, see e.g. [6]) suffer from the locking effect under 'normal' integration of the stiffness matrix ( a 'normal' integration implies the minimum order numerical (usually Gauss) quadrature rule that produces exact integrals taken over a rectangular (or triangular) plan). At first glance, it appears paradoxical that such an approach leads to worthless results even though all of the fundamental convergence requirements are fulfilled (i.e. rigid body modes, constant strain, compatibility and differentiability criteria, [6]). However, upon examination of the penalty strains while under Kirchhoff constraints, the cause of the spurious constraining becomes apparent.

To illustrate this point, consider an isoparametric , linear triangular element located in the x-y Cartesian framework whose origin is at the element centroid. The kinematic interpolations may be written as

$$w = c_{00} + c_{10}x + c_{01}y \qquad \text{(a)}$$

$$\theta_i = d_{00}^{(i)} + d_{10}^{(i)}x + d_{01}^{(i)}y \ , \quad (i=x,y) \qquad \text{(b)} \qquad (2.3)$$

where $c_{kl}$ and $d_{kl}^{(i)}$ are generalized constants dependent upon $w_j$ and $\theta_{ij}$ ( $1 \le j \le 3$ ) degrees-of-freedom (dof), respectively. The Kirchhoff constraints

$$0 = \gamma_{xz} = w_{,x} + \theta_y$$

$$= (c_{10} + d_{00}^{(y)}) + d_{10}^{(y)}x + d_{01}^{(y)}y \qquad (2.4)$$

and

$$0 \leftarrow \gamma_{yz} = w_{,y} + \theta_x$$

$$= (c_{01} + d_{00}^{(x)}) + d_{10}^{(x)} x + d_{01}^{(x)} y \qquad (2.5)$$

are enforced in the limit as h-->0. If the shear strain energy is integrated exactly, which is in accordance with variational requirements, then all six constraints resulting from (2.4,5) are approximately enforced. Among these, the four constraints,

$$d_{10}^{(y)}, d_{01}^{(y)}, d_{10}^{(x)}, d_{01}^{(x)} \rightarrow 0 \qquad (2.6)$$

exclusively dependent on the rotational dof, are the spurious (nonphysical) ones. Upon their enforcement, normal rotations (2.3b) can assume either zero or constant values, and this causes the bending energy to vanish (see (2.1)). The practical implication is that excessively stiff results are obtained ,i.e, the solution is 'locked' due to spurious shear constraining.

To alleviate the spurious locking phenomenon within the framework of the isoparametric assumptions, a one-point centroidal quadrature on the shear strain energy might be used, which effectively 'drops' the spurious constraints (i.e., (2.4, 5) are evaluated at x=y=0.) Such reduced integration schemes, though often plagued by one or more spurious zero energy modes, have been used widely due to their simplicity accompanied with significant improvements in results (e.g., refer to [7,8]).

The locking phenomenon can also be linked to the inability of the isoparametric, exactly integrated elements to represent the state of constant shear, even though a constant term is present in each shear strain. To observe this, consider a simple one-dimensional example -- a cantilevered beam loaded by a shear force at the tip. Assuming the beam axis coincides with the x-direction and setting y=0 (i.e., considering a beam element), it becomes apparent that the theoretically constant shear state ($\gamma_{xz}$ = const.) can only be achieved if $d_{10}^{(y)}=0$. However, in the linear isoparametric element,

$$d_{10}^{(y)} \sim \theta_{y1} - \theta_{y2} \qquad (2.7)$$

and for $d_{10}^{(y)}$ to vanish, the two rotational dof must be equal or be identically zero. But in this case , again, the bending energy vanishes and locking is encountered. On the other hand, a non-locking solution can be obtained in the thick regime [7], in which case $\theta_{y1} \neq \theta_{y2}$, yielding a linear (not constant!) shear.

It then follows that the isoparametric approach should be regarded inadequate for the treatment of the penalty functional of (2.1).

3. AUXILIARY CRITERIA. In order to produce effective non-locking solutions, it is proposed that the kinematic field adheres to the auxiliary, penalty-constraint impelled requirements:

Criterion 1. Constant penalty strains must be accommodated when such deformation states are imposed upon a finite size element.

Criterion 2. The classical limiting condition of vanishing shear strains must be attainable without engendering any spurious constraining.

These criteria point toward the notion of 'proper' shear strain polynomials [13]. In these polynomials, each generalized coordinate is represented by a linear combination of the deflection and normal rotation dof, i.e., they possess no spurious terms dependent on a single independent variable. It can be readily ascertained that the proper, complete polynomials are perfectly suited for triangular elements (also one-dimensional elements such as beams and axisymmetric shells , see [11,12] ). Such shear strain polynomials of degree P may be expressed as:

$$\gamma_{jz}(w_{,j}\ ;\ \theta_i) \overset{\text{def}}{=} \sum_{\alpha=0}^{p}\sum_{\beta=0}^{p} a^{(j)}_{\alpha\beta} x^{\alpha}y^{\beta} \ni a^{(j)}_{\alpha\beta} = 0 \ \forall\ (\alpha+\beta) > p,$$

$$a^{(j)}_{\alpha\beta} \neq 0 \ \forall\ (\alpha+\beta) \leqslant p\ , \tag{3.1}$$

$$a^{(j)}_{\alpha\beta} = a^{(j)}_{\alpha\beta}(w_k,\ \theta_{ik}),\quad (j=x,y\ ;\ i=y,x\ ;\ i\neq j\ ;\ p=1,2,\ldots)$$

where $w_k$ and $\theta_{ik}$ denote the appropriate nodal dof.

The form of (3.1) suggests that the polynomial for w must be one degree higher than that for $\theta_i$. Hence, the desired triangular element interpolations can be written as

$$w = \sum_{\alpha=0}^{p+1}\sum_{\beta=0}^{p+1} c_{\alpha\beta} x^{\alpha}y^{\beta}\ ,\quad w_{,j} = \sum_{\alpha=0}^{p}\sum_{\beta=0}^{p} c^{(j)}_{\alpha\beta} x^{\alpha}y^{\beta}\ ,\quad \theta_i = \sum_{\alpha=0}^{p}\sum_{\beta=0}^{p} d^{(i)}_{\alpha\beta} x^{\alpha}y^{\beta},$$

$$\gamma_{jz} = w_{,j} + \theta_i = \sum_{\alpha=0}^{p}\sum_{\beta=0}^{p} a^{(j)}_{\alpha\beta} x^{\alpha}y^{\beta},\quad a^{(j)}_{\alpha\beta} = c^{(j)}_{\alpha\beta} + d^{(i)}_{\alpha\beta}$$

$$(j=x,y\ ;\ i=y,x\ ;\ i\neq j) \tag{3.2}$$

where conditions on $c_{\alpha\beta}$, $c^{(j)}_{\alpha\beta}$ and $d^{(i)}_{\alpha\beta}$ are the same as those on $a^{(j)}_{\alpha\beta}$ in (3.1). Consequently, the deflection slopes $w_{,j}$ are represented by the same complete polynomials as the normal rotations $\theta_i$'s.

It must be remarked that fulfillment of Criterion 1 will generally guarantee satisfaction of Criterion 2 as well, except when certain 'overrestrained' boundary conditions are prescribed. The interested reader is referred to [15] for a detailed discussion on this subject.

4. ANISOPARAMETRIC INTERPOLATIONS. Interpolations (3.2) once expressed in terms of the element parametric coordinates will be referred to as anisoparametric. The term is meant to emphasize the distinct order of the kinematic interpolations. The well-established isoparametric approach would then constitute a sub-class of the anisoparametric strategy.

In the derivation of a triangular element, it is convenient to utilize parametric coordinates expressed in terms of subtriangle areas (refer to Fig.2), i.e.,

$$( \zeta_1, \zeta_2, \zeta_3 ) = 1/A \; ( A_1, A_2, A_3), \quad \text{with} \quad \sum_{i=1}^{3} A_i = A \tag{4.1}$$

The unique linear relation between Cartesian and parametric coordinates

$$(1, x, y) = \left( \sum_{i=1}^{3} \zeta_i, \sum_{i=1}^{3} \zeta_i x_i, \sum_{i=1}^{3} \zeta_i y_i \right) \tag{4.2}$$

permits an inverse relation as well. The simple integration formula

$$\iint \zeta_1^a \, \zeta_2^b \, \zeta_3^c \, dxdy = \frac{a!\, b!\, c!}{(a+b+c+2)!} \, 2A \tag{4.3}$$

allows for the explicit exact quadratures of all element matrices.

The initial kinematic assumptions for the lowest interpolation order element (p = 1) are taken as (refer to Fig.3 for the nodal pattern):

quadratic deflection

$$w = \underline{N} \, \underline{w}_v \tag{4.4}$$

where the row vector of quadratic trial functions and the associated vector of nodal dof are:

$$\underline{N} = \{N_i\} \quad , \quad \underline{w}_v^T = \{w_i\} \quad , \quad (i = 1,\ldots,6)$$

with

$$N_i = \zeta_i(2\zeta_i - 1), \quad N_{i+3} = 4\zeta_i\,\zeta_k, \quad (i = 1,2,3;\ k = 2,3,1) \tag{4.4a}$$

linear rotations

$$\theta_i = \underline{\zeta}\,\underline{\theta}_i, \quad i = x,y \tag{4.5}$$

where the row vector of linear trial functions and the associated vector of nodal dof are:

$$\underline{\zeta} = \{\zeta_k\}, \quad \underline{\theta}_i^T = \{\theta_{ik}\}, \quad k = 1,2,3 \tag{4.5a}$$

Equations (4.4, 5) can be directly used in formulating element matrices. However, it may be advantageous, from the standpoint of nodal simplicity, to condense out the mid-edge deflection dof. This can be accomplished in the following manner: enforce continuous shear constraints at every element edge as given by the differential relation

$$\gamma_{sz,s} = (w_{,s} + \theta_n)_{,s}\big|_{\zeta_i=0} \equiv 0, \quad (i = 1,2,3) \tag{4.6}$$

where s denotes the edge coordinate and $\theta_n$ is the tangential edge rotation (refer to Fig. 3). The enforcement of constraints (4.6) at the three element edges yields:

$$w_{i+3} = 1/2\,(w_i + w_j) + 1/8\,[b_k(\theta_{xi} - \theta_{xj}) + a_k(\theta_{yj} - \theta_{yi})] \tag{4.7}$$

$$(i = 1,2,3;\ j = 2,3,1;\ k = 3,1,2)$$

Upon substituting (4.7) into w in (4.4), there results a constrained deflection field exclusively in terms of vertix dof (refer to Fig. 3 ), i.e.,

$$w = \underline{\zeta}\,\underline{w} + \underline{L}\,\underline{\theta}_x + \underline{M}\,\underline{\theta}_y,$$

$$\underline{L} = \{L_i\}, \quad \underline{M} = \{M_i\}, \quad \underline{w} = \{w_i\}, \quad \underline{\theta}_x = \{\theta_{xi}\}, \quad \underline{\theta}_y = \{\theta_{yi}\}, \tag{4.8}$$

$$L_i = 1/8\,(b_k N_{i+3} - b_j N_{k+3}), \quad M_i = 1/8\,(a_j N_{k+3} - a_k N_{i+3}),$$

$$(i = 1,2,3;\ j = 2,3,1;\ k = 3,1,2)$$

where the conditions

$$\sum_{i=1}^{3} L_i \equiv 0, \quad \sum_{i=1}^{3} M_i \equiv 0, \quad \sum_{i=1}^{3} \zeta_i \equiv 1 \tag{4.9}$$

guarantee satisfaction of the fundamental constant strain criterion.

The deflection given by (4.8) possesses continuity of order $C^0$ within the element and across its edges. Together with the linear rotation assumptions (4.5), this constrained deflection field of the second degree is readily employed in the formulation of the element matrices. Further formulation details of this element, designated as MIN3, may be found in [15].

5. ELEMENT-APPROPRIATE SHEAR CORRECTION. The major impetus contributing to the performance enhancement of the anisoparametric elements [14,15] is the notion of the element-appropriate shear correction factors. This idea is based on the assertion that, in the context of finite element (in-plane) approximations, additional and often unwanted kinematic constraining commonly occurs. It may, therefore, be desirable to introduce an element (in-plane) correction , $\phi$ , to complement the classical (through-the-thickness) $\kappa$-factor. Such a compounded factor is element dependent and can be expressed in a separable form, i.e., for the n-th element in the discretization, we have

$$k_n^e = \phi_n^e \, k \tag{5.1}$$

whereas the classical correction usually pertains to the whole plate. In [15], by employing an energy matching procedure, $\phi_n^e$ emerged in the form

$$\phi_n^e = [1 + (C_b - 1)\psi_n^2]^{-1} \tag{5.2}$$

where $C_b$ is an invarient element constant established by a numerical experiment; $\psi_n^2$ represents a characteristic material-geometric parameter which can be linked to the element stiffness penalty parameter. Two distinct approaches for determining $\psi_n^2$ are proposed in [14,15]. The interested reader may consult these references for further details.

6. NUMERICAL EXAMPLES. Several numerical experiments dealing with isotropic, homogeneous plates are presented. The examples are chosen to demonstrate MIN3's performance in the critical patch tests and standard convergence problems.

Kirchhoff Patch Test The popular thin plate patch test [16] is undertaken. A thin, rectangular plate is subjected to the state of bending such that all three moment resultants are constant throughout the plate domain. The plate is discretized with four geometrically distinct MIN3 elements as depicted in Fig.4. The

model apparently recovers exactly the three constant moments in each of the elements.

Shear Patch Test Since the present refined theory involves transverse shear stress resultants, it may be desirable to conduct an appropriate shear patch test. To do this, we propose to examine an infinitely long clamped plate subjected to an uniform transverse line load , $Q_z$, along the long free edge (see Fig.5a). Here, the state of constant $Q_{xz}$-shear exists (i.e., $Q_{xz}=Q_z$), while $Q_{yz}$ is zero everywhere in the plate. The narrow strip four element model subject to symmetric boundary conditions (Fig.5b) produced exact constant values of $Q_{xz}$ and $Q_{yz}$ in every element.

Thin Square Plates Thin square plates are commonly used to assess convergence characteristics of plate bending elements. Figure 7 depicts convergence curves for problems solved with different mesh patterns, loadings and boundary conditions (refer to Fig. 6), where MIN3 solutions are compared with those of DKT (non-conforming, thin triangular element, [17]). The present element performance is seen to be very competitive.

7. CONCLUDING REMARKS. The anisoparametric finite element method, requiring two auxiliary kinematic field criteria, has demonstrated to be effective for the treatment of the penalty-constraint flexure formulations [11-15]. The method relies on an exact (normal) integration and , hence, assures consistent and kinematically reliable (correct rank) models. It takes advantage of an explicit dof reduction technique, achieved via 'continuous' shear edge constraints, which generates compatible kinematic fields and simple isoparametric-like nodal configurations. There exists no ambiguity regarding derivation of consistent load vectors and, in dynamics, consistent mass matrices. The anisoparametric elements suffer no deficiencies in either the classical (thin) or shear-weak (thick) flexure regimes and at the same time produce rapidly convergent solutions.

REFERENCES.


1. J. N. Reddy, "The Penalty Function Method in Mechanics. A Review of Recent Advances," ASME, Penalty-Finite Element Methods in Mechanics (Ed. J. N. Reddy) 51 (1982) 1-28.

2. S. P. Timoshenko, "On the Correction for Shear of the Differential Equation for Transverse Vibrations of Prismatic Bars," Philosoph. Magazine, 41 (1921) 744-746.

3. E. Reissner, "The Effect of Transverse Shear Deformation on the Bending of Elastic Plates," ASME J Appl. Mech. 67 (1945).

4. R. D. Mindlin, "Influence of Rotatory Inertia and Shear on Flexural Motions of Isotropic, Elastic Plates," ASME J. Appl. Mech. 18 (1951) 31-38.

5. P. M. Naghdi, "Foundations of Elastic Theory," Progress in Solid Mechanics ( Ed. I.N. Sneddon and R. Hill), IV, Chapter 1, North-Holland, Amsterdam (1963).

6. O. C. Zienkiewicz, The Finite Element Method (McGraw-Hill, London, 3rd ed., 1977).

7. T. J. R. Hughes, R. L. Taylor and W. Kanoknukulchai, "A Simple and Efficient Element for Plate Pending," Internat. J. Numer. Meths. Engrg. 11 (1977) 1529-1543.

8. E. D. L. Pugh, E. Hinton, and O. C. Zienkiewicz, "A Study of Quadrilateral Plate Pending Elements with 'Reduced' Integration," Internat. J. Numer. Meths. Engrg. 12 (1978) 1059-1079.

9. A. H. MacNeal, "A Simple Quadrilateral Shell Element," Computers and Structures 8 (1978) 175-183.

10. T. J. R. Hughes and R. L. Taylor, "The Linear Triangular Bending Element," in IV-MAFELAP 1981 (Ed. J. R. Whiteman), Academic Press, London (1982).

11. A. Tessler and S. B. Dong, "On a Hierarchy of Conforming Timoshenko Beam Elements," Computers and Structures, 14 (1981) 335-344.

12. A. Tessler, "An Efficient, Conforming Axisymmetric Shell Element Including Transverse Shear and Rotary Inertia," Computers and Structures 15 (1982) 567-574.

13. A. Tessler, "On a Conforming, Mindlin-Type Plate Element," in IV-MAFELAP 1981 (ed., J. R. Whiteman), Academic Press, London (1982) 119-126.

14. A. Tessler and T. J. R. Hughes, "An Improved Treatment of Transverse Shear in the Mindlin-Type Four-Node Quadrilateral Element," Comput. Meths. Appl. Mech. Engrg. 39 (1983) 311-335.

15. A. Tessler and T. J. R. Hughes, "Three-Node Mindlin Plate Element with Improved Transverse Shear," (to appear in Computer Meths. Appl. Mech. Engrg.).

16. J. Robinson, "Element Evaluation. A Set of Assessment Points and Standard Tests," Proc. F.E.M. in the Commercial Environment, 1, (1978) 217-248.

17. J. L. Batoz, "An Explicit Formulation for an Efficient Triangular Plate-Bending Element," Internat. J. Numer. Meths. Engrg. 18 (1982) 1077-1089.

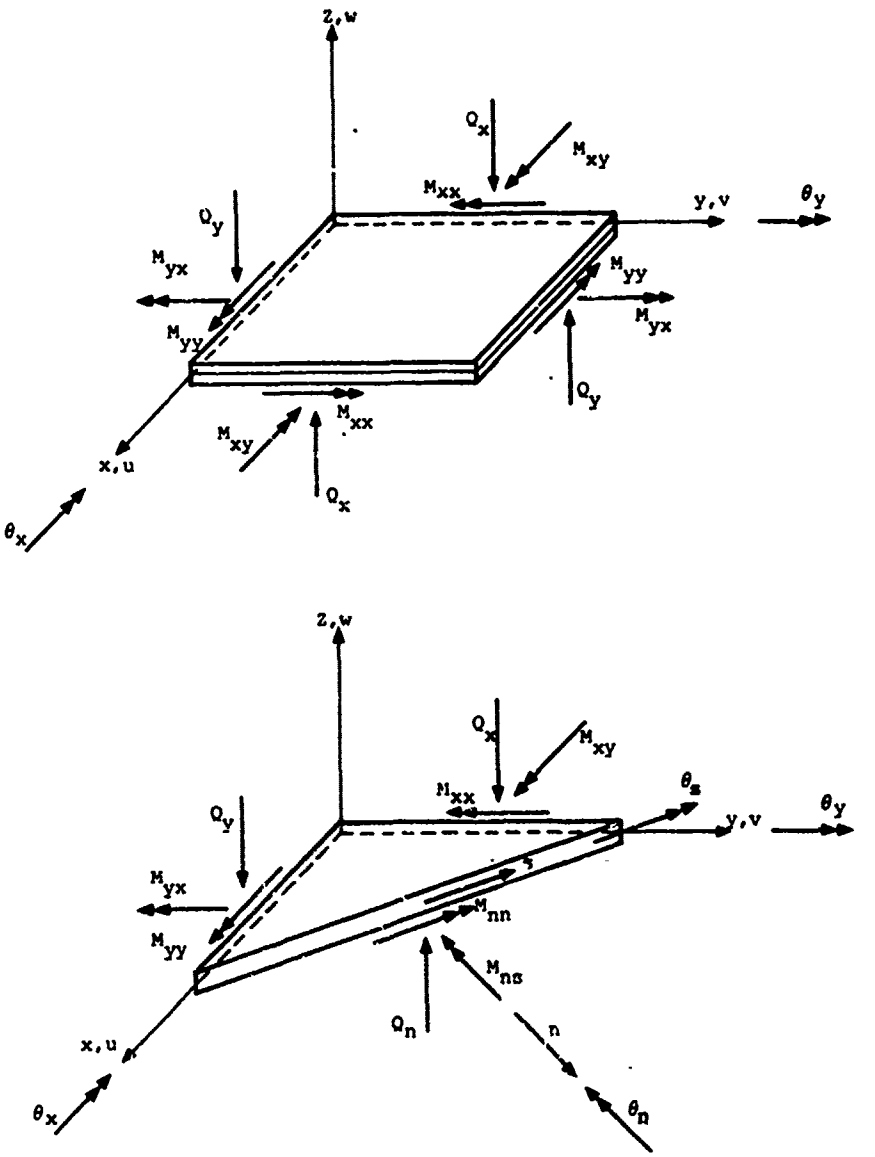
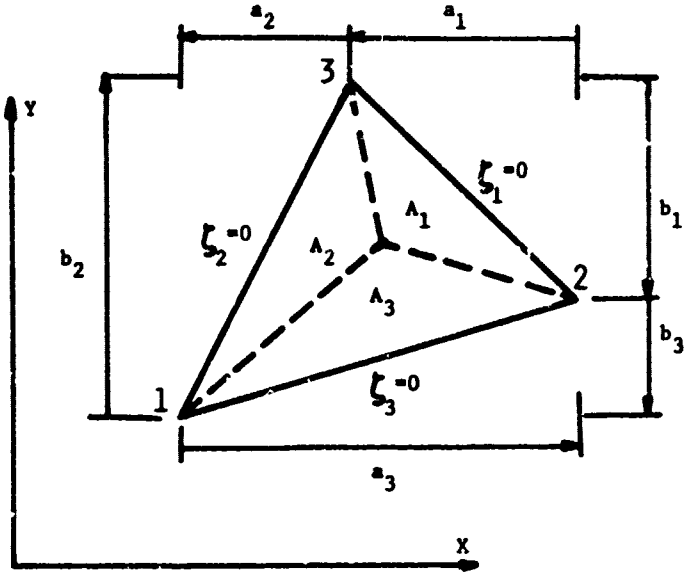


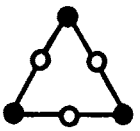
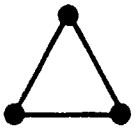

Fig. 1 Plate sign convention




$a_i = x_j - x_k$, $b_i = y_k - y_j$,

$(i = 1,2,3;\ k = 2,3,1;\ j = 3,1,2)$

Fig. 2 Triangular coordinate description

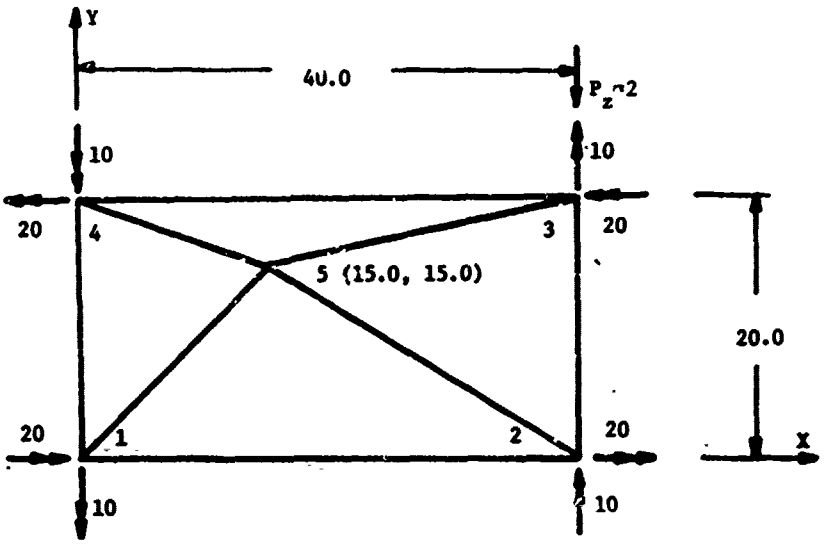
| SHAPE FUNCTIONS | | INITIAL NODAL CONFIGURATION | CONTINUOUS SHEAR EDGE CONSTRAINTS: $(w_{,s}+\theta_N)_{,s}=0$ | CONSTRAINED NODAL CONFIGURATION |
|---|---|---|---|---|
| $w$ | $\theta_x, \theta_y$ | | | |
| QUADRATIC | LINEAR |  | THREE EDGE CONSTRAINTS  | "MIN3"  |

KEY:

● $w, \theta_x, \theta_y$ DEGREES OF FREEDOM

○ $w$ DEGREES OF FREEDOM

Fig. 3 Anisoparametric triangular element



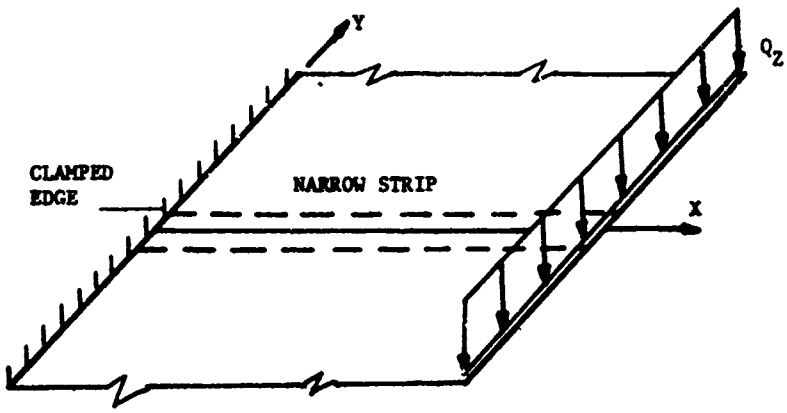

DATA: E = 1000.0, $\upsilon$ = 0.3, h = 1.0

BOUNDARY RESTRAINTS: w = 0 at nodes 1, 2 and 4.

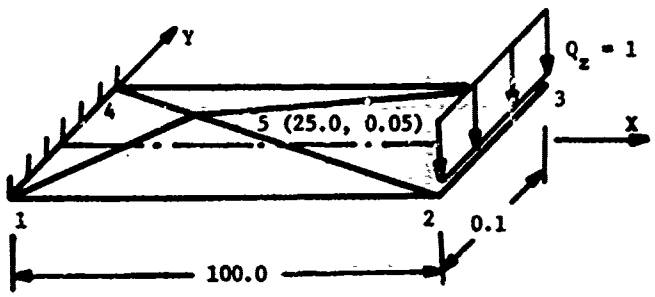
ANALYTIC SOLUTION: $M_{xx} = M_{yy} = M_{xy} = 1.000$ throughout the plate.

Fig. 4 Thin plate patch test

a. INFINITELY LONG CANTILEVERED PLATE




b. NARROW STRIP MODEL



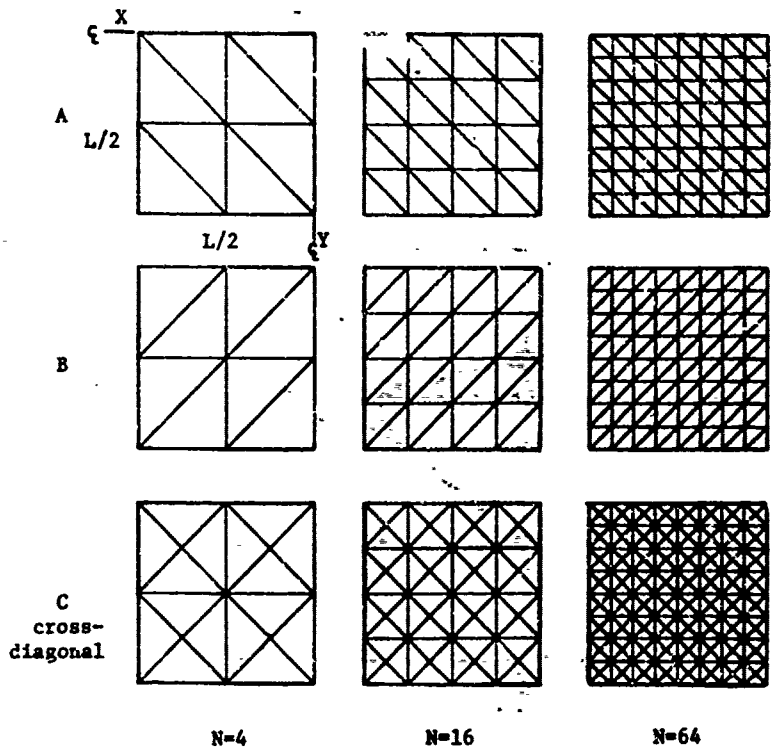

DATA: E = 1000.0, $\upsilon$ = 0.3, h = 0.1

BOUNDARY RESTRAINTS: $w = \theta_x = \theta_y = 0$ at nodes 1, 4;

$\theta_x = 0$ at nodes 2, 3.

ANALYTIC SOLUTION: $Q_{xz} = 1.0$, $Q_{yz} = 0.0$ throughout the plate.

Fig. 5 Transverse shear patch test




BOUNDARY RESTRAINTS:

SS -- simply supported boundary:

$$w(L/2, y) = w(x, L/2) = \theta_x(L/2, y) = \theta_y(x, L/2) = 0.$$

CL -- clamped boundary:

$$w(L/2, y) = \theta_y(L/2, y) = \theta_y(L/2, y) = 0,$$

$$w(x, L/2) = \theta_x(x, L/2) = \theta_y(x, L/2) = 0.$$

LOADING:

U -- uniform load; C -- center load.

Fig. 6 Square plate (one quadrant) discretization

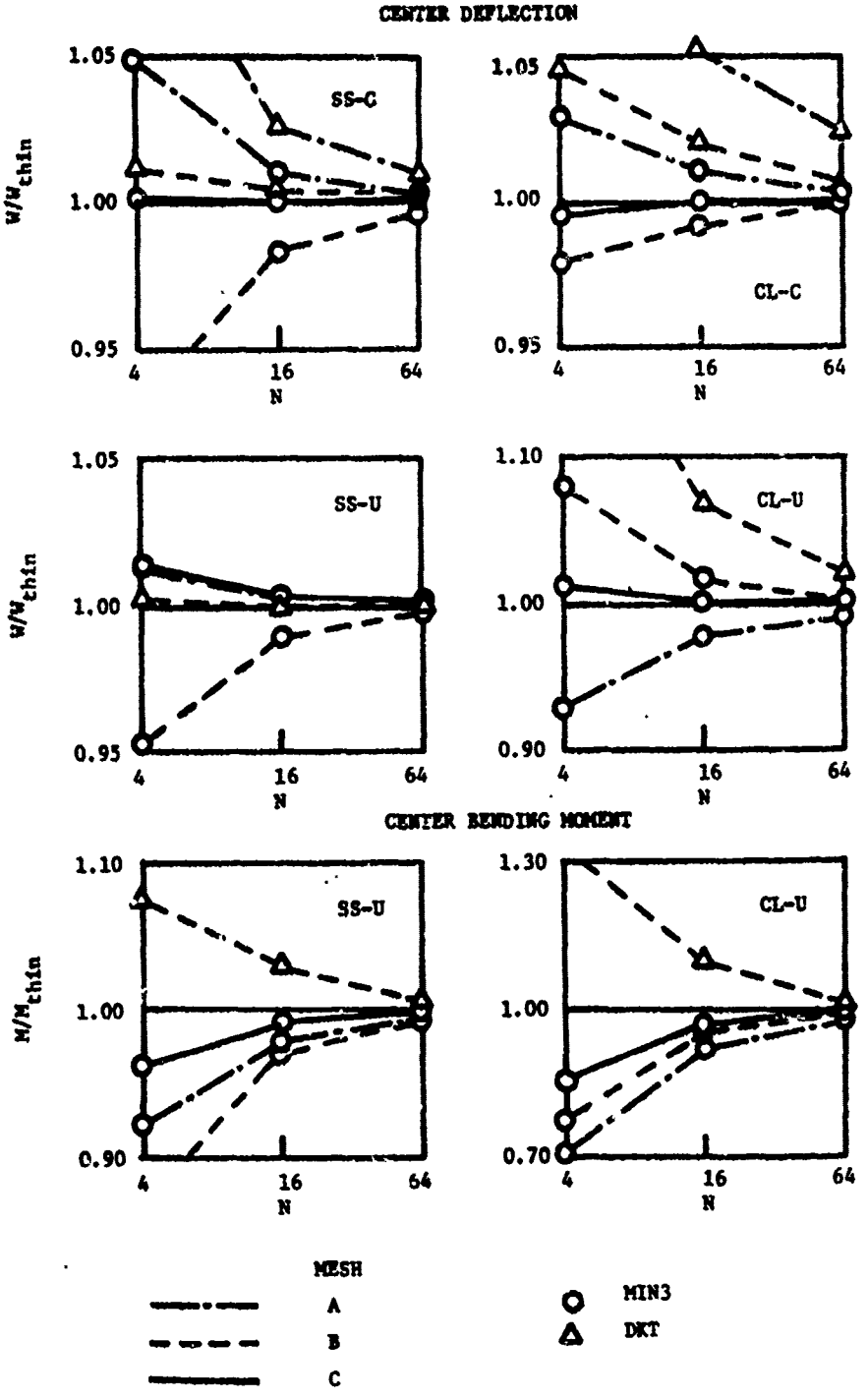



Fig. 7 Convergence study of thin square plates (L/h = 1000, $\upsilon$ = 0.3)

# IMPORTANCE OF CRACK TIP SHAPE IN ELASTIC-PLASTIC FRACTURE ANALYSIS

Dennis M. Tracey and Colin E. Freese

Mechanics of Materials Branch
Army Materials and Mechanics Research Center
Watertown, Massachusetts 02172

ABSTRACT: Elastic-plastic solutions for two crack problems are discussed. The problems involve blunt cracks. In one case, the crack tip geometry is semi-circular with parallel flanks; in the other case, a circular tip geometry is considered. We refer to them as the "U-tip" and "keyhole" crack problems. The cracks have finite length and are within an infinite domain under plane strain constraint. Far from the crack a uniform tensile stress state is specified. The solution history to a remote stress level equal to 28% of the yield stress was obtained for each problem. Slipline analysis suggests that the near tip stress distribution for these cases and others such as the narrow ellipse and perfectly sharp crack are significantly different. At least for the acuity and load level considered, the elastic-plastic analyses show that the U-tip and keyhole crack solutions are virtually identical. The slipline maximum stress value which is commonly used in fracture analysis is not achieved in our solutions. This is seen to follow from the fact that the top of the crack tip surface does not reach fully plastic conditions. Although the elastic-plastic boundary reaches the top, plastic straining occurs there only temporarily. Local unloading develops with the result that only the surface in the angular range $\pm 75^{\circ}$ is included in the zone of plastic flow at the later stages of loading.

INTRODUCTION: Criteria for ductile crack extension are often based upon the stress and deformation states suggested by continuum crack tip analysis. The shape of the crack tip can be an important aspect of such analysis, as can be readily seen from slipline solutions. However, in many cases, the sharp crack model is appropriate. Solutions then are singular at the crack tip and fracture criteria utilize singularity amplitudes such as K or J. When the opening at

the crack tip is significant on the microscale of the fracture investigation, the sharp crack singular solutions do not provide the necessary local detail, so that then solutions which account for the crack tip opening and shape are necessary.

There has been very little work done on the elastic-plastic blunt crack problem. Plasticity effects are generally discussed in terms of the local slipline predicitions, even though the limits of applicability of these predictions are often unknown. There has been noteworthy work published in the current literature on the solution ahead of cracks which have blunted as a result of large plastic deformations incurred during loading. In the present work, we are concerned with the flaw which is blunted in the stress-free state and which is not so sharp that finite deformation blunts it further in contained plasticity situations.

With numerical approaches such as the finite element method, the single most difficult feature of the blunt flaw problem is accommodating the relative smallness of the root radius in the discretization. In the analysis of engineering notches or even sharp cracks, the controlling dimensions are usually of the same order of magnitude. In contrast, for the cases that we discuss below, the root radius is smaller than the crack length by a factor of 2000. Since details of the stress variations in the region influenced by the root geometry is desired, grid dimensions must be fractions of the root radius.

We have developed a numerical formulation, Ref.(1), which can readily accommodate the elastic-plastic blunt crack problem, especially those cases which have plastic zone dimensions small compared to crack length. The formulation naturally treats infinite domains so that solutions free of extraneous boundary effects can be obtained. In a region surrounding the crack tip finite element methodology is used. The region size is chosen according to the plastic zone extent that is anticipated. Over the remainder of the domain, the planar elasticity complex stress function method is employed. Boundary collocation techniques are used to couple the equations governing in the two regions. At each step of the analysis, discrete unknownsare nodal displacement increments in the finite

element region and coefficients of a power series approximation to the governing stress function in the remainder of the domain. The unique value of the formulation lies in its capability to accurately treat an infinite elastic region with a single low order power series.

The elastic-plastic solution for a cracklike elliptical flaw under cylic loading was presented in Ref. (1). The non-hardening Prandtl-Reuss elastic-plastic constitutive law was used in the formulation. The flaw is cracklike in the sense that its length is much greater than the radius of curvature of its ends, $\rho$. The small finite element region used in the analysis restricted the remote stress level to 12% of the yield stress Y and the maximum plastic zone extent to $7.5\rho$. Near the crack tip the results were in excellent agreement with the appropriate slipline solution. Departures from the solution and the approach to the purely elastic distribution were discussed. As an example, at peak load with the elastic-plastic boundary at a distance $3.3\rho$ ahead of the flaw, the stress solution agreed with the slipline distribution over roughly one-half this plastic zone extent. While the agreement in this inner region was within 1%, at the elastic-plastic boundary the slipline stress result of 2.36Y was 13% too high. The results graphically demonstrate that slipline predictions are valid only over a portion of the plastic zone, essentially that portion far enough away from the elastic-plastic boundary so that fully plastic flow is kinematically possible.

Slipline analysis suggests that the stress gradient near a traction free circular boundary is more severe than that near an elliptical boundary. If a crack tip is formed by a semi-circle with parallel flanks at $\pm 90^{\circ}$ (U-tip), the normal stress rises from the surface value of 1.15Y to a stress plateau near 3Y, at a distance $3.8\rho$ ahead of the crack. At this location the stress value is 2.4Y for the cracklike ellipse. If the crack tip is a complete circle (keyhole crack), the stress can increase to a maximum of 4.78Y, the exact result depending upon the portion of the tip surface and surrounding material which yields and reaches fully plastic conditions. The slipline stress distributions for the cracklike ellipse, U-tip and keyhole cracks are compared in Fig.(1). In the following, we present

results from the elastic-plastic analysis of the latter two problems, delineating those aspects which define the limits of applicability of the slipline predictions.

FORMULATION: The numerical method used in this work has been discussed in Ref. (1). It combines features of the elastic complex variable stress function approach discussed by Bowie and Freese, Ref. (2), and the elastic-perfectly plastic incremental finite element approach discussed by the authors, Ref. (3). A very important aspect of the method for our cases of long, narrow flaws involves conformal mapping and analytic continuation. This allows the traction free crack boundary condition to be implicitly satisfied so that discretization and collocation are necessary only at the ends of the flaw. This very powerful aspect can be used if a mapping function $z = \omega(\zeta)$ is known which maps the unit circle onto the flaw. The mapping function for the elliptical flaw (semi-axis a,b) is of course known and this function applies to the limiting case of a slit or sharp crack with $b=0$. We use the slit in the present work to represent the crack in the elastic region. The desired blunt shape is taken to be connected to the slit in the finite element region. Whereas the general planar elasticity problem requires determination of two analytic stress functions, continuation reduces the problem to finding a single function $\phi(\zeta)$ which satisfies the remote stress condition and the equilibrium and compatibility coupling conditions along the finite element interface. While formally there is an interface around each end, symmetry allows us to treat a single quadrant of the plane in our analysis.

The interface is defined in the auxiliary plane as a circle of radius R centred at the end of the slit, so that the problem in the elastic region is to find $\phi$ outside the disks $|\zeta \pm 1| < R$. The approximation used had $\phi$ as the sum of the purely elastic function and a 15 term Laurent series expanded about the ends of the slit

$$\phi = T\,[a\zeta - 3a/\zeta\,]/8 + \sum_{n=1}^{15} \alpha_n \zeta\,/(\zeta^2 - 1)^n$$

The radius R was chosen in the present analysis so that the finite element region extended a distance of $75\rho$ from the flaw tip at $70^\circ$

above the flaw direction. In the physical plane the interface is a non-circular, smooth contour ending on the faces of the slit. The piecewise linear interface shown in Fig. (2) corresponds to the finite element edges there.

The finite element grids consisted of quadrilaterals, each divided by its diagonals into four constant state triangles to accommodate incompressible deformation. There were roughly 1000 nodes and 2000 triangles in each grid. The finite elements defined the blunt ends of the cracks. In the U-tip problem, the slit of the elastic formulation opens to a parallel faced, semi-circular ended slot of length $2\rho$, while in the keyhole problem a split circular boundary is connected to the slit. In both cases the tip radius was taken to be $a/1000$. While a more natural U-tip problem would have a uniform opening along the entire crack length, there is no known mapping function for this case. There is little reason to expect that the geometry used will have a near tip solution significantly different than this case.

The coefficients in the series are real numbers and they are determined by matching force and displacement conditions at the nodal points of the interface. Actually in the incremental analysis we deal with increments of force and displacements corresponding to remote stress changes $\Delta T$. Force increments are expressed in terms of $\alpha_n$ at the nodes and this represents the load transfer across the interface. The usual stiffness approach is followed to develop the governing equations in the finite element region. Condensation results in a system of equations relating the interface force and displacement increments. By expressing the latter in terms of $\phi$, a system is obtained exclusively in terms of $\alpha_n$. A series with fewer terms than finite element interface degrees of freedom is chosen so that the system is over-determined and the solution is by the least squares method. With knowledge of $\alpha_n$, displacement, strain and stress increments are computed throughout the finite element mesh. The value of $\Delta T$ at each step in the incrementation is adaptively established in an iterative fashion to accurately trace the details of the plastic yielding and flow.

ELASTIC-PLASTIC RESULTS: The adaptive incrementation technique determines load step size on the basis of a specified deviatoric stress change maximum which we chose to be equal to 0.05Y. The result was that 51 steps were taken for the plastic zone to extend to the interface. The 50th step is thus the last reliable one and it corresponded to a final remote stress of 0.28Y. First yield was at the triangle bordering the tip surface at the x-axis. The load at first yield suggests on elastic stress concentration factor of 68.5. The results for the two problems differed by only 0.2%. This factor can be compared to the value 64.2 for an elliptical flaw with the same $a/\rho$ ratio.

The grids used in the two analyses are shown in Fig. (3) along with the outlines of the plastic zones at maximum load. The grids are shown as quadrilaterals but of course each of these consists of four triangles. The elements which satisfy the yield condition are drawn in the first quadrants. There is remarkably close agreement between the results. Although the keyhole plastic zone extent appears greater in the 52.5 - 67.5° range, actually the two triangles which differ are within 1% of yield in the U-tip problem, corresponding to the fact that the load level is 0.7% lower in this case. Smoothing suggests that the maximum extent of the plastic zone is at 60° and approximately equal to $64\rho$. The plastic zone has the form associated with sharp cracks under small scale tensile yielding conditions. Larsson and Carlsson, Ref. (1), considered the problem of a sharp crack in a finite plate and found the maximum extent at 65° and equal to 0.23 $(K/Y)^2$. K is the elastic stress intensity factor which would equal $T\sqrt{\pi a}$ for our problem. This result suggests an extent of $57\rho$, 11% less than our result. Whereas the plastic zone extends a distance $8.8\rho$ along the x-axis in our problem, the sharp crack result suggests $5.5\rho$.

A very significant aspect of the solution involves elastic unloading near the top of the crack tip. In the early stages of loading there is a steady advance of the elastic-plastic boundary along the tip surface. Yielding reaches the top (half of the quad between 82.5 and 90° yields), plastic straining temporarily occurs near the top and then elements between 67.5 and 90° unload. It appears that the yielding away from the surface, beyond the 90° ray, effectively serves as a load shedding mechanism. The small unloaded region can be seen in Fig. (4) where plastic zones at T = .14Y (prior to unloading) and maximum load are compared. The scale is such that only the local plastic zone is included for the maximum load. The keyhole results are identical to those shown in the figure.

The near tip elastic-plastic stress distributions for the U-tip and keyhole problems are displayed in Fig. (5) for six stages of loading. The results are virtually identical for the two problems. The plot shows $\sigma_{yy}/Y$ vs. $x/\rho$ where x is measured from the circular surface. The solid curve in the figure is the logarithmic spiral slipline result,

$$\sigma_{yy}/Y = 1.15\ (1 + \ln\ (1 + x/\rho))$$

which is plotted out to $x/\rho = 3.81$ as is appropriate for the U-tip flaw when yielding reaches the flanks. If yielding spreads along the flanks, the stress beyond $3.81\rho$ remains constant at 2.97Y. The log spiral from $90^\circ$ is drawn in the figure.

The position of the elastic-plastic boundary on the x-axis is indicated by hatch marks in the figure. As load increases from the incipient yield value of 0.17Y to 0.15Y, it is seen that the slipline prediction applies over an increasing region. Beyond 0.15Y this behavior ceases and the slipline distribution applies only out to the fixed location $x = 2.18\rho$. The log spiral which intersects the x-axis there is drawn and it is seen to intersect the crack tip at $66^\circ$. This behavior reflects the unloading from $90^\circ$ to the $67.5 - 75^\circ$ range that has been discussed. The slipline drawn effectively defines the limit to the log spiral region for our case of small scale yielding. Fully plastic conditions can develop beyond this slipline, of course, but the log spiral distribution cannot apply there. At higher load levels the solution beyond $x = 2.18\rho$ appears to be approaching a steady state, with a maximum stress of 2.57Y at $x = 2.76\rho$.

CONCLUSIONS: The results have important implications in many areas of fracture analysis. For instance, the prediction of cleavage ahead of a crack is often based upon the characteristic fracture stress of a material. The results which show a maximum achievable stress different than commonly thought is very significant to this work. Another area of application of the results is in the assessment of fracture ahead of machined flaws in fracture testing. Analytically, it remains to be seen if the crack tip unloading phenomenon influences behavior once finite strains develop at the tip. The effect of strain-hardening on the observed behavior is another area for investigation.

## REFERENCES


1. Tracey, D. M. and Freese, C. E. (1982), "Cyclic Plasticity Near A Cracklike Elliptical Flaw," Mechs. Materials, 1, 151-159.

2. Bowie, O. L. and Freese, C. E. (1978), "Analysis of Notches Using Conformal Mapping," in Mechanics of Fracture 5: Stress Analysis of Notch Problems, ed. Sih, G. C., Noordhoff, 69-134.

3. Tracey, D. M. and Freese, C. E. (1981), "Adaptive Load Incrementation in Elastic-Plastic Finite Element Analysis," Computers and Structures, 13, 45-53.

4. Larsson, S. G. and Carlsson, A. J. (1973), "Influence of Non-Singular Stress Terms on Small Scale Yielding at Crack Tips in Elastic-Plastic Materials," Jour. Mechs. Phys, Solids, 21, 263-277.

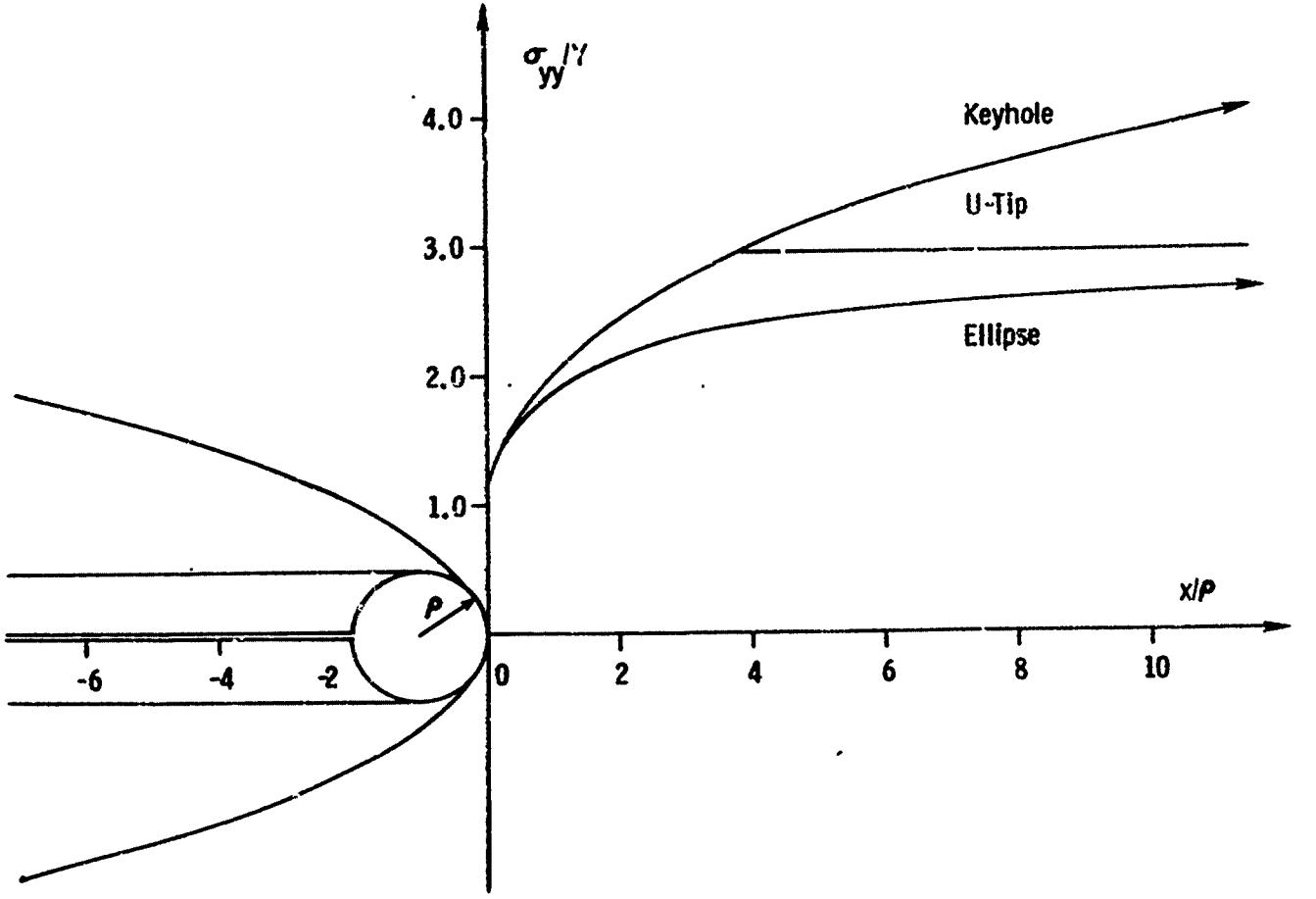
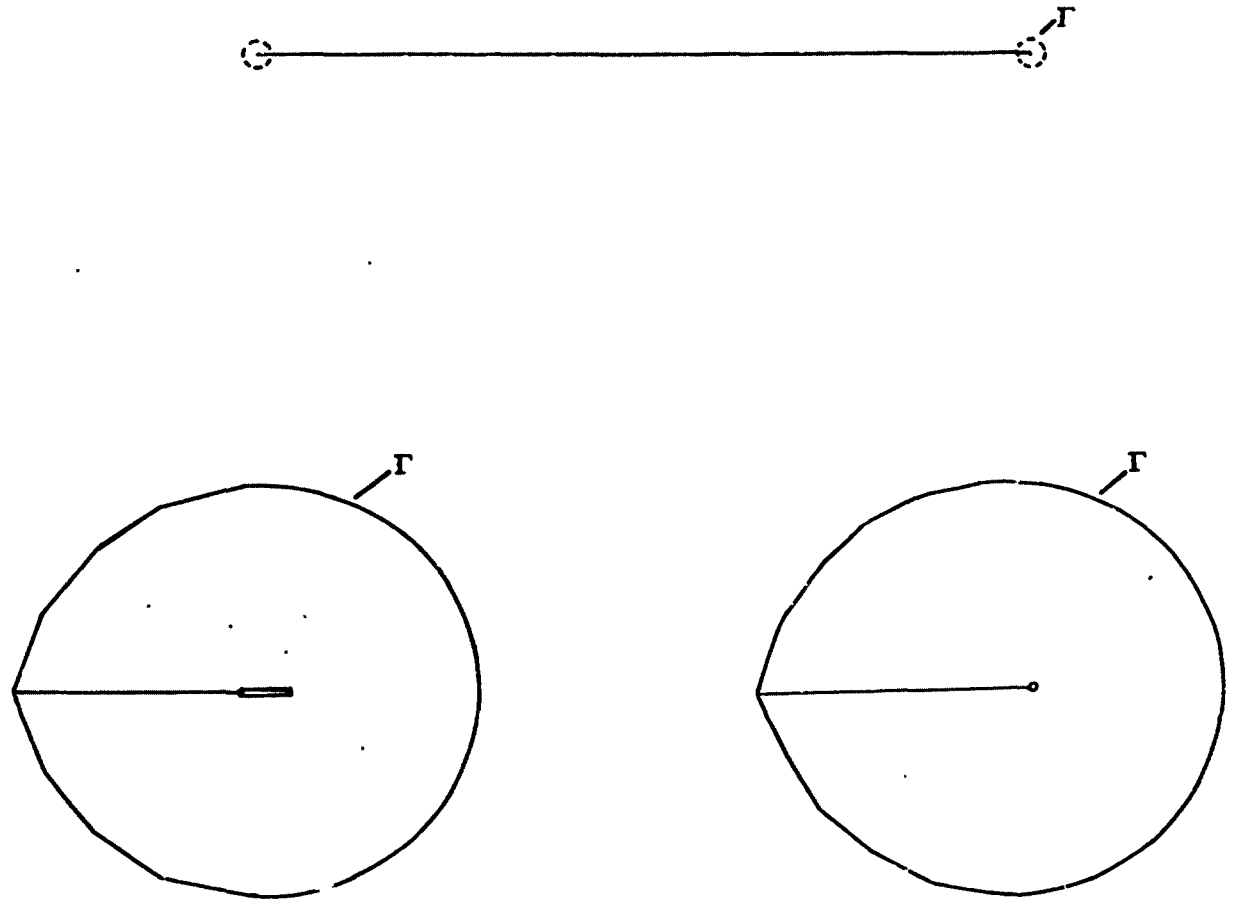


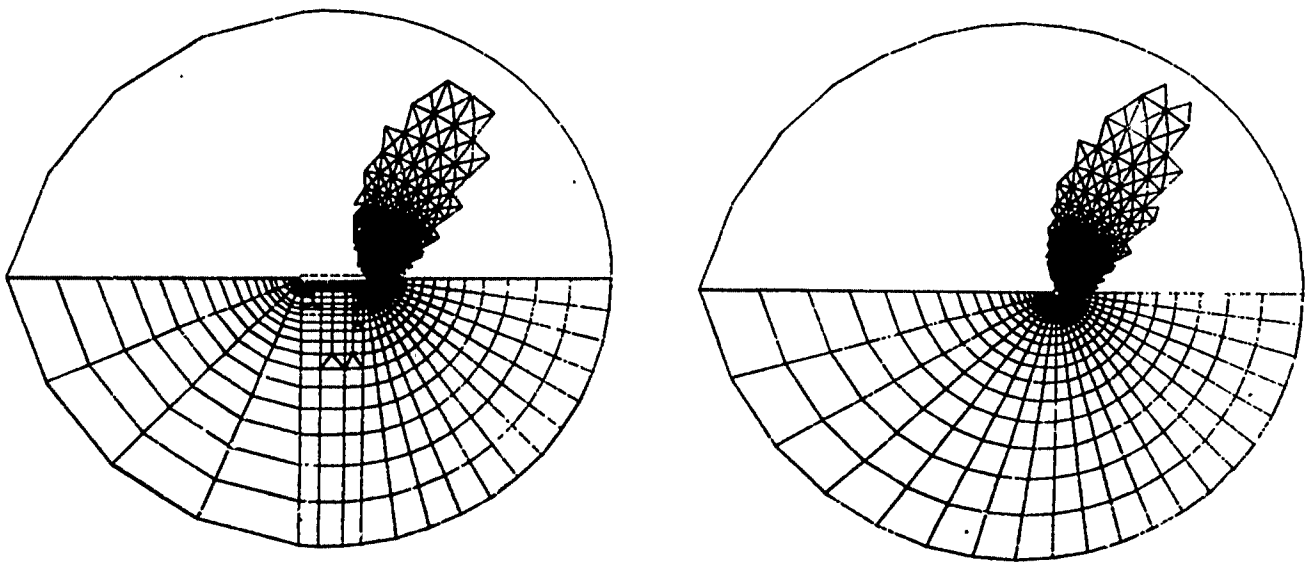

FIG. 1 NEAR TIP STRES_ DISTRIBUTIONS FOR THREE CRACK MODELS, FROM SLIPLINE THEORY




U-TIP MODEL

KEYHOLE MODEL

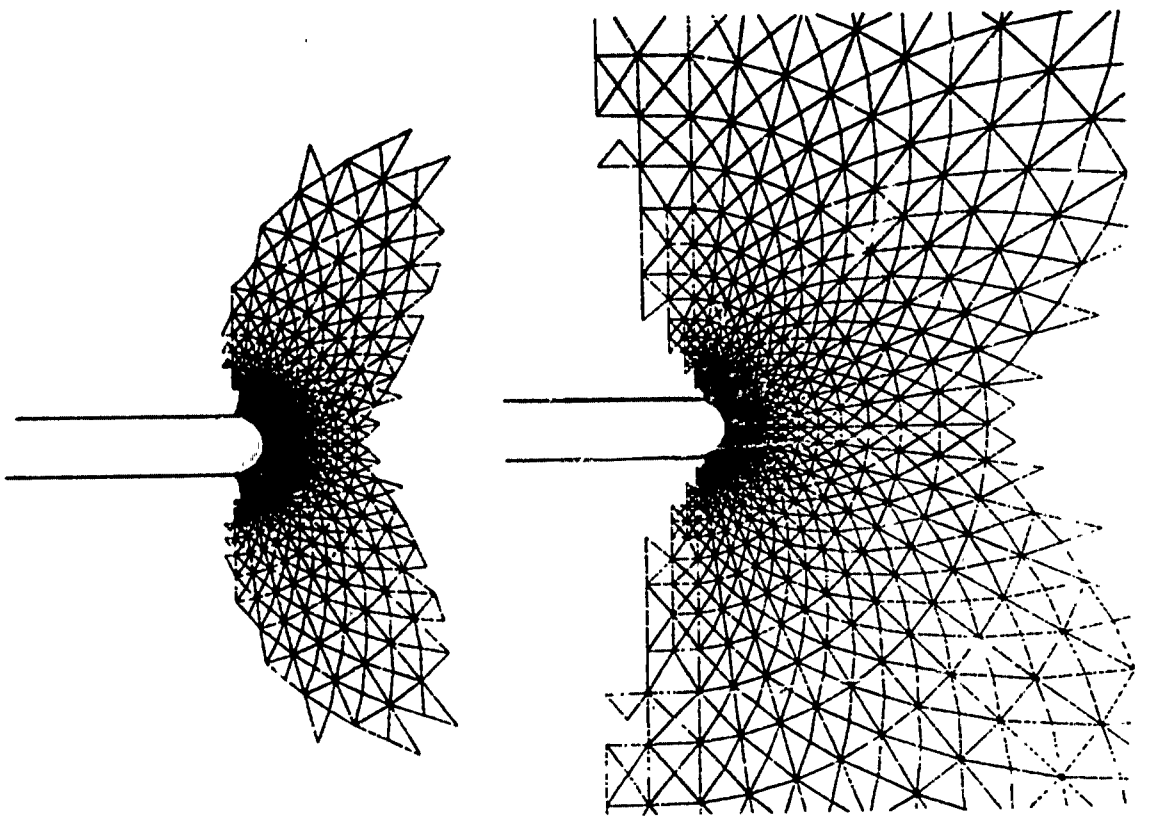
Figure 2. FINITE ELEMENT REGIONS FOR U-TIP AND KEYHOLE CRACK PROBLEMS



U-TIP CRACK

KEYHOLE CRACK

FIGURE 3. PLASTIC ZONES AT A REMOTE STRESS LEVEL 28% OF YIELD STRESS



PLASTIC ZONE T=.14Y

PLASTIC ZONE T=.28Y

FIGURE 4. ELEMENTS SATISFYING YIELD INDICATING UNLOADING NEAR TOP SURFACE

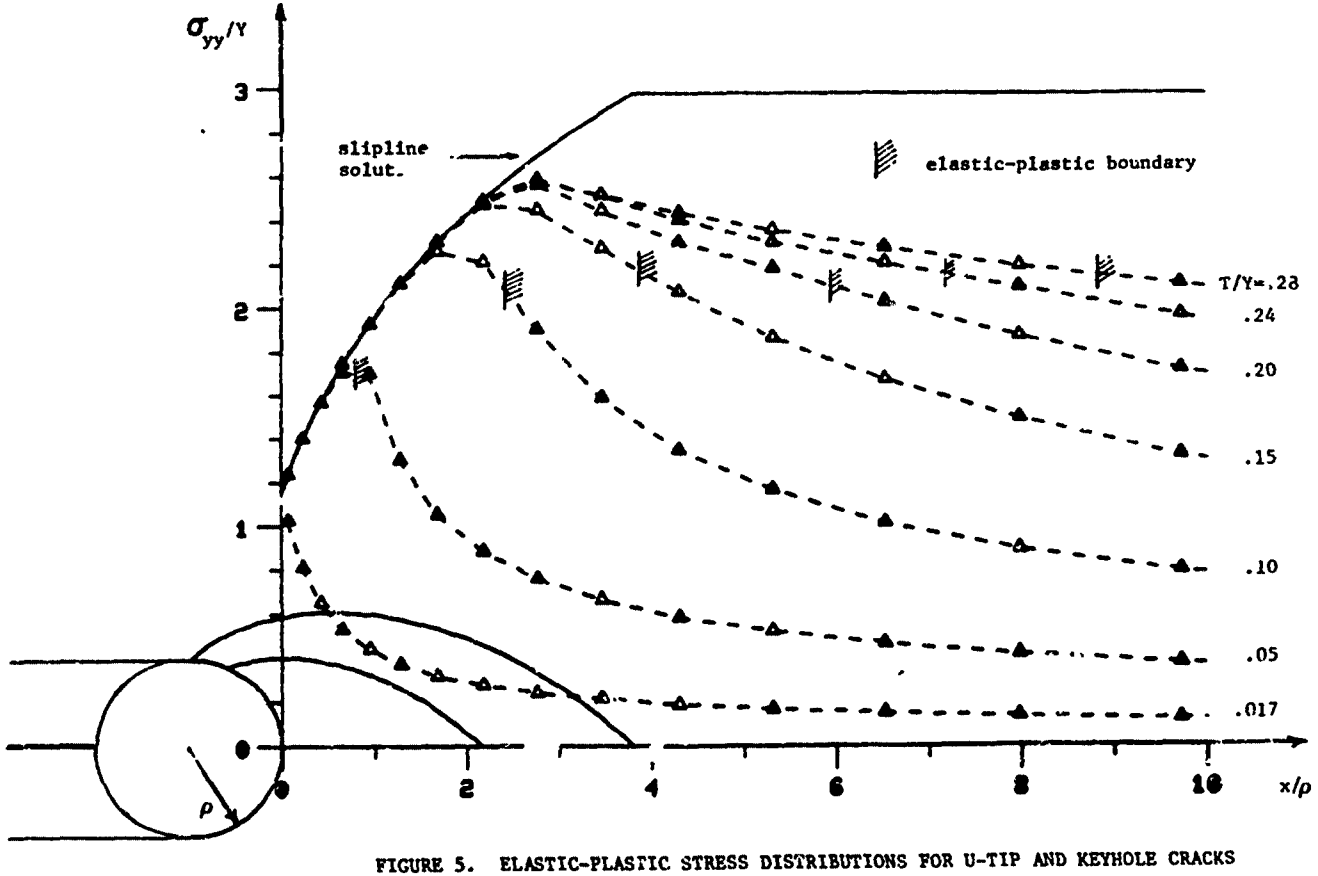



FIGURE 5. ELASTIC-PLASTIC STRESS DISTRIBUTIONS FOR U-TIP AND KEYHOLE CRACKS

PREVIOUS PAGE IS BLANK

# KINEMATIC HARDENING APPLIED TO NON-PROPORTIONAL LOADING

Charles S. White
First Lieutenant

Army Materials and Mechanics Research Center
Watertown, Massachusetts 02172

## Abstract

A series of critical experiments in the determination of a plastic flow rule is re-examined using kinematic hardening. Non-proportional loading experiments on thin-walled, aluminum tubes were conducted by Budiansky, Dow, Peters, and Shepherd in 1951 to determine whether plastic flow exhibits behavior consistent with the physical, slip theory of plasticity or with the phenomenological, $J_2$ flow and deformation theories. Their results were mixed since none of these theories predict the full range of exhibited, material behavior. Pan and Rice have sparked recent interest in these experiments by introducing a slight rate dependence into slip theory. Through a judicious choice of a strain rate sensitivity parameter they match the experiments reasonably well.

This note reports on the comparison of these experiments against the predictions of a flow rule based on the Prager/ Ziegler kinematic hardening theory. Both shear and axial strains are predicted for a variety of load histories. Results show good agreement between theory and experiment. The implications to buckling and instability analysis are briefly discussed.

## I. INTRODUCTION

In this note a brief reassessment of some experiments [1] on plastic flow rules will be made in light of the results predicted using Prager/Ziegler kinematic hardening. First the experiments will be described with the original comparisons using the models of $J_2$ deformation and flow theories (isotropic hardening) and slip theory. Then the recent paper by Pan and Rice [2] employing rate sensitivity in the slip theory will be shown to improve the modeling. Finally some recent calculations using simple kinematic hardening will be presented and discussed.

At the First National Congress of Applied Mechanics in 1951 the results of some nonproportional loading experiments on thin wall tubular test specimens were presented. Budiansky, Dow, Peters, and Shepherd [1] conducted tests at NACA labs at Langley Field to investigate the behavior of

plastic flow near the point of a change in loading direction. They compressed tubes of 14S-T4 aluminum alloy into the plastic range to strains of about 0.5% then abruptly changed the loading path and continued loading at a fixed ratio of axial stress increment to shear stress increment, $d\sigma/d\tau$ . Their intent was to look at shear and axial strain response just after this loading corner. The simple plasticity theories in use at that time predict quite different strain behavior. The $J_2$ isotropic, flow theory contains a smooth yield surface so it predicts that the initial shear strain response would be elastic at a change in the loading direction.

The $J_2$ deformation theory and the then recently proposed slip theory predict the immediate accumulation of plastic flow. They both predict a tangent modulus which is reduced from its elastic value by the formation of a corner on the yield surface. By determining which theory better approximated the experiments, the authors hoped to explain why plastic buckling experiments agreed better with calculations using a reduced tangent modulus while the body of experimental evidence had supported the model of a smooth yield surface.

Their results are not repeated here in detail except to describe the general trends and the authors conclusions. In each specimen the initial shear response was elastic. For all ratios of $d\sigma/d\tau$ the elastic shear strain accounted for all the measured shear strain just after the loading corner. This observation is in accordance with $J_2$ isotropic, flow theory or any flow theory having a smooth yield surface.

For continued straining the results did not favor one theory over another. The experiments showed shear response that was "softer" than predicted by isotropic hardening flow theory but was "stiffer" than predicted by slip or deformation theories. The experimental results fell between the predictions.

For most of the cases, the continued accumulation of plastic strain after the loading corner is underestimated by these theories. One explanation lies with the treatment of the behavior of this aluminum alloy as rate independent at room temperature. The tests were run at a very slow strain rate ( $\sim 10^{-6}$ sec $^{-1}$ ) and a component of creep strain might be expected. This would lead to an increase in the axial strain over the predictions of the rate independent theory.

## II. RATE DEPENDENT SLIP THEORY

In a 1983 paper by Pan and Rice [2], recent interest was shown in these experiments. Rate dependence was used to improve the predictions of slip theory. Pan and Rice investigated the implications of introducting a slight rate

dependence into the simple slip theory of Batdorf and Budiansky [3]. The original assumption was that the shear strain on any slip system in a crystal is a function only of the maximum resolved shear stress on that system over the loading history. This leads to a rate independent theory for the macroscopic constitutive behavior when integrated over all slip directions and slip systems.

Pan and Rice assumed that the microscopic behavior is slightly rate dependent through a non-linear viscous relation:

$$\dot{\gamma} = \dot{a}\left(\frac{\tau}{g(\gamma)}\right)^{1/m} \qquad (1)$$

where $\gamma$ is the plastic shearing rate, m is the plastic strain rate sensitivity, $\dot{a}$ is the reference plastic shearing rate and $g(\gamma)$ is a function of the current state. Note that $g(\gamma)$ is just the function for $\tau$ when $\dot{\gamma} = \dot{a}$.

Several values for m were chosen since separate tests for strain rate sensitivity had not been conducted. The value of m which gave the best matching with the nonproportional tests was Ø.Ø3. This value is a little higher than one would normally expect for aluminum at room temperature [21].

Pan and Rice show results for 3 of the 6 experiments conducted by Budiansky et. al. In each case they show that the introduction of rate dependence can greatly increase the agreement of slip theory with the experiments. Their results are repeated here in Figures 1-3. Note that the original results of Budiansky et. al. are also plotted. An initial elastic shear stress-strain response is predicted at the loading corner in accordance with observations. The continued deformation is also predicted quite well although there is some divergence between theory and experiment at higher strain. By judicious choice of the strain rate sensitivity parameter, rate dependent slip theory can be shown to give a good description of these nonproportional loading experiments.

For detailed calculations on buckling and other instability phenomena the information obtained from these types of tests are crucial. The predicted loads are very sensitive to the transverse stiffness after longitudinal plastic straining. The rate dependent slip theory is shown to provide a good model which can match experiments quite well by adjusting the strain rate sensitivity. This sort of a microscopically based model is useful when considering simple geometries and homogenous stress states but is far too computationally expensive for use in general analysis such as might be conducted using a finite element code. It is for this reason that this author has examined simple, kinematic hardening in the context of nonproportional loading.

## III. KINEMATIC HARDENING

Budiansky et. al. [1] remarked that their data might be best correlated by a linear flow theory whose loading function gives a higher curvature to the loading surface at the prestress point that does isotropic hardening. The simple kinematic hardening model proposed by Prager [4] satisfies just such a set of conditions, the curvature being given by that of the initial yield surface. Prager first introduced the concept of a translating yield surface in 1955 so the model was not available to Budiansky et. al. at the time they analyzed these experiments. It appears that nonproportional loading experiments of this type have never been examined with the kinematic hardening model. After the early 1950's, the experimental emphasis in biaxial plasticity turned away from studying flow rules to plotting yield loci. Kinematic hardening concepts have been successfully used to describe some of the phenomena associated with yield surface movement but as a flow rule the theory has not been subject to the same experimental scrutiny.

Without going into a detailed discussion of the development of this phenomenological theory, a few remarks are appropriate. The theory considered here is that proposed by Prager [4] and later modified by Ziegler [5]. Restricting ourselves to small strains we consider an initial yield surface of the von Mises type which retains its size and shape but translates without rotation during plastic straining. The flow rule is associative and the evolution law for the position, in stress space, of the yield surface center is given by

$$\dot{\underset{\sim}{\alpha}} = \mu(\underset{\sim}{S}-\underset{\sim}{\alpha}) \tag{2}$$

where $\underset{\sim}{S}$ is the stress deviator and $\mu$ is the scalar function, derivable from the consistency condition, which describes the hardening behavior. This theory was applied to the experiments of Budiansky et. al. A power law form was applied to match the standard uniaxial stress strain curve in compression given in [2].

$$\varepsilon^{p} = c\,(\frac{\sigma}{\sigma_{o}} - 1)^{n}$$

$$\text{where} \quad \begin{aligned} \sigma_{o} &= 25 \text{ ksi} \\ n &= 3.33 \\ c &= 0.0317 \end{aligned} \tag{3}$$

These values gave a very good match to the compression experiment and provided easy evaluation of the stiffness at any strain level during the nonproportional test. In order to account for slight differences in material properties between specimens, we adopt the same approach as Budiansky et. al. During the pure compressive loading portion of each

test the uniaxial stress strain curve was compared with of the standard curve. The ratio, denoted by $\lambda$, of the stress given by the standard curve to that of the individual specimens during the compressive loading was used to modify the expression above. They assumed that the plastic strain would be a function of $\lambda$ times $\sigma$.

$$\varepsilon^p = c\left(\frac{\lambda\sigma}{\sigma_o} - 1\right)^n \qquad (4)$$

where

$$\lambda = \frac{(\sigma_x)\ \text{standard curve}}{(\sigma_x)\ \text{specimen}}$$

This allowed the same uniaxial stress strain relation to be used for all the specimens even though small differences in the flow stress level were exhibited between specimens. The values of $\lambda$ were determined from the compressive loading portion of the tests. They are tabulated in [1].

The kinematic hardening relations were coded using a one-step, Euler explicit integration scheme. The step size was varied to study error accumulation. Increasing the number of equal steps from 100 to 1000 changed the final plastic strain by less than 1%. Since error varies inversely with step size in a linear fashion for explicit Euler 1000 steps was considered sufficient for predictions within experimental accuracy. The same procedure was also used to integrate small strain, isotropic hardening relations for comparison.

The six loading histories tested by Budiansky et. al. were considered. They are shown schematically in Figure 4. Notice that the ratios of $d\sigma/d\tau$ for continued loading varied from +1.91 to -1.13. This covered the range from total loading to elastic unloading.

Figures 5-10 show the results for the isotropic and kinematic models compared to the experimental data. Shear stress versus plastic shear strain and versus the increase in plastic axial strain are plotted. Notice in each case that the kinematic hardening model matches the shear strain response very well. The kinematic model accurately predicts the softer response for shear following axial extension. The most interesting point is that the kinematic model does such a good job of predicting when plastic flow will recommence for the two cases when the loading trajectories go back through the elastic zone of the kinematic and isotropic models ($d\sigma/d\tau = -0.656, -1.13$). This is clearly seen from Figure 4 where the shear stress levels for the intersection of the loading path with the yield surface is much different for the two models. The kinematic model yields results much closer to experimental observation. This is an example of how non-proportional loading tests are valuable and necessary in constructing a flow rule.

An interesting behavior is predicted for the case $\frac{d\sigma}{d\tau} = -1.13$. Figure 10 shows that the axial strain changes direction for the kinematic hardening model. This is a result of the yield surface translating far enough to the right that the loading point has moved around to the left half of the yield surface. Unfortunately, the experiments were not run far enough to show whether this behavior would occur. The kinematic model does a good job of predicting these axial strains. The kinematic hardening model does better than the other theories applied to this problem. It predicts more axial straining than the other theories and provides a good overall match with experiments.

## IV. CONCLUSIONS

The calculations presented here demonstrate that although most plasticity theories yield identical results when applied to proportional load histories, the change in loading direction can greatly affect the predicted material response. In particular, simple kinematic hardening was shown to provide much better agreement with experiment than isotropic hardening. In light of the overwhelming use of isotropic hardening in even this small strain regime the analyst must use care in applying a particular hardening model. The results presented here indicate that kinematic hardening should be a more suitable model for buckling or bifurcation studies. In fact, Tvergaard [6] showed that large strain, kinematic hardening provided good results for biaxial necking.

## References:


1. Budiansky, B., Dow, N., Peters, R. and R. Shepherd, "Experimental Studies of Polyaxial Stress-Strain Laws of Plasticity", Proc. 1st. U. S. Natl. Congr. Appl. Mech., 1951, pp. 503-512.

2. Pan, J., and J. R. Rice, Int. J. Solids Structures, Vol. 19, 1983, pp. 973-987.

3. Batdorf, S. B. and B. Budiansky, NACA TN 1871, April 1949.

4. Prager, W., J. Appl. Mech., Vol. 23, 1956, pp. 493-496.

5. Ziegler, H., Quar. Appl. Math., Vol. 17, 1959, pp. 55-65.

6. Tvergaard, V., Int. J. Mech. Sci., Vol. 20, 1978, pp. 651-658.

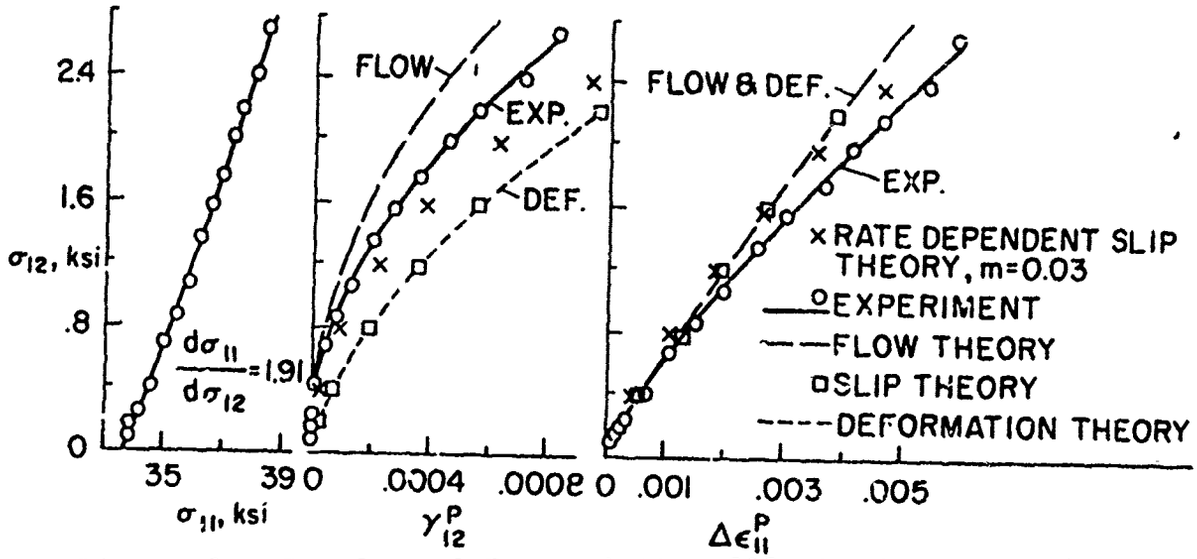



Figure 1. Results of Pan and Rice [2] showing comparison of the various theories with the experiments of Budiansky et. al. [1] for $d\sigma/d\tau = 1.91$ . Note that they use the notation of $\sigma_{11} = \sigma$, and $\sigma_{12} = \tau$.

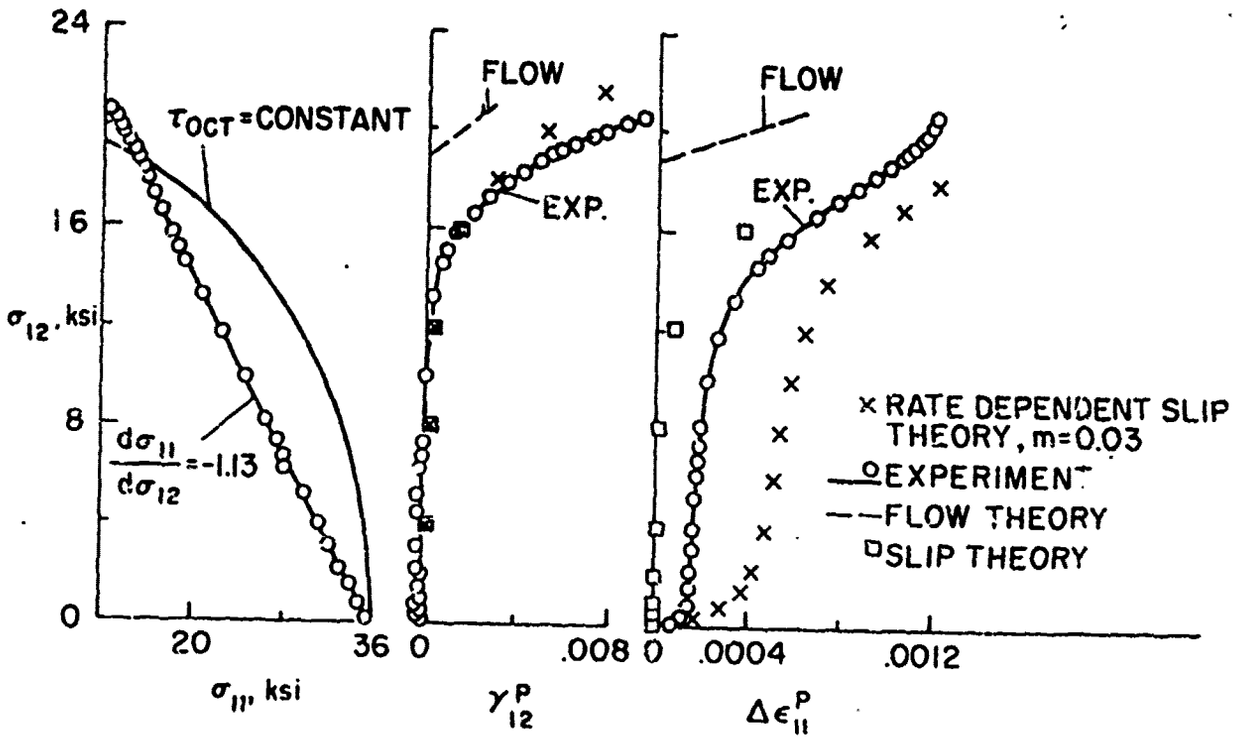



Figure 2. Results of Pan and Rice [2] showing comparisons among the various theories for $d\sigma/d\tau = -1.13$ .

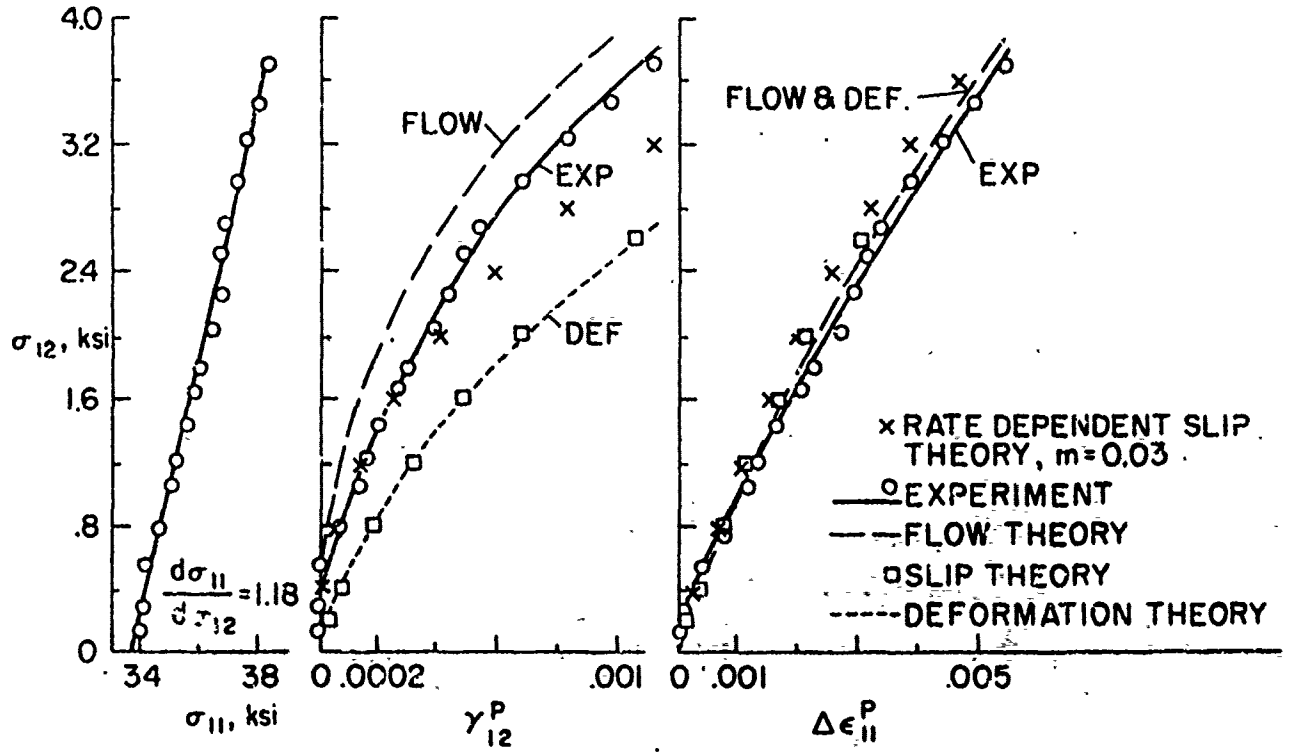



**Figure 3. Results of Pan and Rice [2] showing comparisons among the various theories for $d\sigma/d\tau = 1.18$ .**

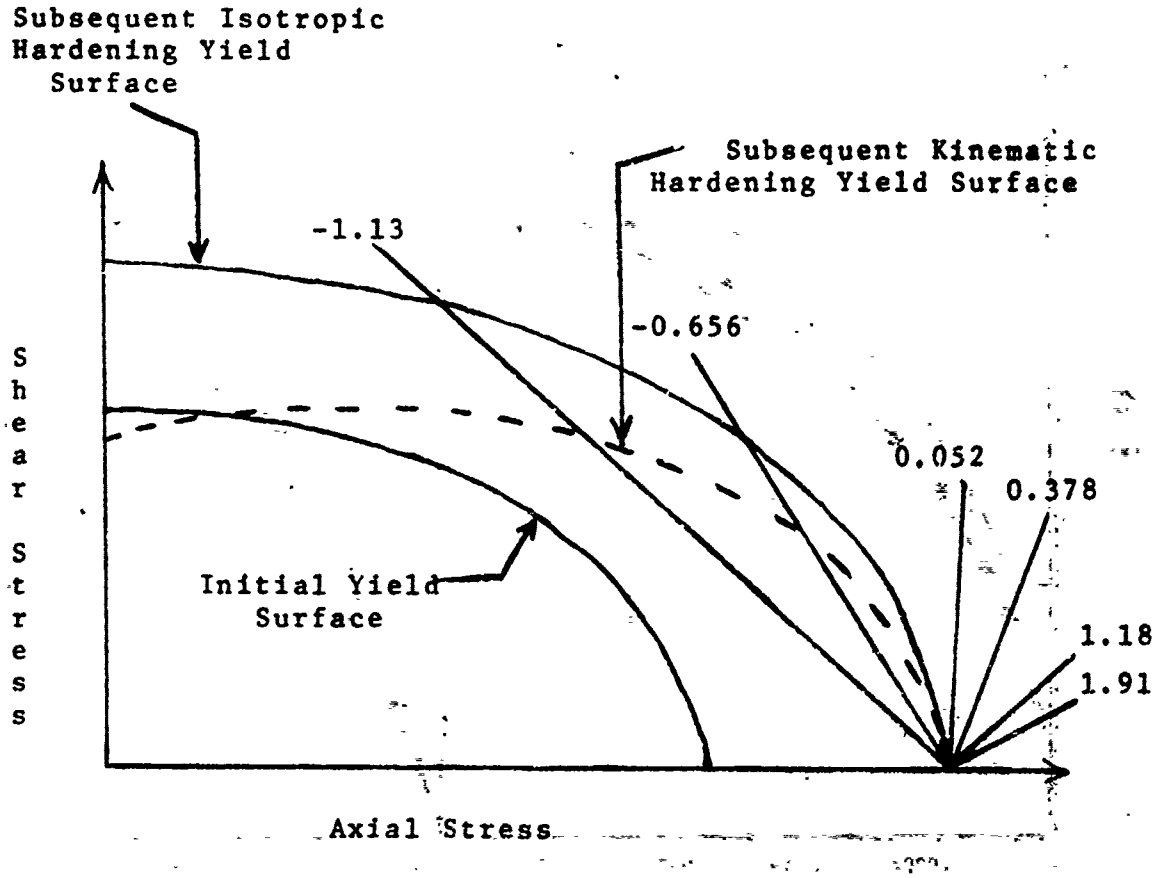



Figure 4. Comparison of initial yield surface with that predicted after initial axial compression using both isotropic and kinematic hardening. The subsequent loading trajectories are also shown with the corresponding values of $d\sigma/d\tau$.

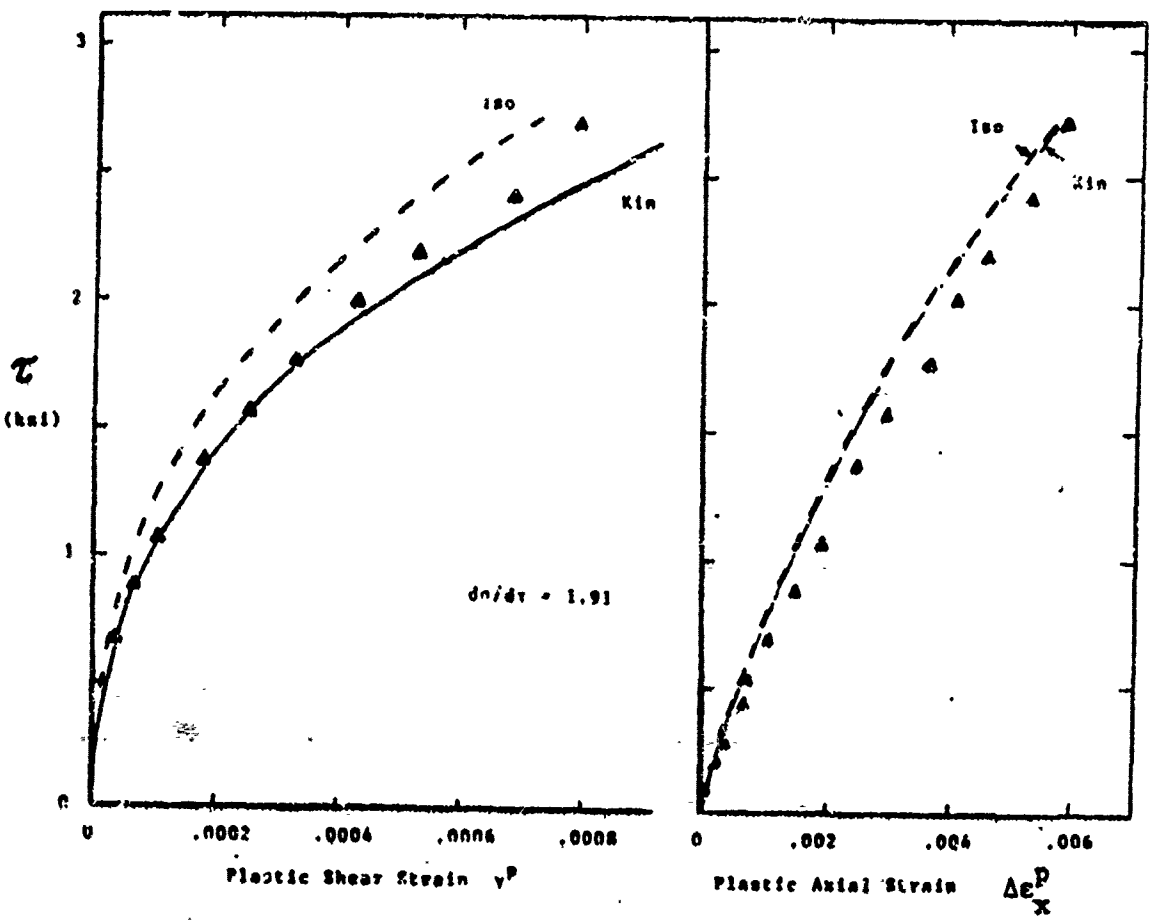



Figure 5. Comparison of isotropic and kinematic hardening flow theories with experiments of Budiansky et. al. [1]. Filled triangles indicate measurements of plastic flow along $d\sigma/d\tau = 1.91$ after initial axial compression.

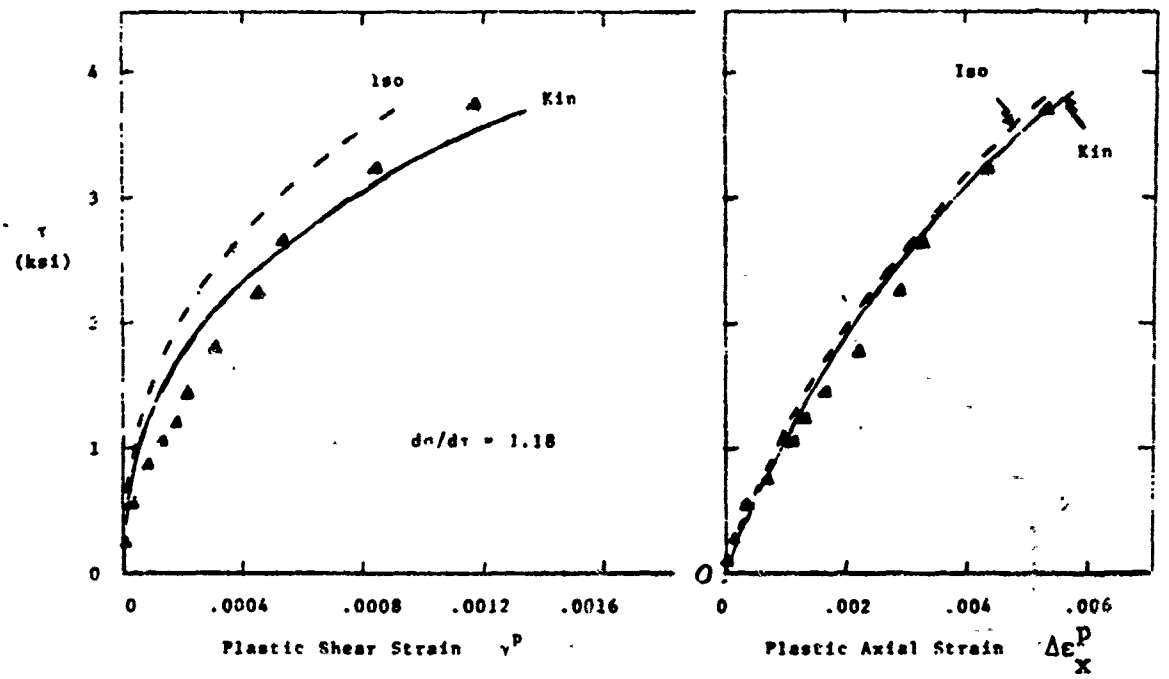



Figure 6. Comparison of isotropic and kinematic hardening flow theories with experiments of Budiansky et. al. [1]. Filled triangles indicate measurements of plastic flow along $d\sigma/d\tau = 1.18$ after initial axial compression.

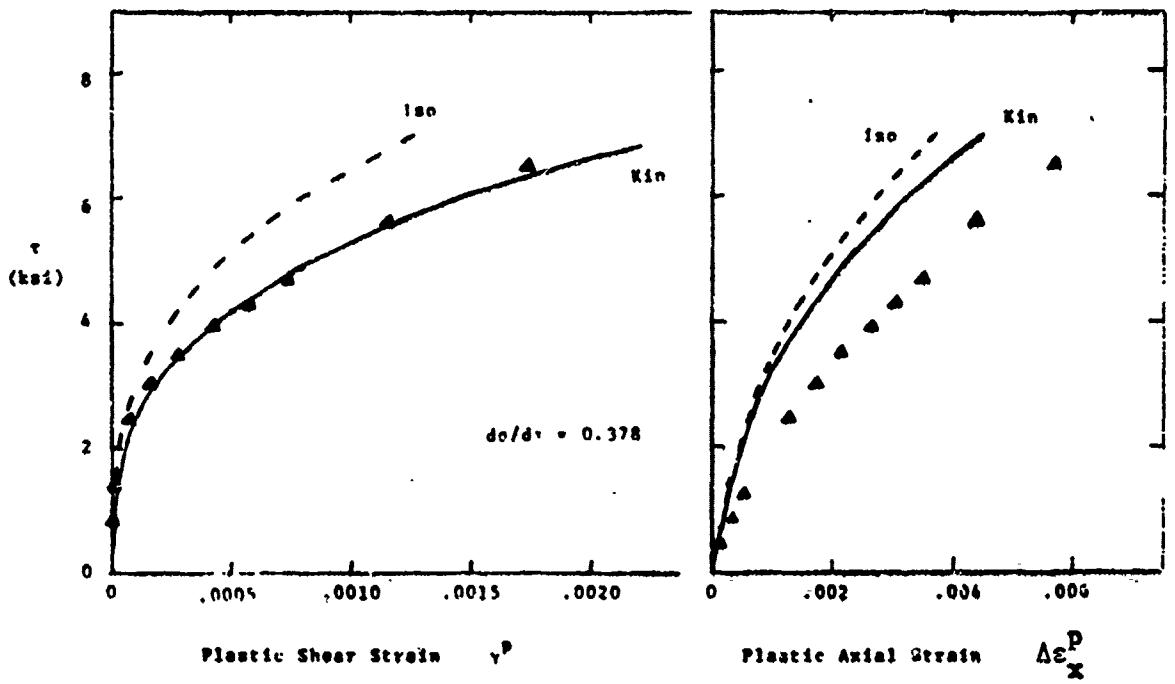



Figure 7. Comparison of isotropic and kinematic hardening flow theories with experiments of Budiansky et. al. [1]. Filled triangles indicate measurements of plastic flow along $d\sigma/d\tau = 0.378$ after initial axial compression.

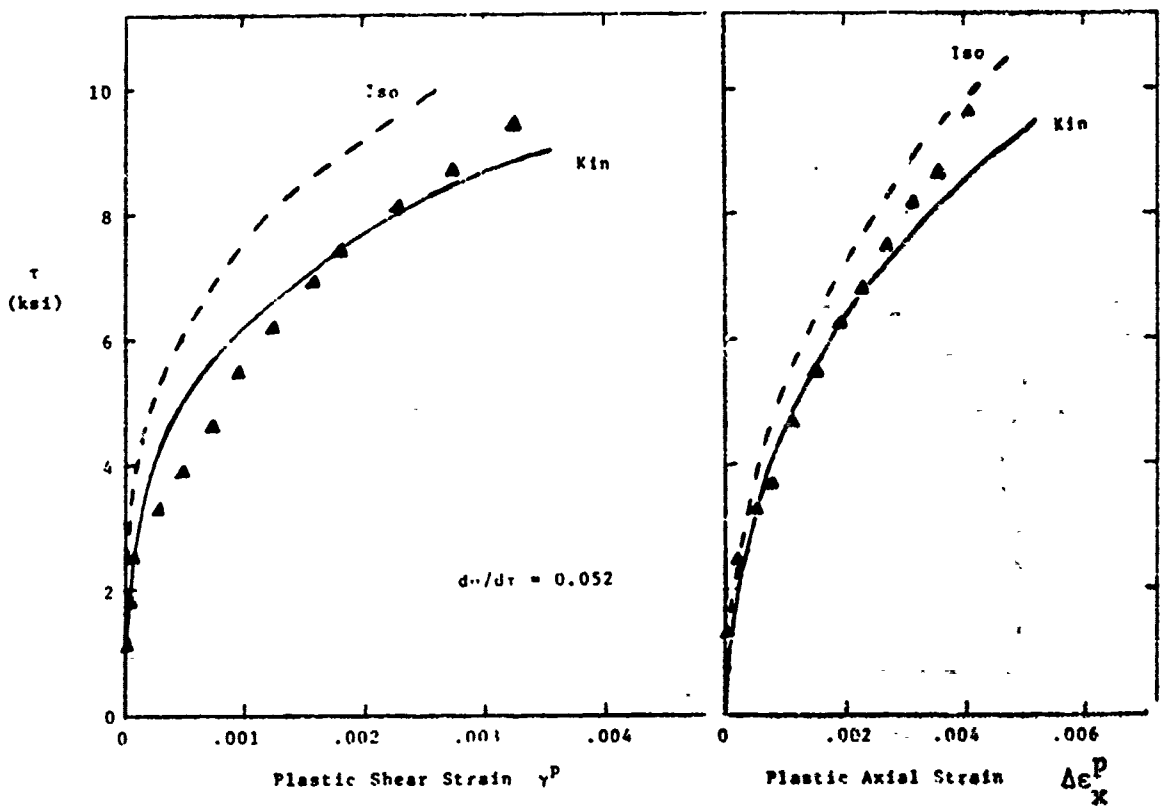



Figure 8. Comparison of isotropic and kinematic hardening flow theories with experiments of Budiansky et. al. [1]. Filled triangles indicate measurements of plastic flow along $d\sigma/d\tau = 0.052$ after initial axial compression.

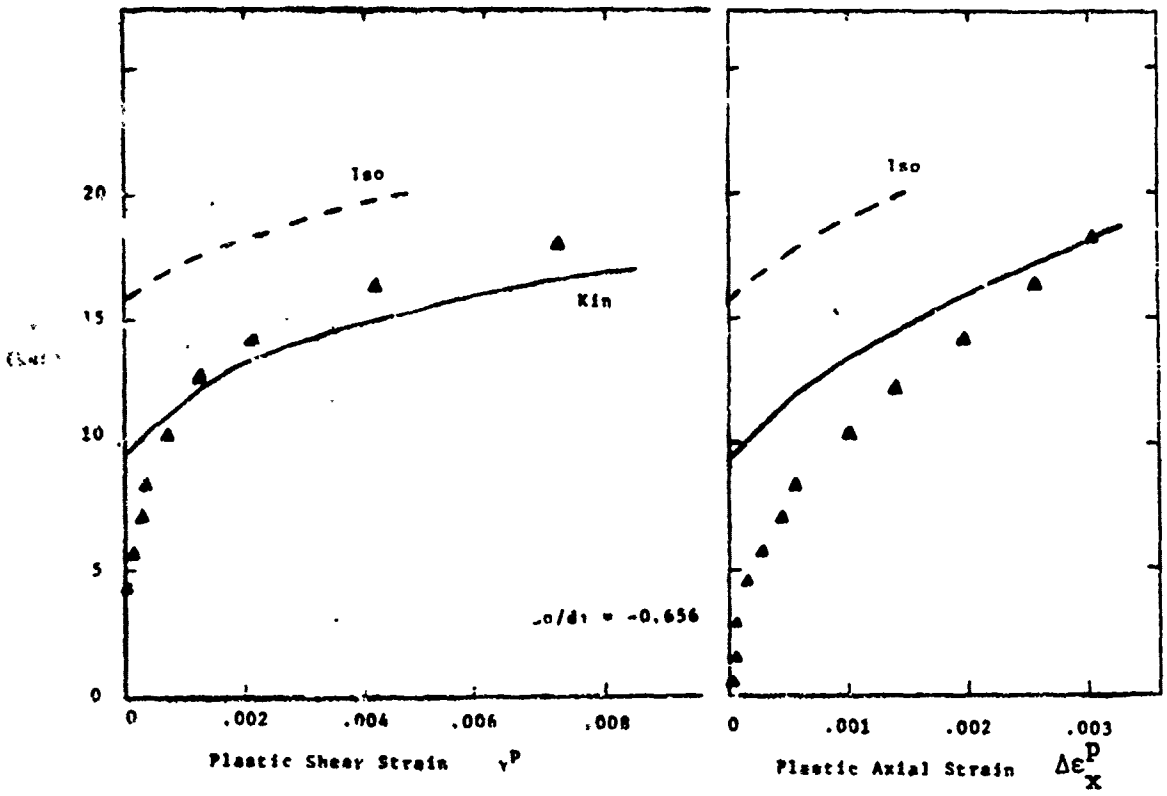



Figure 9 . Comparison of isotropic and kinematic hardening flow theories with experiments of Budiansky et. al. [1]. Filled triangles indicate measurements of plastic flow along $d\sigma/d\tau = -0.656$ after initial axial compression.

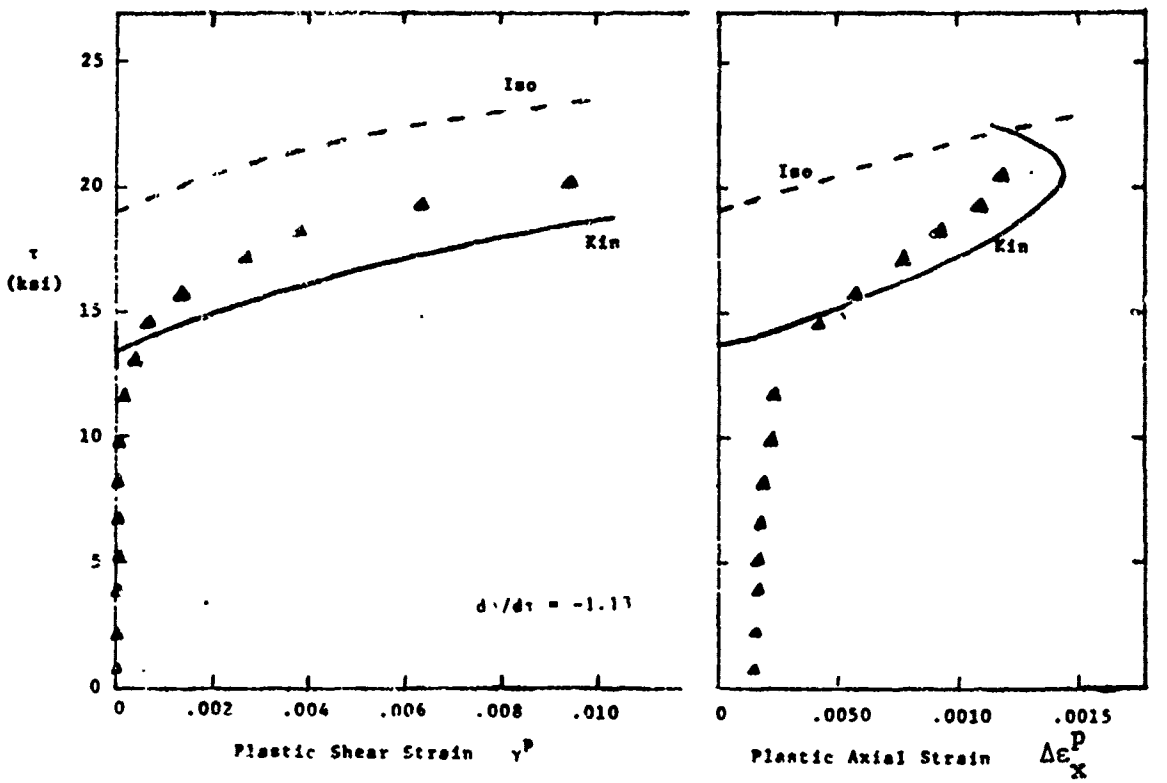



Figure 10. Comparison of isotropic and kinematic hardening flow theories with experiments of Budiansky et. al. [1]. Filled triangles indicate measurements of plastic flow along $d\sigma/d\tau = -1.13$ after initial axial compression.

# THE INVERSE GAUSSIAN PULSE IN THE EXPERIMENTAL DETERMINATION OF LINEAR SYSTEM GREEN'S FUNCTIONS

Alfred S. Carasso
Center for Applied Mathematics
and
Nelson N. Hsu
Center for Manufacturing Engineering
National Bureau of Standards
Gaithersburg, Maryland '20889

## Abstract

We present a new time domain deconvolution method for determining the "impulse response" of linear time invariant systems. The method is based on the use of the one-sided, causal, inverse Gaussian pulse as an approximation to the Dirac δ-function. Deconvolution of that kernel is equivalent to an inverse heat conduction problem. The method is particularly useful in cases where the Green's function for the linear system has singularities such as jumps, cusps, spikes, and the like. Computational reconstructions of singularities, from smooth synthetic data, are presented in the context of Acoustic Emission Green's functions.

## 1. Introduction

The problem of determining the "impulse response" of a linear time invariant system occurs in many areas of measurement science. If g(t) is a causal time signal representing the system's response to a Dirac δ-function input at t = 0, then g(t) is the causal Green's function or impulse response of the system. If g(t) is known, the system may be considered a "black box"; the output y(t) for any given input x(t), is the convolution of x with g,

$$(1.1)\quad y(t) = \int_0^t x(t-\tau)\, g(\tau)\,d\tau = \int_0^t g(t-\tau)\, x(\tau)\,d\tau.$$

The response to a Heaviside input H(t) is sometimes preferred. This response, also called the Green's function, will be denoted by G(t). Clearly,

Research supported in part by the U.S. Army Research Office under MIPR No. ARO 63-82.

PREVIOUS PAGE IS BLANK

$$(1.2)\quad G(t) = \int_0^t g(s)\, ds,$$

and from (1.1),

$$(1.3)\quad y(t) = \int_0^t G(t-\tau)\, \dot{x}(\tau) d\tau.$$

In many important cases g(t) or G(t) are not smooth functions of t. Rather, they exhibit singularities such as jumps, spikes, cusps, and other strikingly sharp features. In fiber optics measurements, an optical time domain reflectometer system, (OTDR), is often used for fiber parameter estimation; see e.g. [1, p. 236], [2, p. 391], [3], and the references therein. This non-destructive technique uses short pulses of light which are launched into one end of the fiber, and the reflected signal is observed, as a function of time, at the same input end. The instrument is based on the fact that fluctuations in the refractive index along the fiber, as well as defects and imperfections, cause light to be backscattered. Isolated imperfections produce sharp spikes superimposed on an otherwise smooth monotone profile. The impulse response of the fiber is particularly useful in characterizing the fiber, as well as in locating flaws, [4, p. 61], [5, p. 1]. Another example in nondestructive testing is the field of acoustic emission, (AE). Ultrasonic waves emitted by stressed regions in elastic materials are detected by means of transducers, and these time signals are used to locate structural flaws. See [6], [7], [8], [9]. In such studies, the dynamic Green's function for the structure plays an important role as a general representative solution to the problem of elastic wave generation and propagation caused by a localized source. In the case of an infinite plate, a comprehensive treatment of the Green's tensor was recently given in [10]. In typical configurations, see

[10], the Green's function is neither positive nor differentiable; the singularities in the impulse response carry valuable information regarding the times of arrival of various wave components of the signal.

Ideally, by using probe waveforms in the form of the $\delta$-function or the Heaviside function, the impulse response can be obtained experimentally. However, the "band-limited" nature of signals which can be produced in the laboratory preclude such infinitely sharp probe waveforms. Rather, $C^\infty$ approximations to $\delta(t)$ or $H(t)$ are synthesized and used. The result is to smooth out or blur some of the important singularites in the system's Green's function. In order to recover the sharp features, an ill-posed deconvolution problem must be carefully solved. This requires constraints on the unknown Green's function in order to stabilize the numerical procedure. Two widely used constraints in the time domain, consist in prescribing an a-priori bound on the second derivative of $g(t)$, or requiring $g(t)$ to be non-negative. While such regularization techniques often succeed in reconstructing smooth or non-negative functions, they are not applicable to cases such as the plate Green's function. In this paper, we outline an alternative time domain deconvolution technique, based on the "inverse Gaussian" pulse. This pulse is intimately related to a heat conduction problem, and this connection can be exploited to impose weaker constraints on the unknown solutions, such as a bound in the $L^2$ -norm. In addition, the deconvolution can be implemented as a Cauchy problem for a simple second order partial differential equation. This approach allows for <u>continuous</u> deconvolution, a useful option which permits monitoring the development of suspected artifacts. A more detailed discussion of this technique is given in [11].

## 2. The Inverse Gaussian Pulse

Let $\sigma > 0$ be a fixed parameter, and consider the function of time, $k(\sigma,t)$, defined by

$$(2.1)\quad k(\sigma,t) = H(t)\left(\frac{\sigma}{2\sqrt{\pi t^3}}\right)\exp\left(\frac{-\sigma^2}{4t}\right).$$

This $C^\infty$ function of t on $(-\infty,\infty)$ is called an "Inverse Gaussian" distribution in statistics and is related to Brownian motion. See [12], [13], [14, p. 221] and their references. For small $\sigma > 0$, $k(\sigma,t)$ approximates $\delta(t)$; for larger $\sigma$, $k(\sigma,t)$ is a signal of short duration. Convolution with $k(\sigma,t)$ smooths out sharp features at an increasing rate the larger $\sigma$ is chosen. The indefinite integral of $k(\sigma,t)$ is a causal $C^\infty$ approximation to $H(t)$, and can be expressed in terms of the complementary error function. We have,

$$(2.2)\quad e(\sigma,t) = \int_{-\infty}^{t} k(\sigma,s)ds = \operatorname{erfc}\left(\frac{\sigma}{2\sqrt{t}}\right).$$

Both $k(\sigma,t)$ and $e(\sigma,t)$ are depicted in Figure 1.

Let $g(t)$ be the impulse response of the time invariant system under study, and let $G(t)$ be as in (1.2). The response of the system to the probe waveform $e(\sigma,t)$ in (2.2) is given by,

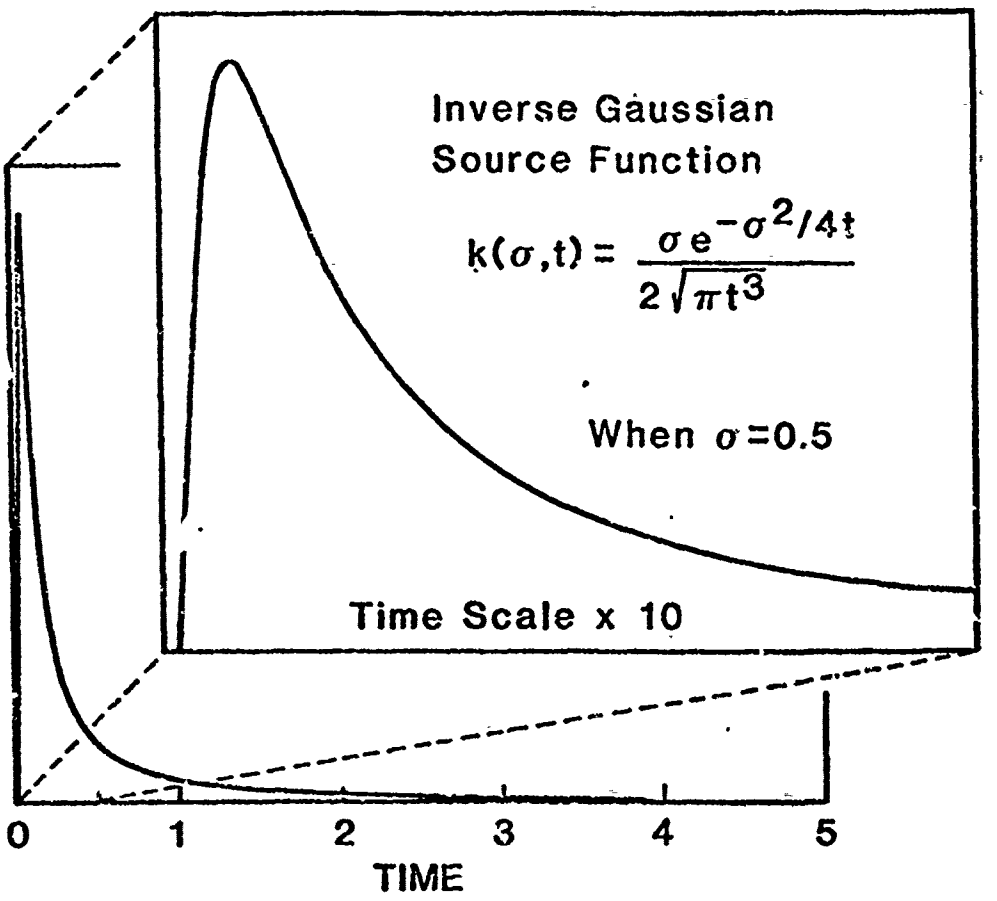
$$(2.3)\quad h(t) = \int_0^t e(\sigma,t-\tau)\, g(\tau)d\tau = \int_0^t k(\sigma,t-\tau)\, G(\tau)d\tau,$$

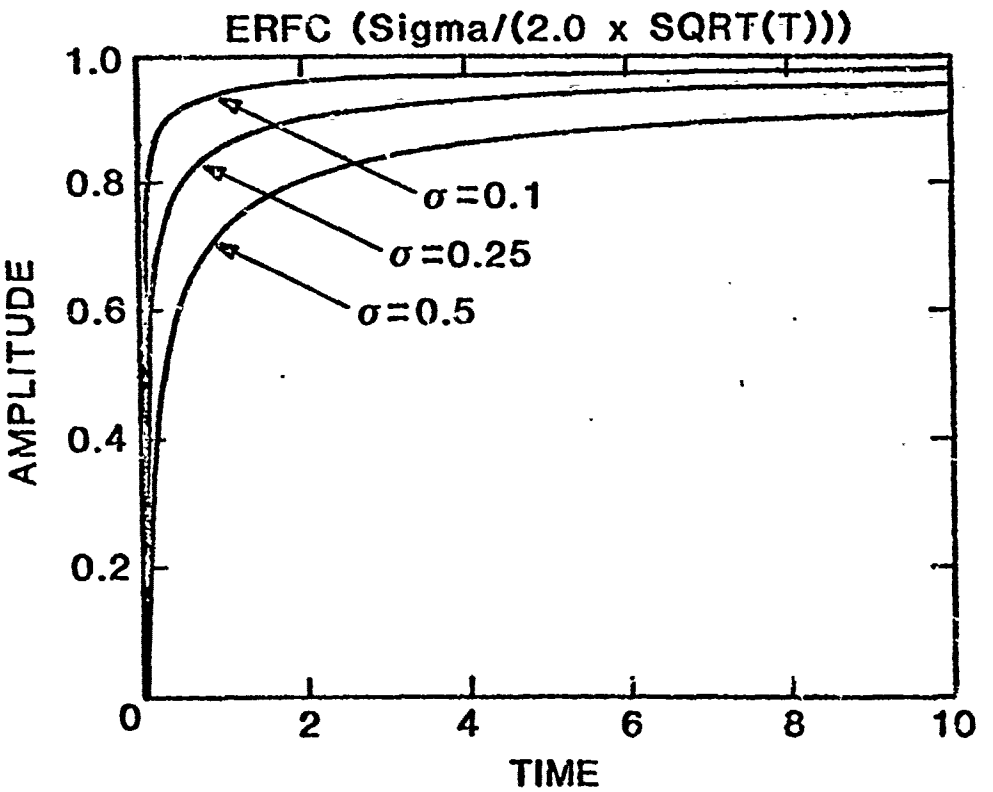
while that due to the waveform $k(\sigma,t)$ is given by,

$$(2.4)\quad w(t) = \int_0^t k(\sigma,t-\tau)\, g(\tau)d\tau.$$

Thus, either probe leads to a time domain deconvolution problem for determining the Green's function, in which the kernel is the inverse Gaussian

FIGURE 1.


Inverse Gaussian
Source Function
$k(\sigma,t) = \frac{\sigma e^{-\sigma^2/4t}}{2\sqrt{\pi t^3}}$
When $\sigma$=0.5
Time Scale x 10
0
1
2
3
4
5
TIME


ERFC (Sigma/(2.0 x SQRT(T)))
1.0
0.8
0.6
0.4
0.2
AMPLITUDE
σ=0.1
σ=0.25
σ=0.5
0
2
4
6
8
10
TIME

pulse (2.1). Some of the properties of the convolution with $k(\sigma,t)$ are listed below.

I. $w(t)$ in (2.4) is a $C^\infty$ function of $t$ on $(0,\infty)$ even if $g(t)$ is in $L^2(0,\infty)$. Moreover, using Schwarz's inequality in (2.4), derivatives of $w(t)$ can be bounded in the supremum norm, in terms of the $L^2$ norm of $g(t)$, if $\sigma > 0$.

II. As a map from $L^2(0,\infty)$ into $L^2(0,\infty)$ the convolution in (2.4) is a holomorphic semi-group in the variable $\sigma$ for $\sigma > 0$. See [15, p. 145].

III. The Laplace transform of $k(\sigma,t)$ is $\exp(-\sigma\sqrt{s})$. The inverse kernel to $k(\sigma,t)$ has the Laplace transform $\exp(\sigma\sqrt{s})$, $\sigma > 0$. This inverse kernel cannot exist as a Schwartz distribution, since the Laplace transform of a Schwartz distribution has polynomial growth at infinity. See [16, p. 189]. Thus, "direct" deconvolution in (2.4) is not feasible. This is always the case with $C^\infty$ probes. See [17, p. 137, p. 145].

IV. Let $\ell$ and $a$ be positive constants such that $\sigma = (\ell/a)$. Let $u(x,t)$ be the unique bounded solution of the heat conduction problem for the semi-infinite rod,

$$(2.5)\quad \begin{cases} u_t = a^2 u_{xx},\ 0 < x < \infty,\ t > 0, \\ u(x,0) = 0,\ 0 < x < \infty, \\ u(0,t) = g(t),\ \lim_{x\to\infty} u(x,t) = 0 \end{cases}$$

Then, see e.g. [18, p. 172], $w(t)$ in (2.4) is $u(x,t)$ evaluated at $x = \ell$.

As a consequence of IV, deconvolution in (2.3) or (2.4) is equivalent to solving the "inverse heat conduction" problem whereby the boundary temperature history g(t) (or G(t)), at x = 0, is reconstructed from knowledge of the temperature history at the location x = $\ell$. This is an ill-posed problem in partial differential equations which is discussed in detail in [19].

## 3. The Deconvolution Procedure

The use of probe waveforms of the specific type given in (2.1), (2.2) is advocated, partly because such probes are smooth functions of time whose "sharpness" can be controlled by means of a single parameter, namely $\sigma$. (It is advantageous to synthesize such probes with as small a value of $\sigma$ as is experimentally feasible). However, an equally important feature of (2.1) is the connection with the heat conduction problem (2.5) which can be exploited to devise a deconvolution procedure with certain desirable properties. Given the probe parameter $\sigma$, we interpret the recorded signal h(t) in (2.3) (or w(t) in (2.4)) as the solution of (2.5) at x = $\ell$. By the semi-group property II above, we may use h(t) to continue the solution to the right of x = $\ell$ by convolution, and obtain both u(x,t) and $u_x(x,t)$ along the line x = $\ell' > \ell$. These two functions of time comprise the necessary initial data for integrating the heat equation "sideways"; i.e. as an initial value problem in the x-variable, in the direction of decreasing x, from x = $\ell'$ to x = 0. See [19, Section 4].

Continuous dependence in the $L^2$ norm, can be restored to the sideways heat equation by imposing an a- priori bound, M, on the $L^2$ norm of the unknown boundary data at x = 0, i.e. on the desired system Green's function G(t). In addition, an a- priori estimate, $\varepsilon$, is assumed known for the $L^2$ norm of the

noise in the recorded data h(t). Let $\omega$ be the $L^2$ "noise to signal" ratio,

$$(3.1) \quad \omega = \left(\frac{\varepsilon}{M}\right).$$

The regularization is accomplished by a preliminary smoothing of the initial data in the frequency domain, using FFT algorithms. The smooth initial values are then used in a discrete numerical step by step marching procedure in the x-variable. The filtering function uses $\omega$ in (3.1) as the only a- priori information, and the regularization is exactly equivalent to minimizing the appropriate Tikhonov functional. See [19] for full details.

## 4. Continuous Deconvolution

We may picture the marching procedure described above as a systematic continuous unsmoothing of the recorded data as we approach the boundary $x = 0$ in the associated fictitious heat conduction problem (2.5). By outputting the intermediate results at positive values of x as x tends to zero, one can monitor this unsmoothing process. Any such intermediate result is a partial deconvolution, while total deconvolution corresponds to the solution at the boundary $x = 0$.

Consider now the solution of the heat flow problem (2.5) at some small positive value of x, $x = x_0$ say. Even if the boundary data g(t) lies in $L^2$, $u(x_0, t)$ is a $C^\infty$ function of t which faithfully approximates g(t) if $x_0$ is sufficiently small. Moreover, constraints on $u(x_0, t)$, together with all its derivatives, have been implicitly placed by virtue of the a- priori bound M on the $L^2$ norm of g(t). See property I above. Thus, at $x = x_0$, the regularized marching procedure is approximating a well-constrained smooth function of t which in turn approximates the non-smooth boundary data. This is the

interpretation which should be placed on the notion of partial deconvolution at $x_0$. In this sense, the heat flow problem associated with the inverse Gaussian pulse allows for a deconvolution procedure in which relatively weak $L^2$ constraints can be imposed. Error bounds, of logarithmic convexity type, for the reconstruction at any x are given in [19, Theorem 1]. These estimates show that the $L^2$ norm of the error at x is $O\left(\varepsilon^{x/\ell}\right)$, if $\varepsilon$ is the $L^2$ error in the recorded data at $x = \ell$.

## 5. Numerical Deconvolution Experiments with Synthetic Data

We present two examples of Green's function reconstruction for the case of an infinite elastic plate. These examples, from the field of acoustic emission, have singularities which are similar to those encountered in optical fiber backscattering measurements in the presence of imperfections. See [3, Figures 3-2, 3-88], [10, Figure 15] and [9, Figures 6B, 11]. Hence, the applicability of the technique in two different physical contexts can be demonstrated simultaneously.

Recent work, [20], has made it possible to develop computational software capable of producing the exact Green's function for an infinite elastic plate, given the test configuration. Use was made of this software to obtain several typical examples of plate Green's functions. By careful numerical convolution with the probe waveform (2.1), using adaptive quadrature routines, synthetic smoothed out data were created, simulating the recorded signals. These data were then used as input initial values into the deconvolution algorithm of Section 3. By comparing the deconvolution results with the known exact Green's functions, the performance of the procedure can be assessed.

The numerical experiments deal with the $G_{33}$ component of the Green's tensor, describing the normal displacement due to a normal applied force. One example studied is the epicentral response to a $\delta$-function input. The other example is the response to a Heaviside input source, located two plate thickness away and on the same side of the plate as the receiver. Our figures below display normalized displacement versus normalized time, where

(5.1) Normalized Displacement = $\pi$* Shear Modulus * Plate Thickness* Displacement/Force

(5.2) Normalized Time = Time* Shear Wave Speed/Plate Thickness.

Our experiments take place on the time interval [0,5] in normalized units, using synthetic data at 550 equispaced points in that interval. The probe in (2.1) was used with a parameter value of $\sigma = .5$, time being measured in normalized units. The related heat flow problem (2.5) was given the constant diffusivity $a^2 = .25$, while $\ell$ was fixed at the value .25, so that $\sigma = (\ell/a)$. The marching scheme in the x-variable was implemented using 1000 equispaced mesh points on the x-interval $[0,\ell]$. Because of discretization errors in the numerical computation of the exact Green's functions as well as in the subsequent convolutions, the synthetic input data are not free from noise. Accordingly, a value of $\omega = 10^{-4}$ was selected in (3.1), reflecting an estimate of the probable $L^2$ norm of the error in the recorded data. When the noise level is of the order of 0.1% or lower, the deconvolution procedure is not found to be overly sensitive to the value of $\omega$, beyond the general order of magnitude of that quantity.

In Figure 2, the epicentral response is considered. The exact Green's function has a series of sharp peaks, while the simulated convolved signal

FIGURE 2.

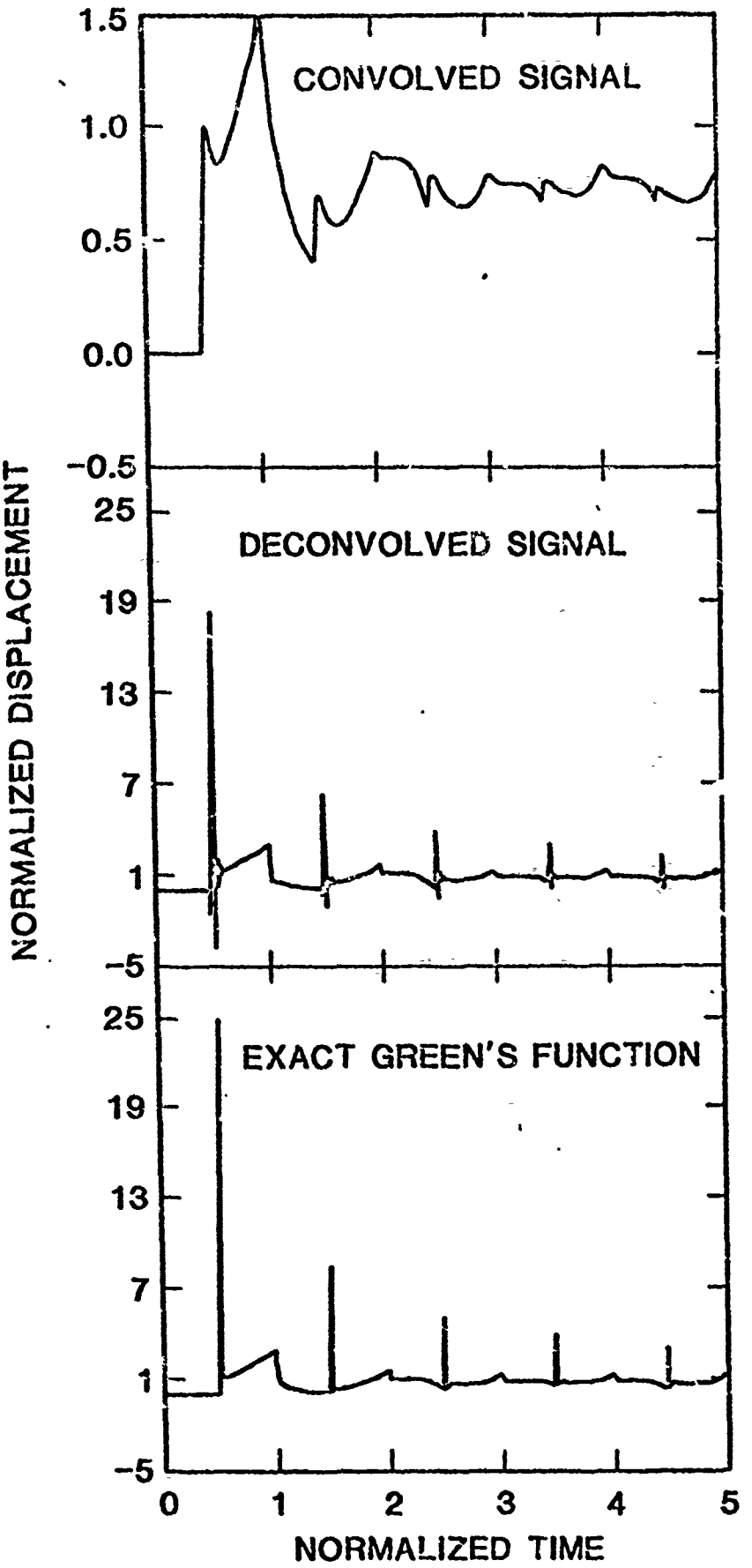

CONVOLVED SIGNAL
DECONVOLVED SIGNAL
EXACT GREEN'S FUNCTION
NORMALIZED DISPLACEMENT
NORMALIZED TIME
1.5
1.0
0.5
0.0
-0.5
25
19
13
7
1
-5
0
1
2
3
4
5

FIGURE 3.

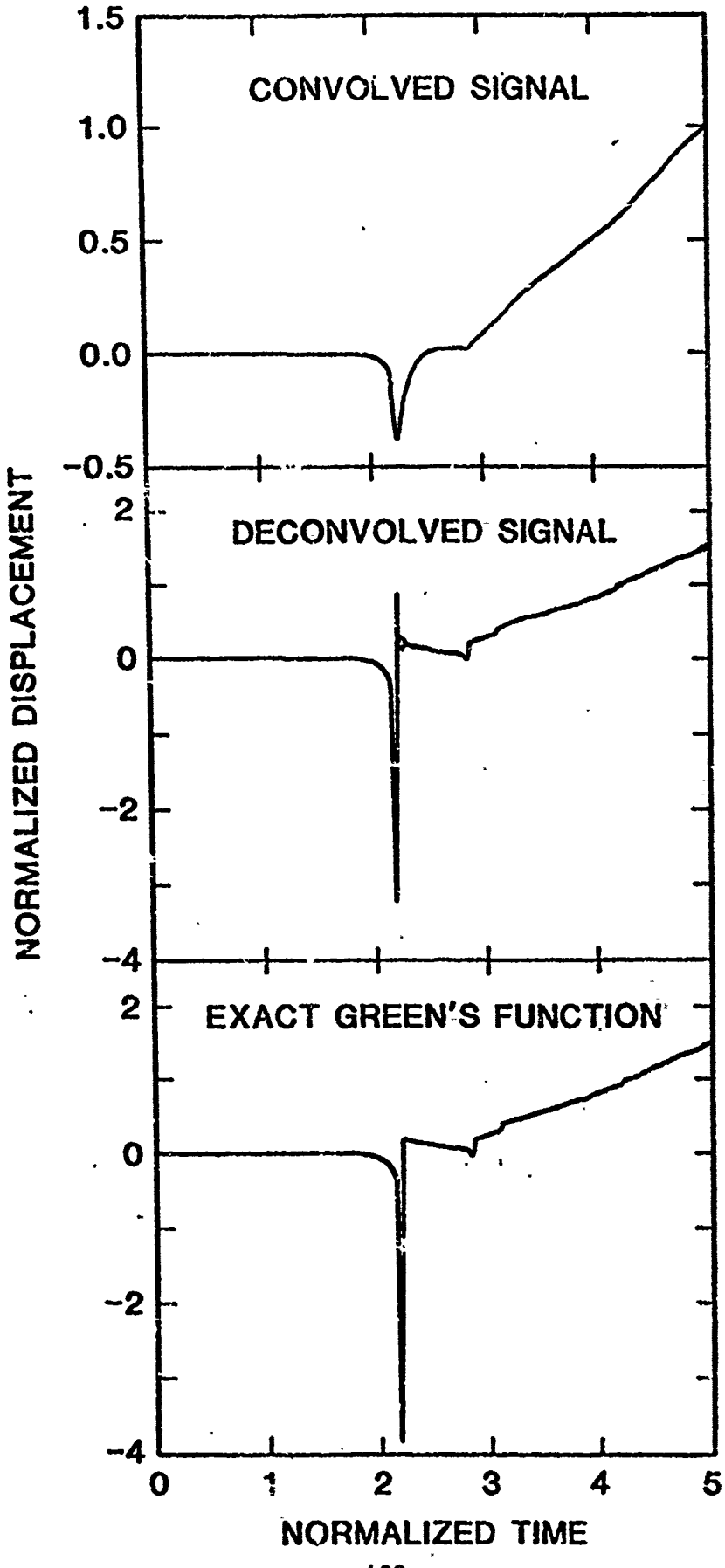
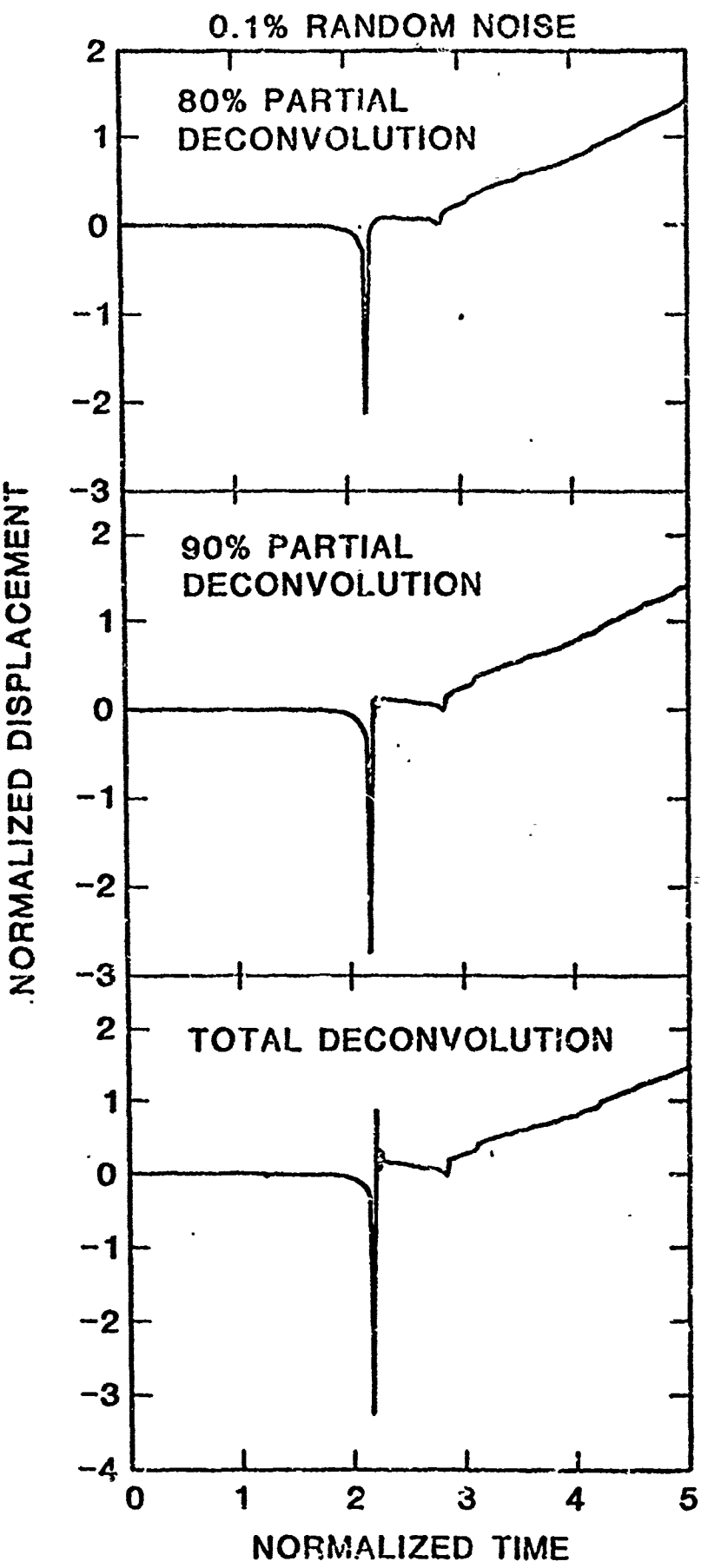

CONVOLVED SIGNAL
DECONVOLVED SIGNAL
EXACT GREEN'S FUNCTION
NORMALIZED DISPLACEMENT
NORMALIZED TIME
1.5
1.0
0.5
0.0
-0.5
2
0
-2
-4
2
0
-2
-4
0
1
2
3
4
5

FIGURE 4.


0.1% RANDOM NOISE
80% PARTIAL DECONVOLUTION
90% PARTIAL DECONVOLUTION
TOTAL DECONVOLUTION
NORMALIZED DISPLACEMENT
NORMALIZED TIME
2
1
0
-1
-2
-3
-4
0
1
2
3
4
5

bears little resemblance to the exact solution. Total deconvolution was attempted with the calculation pursued up to the boundary $x = 0$. While "ringing" phenomena are visible in the reconstruction, the ability of the scheme to restore and localize these sharp peaks is significant. In Figure 3, with source and receiver on the same side of the plate, the exact Green's function has a sharp trough and a "corner" near $t = 2.2$. These sharp features are not visible in the smooth recorded data. Total deconvolution restores many of these features although an artifact in the form of a Gibbs phenomenon obscures the sharp corner near $t = 2.2$. This last example was also studied from the standpoint of continuous deconvolution. Random noise of magnitude 0.1% was added to each of the 550 synthetic data values prior to deconvolution, and the partial deconvolutions at 80% and 90% of the way back from $x = \ell$, were obtained, in addition to the total deconvolution at $x = 0$.

The results are depicted in Figure 4. The genesis of the Gibbs phenomenon near $t = 2.2$ is apparent; substantial tightening of the bend to the right of the trough occurs prior to the onset of ringing, as the algorithm attempts to reconstruct the corner in the exact solution. Used in this way, continuous deconvolution is a desirable option which can reveal important features in the solution before these features become obscured by artifacts.

## 6. Concluding Remarks

The usefulness of the above procedure rests on the feasibility of an appropriate device for producing the proposed waveforms. At the present time, input electrical voltages with a prescribed waveform can be synthesized using a digital to analog converter. To produce a mechanical pulse with a prescribed time dependence requires a high fidelity transducer to convert the

synthesized voltage into an impact force. Work in this direction is currently under way at the National Bureau of Standards. See [21], [22]. On the other hand, the probes (2.1), (2.2) are similar in shape to waveforms commonly found in the experimental literature. By attempting to fit such probes with the desired expressions (2.1) or (2.2), for suitable values of $\sigma$, one can consider using the above procedure in a variety of linear system contexts. Future work should explore the feasibility of such "inexact" deconvolution.

## REFERENCES


1. D. Marcuse, "Principles of Optical Fiber Measurements," Academic Press, New York, (1981).

2. S. E. Miller and A. G. Chynoweth, "Optical Fiber Telecommunications," Academic Press, New York, (1979).

3. B. L. Danielson, "An Assessment of the Backscatter Technique as a means for Estimating Loss in Optical Waveguides," NBS Technical Note 1018, February 1980, U.S. Department of Commerce/National Bureau of Standards, Washington, DC 20234.

4. G. E. Chamberlain, G. W. Day, D. L. Franzen, R. L. Gallawa, E. M. Kim and M. Young, "Optical Fiber Characterization," NBS Special Publication 637, October 1983, U.S. Department of Commerce/National Bureau of Standards, Washington, DC 20234.

5. B. L. Danielson, "Backscatter Signature Simulations," NBS Technical Note 1050, December 1981, U.S. Department of Commerce/National Bureau of Standards, Washington, DC 20234.

6. Y. H. Pao, "Theory of Acoustic Emission; in Elastic Waves and Non-Destructive Testing of Materials," AMD-Vol. 29, Y. H. Pao, Editor, The American Society of Mechanical Engineers, New York, pp. 107-128, (1978).

7. Y. H. Pao, R. R. Gajewski and A. N. Ceranoglu, "Acoustic Emission and Transient Waves in an Elastic Plate," J. Acoust. Soc. Am., 65, pp. 96-105, (1979).

8. N. N. Hsu, J. A. Simmons and S. C. Hardy, "An Approach to Acoustic Emission Signal Analysis -- Theory and Experiment." Materials Evaluation, 35, pp. 100-106, (1977).

. N. N. Hsu and S. C. Hardy, "Experiments in Acoustic Emission Waveform Analysis for Characterization of AE Sources, Sensors and Structures; In Elastic Waves and Non-Destructive Testing of Materials," AMD-Vol. 29, Y. H. Pao, Editor, The American Society of Mechanical Engineers, New York pp. 85-106, (1978).

0. A. N. Ceranoglu and Y. H. Pao, "Propagation of Elastic Pulses and Acoustic Emission in a Plate," ASME Journal of Applied Mechanics, 48, pp. 125-147, (1981).

1. A. S. Carasso and N. N. Hsu, "Probe Waveforms and Deconvolution in the Experimental Determination of Elastic Green's Functions," March 1984, Submitted for Publication.

12. N. L. Johnson and S. Kotz, "Continuous Univariate Distributions," Vol. 1, Houghton-Mifflin Company, Boston, (1970).

13. M. T. Wasan, "On an Inverse Gaussian Process," Skandinavisk Aktuerietidskrift, 51, pp. 69-96, (1968).

14. D. R. Cox and H. D. Miller, "The Theory of Stochastic Processes," John Wiley and Sons, New York, (1965).

15. P. L. Butzer and H. Berens, "Semi-Groups of Operators and Approximation," Springer-Verlag, New York, (1967).

16. G. Doetsch, "Introduction to the Theory and Application of the Laplace Transformation," Springer-Verlag, New York, (1974).

17. L. Schwartz, "Methodes Mathematiques Pour les Sciences Physiques," 2eme Edition, Hermann, Paris, (1965).

18. A. G. Webster, "Partial Differential Equations of Mathematical Physics," Dover Publications, New York (1955).

19. A. Carasso, "Determining Surface Temperatures From Interior Observations," SIAM J. Appl. Math, 42, pp. 558-574, (1982).

20. J. R. Willis, N. N. Hsu and J. A. Simmons, "The Dynamic Green's Tensor for an Elastic Plate." In preparation.

21. T. M. Proctor, Jr., "An Improved Piezoelectric Acoustic Emission Transducer," J. Acoust. Soc. Am., 71, pp. 1163-1168, (1982).

22. D. Eitzen, N. Hsu, A. Carasso and T. Proctor, "Deconvolution by Design-An Approach to the Inverse Problem of Ultrasonic Testing." In Review of Progress in Quantitative NDE, University of California-San Diego, LaJolla, California, July 1984. D. O. Thompson, Editor. To appear.

# SIMULATIONS OF SPECIAL INTERIOR BALLISTIC PHENOMENA WITH AND WITHOUT HEAT TRANSFER TO THE GUN TUBE WALL

Rudi Heiser
Fraunhofer-Institut für Kurzzeitdynamik
Ernst-Mach-Institut, Abteilung für Ballistik
Hauptstrasse 18, D-7858 Weil am Rhein
West-Germany

and

James A. Schmitt*
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

ABSTRACT. The computer code DELTA uses a linearized Alternating Direction Implicit (ADI) scheme to provide a numerical approximation of the solution of the averaged two-phase (gas-solid) two-dimensional (axisymmetric) equations governing viscous interior ballistic flows within conventional guns. To further the understanding of phenomena affecting gun tube life as well as gun performance, a heat transfer model, which simulates the interactions of fluid dynamics and thermal profile in the gun tube wall, has been incorporated in the DELTA code. The same linearized ADI method is utilized to obtain the numerical solution of the two-dimensional nonlinear heat conduction equation in the gun tube. Our model of the heat transfer process couples completely and simultaneously all three controlling events without any approximations to the governing equations; that is, the axisymmetric viscous flow within the entire gun tube which naturally gives a precise definition of the gas thermal boundary layer, the time-dependent balance of the heat fluxes at the inner wall surface, and the two-dimensional temperature calculation within the gun tube wall. The nonlinear heat flux boundary conditions at the inner and outer tube surfaces are linearized as to be compatible with the solution scheme.

Results computed with DELTA compare the flow patterns for two different types of idealized one-phase interior ballistics environments. The first is a pure expansion flow, and the other includes mass and heat sources. Besides laminar flow calculations the effects of two algebraic turbulence models are considered. Finally, simulations for an adiabatic tube and for a tube that allows heat transfer are reported.

---

* Current address is AT&T Bell Laboratories, Crawford Corners Road, Holmdel, NJ 07733

## I. INTRODUCTION

The interior ballistics flows of conventional charges in gun tubes show complex flow patterns. This is due to both the heterogeneous structure of the charge itself, and the rapidly changing flow conditions within a few milliseconds. The fast rise in pressure and temperature caused by the burning propellant initiates turbulent, multidimensional, multiphase flow which is coupled with the accelerating projectile motion. Up to now a mathematical model that describes all the physics occurring in a complete interior ballistics cycle is not available. However, several models that simulate some of the phenomena exist or are being developed [2].

On the other hand, detailed experimental measurements of the total interior ballistic cycle are not possible. Some standard techniques as well as some new special techniques under development determine only specific quantities in real weapons or in simulators under simplified flow conditions. Commonly measured quantities are the gas pressure and projectile motion. However, quantities like the temperature distribution in the gas and in the gun tube wall, velocity distributions of the gas or solid particulates inside or outside of boundary layers, the particle distribution, or the turbulent pattern cannot yet be determined accurately by experiment. Thus, the need for modelling the interior ballistic cycle exists so that the dynamic development of these quantities can be studied, and their impact on ballistic problems can be evaluated.

One new computational capability for investigating interior ballistics flows is the DELTA code which is under development at BRL. The purpose of this code is to address special ballistic problems, related to boundary layer development, heat transfer to the tube wall, turbulence and time-dependent distribution of additive particles. In the following the basis of DELTA and its applications will be discussed.

## II. REVIEW OF THE MODELS IN DELTA

The flow to be modeled by DELTA is the multidimensional, two-phase flow inside of a gun tube. Presently, the geometry of the flow is simplified to axisymmetry. At the rear end, the so-called breech, the tube is closed by a stationary flat plate while the front boundary is represented by the moving flat based projectile. The flow is always assumed to be viscous and heat conducting, but can be either laminar or turbulent. The walls may be adiabatic or allow heat transfer from the gas to the surrounding metal surfaces. Heat transfer is restricted to the tube wall. The core flow is fully coupled to the moving projectile, to the boundary layer development, and if desired, to the heat conduction in the tube wall. By fully coupled we mean that each of these phenomena can affect any other one; for example, the boundary layer development can alter the details of the core flow. This situation is not the case in a boundary layer type model.

The mathematical model in DELTA, the balance equations for the gas-phase and one solid phase, are based on the unsteady volume-averaged formulation. The gas phase is described by the averaged equations corresponding to the full Navier-Stokes equations for a compressible fluid. The averaged coefficients of viscosity and heat conduction, the averaged viscous stress tensor, the

averaged dissipation function and averaged heat conduction function are included. Since interior ballistics flows usually produce high gas pressure, the Noble-Abel equation of state is used so that some real gas effects can be included. The turbulence is considered by algebraic mixing length models. The solid phase is described by the averaged equations corresponding to an incompressible fluid which can undergo deformation. The derivation of these equations is given in Ref. [3]. The equations are listed in Ref. [1].

The set of partial differential equations for the axisymmetric two-phase flow region is solved by a linearized Alternating Direction Implicit (ADI) scheme. This scheme transforms the differential equations to a set of linear algebraic equations. The corresponding matrix has a block tridiagonal structure, and thus, the solutions at each new time-step can be efficiently determined. Detailed information about the derivation of the scheme is presented in Ref. [1].

## III. HEAT TRANSFER TO AND TEMPERATURE DISTRIBUTION IN THE TUBE WALL

The heating of the gun tube wall caused by convection, heat conduction and radiation of the hot propellant gas along its inner surface enhances the gun tube wear and erosion, and therefore, affects the lifetime of gun tubes. The experimental determination of the inner wall surface temperature is quite difficult. The commonly used thermocouples are restricted in interior ballistic applications due to the demands of an accurate measurement in a very short time interval, of the close contact to the flow, and of the special thermal properties of the gauge itself.

Currently, there exist several different models for the heat transfer to the gun tube wall. They can be divided in four categories according to their complexity. The first type, at most, uses a very simplistic boundary layer calculation and a heat transfer correlation, e.g., Colburn's analogy, to obtain the heat transfer [4-6]. Heat losses in the core flow are considered. These empirical heat transfer correlations which are derived for fully developed, steady, one-phase pipe flow are used to predict the highly unsteady flow and heat flux in a gun. The main feature of this type of model is the emphasis on the calculation of the core flow. In contrast another category tries to model more exactly the boundary layer by using more general boundary layer equations [7,8]. The heat transfer to the tube wall again is described by correlations. The boundary layer edge is not coupled with any computation of the core flow. Instead, simplified qualitative values are assumed which represent approximately the core flow conditions. The third category makes use of the general boundary layer equations and of a balance of heat fluxes from the hot gas to the gun tube wall at the inner wall surface [9]. The conditions at the boundary layer edge are comparable to those in the second category. These approaches do not include all the feedback mechanism to the core flow, and therefore, to the projectile motion. That is, the phenomena occurring in the boundary layer regions are not fully coupled to the phenomena occurring in the regions away from the boundary layer, the core flow region. The fourth category is the fully coupled approach wherein the phenomena associated with the core flow is directly linked to the projectile motion, the boundary layer development, the heat transfer to and the heat conduction in the tube wall, and vice versa. This is achieved by using a single but general set of equations which is valid everywhere in the gas region. The solution of

this set automatically provides the boundary layer solution in the boundary layer region, the core flow solution in the core flow, and all the necessary coupling that naturally occurs in the flow. Although this category of solution possesses the most complexity, it provides the solution with the least number of assumptions and approximations. In light of the scarcity of experimental measurements with which to compare the calculations, we feel the fourth category is best. The DELTA approach is an example of the fourth category.

The heat transfer model in DELTA consists of the equations governing the heat conduction in the tube wall, and the boundary conditions which couple the temperature in the wall to the flow inside and outside the tube. The heat conduction in the tube wall is described by the two-dimensional, nonlinear axisymmetric equation for the wall temperature $T_w(t,r,z)$:

$$\rho_w(T_w)\ c_w(T_w)\ \frac{\partial T_w}{\partial t} = \frac{\partial}{\partial z}\left[\lambda_w(T_w)\ \frac{\partial T_w}{\partial z}\right] + \frac{1}{r}\ \frac{\partial}{\partial r}\left[r\ \lambda_w(T_w)\ \frac{\partial T_w}{\partial r}\right] .$$

The variables t, z and r denote the time, axial coordinate in the wall and radial coordinate in the wall, respectively. The specific heat $c_w$ and the thermal conductivity $\lambda_w$ of gun tube steel are non-constant and depend remarkably on temperature [10]. Although, the density of the steel $\rho_w$ depends on temperature, its variation is small.

The most important boundary condition is at the inner tube wall surface where the coupling of the flow region and the wall occurs. At this surface, the balance of heat fluxes with a radiation effect

$$-\lambda_g\ \frac{\partial T_g}{\partial r_g} + \varepsilon\ \sigma\ (T_{g_o}^4 - T_g^4) = -\lambda_w\ \frac{\partial T_w}{\partial r}$$

and the temperature equilibrium

$$T_g = T_w$$

are enforced. The variables $\lambda_g$, $\varepsilon$, $\sigma$, $T_g$ and $T_{g_o}$ denote the thermal conductivity of the gas, the emissivity of the wall surface, the Stefan-Boltzmann constant, the gas temperature, and the maximum gas temperature in a given cross-section (z = constant), respectively. We emphasize that both conditions are used only at the inner tube wall surface. The first term on the left hand side of the first condition represents the heat flux on the gas side towards the wall by conduction, while radiation is included by the second term. The right hand side gives the heat flux into the tube wall. The boundary condition at the outer tube wall surface is chosen as a simple engineering condition

$$-\lambda_w\ \frac{\partial T_w}{\partial r} = h\ (T_w - T_{amb}) ,$$

where we are using a heat transfer number h and an outer ambient temperature $T_{amb}$. A more sophisticated condition, at least for a single shot weapon, is not necessary because the heat usually does not reach the outer surface during the ballistic cycle due to the wall thickness. Two additional boundary conditions are needed in axial direction. At the projectile base, we set

$$T_w = T_{amb}$$

across the wall thickness, that is, we assume the projectile is moving into a cold area with ambient temperature. At the breech, an adiabatic condition

$$\frac{\partial T_w}{\partial z} = 0$$

seems to be adequate.

These equations governing the temperature distribution in the tube wall are solved using the same linearized ADI method as are the equations for the gas flow region; that is, the equations are linearized in time, and are split along coordinate directions. At each new time level, we first update the temperature distribution in the wall, and then update the governing variables in the flow region. This is performed by the following sequence of sweeps along coordinate directions: an axial sweep followed by a radial sweep in the wall; a radial sweep followed by an axial sweep in the gas region, and finally an adjustment of the dependent variables along the inner wall surface to the flux boundary condition. We omit a discussion of most details of the numerical procedure because they are identical to those explained in Ref. [1]. Because the thermodynamic dependent variables in the gas region are the specific gas entropy (s) and logarithm of gas pressure (q), the heat flux boundary condition expressed in terms of $T_g$ must be reformulated as a linear equation in terms of s and q at the new unknown time level for the radial sweep in the gas region. To this end, we transform the heat flux term on the gas side via the chain rule:

$$\lambda \frac{\partial T}{\partial r} = \lambda(T(s,q)) \frac{\partial T(s,q)}{\partial r} = \lambda(T(s,q)) \left[T_s \frac{\partial s}{\partial r} + T_q \frac{\partial q}{\partial r}\right] ,$$

where $T_s$, $T_q$ denote the partial deriviatives of T with respect to s and q, respectively. For simplicity, we dropped the index g. The linearization in time gives a relation between the unknown new time level (n) and the known current time level (c)

$$\lambda^n \left(\frac{\partial T^n}{\partial r}\right) = \lambda^c \left(\frac{\partial T^c}{\partial r}\right) + \frac{d}{dt}\left[\lambda \frac{\partial T}{\partial r}\right]^c \Delta t + O(\Delta t^2)$$

with

$$\frac{d}{dt}\left[\lambda \frac{\partial T}{\partial r}\right]^c \Delta t = \left(\frac{d\lambda}{dT}\right)^c \left[T_s^c \frac{ds}{dt} + T_q^c \frac{dq}{dt}\right] \left(\frac{\partial T}{\partial r}\right)^c \Delta t$$

$$+ \lambda^c \left[\left(T_{qq} \frac{\partial q}{\partial r} + T_{sq} \frac{\partial s}{\partial r}\right)^c \frac{dq}{dt} + \left(T_{ss} \frac{\partial s}{\partial r} + T_{qs} \frac{\partial q}{\partial r}\right)^c \frac{ds}{dt}\right] \Delta t .$$

The time-deriviatives are approximated by

$$\frac{ds}{dt}\,\Delta t \approx s^n - s^c,$$

$$\frac{dq}{dt}\,\Delta t \approx q^n - q^c,$$

so that we get a linear equation in $s^n$ and $q^n$. This linear equation is compatible with the set of finite differenced linear equations derived from the flow equations.

## IV. TURBULENCE MODEL

For studying the influence of turbulence on the flow pattern, two different turbulence models were considered. Both are equilibrium algebraic eddy viscosity models based on Prandtl's mixing length hypothesis. A turbulent eddy viscosity $\mu_t$ as well as a turbulent thermal conductivity $\lambda_t$ are calculated such that molecular viscosity $\mu$ and thermal conductivity $\lambda$ in the equations describing a laminar flow are replaced by the effective values

$$\mu_{eff} = \mu + \mu_t ,$$

$$\lambda_{eff} = \lambda + \lambda_t .$$

The two models differ in the underlying assumption that the boundary layer is composed of one region or two regions.

The one-layer model expresses the turbulent eddy viscosity as

$$\mu_t = \rho\, \ell^2 \left| \frac{\partial w}{\partial r} + \frac{\partial u}{\partial z} \right|$$

where $\rho$ is the local density, $\ell$ is Prandtl's mixing length, w and u are the velocity in axial and radial direction, respectively [11]. Density and velocity gradients are determined by the solution of the partial differential equations. The mixing length is given by a correlation. For a steady incompressible flow in a tube Nikuradse [11] experimentally determined that

$$\ell = R \left[ 0.14 - 0.08 \left(1 - \frac{y}{R}\right)^2 - 0.06 \left(1 - \frac{y}{R}\right)^4 \right] .$$

In using this correlation in the unsteady compressible flow simulations, we assume that the correlation still models the turbulence. We try to test this assumption by comparing the simulations with another turbulence model, a so-called two-layer model.

The two-layer model separates the boundary layer in an inner and outer region with different formulations for each region [12,13]. The expression for the eddy viscosity of the inner region is

$$\mu_{t_{in}} = \rho \ell^2 \left| \frac{\partial w}{\partial r} + \frac{\partial u}{\partial z} \right| ,$$

which is the same as the one-layer formulation. The difference is in the definition of the mixing length $\ell$, that is,

$$\ell = \hat{k}\, y\, D,$$

where the von Kármán constant $\hat{k} = 0.4$, y is the distance from the wall and D is the van Driest damping factor. Van Driest suggested that

$$D = 1 - e^{-\frac{y^+}{A^+}} ;$$

where

$$y^+ = \frac{y}{\mu_w} \sqrt{\rho_w \tau_w}$$

with the van Driest constant $A^+ = 26$. The subscript w indicates that these quantities are to be evaluated at the wall surface (y=o). The wall shear stress $\tau_w$ is expressed by

$$\tau_w = \mu_w \frac{\partial w}{\partial y} \Big|_{y=o} .$$

In the outer layer we use

$$\mu_{t_{out}} = \frac{0.0168\, \rho\, w_e\, \delta^*}{[1 + 5.5\, (y/\delta)^6]},$$

where $w_e$ denotes the axial edge velocity, $\delta^*$ the kinematic boundary layer displacement thickness and $\delta$ the boundary layer thickness. In our case $w_e$ is the maximum axial velocity in the cross-section z=constant, this is, the velocity on the axis of symmetry.

Now the two-layer eddy viscosity is expressed as

$$\mu_t = \begin{cases} \mu_{t_{in}} & \text{if } y < y_c \\ \mu_{t_{out}} & \text{if } y > y_c \end{cases} ,$$

where $y_c$ is the first point at which $\mu_{t_{in}}$ exceeds $\mu_{t_{out}}$.

In both cases the turbulent thermal conductivity is calculated by

$$\lambda_t = \frac{\mu_t \cdot c_p}{Pr_t},$$

where the turbulent Prandtl number is $Pr_t = 0.9$ and $c_p$ is the specific heat at constant pressure.

## V. RESULTS.

Computational results are presented for two different types of interior ballistics flows. The first simulates a pure gas expansion flow behind a projectile inside of a constant cross-section tube. The tube is closed at one end by a stationary surface called the breech, and at the other end by a movable flat based projectile. The initial states of the gas are uniform and quiescent. Geometrical data, initial conditions as well as the thermodynamic properties of the gas are listed in Table I. We designate this

TABLE I. Lagrange Gun Parameters

| | |
|---|---|
| Bore Diameter | 20 mm |
| Tube Length | 2.0 m |
| Chamber Length | 0.175 m |
| Projectile Mass | 120 g |
| Ratio of Specific Heats | 1.271 |
| Covolume | $1.08 \cdot 10^{-3}$ $m^3/kg$ |
| Molar Mass | 23.8 g/mole |
| Initial Gas Pressure | 300 MPa |
| Initial Gas Temperature | 3000 K |
| Initial Velocities | 0 m/s |

idealization as the Lagrange gun. The flow in the Lagrange gun is very well suited for studying several important features like the numerical procedure, boundary layer development, laminar and turbulent axisymmetric flow patterns, and heat transfer to the tube wall. Since an expansion flow is quite removed from the phenomena occurring during a ballistic cycle, a second type of flow which has the time-dependent pressure and temperature profiles as in a real tube weapon is simulated by adding heat and mass to the one-phase flow via source terms. An empirical burning law for pressure-dependent sources is used. The sources move with the flow. We designate this idealization as the real gun. The essential parameters for the real gun differ from Table I only with respect to the initial conditions. The initial gas pressure is assumed to be ambient pressure (0.1 MPa) and the initial gas temperature to be ambient temperature (293 K). In all cases involving the real gun simulation, the projectile is released from its initial position when the pressure at the projectile base reaches 30 MPa.

Because an implicit finite difference scheme is used, no stability restriction exists on the size of the time step for a given mesh. We have chosen a constant time-step of 10 μs for computing all results shown here. The computational mesh consists of 49 uniformly or nonuniformly spaced mesh points in the axial direction and 19 nonuniformly spaced mesh points in the radial direction. To obtain a finer spatial resolution near the solid wall, the mesh points are more concentrated near the wall than near the center line. An example of a 19x49 computational mesh, which is used in most of the computations, is shown in Figure 1. The smallest grid size in radial direction at the bore surface is 7.7 μm . The mesh for computing the heat conduction in the tube wall is generated in the same way with the same mesh distribution in axial direction as on the gas side, and a corresponding mesh concentration near the inner bore surface. Both the size of time-step, and number of grid points seem to be a reasonable compromise between accuracy and computation time.

The performances of the Lagrange and real gun for laminar flows with adiabatic boundaries are compared in Figures 2-5. The histories of the gas pressures, temperatures at the center of the breech and projectile, and the velocities and displacements of the projectile are shown. The main differences are the temporal distribution of all the quantities, and the final values of the muzzle velocities. For the real gun simulations, the pressure at the projectile base reaches 30 MPa at about 2.3 ms at which time the projectile begins to accelerate down the tube.

Qualitative results for a laminar flow with adiabatic walls in the Lagrange gun are presented in Figures 6-10. We caution the reader of the different view points in the three-dimensional graphs. They were chosen to best display the results. The spatial distribution of the axial velocity is shown in Figures 6 and 7 at 3.75 ms. At this time, the projectile exits the gun tube with a muzzle velocity of 623 m/s. Figure 6 shows the axial velocity field when a uniformly axially spaced mesh in Figure 1 is used. Figure 7 shows the same quantity but computed on a nonuniformly spaced axial mesh. The radial distribution of the mesh points are the same in both figures. For a given axial position the axial velocity is constant across much of the radius of the tube (the core flow region), and decreases to zero only very close to the wall (the boundary layer region). The boundary layer is the result of the no-slip condition (w=o) at the wall. The thickness of the boundary layer is approximately 0.2 to 0.3 mm. In the axial direction, the velocity is distributed linearly in the core region between the zero value at the breech and the muzzle velocity. The three-dimensional temperature and pressure distributions are given in Figures 8 and 9. Due to the adiabatic boundary condition at the wall surface the heat generated by the viscous forces near the tube wall cannot transfer to the tube wall, and the gas temperature rises towards the wall surface. Here again the boundary layer is only 0.2 to 0.3 mm thick. The pressure, however, stays constant in radial direction over the entire cross-section. The assumption in boundary layer theory that the radial pressure gradient is zero would be valid in this example. Figure 10 shows the 3-D graph of the radial gas velocity. In approximately the first 70% of the distance to the projectile, the radial velocity is negative, i.e., the flow is directed towards the center line (C-L). Thereafter, it is positive and increases remarkably towards the projectile base. Only very close to the projectile does the value of the radial velocity drop rapidly from its maximum value to zero. Of course, the radial velocity is small since it is induced

only by molecular viscosity and heat conductivity. The results in Figures 2-10 agree very well with both a one-dimensional solution of the core flow using the method of characteristics, and the two-dimensional numerical calculations of Heiser and Hensel [14,15].

The assumption of a laminar flow may not be realistic for the Lagrange gun considered above. Because the Reynolds number based on the tube's diameter and muzzle velocity is of the order of $10^7$, the effects of turbulence should be examined. In comparison to the laminar flow several important differences occur when the two algebraic turbulence models described in IV are applied. However, the difference in the results between the two turbulence models is insignificant. Consequently, we do not specify the type of model in the results. First of all, we see the difference in the projectile performance between the laminar and turbulent type flow. This is demonstrated in Figure 11 with the pressure histories at the breech, in Figure 12 with the temperature histories at the breech, and in Figure 13 with the projectile velocity histories. In the turbulent flow simulation, the projectile velocity is about 50 m/s less than for the laminar flow at muzzle time. The axial velocity flow field at the time of muzzle clearance is shown in Figure 14. The velocity boundary layer is fully developed between the center line and the tube wall. The axial velocity overshoots the projectile velocity near the center line. This overshoot is related to the radial gas velocity (Figure 15) which is one to two order of magnitude larger than for the laminar flow (Figure 10) near the projectile base. In this region, the radial gas flux is toward the tube wall as before, and transports mass away from center-line which in turn can accelerate the axial flow. The radial variation of axial velocity (Figure 16), radial velocity (Figure 17) and temperature (Figure 18) taken 0.25 m upstream of the muzzle show some details about the differences between the laminar and turbulent flows at the time of muzzle clearance. Boundary layer calculations in Ref. [7] show comparable trends between laminar and turbulent flow. However, the flow patterns computed by different types of turbulence models, e.g., non-algebraic models, may differ. An experiment corresponding to this idealized expansion flow is needed to validate a turbulence model. Such experiments are being attempted at the French-German Institute (ISL) in France and the Ernst-Mach-Institut (EMI) in Germany.

Besides turbulence, heat transfer is important in an interior ballistics cycle because in reality energy is lost from the gas flow, and is gained by the gun tube wall. This heating of the wall's surface does influence gun tube erosion. Our heat transfer model together with the calculation of the heat conduction in the wall couples the unsteady behavior in both media, and is discussed in Section III. We now show results concerning the laminar Lagrange gun expansion flow with heat transfer from the gas to the tube wall. Initially the wall is supposed to be at ambient temperature. The thermal properties of the gun barrel are characterized by the barrel material density $\rho_w = 7.8$ kg/m$^3$, thermal conductivity $\lambda_w = 43$ W/(m K) and specific heat $c_w = 460$ J/(kg K). The pressure and temperature histories at the breech, and projectile velocity history are given in Figures 11-13, and can be compared with the other two simulations. The spatial profile of the axial velocity, radial velocity, pressure and temperature are plotted in Figures 19-22, respectively, at the time of muzzle clearance. When comparing the figures to those of the laminar flow with adiabatic walls, we find with heat transfer that the velocity boundary layer as well as the temperature boundary layer are

thinner, that the radial velocity is about one order of magnitude higher, and that the radial velocity is always directed towards the wall. For example, the velocity boundary layer is only about 0.05 mm thick. Detailed comparisons across the boundary layers regions for axial velocities and temperatures are presented in Figures 23 and 24, and across the tube cross-section for the radial velocities is shown in Figure 25. All these profiles are at 0.25 m from the muzzle at the time of muzzle clearance. Figure 26 shows the history of the wall surface temperature at two different locations along the tube wall. One is taken at 50 mm away from the breech. This wall point belongs to the chamber, and is heated up from the beginning of the flow cycle. The second wall point is 250 mm away from the breech, and is heated up only after the projectile has passed it (at 0.47 ms). The maximum wall temperature occurs early in the 4 ms cycle. The laminar gas layer close to the tube wall cools rapidly because the transport of heat in the tube wall by conduction is much faster than the transport of heat by conduction and convection on the gas side towards the wall. To obtain the heat transfer precisely from the gas to the wall, we need to compute the radial temperature gradient on both sides of the inner bore surface as accurately as possible. The temperature boundary layer, however, is very thin as it is shown in Figures 22 and 24 which implies very small grid sizes must be used close to the wall. The smallest one in radial direction for Figure 26 is 0.73 μm .

The final simulations are for the one-phase flows in the real gun, i.e., where mass and energy are added via source terms. The pressure and temperature histories at the breech and projectile, and the projectile velocity and displacement histories for the laminar flow simulation and for the real gun are given in Figures 2-5. Figures 27-40 are a series of three-dimensional profiles which characterize both the laminar and turbulent flows inside an adiabatic tube. The group of figures associated with the laminar flow are at two times: the first at 3.6 ms when maximum pressure is achieved and the second at 5.3 ms when the projectile exits the tube. For these calculations a nonuniform radial grid is used with the minimum grid size of 0.73 μm because of the thinness of the boundary layer. We excluded the profile of the radial velocity at 3.6 ms because the magnitude of this component was smaller than 0.05 m/s. Three-dimensional profiles show the turbulent flow fields when maximum pressure occurs at 3.6 ms (Figures 34-37), and when the projectile exits the tube at 5.49 ms (Figures 38-40). As in the Lagrange gun, the differences are minimal between the simulations with different turbulence model, but significant between the laminar and turbulence simulations.

## VI. SUMMARY

Special interior ballistic phenomena are investigated by using the DELTA computer code designed to calculate the multidimensional, two-phase flow behind an accelerating projectile inside a gun tube. Details of the heat transfer and turbulence submodels used in the simulations are given in this paper. The general mathematical model and numerical algorithm are described in a companion paper Ref. [1]. Results are given for two types of idealized , one-phase interior ballistics gun simulations: the pure expansion flow, and the flow with moving mass and heat sources. Comparison are made between laminar and turbulent flows as well as between flows in an adiabatic tube and in a tube that allows heat transfer.

## REFERENCES


[ 1 ] J.A Schmitt, "A Numerical Algorithm for the Multidimensional, Multiphase, Viscous Equations of Interior Ballistics," Proceedings of the Second Army Conference on Applied Mathematics and Computing, Rensselaer Polytechnic Institute, Troy, New York, May 22-25, 1984.

[ 2 ] "Fluid Dynamics Aspects of Internal Ballistics," AGARD Advisory Report No. 172, 1982.

[ 3 ] A.K.R. Celmins, J.A. Schmitt, "Modeling of Gas-Solid Phenomena in Interior Ballistics," Proceedings Seventh International Symposium on Ballistics, The Hague, Netherlands, April 19-21, 1983.

[ 4 ] S. Shelton, A. Bergles, P. Saha, "Study of Heat Transfer and Erosion in Gun Barrels," Air Force Armament Laboratory, Eglin Air Force Base, Florida, AFATL-TR-73-69, 1973.

[ 5 ] P. Gough, "Modeling of Rigidized Gun Propelling Charges,"Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland ARBRL-CR-00518, 1983.

[ 6 ] C.W. Nelson, J.R. Ward, "Calculation of Heat Transfer to the Gun Barrel Wall," J. Ballistics 6 (3), pp. 1518-1524, 1982.

[ 7 ] E.P. Bartlett, L.W. Anderson, R.M. Kendall, "Time-Dependent Boundary Layers with Application to Gun Barrel Heat Transfer," Proceedings 1972 Heat Transfer Fluid Mech. Institute, Stanford University, CA, 1972.

[ 8 ] A.C. Buckingham, "Modeling Propellant Combustion Interacting with an Eroding Solid Surface," Lawrence Livermore Laboratory, UCRL-83727, 1980.

[ 9 ] M.J. Adams, H. Krier, "Unsteady Internal Boundary Layer Analysis Applied to Gun Barrel Wall Heat Transfer," Int. J. Heat Mass Transfer, Vol. 24, No. 12, pp. 1925-1935, 1981.

[10 ] Aerospace Structural Metal Handbook, "Ferrous Alloys," 1973.

[11 ] H. Schlichting, "Boundary Layer Theory," McGraw-Hill, 1968.

[12 ] M. W. Rubesin, "Numerical Turbulence Modeling," AGARD-LS-86.

[13 ] M. I. Kussoy, J. R. Viegas, C. C. Horstman, "Investigation of a Three-Dimensional Shock Wave Separated Turbulent Boundary Layer," AIAA J., Vol. 18 (1980), No. 12, pp. 1477.

[14 ] R. Heiser, D. Hensel, "AMI: Ein achsensymmetrisches Modell der Innenballistik, Teil 1: Laminare Einphasenströmung ohne Wärmeübergang (AMI: An Axisymmetric Model of Interior Ballistics, Part 1: Laminar One-Phase Flow without Heat Transfer)," Fraunhofer-Institut für Kurzzeitdynamik, Ernst-Mach-Institut, Abteilung für Ballistik, Weil am Rhein, FRG, Report No. 4/80, 1980.

[15 ] R. Heiser, D. Hensel, "Berechnung der Gasströmung in einem Waffenrohr mit Hilfe des zweidimensionalen AMI-Modells (Calculation of the Gas Flow Inside a Gun Tube Using the Two-Dimensional AMI Model)," Fraunhofer-Institut für Kurzzeitdynamik, Ernst-Mach-Institut, Abteilung für Ballistik, Weil am Rhein, FRG, Report No. E 1/81, 1981.

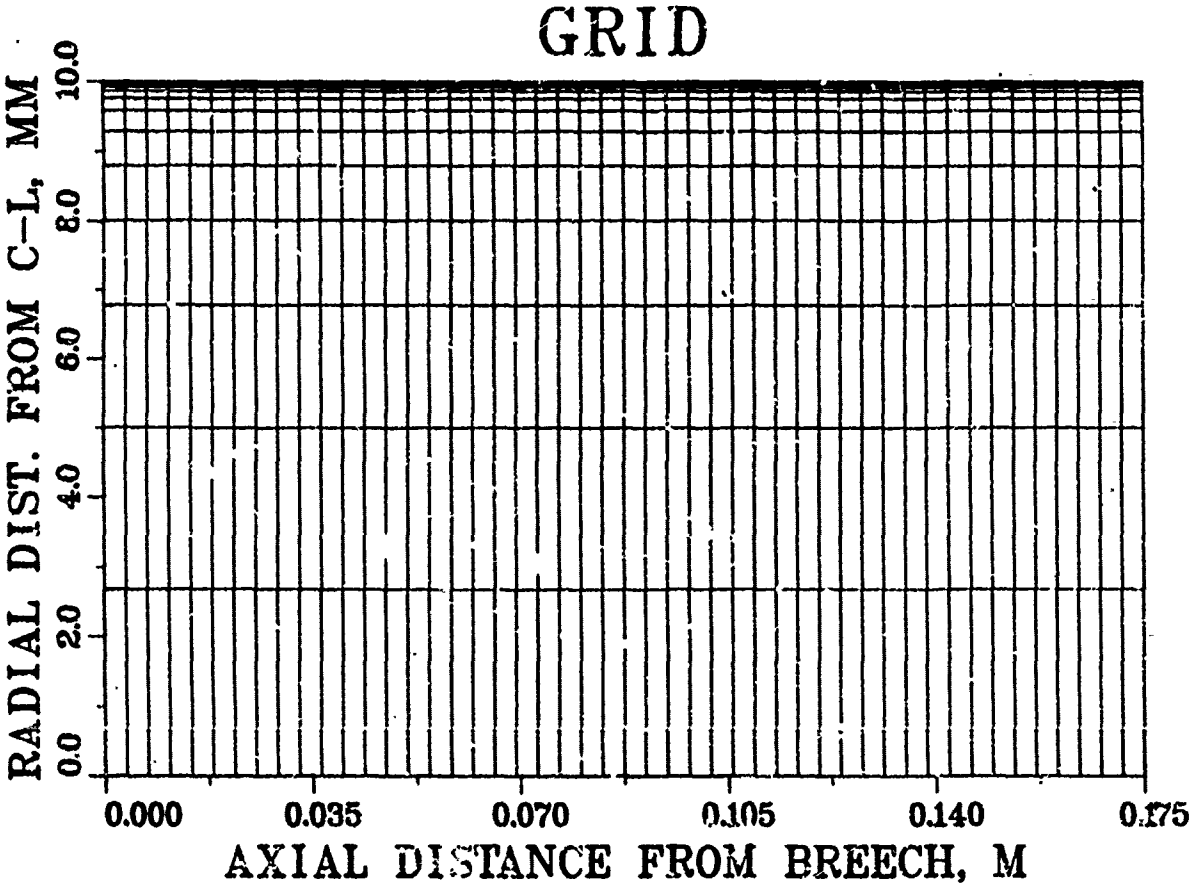



Figure 1. Standard computational mesh 19x49.
Minimum radial grid size is 7.7 μm .

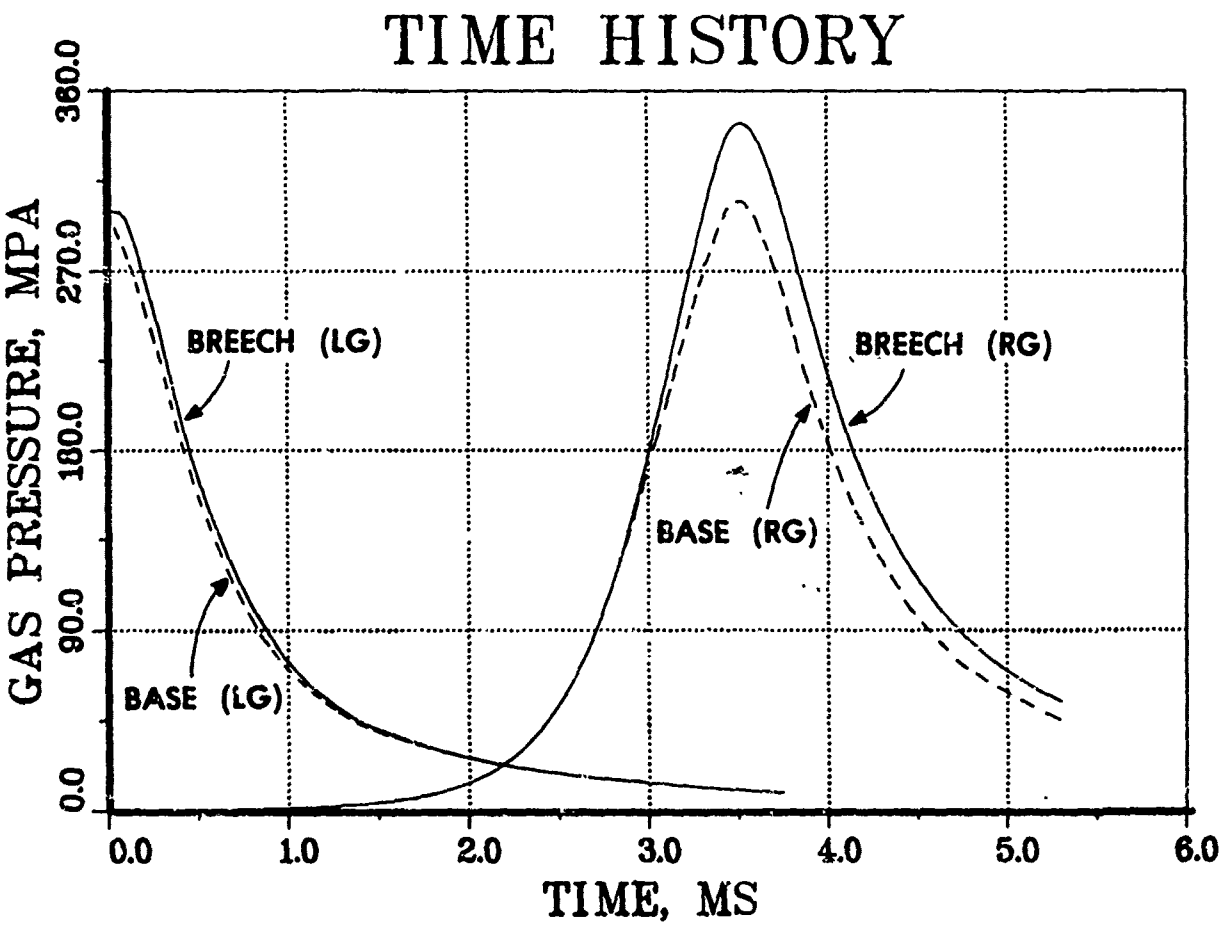



Figure 2. Pressure histories at the center of the breech and the projectile base for the laminar flow with adiabatic walls in the Lagrange gun (LG) and real gun (RG) simulations.

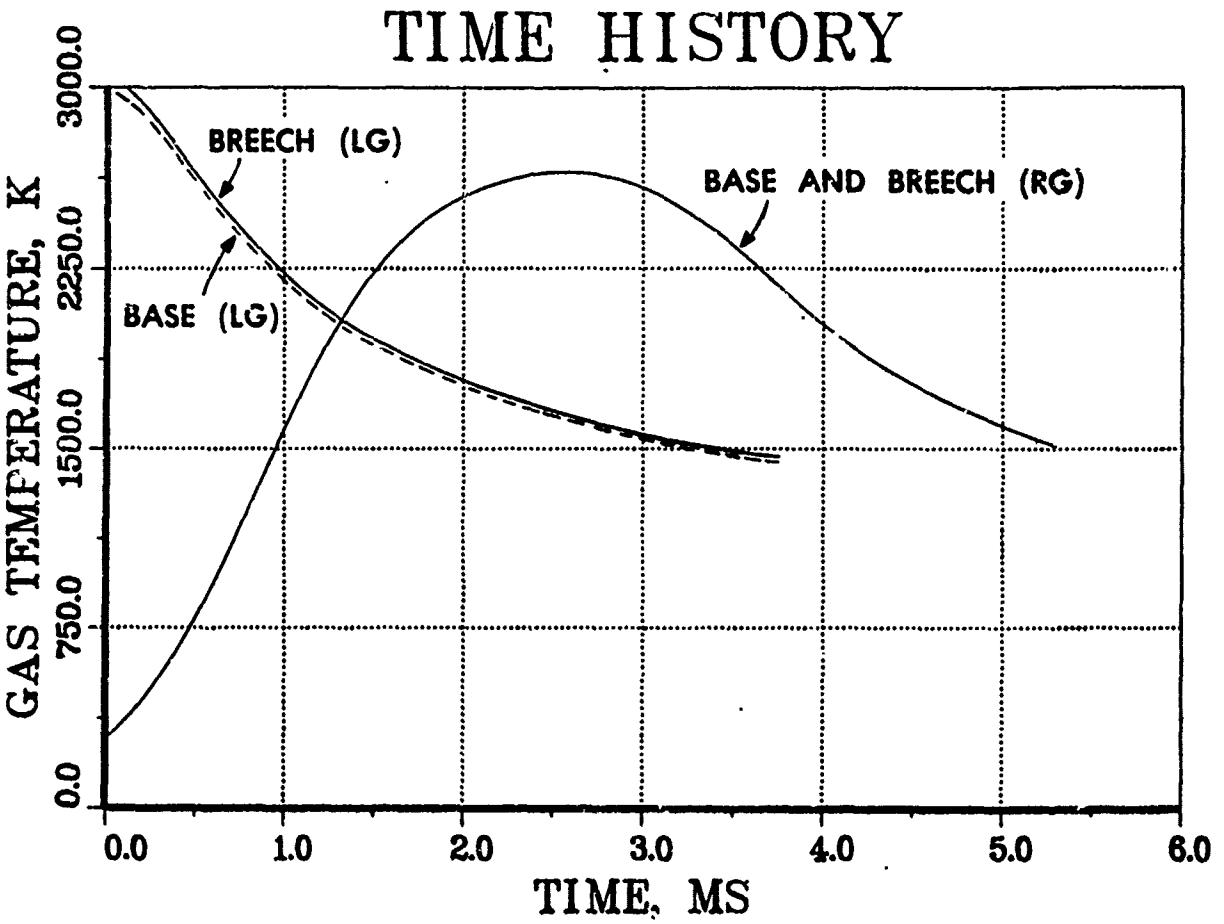



Figure 3. Temperature histories at the center of the breech and the projectile base for laminar flow with adiabatic walls for the Lagrange gun (LG) and real gun (RG) simulations.

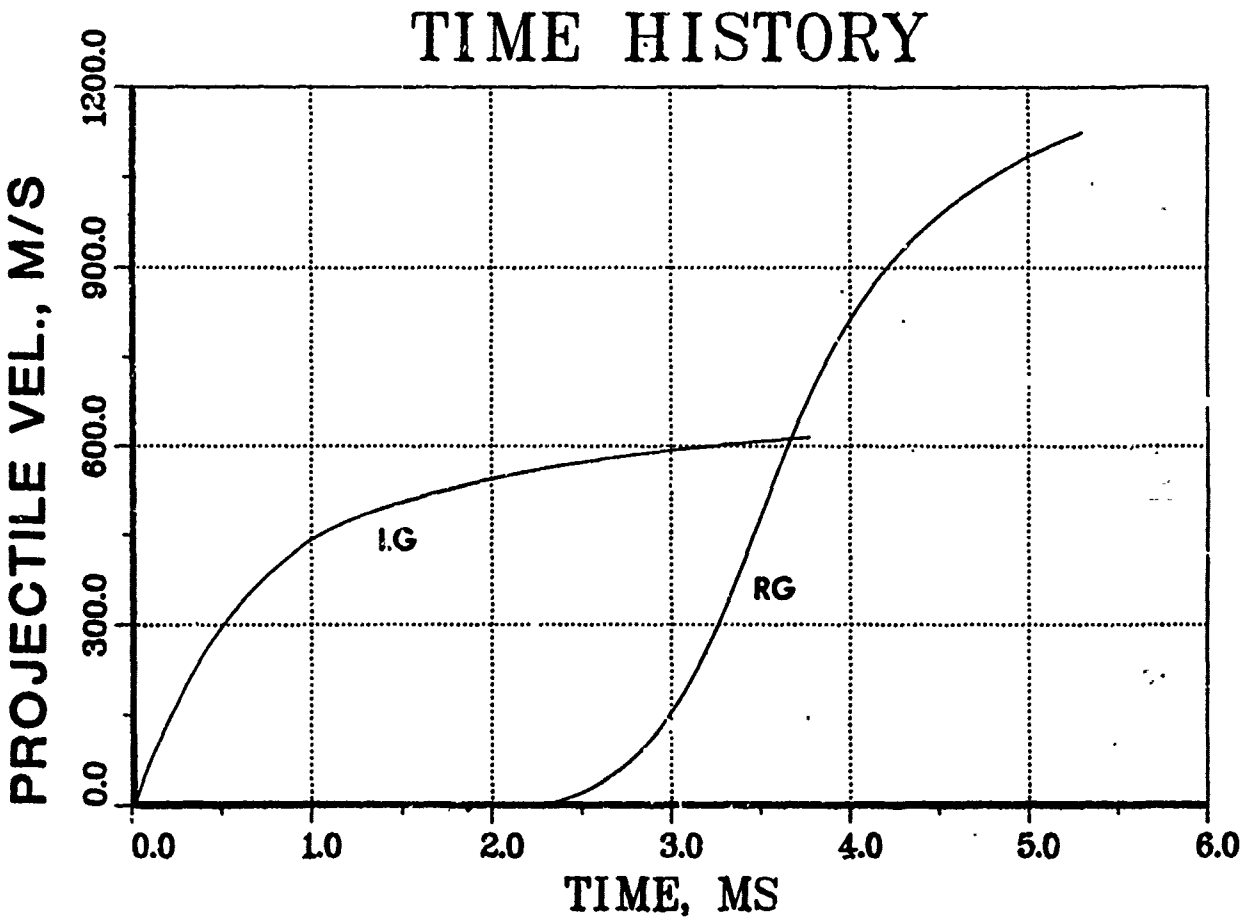



Figure 4. Projectile velocity histories for laminar flow with adiabatic walls in the Lagrange gun (LG) and real gun (RG) simulations.

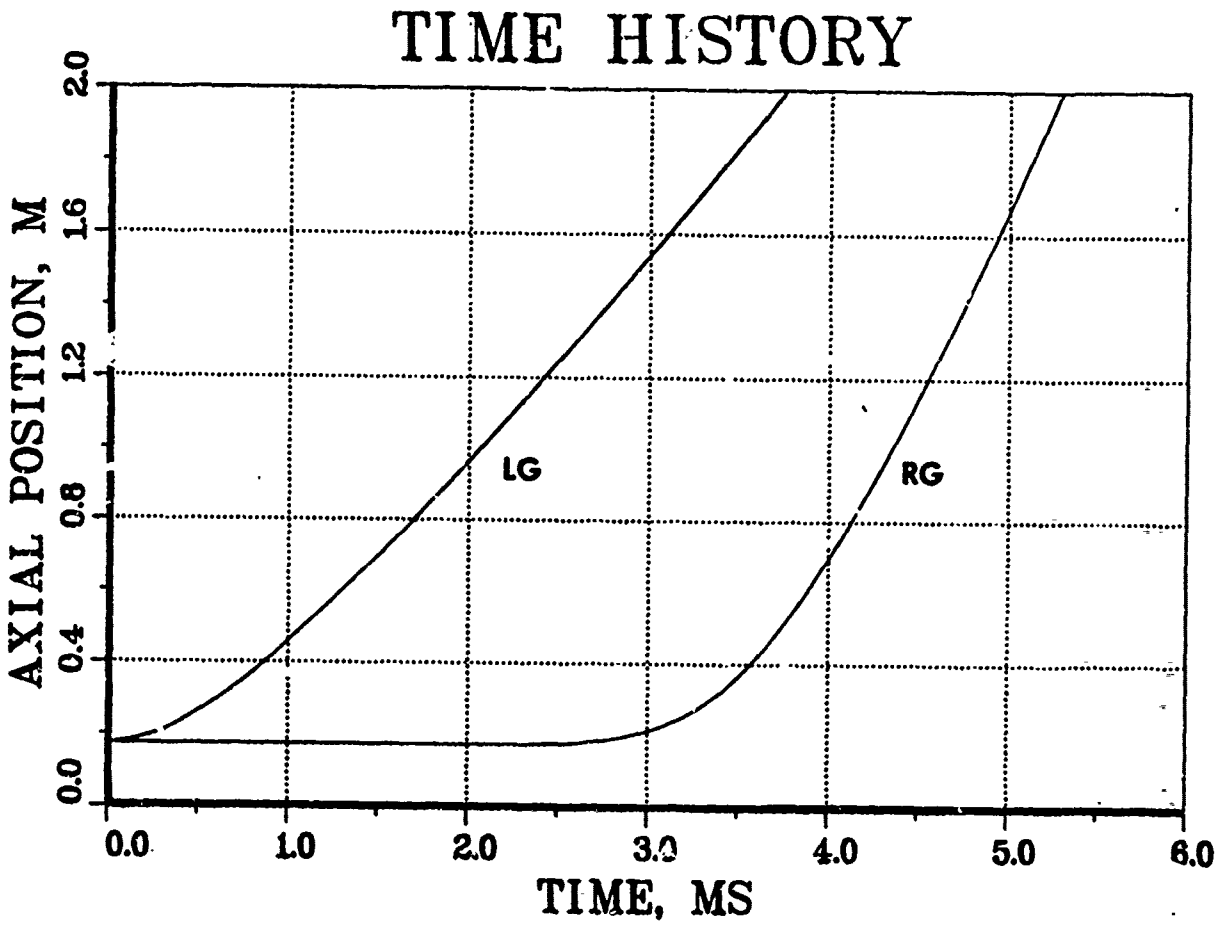



Figure 5. Projectile displacement from the breech for laminar flow with adiabatic walls in the Lagrange gun (LG) and real gun (RG) simulations.

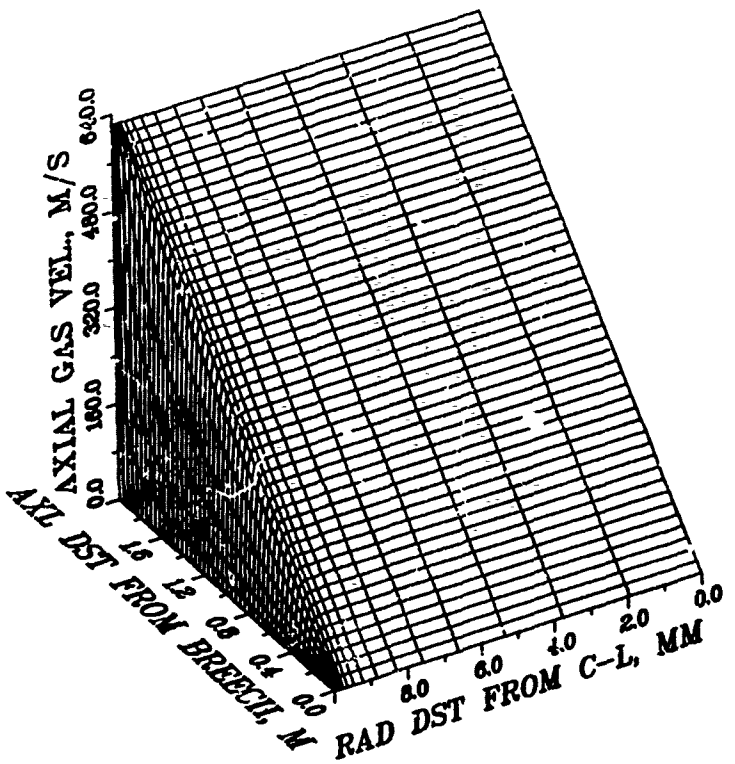



Figure 6. Lagrange gun, laminar flow, adiabatic walls: Spatial distribution of the axial gas velocity at the time of muzzle clearance. Mesh is uniformly spaced in axial direction and nonuniformly spaced in radial direction.

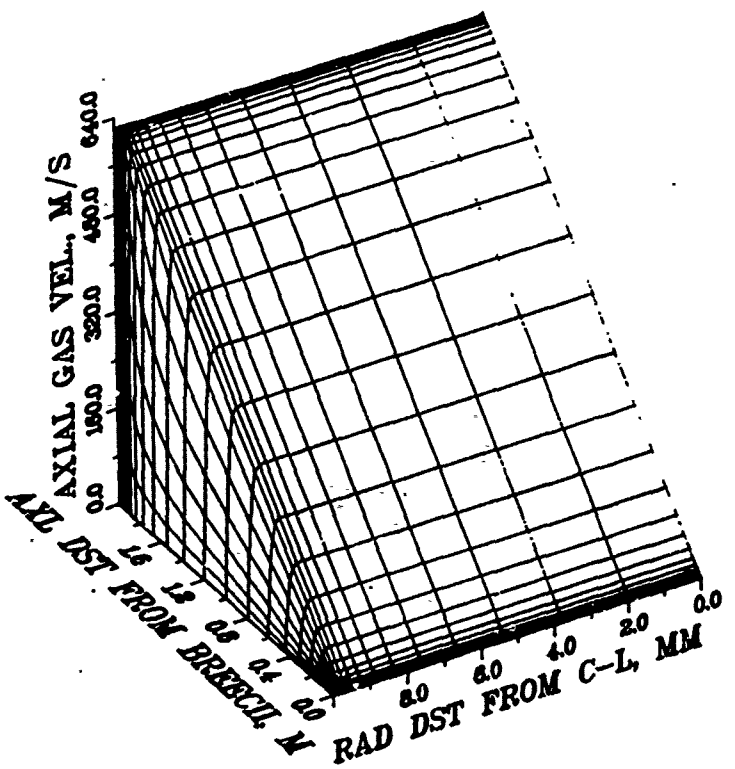



Figure 7. Lagrange gun, laminar flow, adiabatic walls: Spatial distribution of the axial gas velocity at the time of muzzle clearance. Mesh is nonuniformly spaced in both directions.

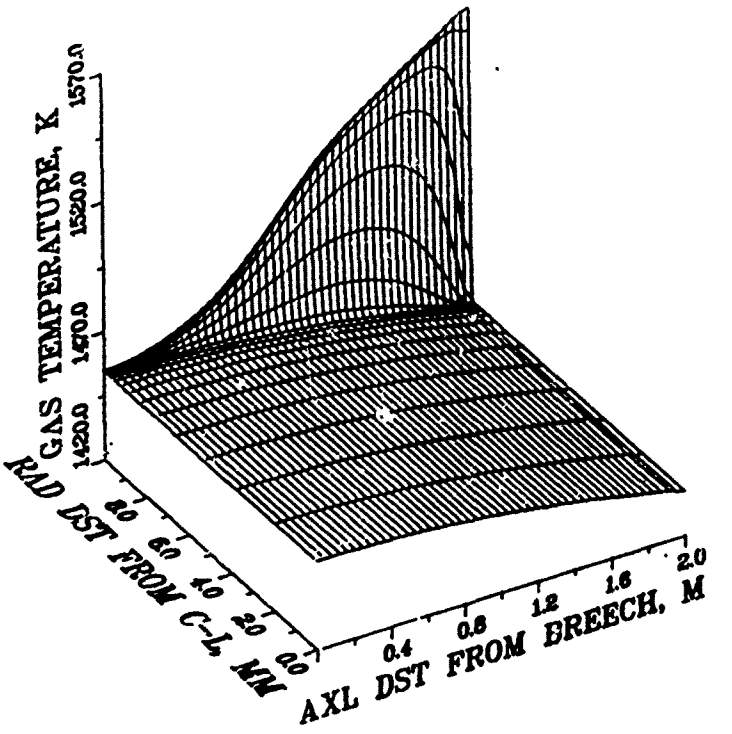



Figure 8. Lagrange gun, laminar flow, adiabatic walls: Spatial distribution of the gas temperature at the time of muzzle clearance.

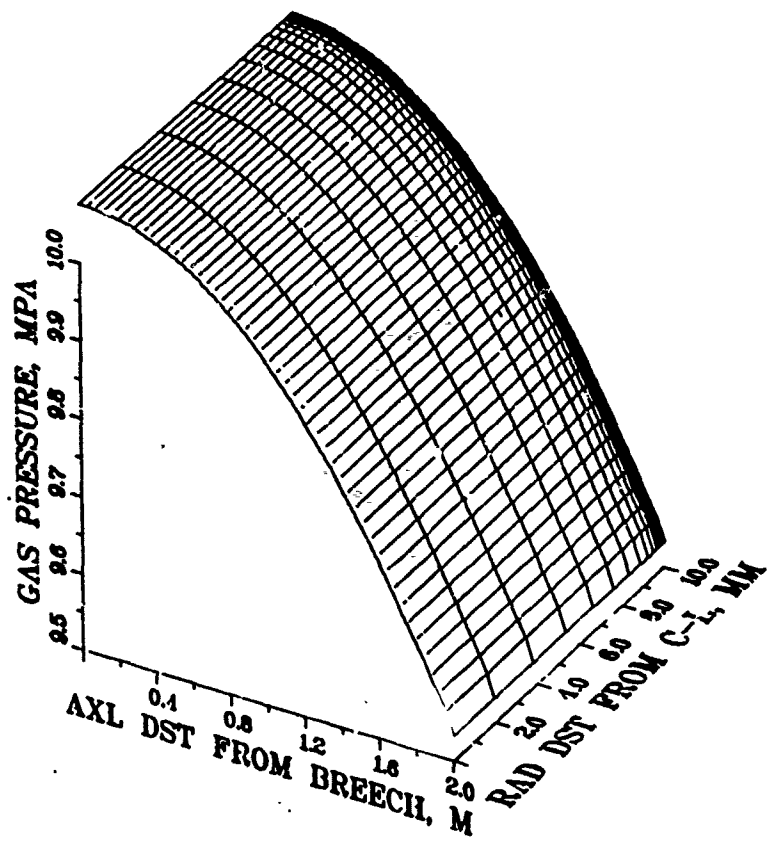



Figure 9. Lagrange gun, laminar flow, adiabatic walls: Spatial distribution of the gas pressure at the time of muzzle clearance.

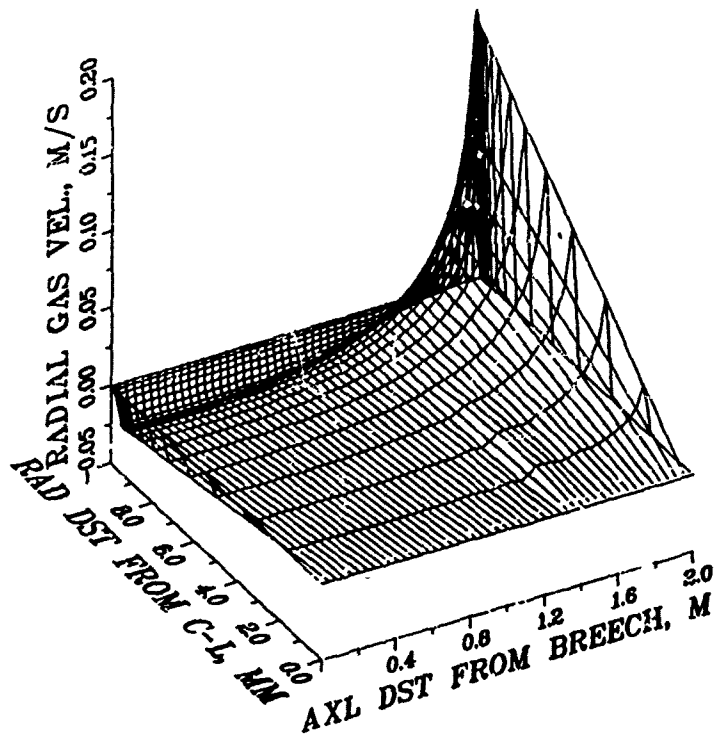



Figure 10. Lagrange gun, laminar flow, adiabatic walls: Spatial distribution of the radial gas velocity at the time of muzzle clearance.

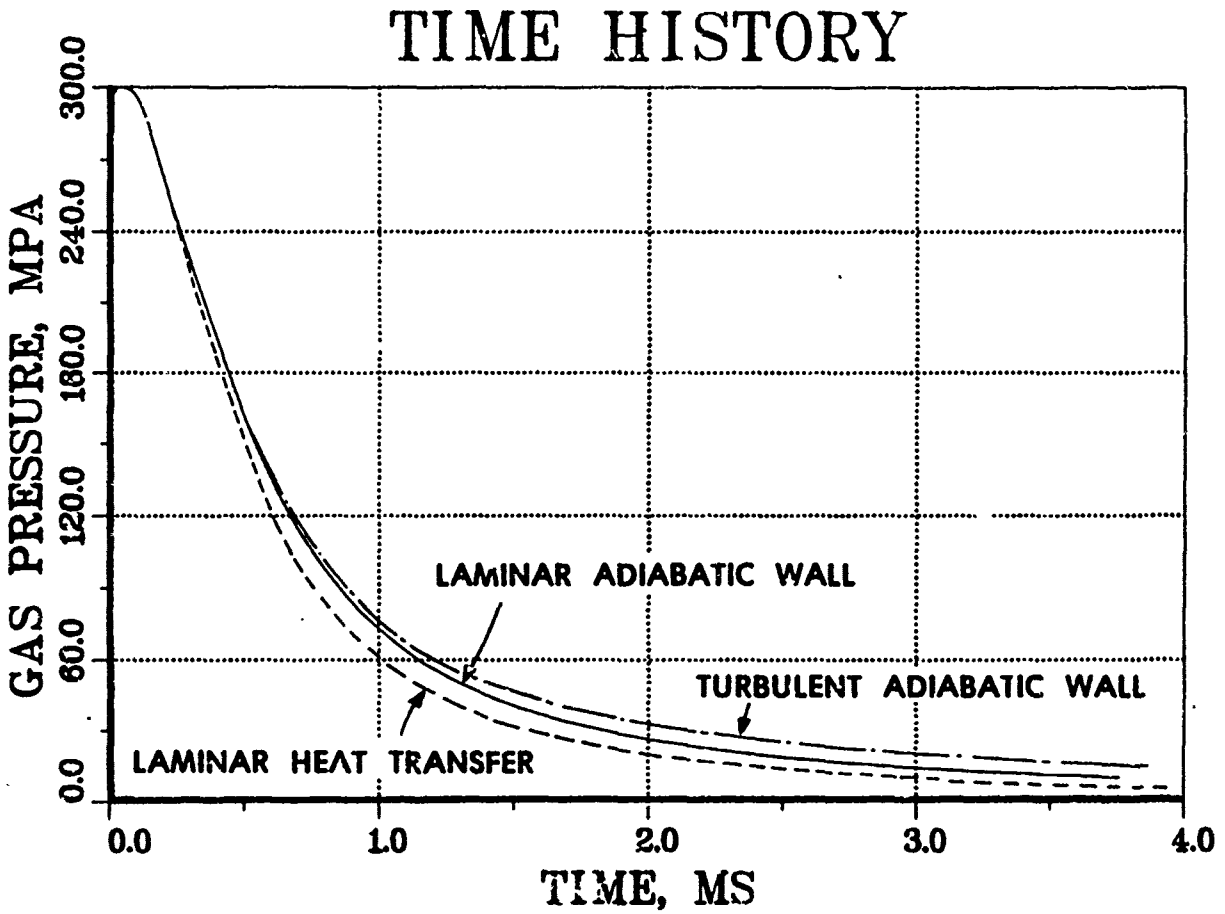



Figure 11. Lagrange gun: Pressure histories at the center of the breech for both laminar and turbulent flows with adiabatic walls, and for laminar flow with heat transfer.

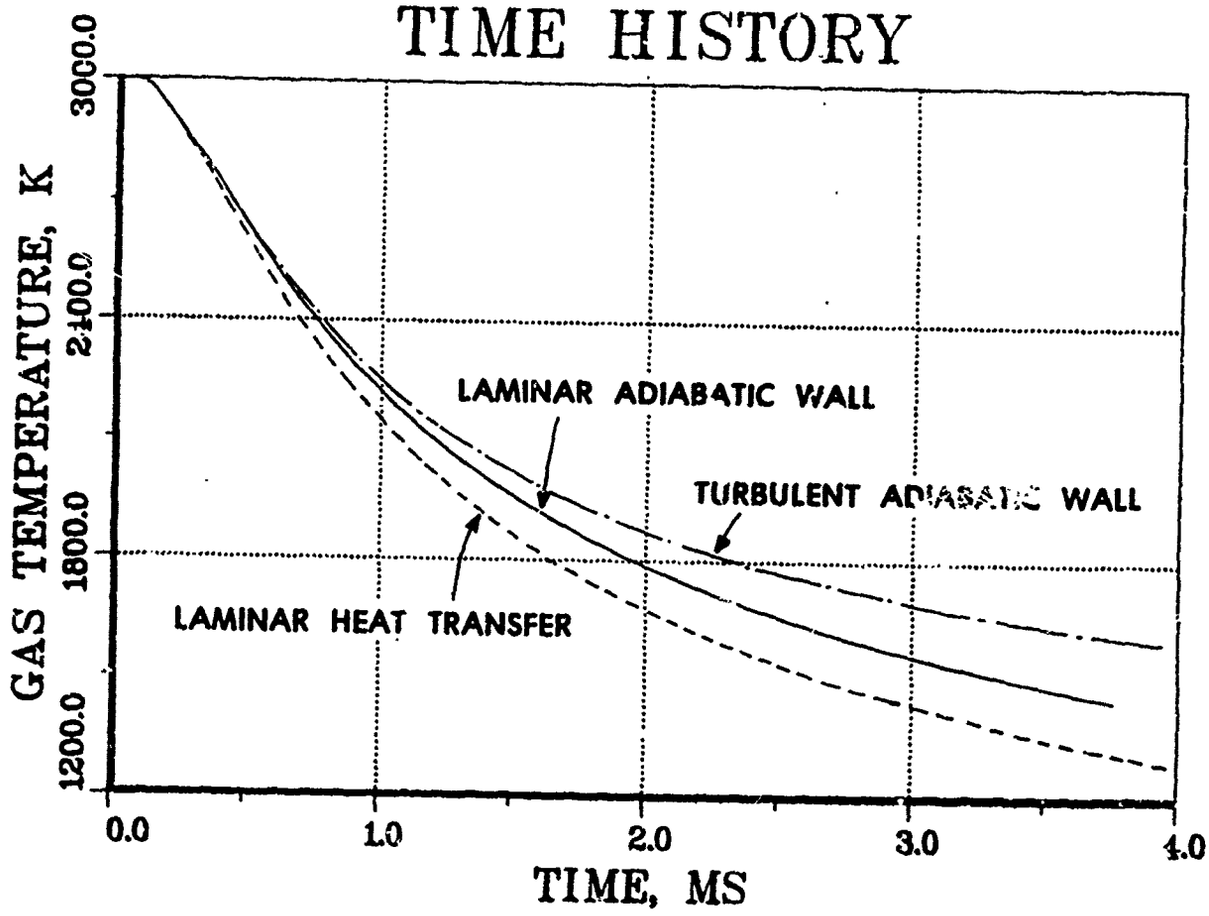



Figure 12. Lagrange gun: Temperature histories at the center of the breech for both laminar and turbulent flows with adiabatic walls, and for laminar flow with heat transfer.

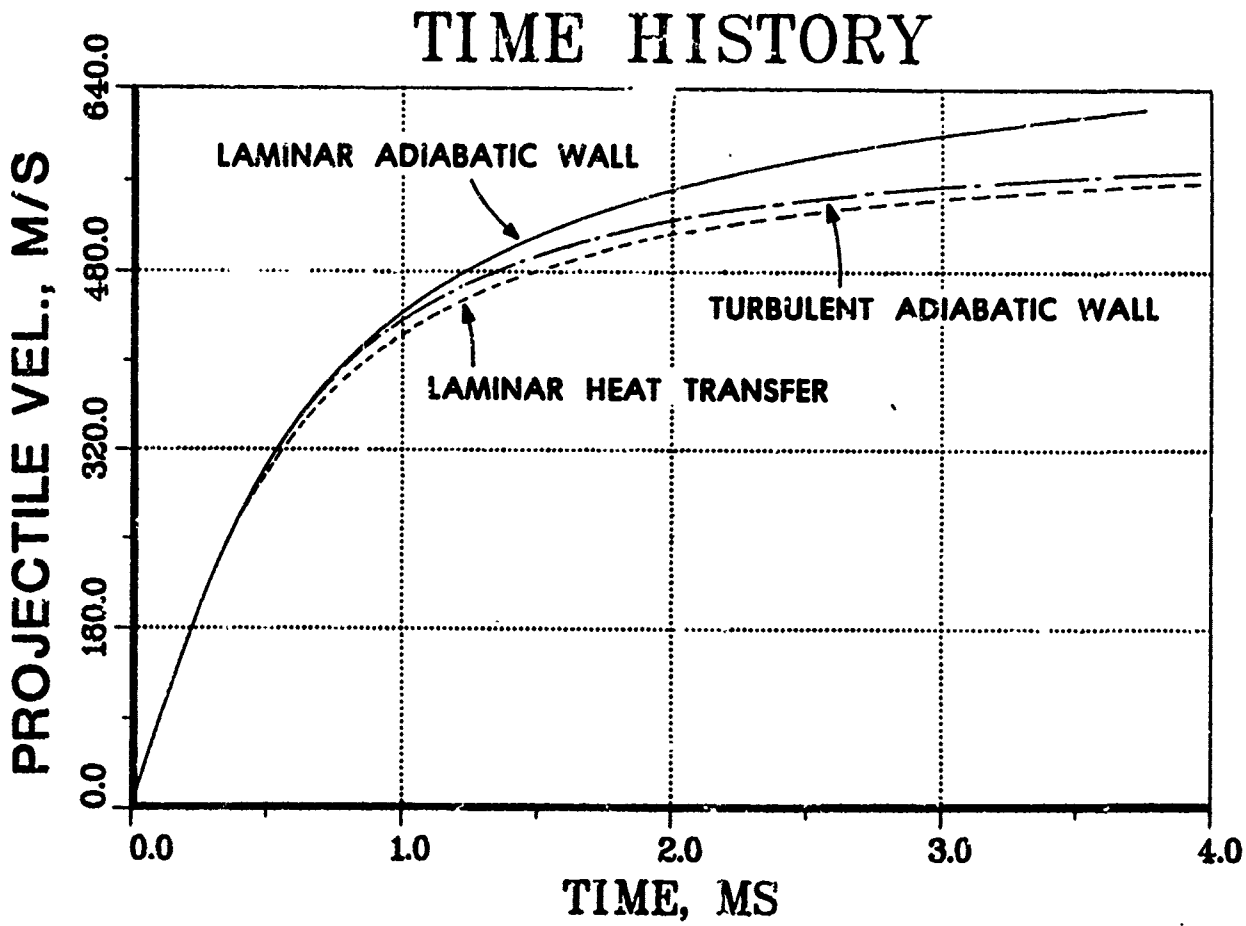



Figure 13. Lagrange gun: Projectile velocity histories for both laminar and turbulent flows with adiabatic walls and for laminar flow with heat transfer.

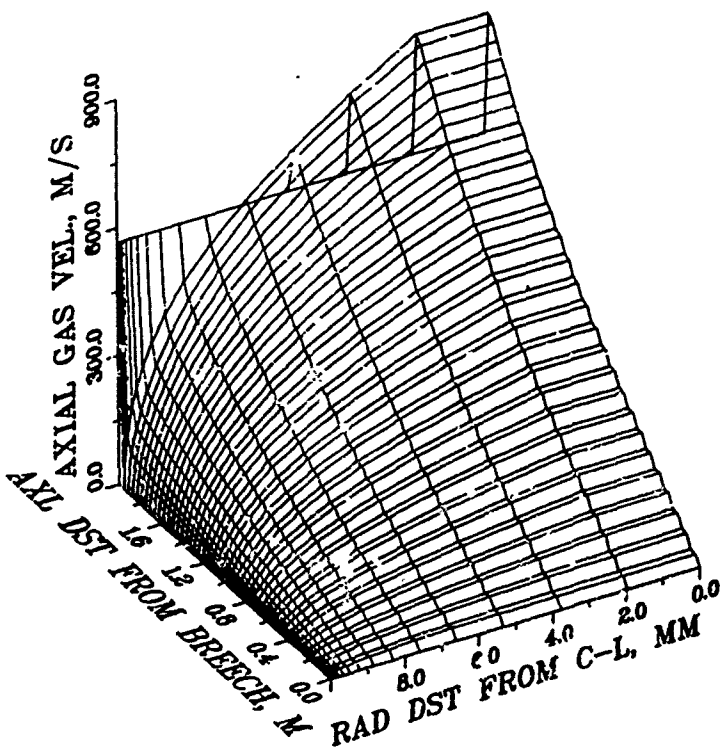



Figure 14. Lagrange gun, turbulent flow, adiabatic walls: Spatial distribution of the axial gas velocity at the time of muzzle clearance.

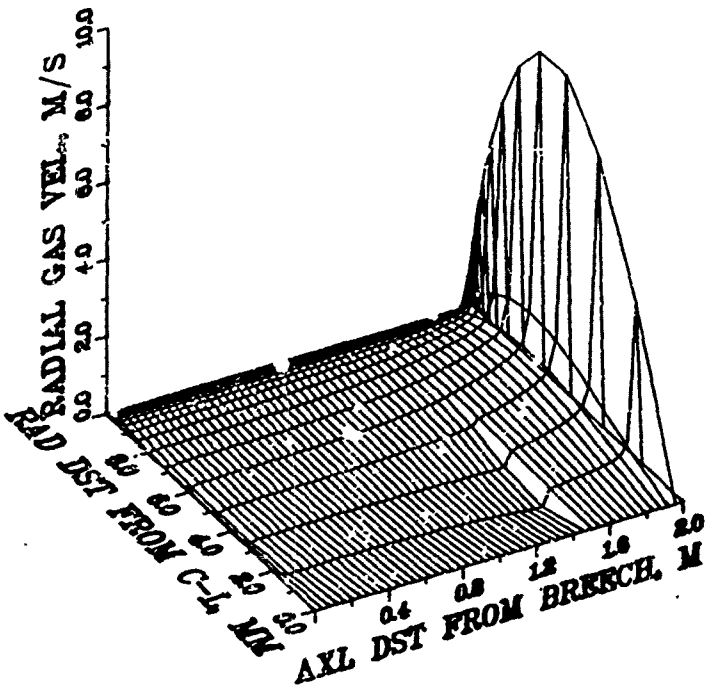



Figure 15. Lagrange gun, turbulent flow, adiabatic walls: Spatial distribution of the radial gas velocity at the time of muzzle clearance.

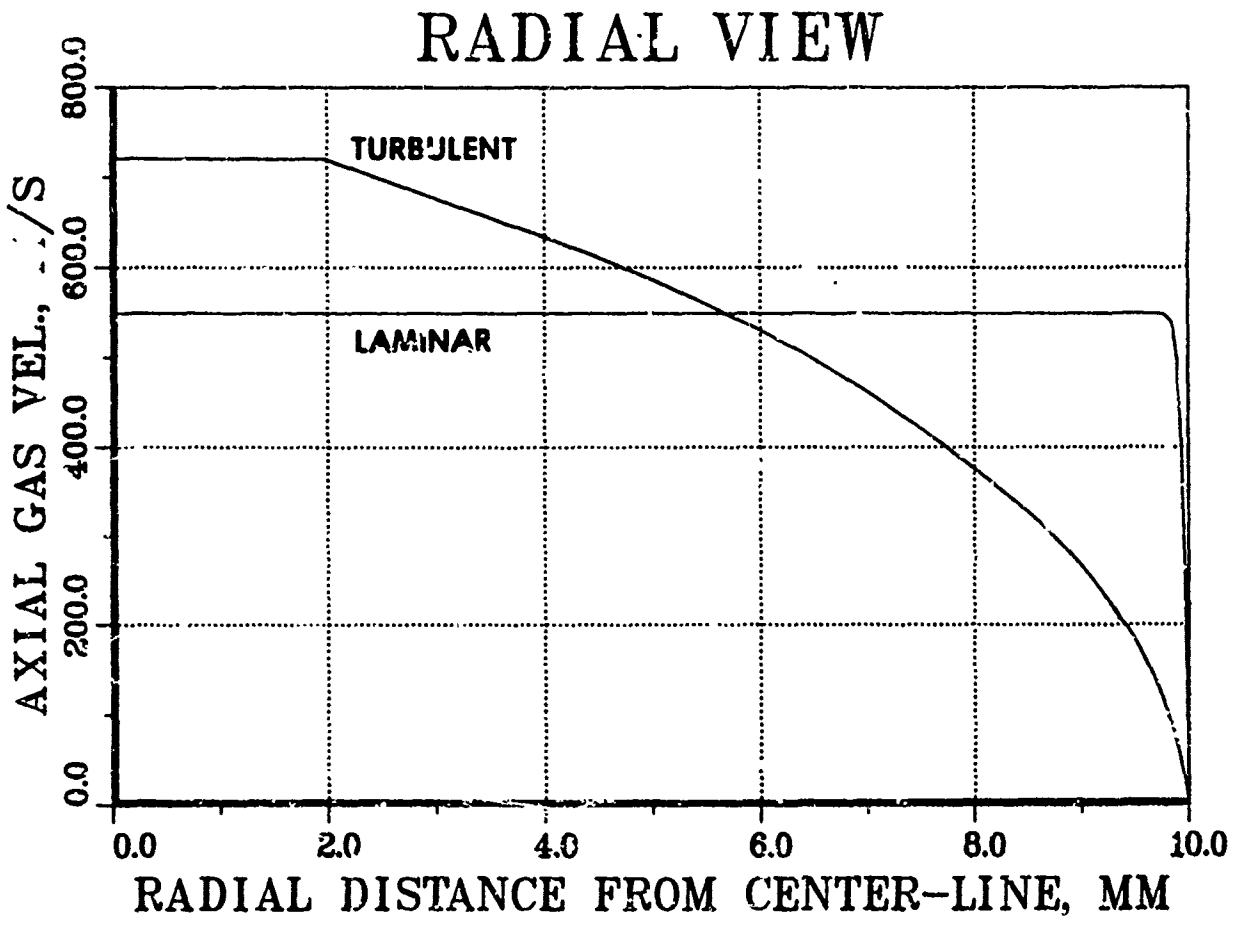



Figure 16. Lagrange gun, adiabatic walls: Radial profiles of the axial gas velocity for both laminar and turbulent flow at the time of muzzle clearance at 0.25 m away from the muzzle.

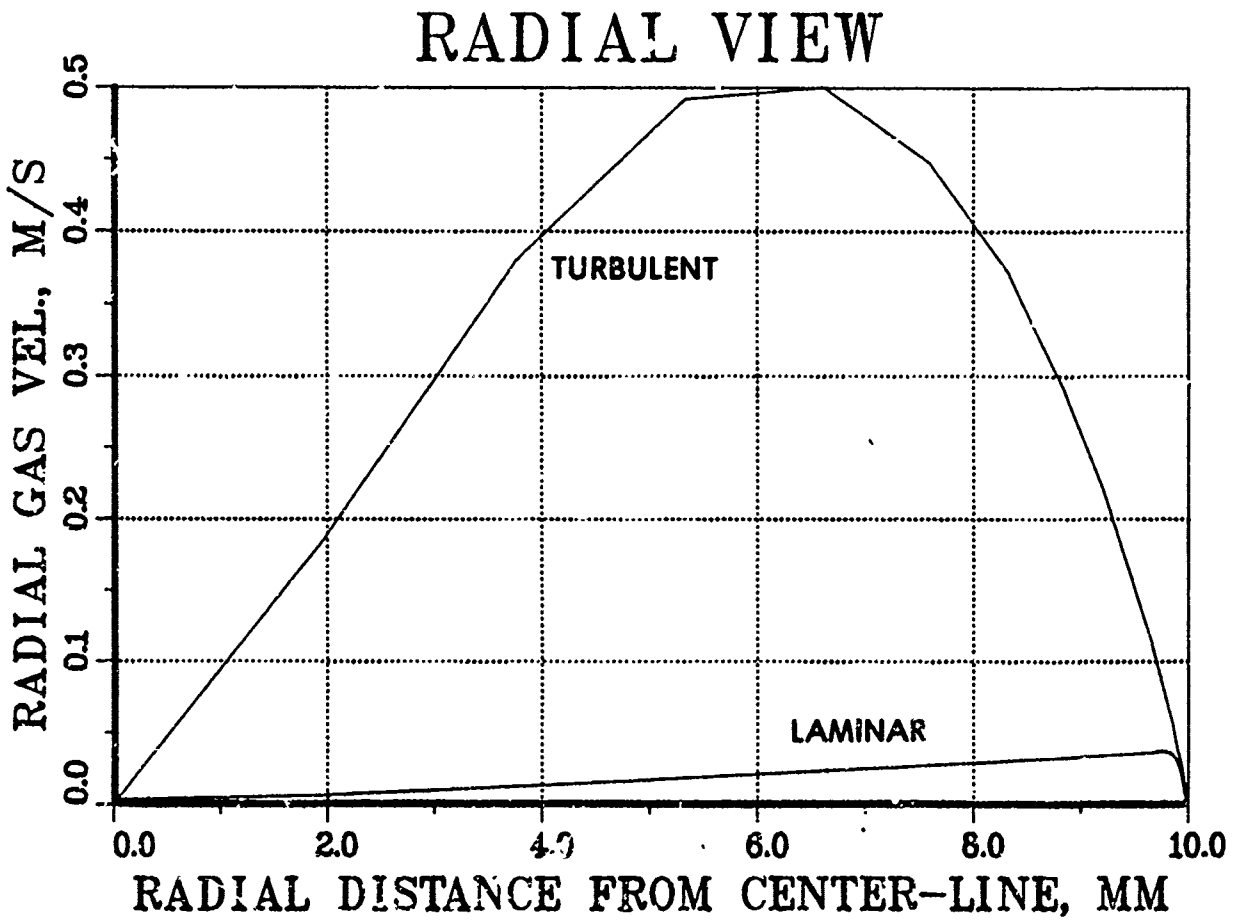



Figure 17. Lagrange gun, adiabatic walls: Radial profiles of the radial gas velocity for both laminar and turbulent flow at the time of muzzle clearance at 0.25 m away from the muzzle.

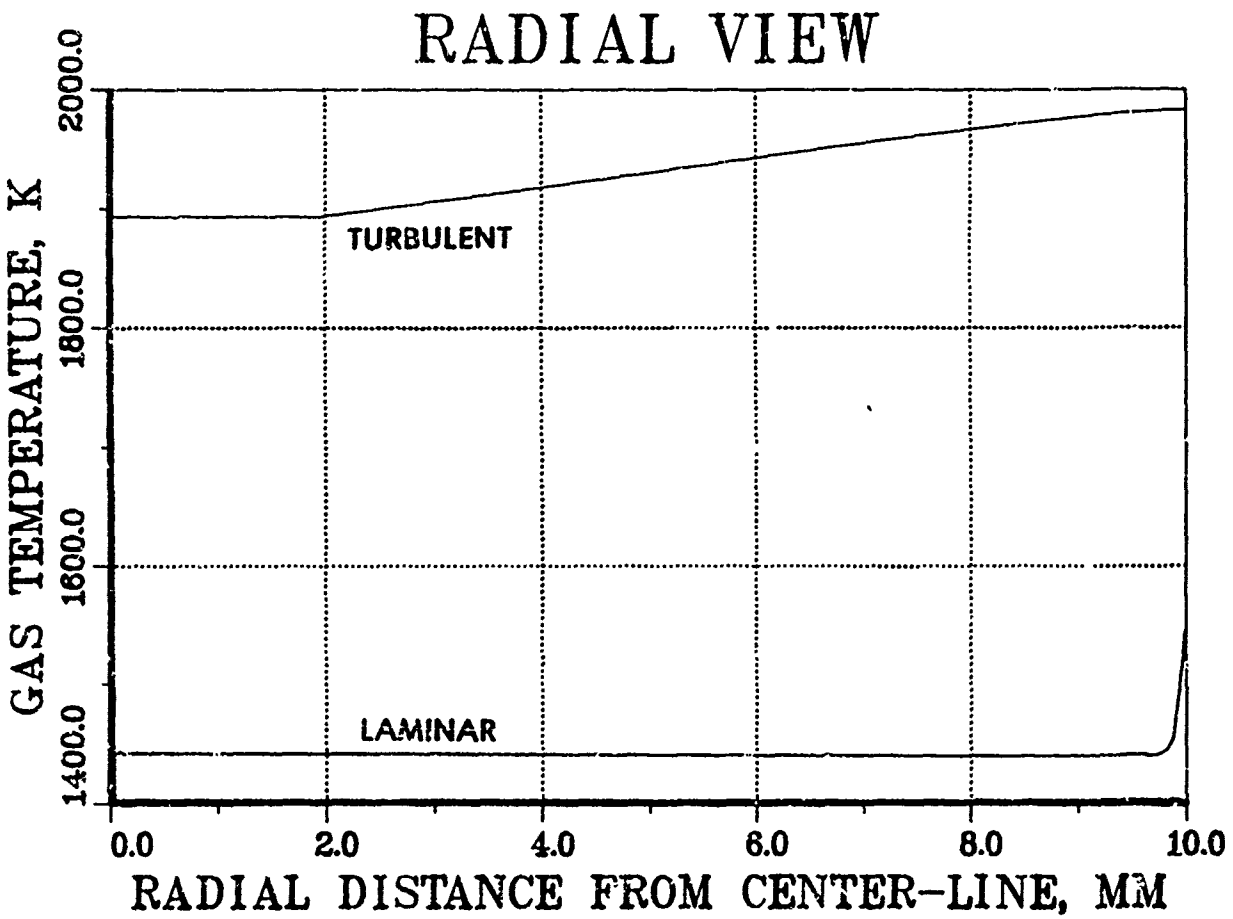



Figure 18. Lagrange gun, adiabatic walls: Radial profiles of the gas temperature for both laminar and turbulent flows at the time of muzzle clearance at 0.25 m away from the muzzle.

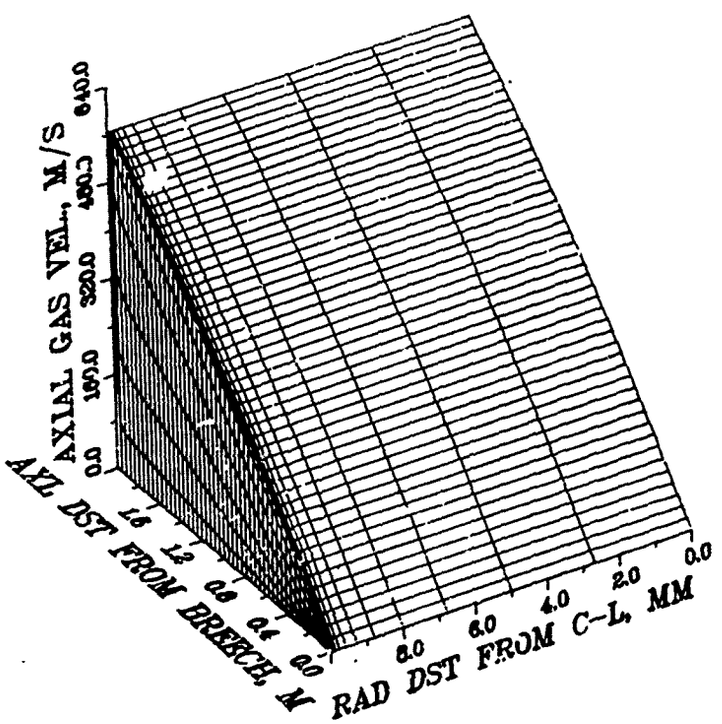



Figure 19. Lagrange gun, laminar flow, heat transfer: Spatial profile of the axial gas velocity at the time of muzzle clearance.

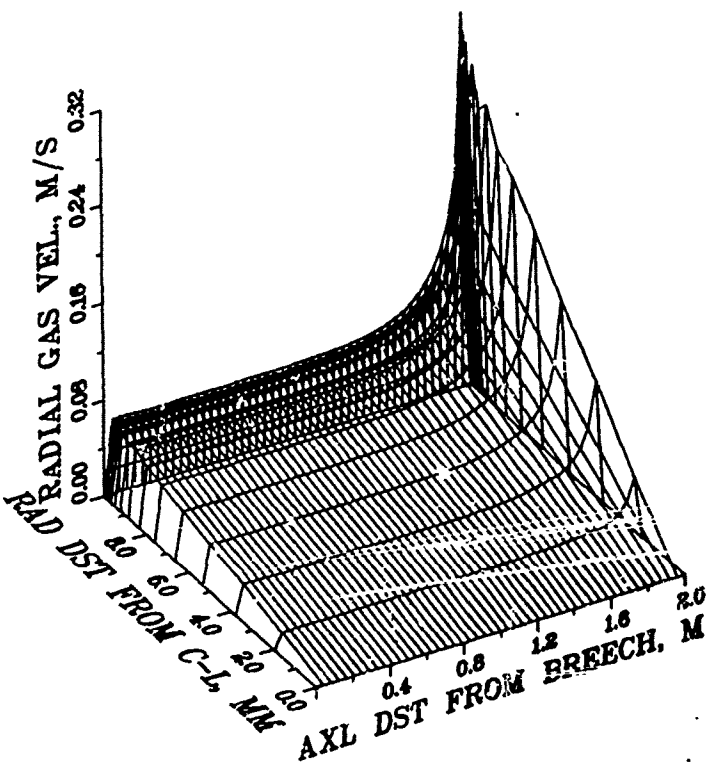



Figure 20. Lagrange gun, laminar flow, heat transfer: Spatial profile of the radial gas velocity at the time of muzzle clearance.

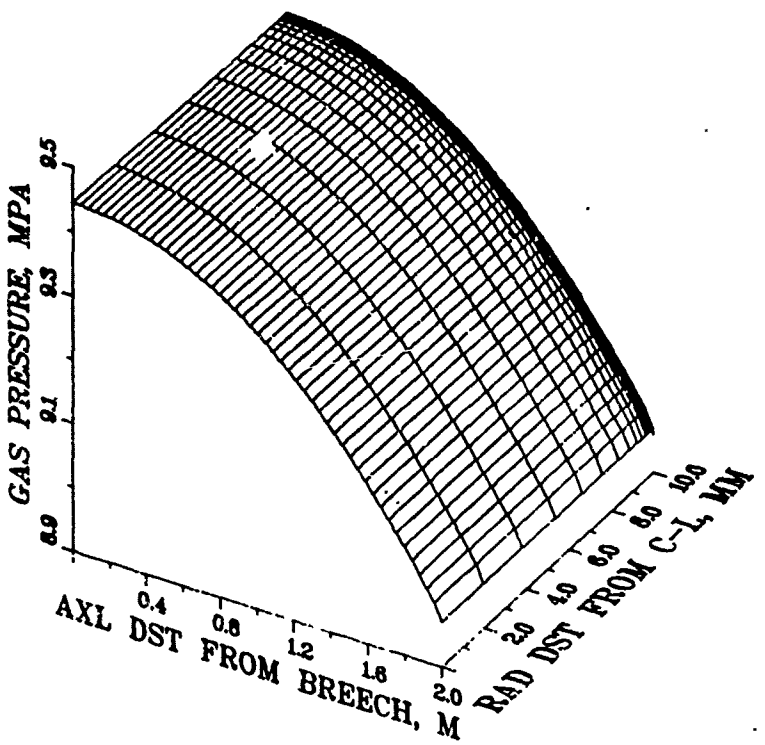



Figure 21. Lagrange gun, laminar flow, heat transfer: Spatial profile of the gas pressure at the time of muzzle clearance.

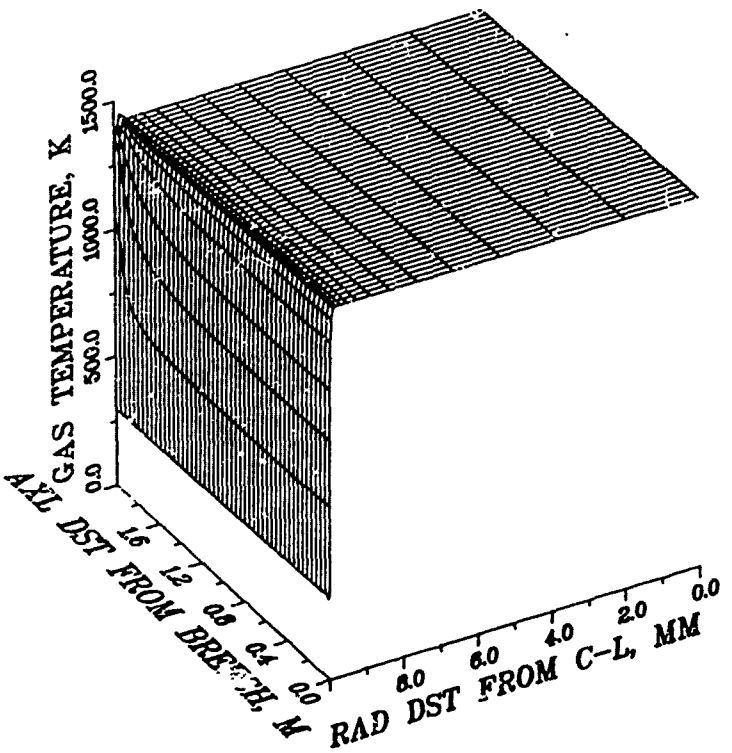



Figure 22. Lagrange gun, laminar flow, heat transfer: Spatial profile of the gas temperature at the time of muzzle clearance.

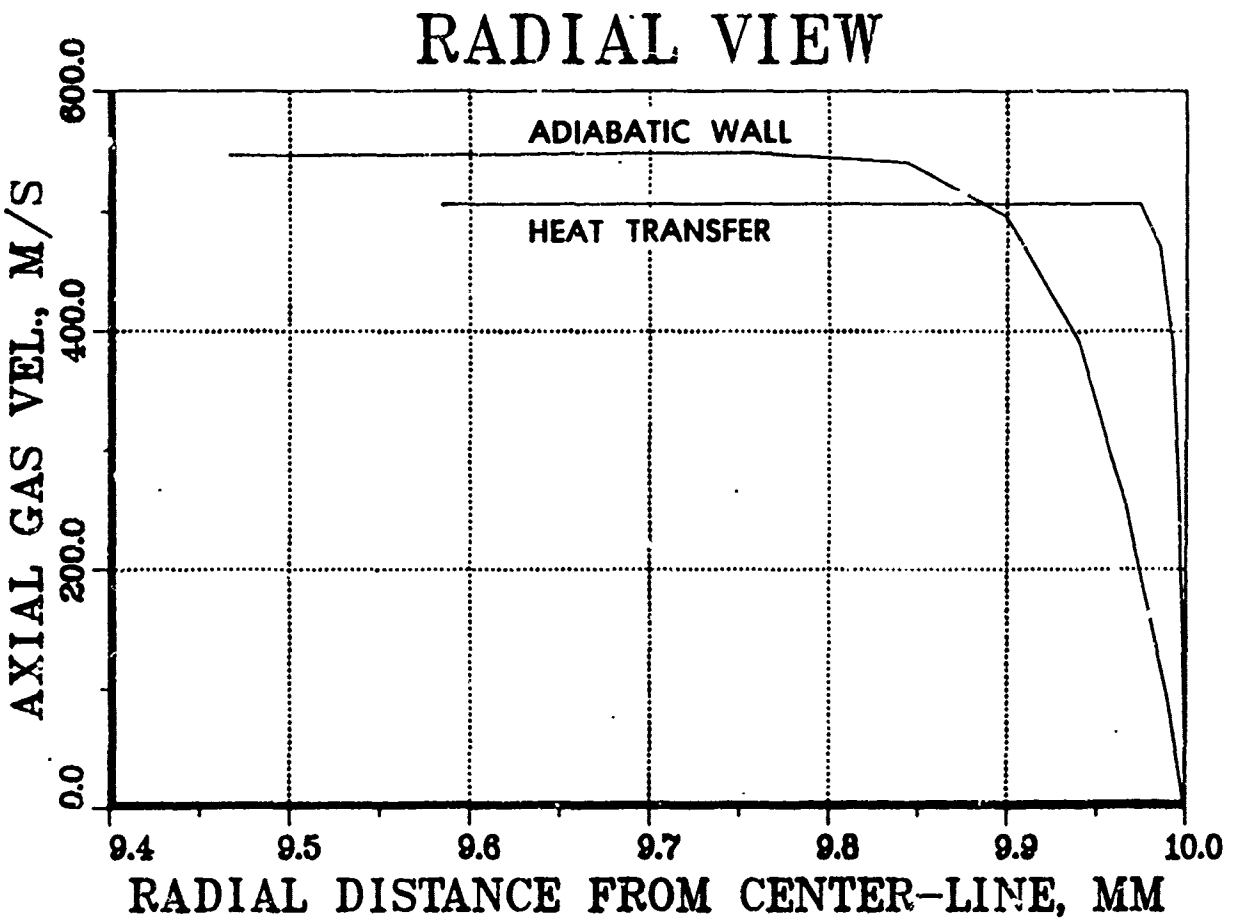



Figure 23. Lagrange gun, laminar flow: Radial profiles of the axial gas velocity for adiabatic and heat permeable walls at the time of muzzle clearance 0.25 m away from the muzzle.

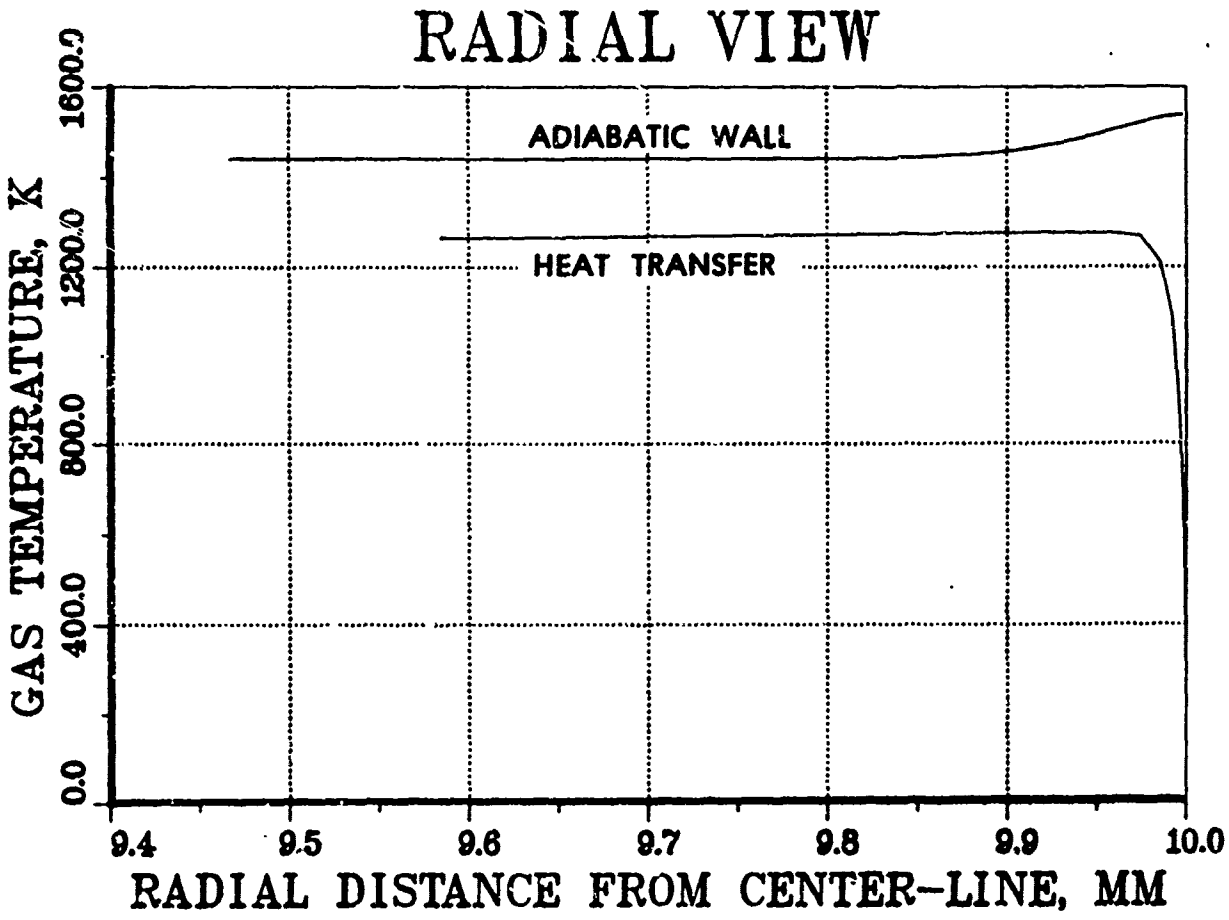



Figure 24. Lagrange gun, laminar flow: Radial profiles of the gas temperature for adiabatic and heat permeable walls at the time of muzzle clearance 0.25 m away from the muzzle.

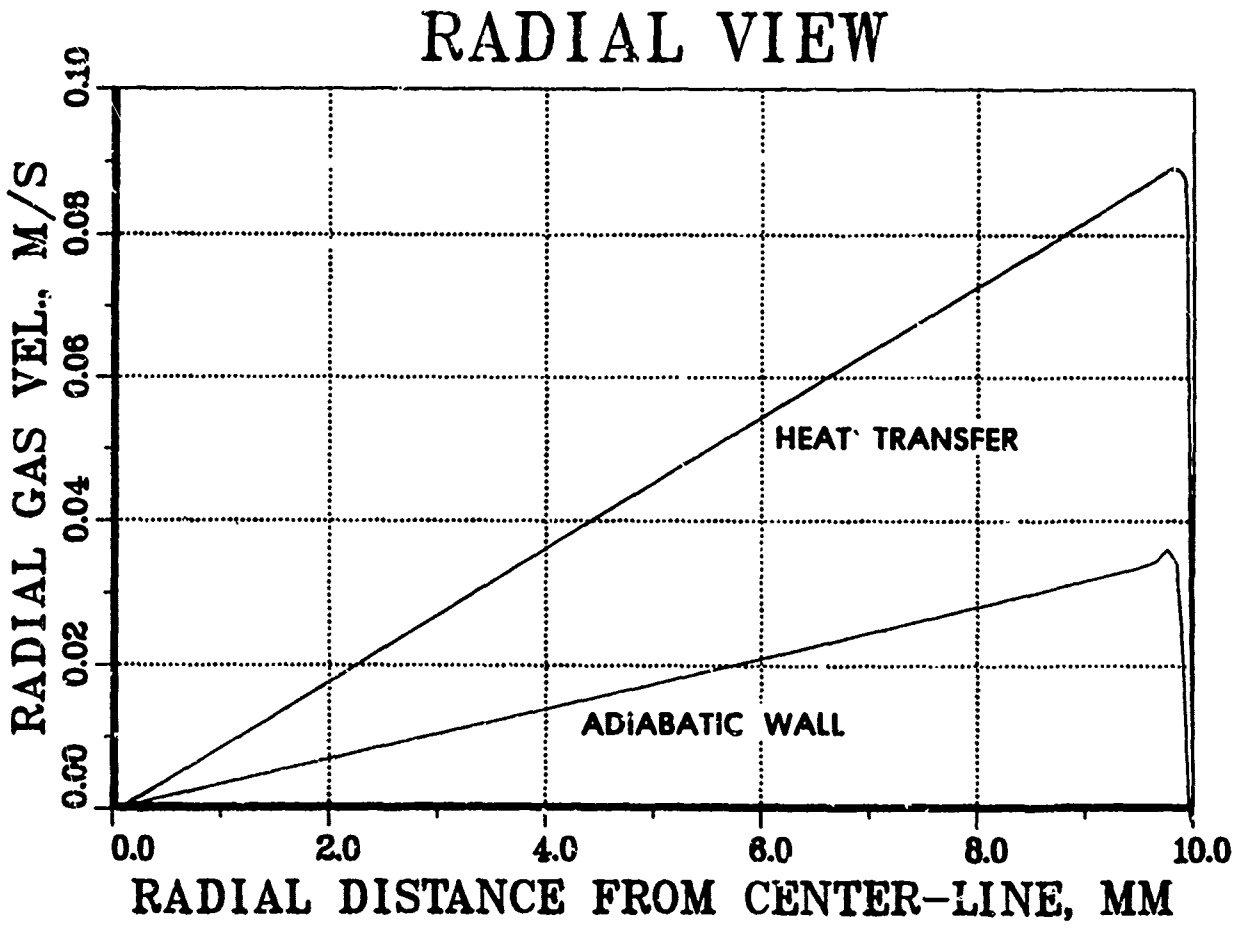



Figure 25. Lagrange gun, laminar flow: Radial profiles of the radial gas velocity for adiabatic and heat permeable walls at the time of muzzle clearance 0.25 m away from the muzzle.

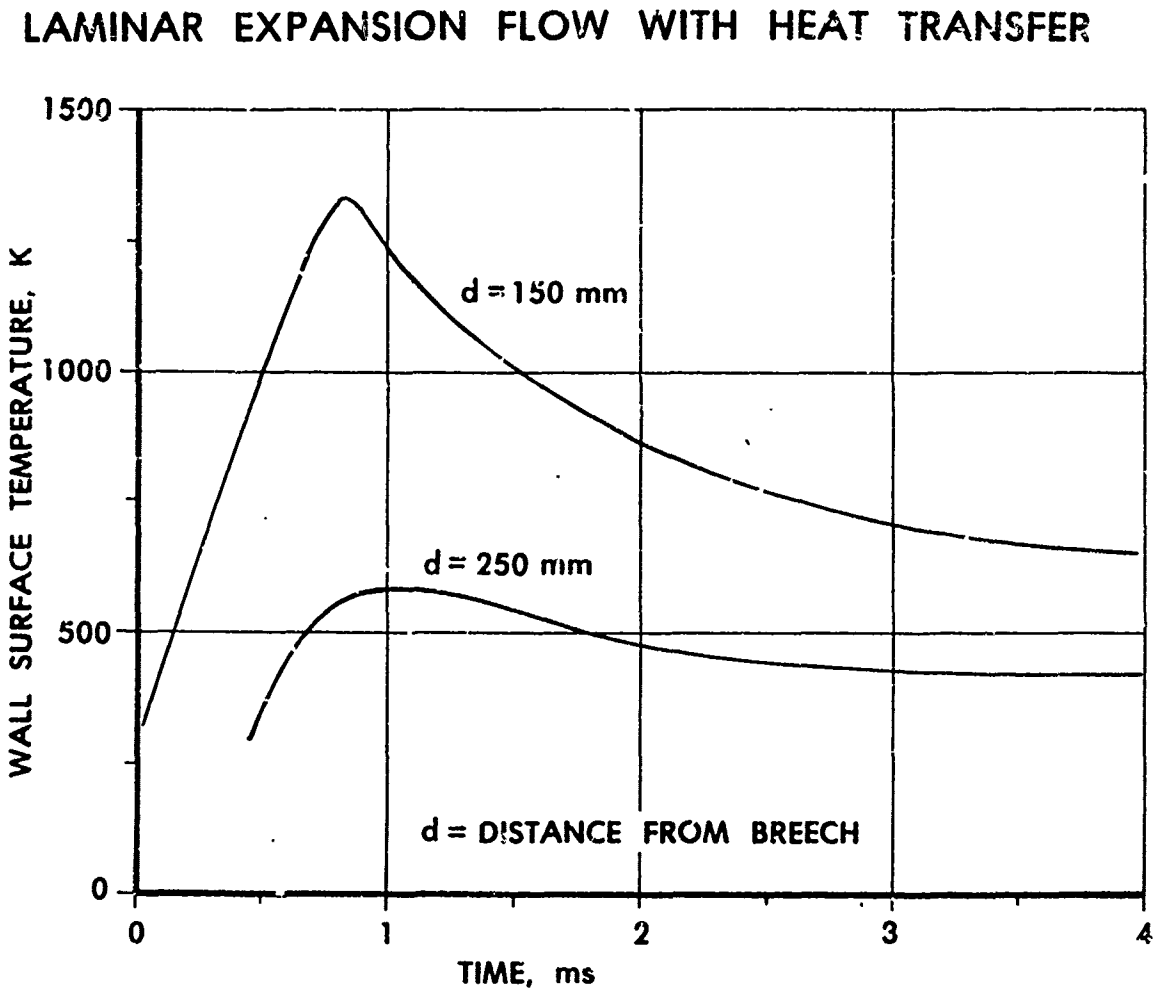



Figure 26. History of the wall surface temperature for a laminar expansion flow (LG) with heat transfer to the tube wall at the locations 150 mm (inside the chamber) and 250 mm (inside the barrel) away from the breech.

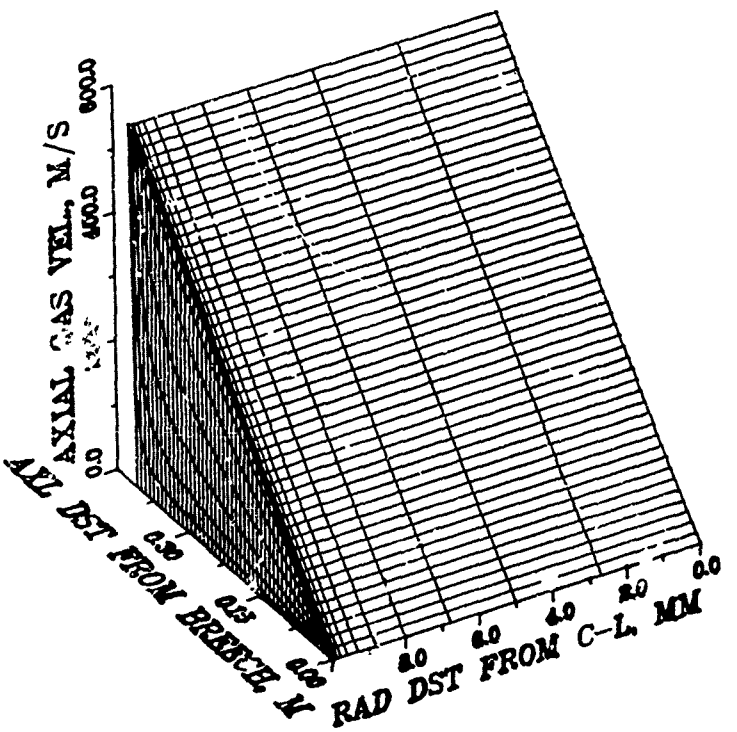



Figure 27. Real gun, laminar flow, adiabatic walls: Spatial distribution of the axial gas velocity at 3.6 ms.

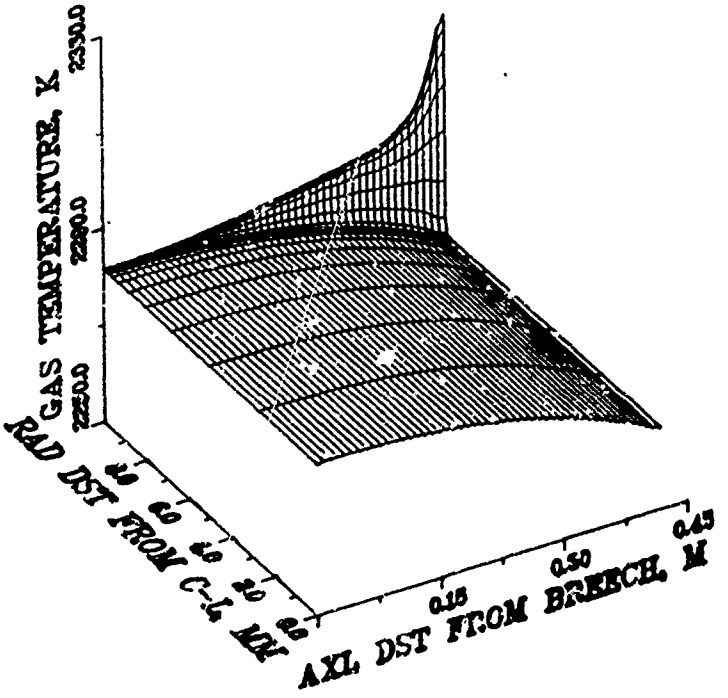



Figure 28. Real gun, laminar flow, adiabatic walls: Spatial distribution of the gas temperature at 3.6 ms.

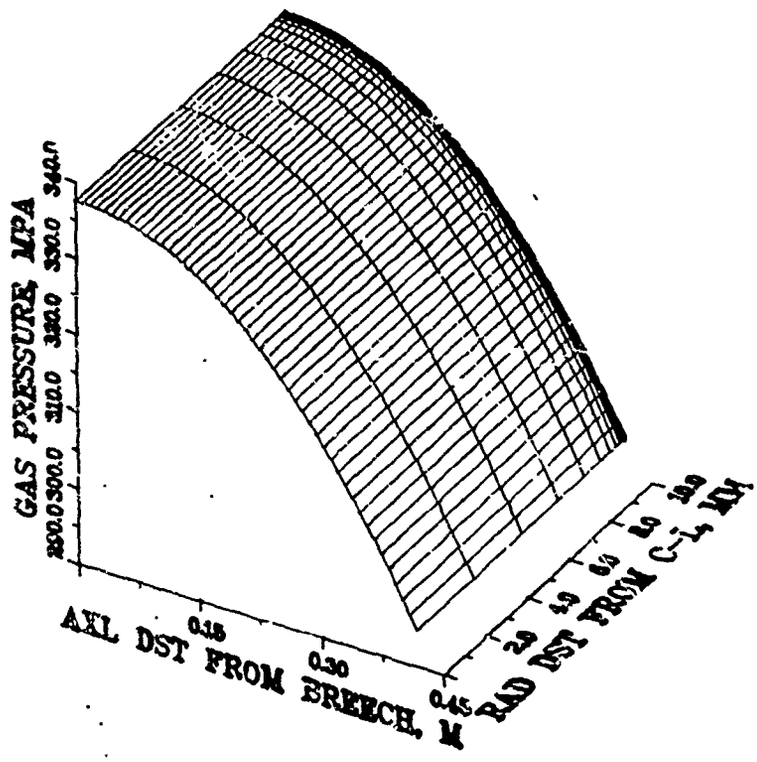



Figure 29. Real gun, laminar flow, adiabatic walls: Spatial distribution of the gas pressure at 3.6 ms.

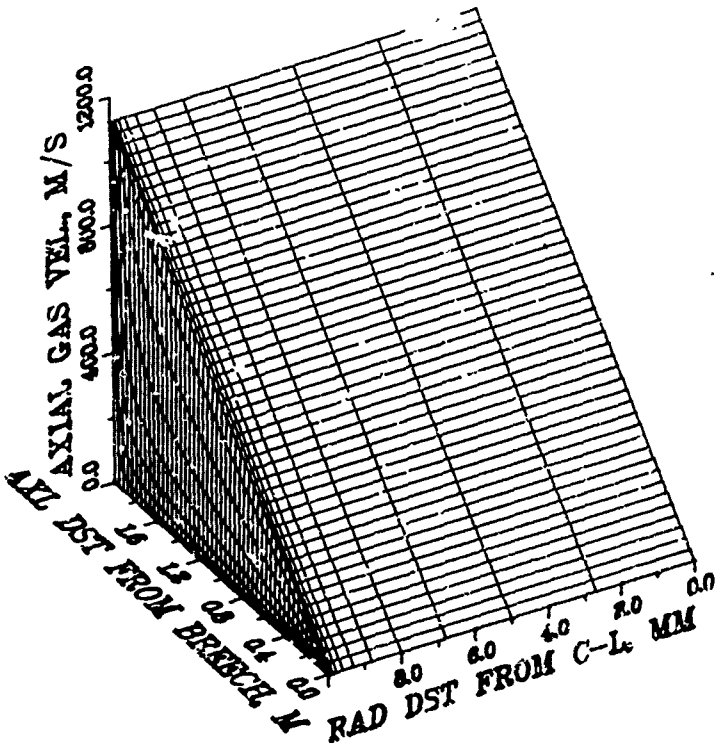



Figure 30. Real gun, laminar flow, adiabatic walls: Spatial distribution of the axial gas velocity at muzzle clearance (5.3 ms).

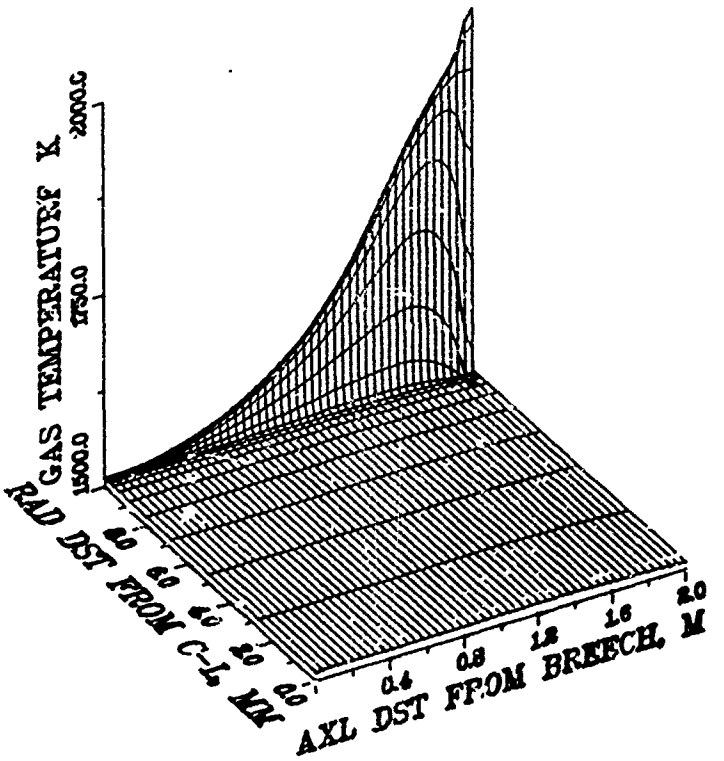



Figure 32: Real gun, laminar flow, adiabatic walls: Spatial distribution of the gas temperature at muzzle clearance (5.3 ms).

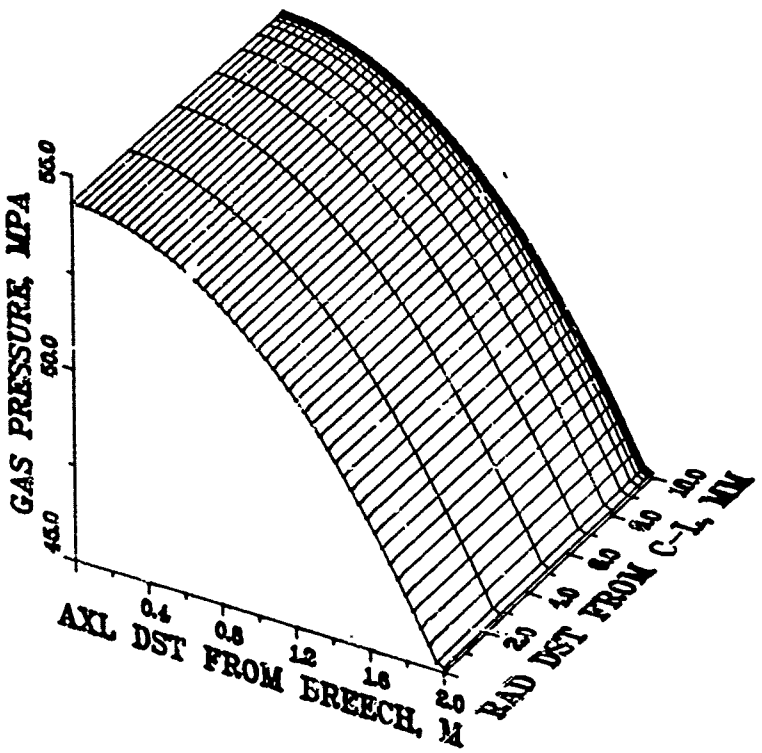



Figure 33. Real gun, laminar flow, adiabatic walls: Spatial distribution of the gas pressure at muzzle clearance (5.3 ms).

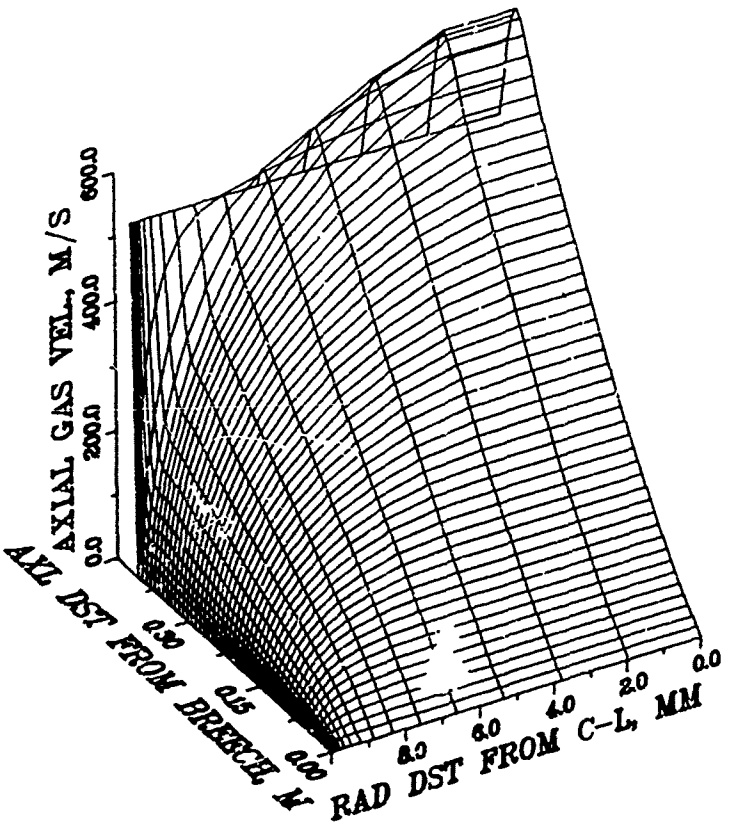



Figure 34. Real gun, turbulent flow, adiabatic walls: Spatial distribution of the axial gas velocity at 3.6 ms.

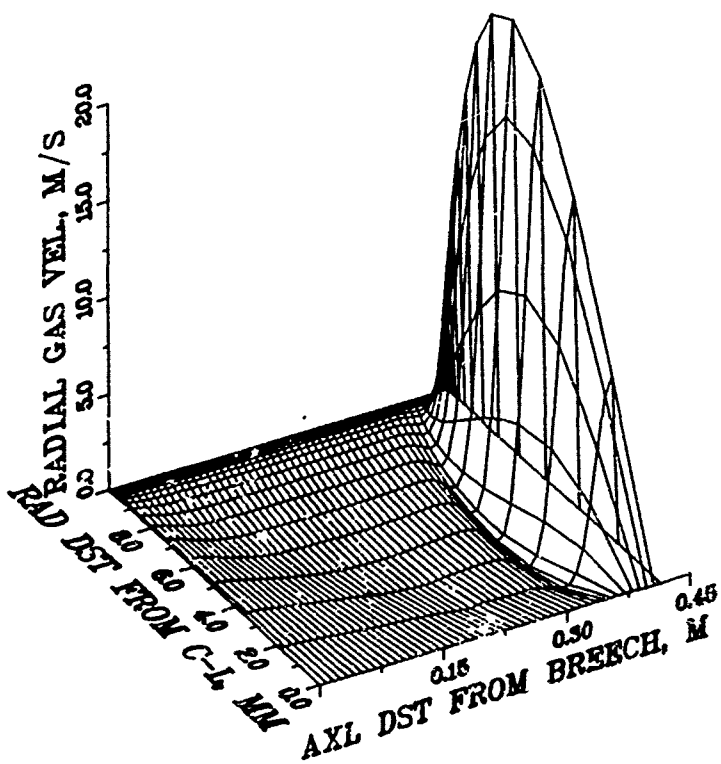



Figure 35. Real gun, turbulent flow, adiabatic walls: Spatial distribution of the radial gas velocity at 3.5 ms.

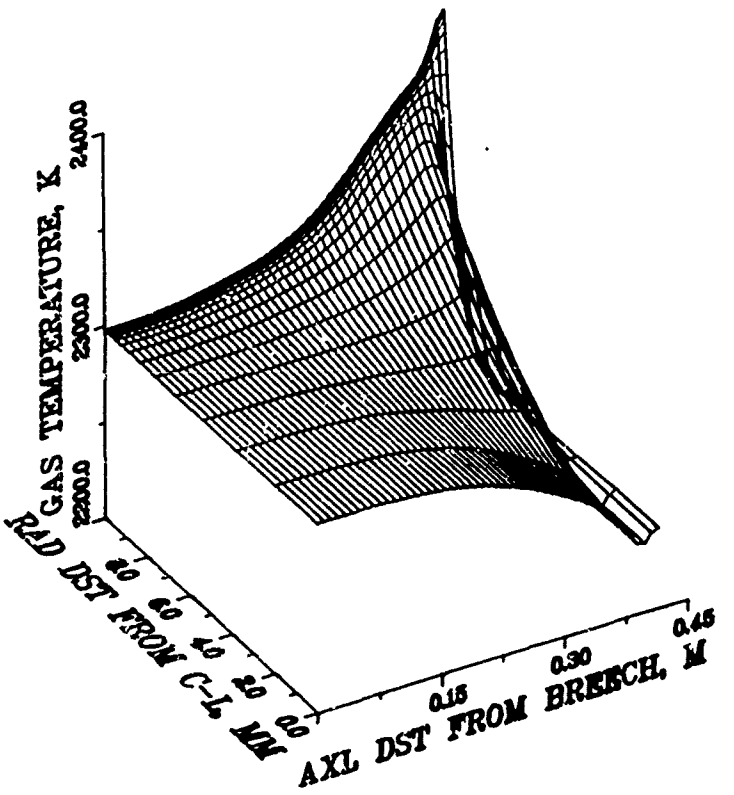



Figure 36. Real gun, turbulent flow, adiabatic walls: Spatial distribution of the gas temperature at 3.6 ms.

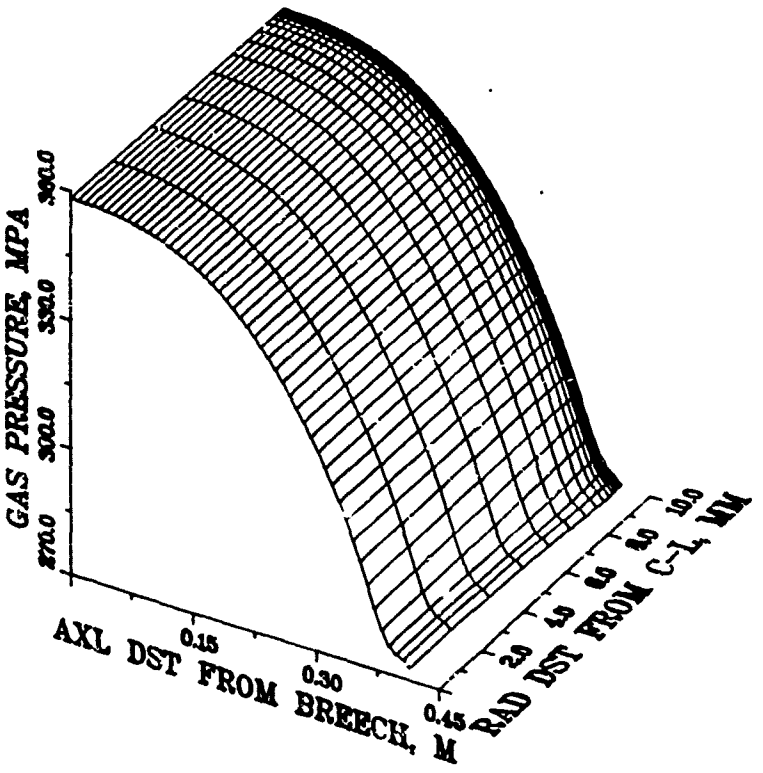



Figure 37. Real gun, turbulent flow, adiabatic walls: Spatial distribution of the gas pressure at 3.5 ms.

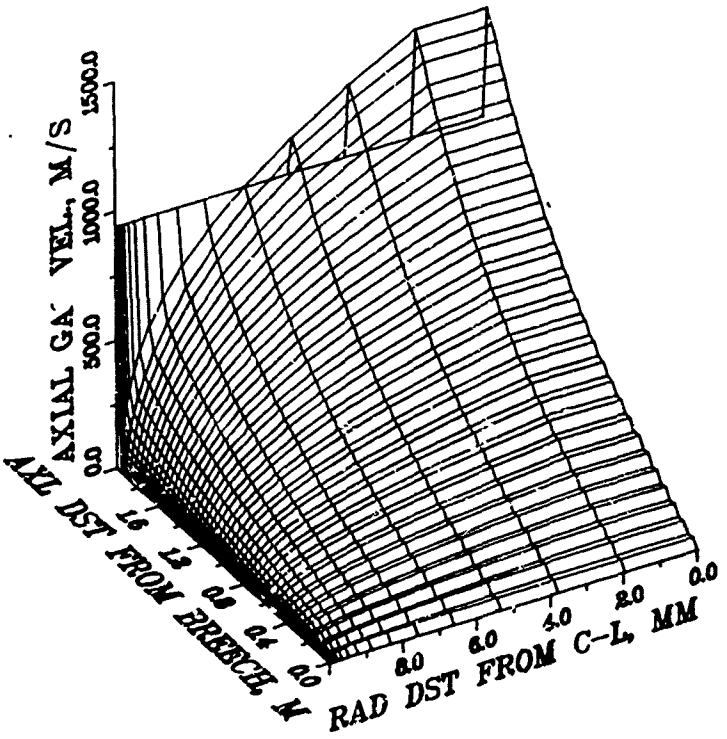



Figure 38. Real gun, turbulent flow, adiabatic walls: Spatial distribution of the axial gas velocity at muzzle clearance (5.49 ms).

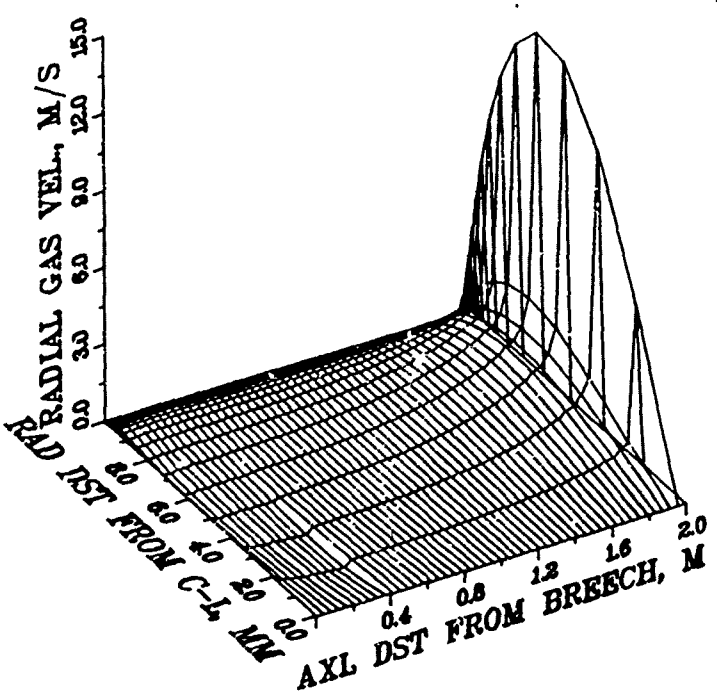



Figure 39. Real gun, turbulent flow, adiabatic walls: Spatial distribution of the radial gas velocity at muzzle clearance (5.49 ms).

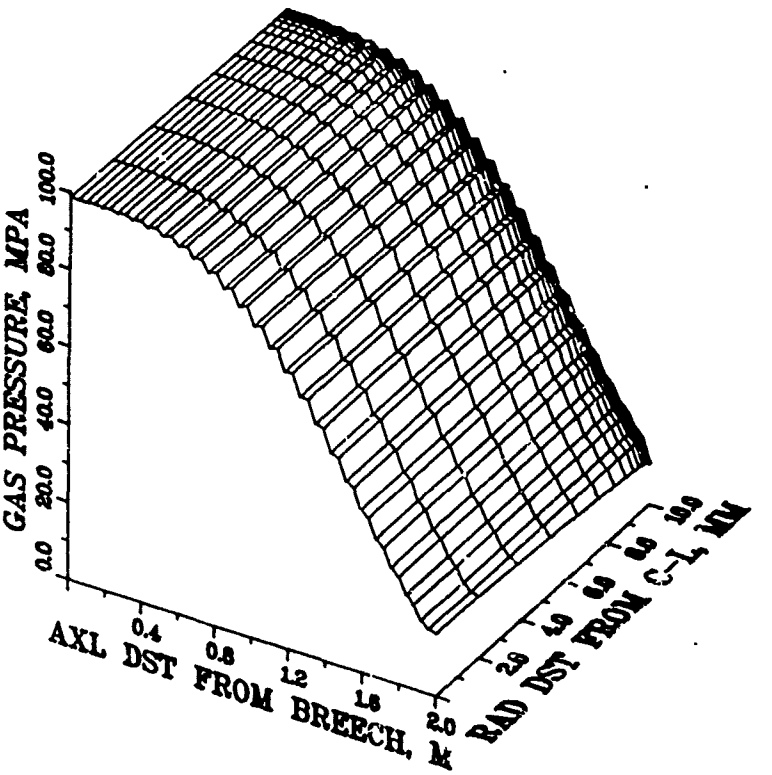



Figure 40. Real gun, turbulent flow, adiabatic walls: Spatial distribution of the gas pressure at muzzle clearance (5.49 ms).

# AN ANALYTICAL MODEL OF PERIODIC WAVES IN SHALLOW WATER -- SUMMARY*

Harvey Segur[+] and Allan Finkel[++]

[+]Aeronautical Research Associates of Princeton, Inc.
50 Washington Road, P. O. Box 2229
Princeton, NJ 08540

[++]Thomas Watson Research Center
IBM
P. O. Box 218
Yorktown Heights, N. Y. 10598

ABSTRACT. An explicit, analytical model is presented of finite amplitude waves in shallow water. The waves in question have two independent spatial periods, in two independent horizontal directions. Both short-crested and long-crested waves are available from the model. Every wave pattern is an exact solution of the Kadomtsev-Petviashvili equation, and is based on a Riemann theta function of genus 2. These bi-periodic waves are direct generalizations of the well-known (simply periodic) cnoidal waves. Just as cnoidal waves are often used as one-dimensional models of "typical" nonlinear, periodic waves in shallow water, these bi-periodic waves may be considered to represent "typical" nonlinear, periodic waves in shallow water without the assumption of one-dimensionality.

EXTENDED SUMMARY. Waves in shallow water are familiar to virtually everyone. The objective of the work presented here is to construct an analytical model of waves in shallow water that is simple and explicit enough to be useful for engineering purposes, without being so simple that it fails to describe realistic wave patterns.

Some of the features that occur in "typical" waves in shallow water may be seen in Figures 1 and 2. Figure 1, taken from an ancient National Geographic (1933), shows a very regular train of one-dimensional waves off the coast of Panama. (By "one-dimensional", we mean that the surface pattern is one-dimensional, with virtually no variation along the wave crests. We call the waves in Figure 2 "two-dimensional", because the surface pattern is two-dimensional. In this terminology, there can be no three-dimensional waves on the two-dimensional water surface, even though the velocity fields may exhibit vertical structures. We emphasize that this is only a semantic convention.)

*Presented at the Second Army Conference on Applied Mathematics and Computing, May 22-25, 1984, Troy, NY



PREVIOUS PAGE IS BLANK

Two features of the waves in Figure 1 are worth noting. The first is the very evident spatial period of the waves. The second is that the periodic wave pattern is far from sinusoidal: the wave crests are localized and rather steep, while the troughs are very broad and flat. The usual linear theory of infinitesimal water waves (e.g., Stoker, 1957) predicts a sinusoidal wave pattern, and the deviation of these waves from a sinusoidal shape is a measure of how nonlinear the waves are. Thus, Figure 1 suggests that a good model of "typical" waves in shallow water should admit waves that are: (i) periodic, and (ii) nonlinear.

Figure 2, taken by Mr. T. Toedtemeier off the coast of Oregon, shows an oblique interaction of two waves in shallow water. As in Figure 1, each wave has a sharp, localized crest and a broad, flat trough. Moreover, each of the interacting waves is part of a periodic wave train, but their wavelengths are so long, relative to the local water depth (about three feet, according to the photographer), that they behave like interacting solitary waves. The dominant effect of the interaction on each of the two wave crests is that each crest experiences a phase shift as a result of the interaction. The localized crests indicate that each (periodic) wave is individually nonlinear; the phase shift indicates that the interaction is also nonlinear.

The waves in Figure 1 happen to be one-dimensional, but those in Figure 2 clearly are not. In general, one would expect "typical" waves on the two-dimensional water surface to be two-dimensional.

Thus, we seek a model of waves in shallow water that are: (i) periodic; (ii) nonlinear; and (iii) two-dimensional. In fact, we want the simplest model possible that has these three properties. The main conclusion of this paper is that such a model now exists. In this summary we exhibit some consequences of the model. Full details will be published elsewhere (Segur & Finkel, 1984).

The Kadomtsev-Petviashvili (KP; 1970) equation,

$$\left(u_t + 6uu_x + u_{xxx}\right)_x + 3u_{yy} = 0, \tag{1}$$

is a scaled, dimensionless equation that describes the evolution of long water waves of moderate amplitude as they propagate primarily in one direction in shallow water of uniform depth, without dissipation. Mathematically, the KP equation generalizes the Korteweg-deVries (KdV) equation,

$$u_t + 6uu_x + u_{xxx} = 0. \tag{2}$$

Physically, the derivation of (1) is very similar to that of (2), except that the waves in (2) are required to be strictly one-dimensional, while those in (1) may be weakly two-dimensional (see Ablowitz & Segur, 1979 for the derivation).

The KP equation also generalizes the KdV equation in the sense that both are completely integrable. In particular, Satsuma (1976) showed that the KP equation admits a two-soliton solution of the form,

$$u\ (x,y,t) = 2\partial_x^2 \ln F, \tag{3a}$$

where

$$F = 1 + \exp\ (\eta_1) + \exp\ (\eta_2) + \exp\ (\eta_1 + \eta_2 + A)\ , \tag{3b}$$

$$\eta_j = \kappa_j \left(x + \rho_j y - c_j t\right), \qquad c_j = \kappa_j + 3\rho_j^2\ , \tag{3c}$$

$$\exp\ (A) = \frac{\left(\kappa_1 - \kappa_2\right)^2 - \left(\rho_1 - \rho_2\right)^2}{\left(\kappa_1 + \kappa_2\right)^2 + \left(\rho_1 - \rho_2\right)^2}\ . \tag{3d}$$

A particular two-soliton solution is shown in Figure 3. The qualitative agreement between the wave patterns in Figures 2 and 3 is clear. Whether there is also quantitative agreement requires more detailed information about the ocean wave than is available.

This 2-soliton solution of the KP equation is nonlinear and two-dimensional, but it is not periodic. To achieve the desired model, we must generalize the 2-soliton solutions of the KP equation to include periodic waves.

Krichever (1976) showed that the KP equation admits periodic and quasi-periodic solutions in the form

$$u\ (x,y,t) = 2\partial_x^2 \ln \theta\ \left(\phi_1, \ldots\ \phi_N\right) \tag{4}$$

where $\theta$ is a Riemann theta function of genus N. In the simplest case, where N = 1,

$$\phi = \mu\ \left(x + \rho y - ct\right)\ , \tag{5}$$

$$\theta(\phi) = \sum_n \exp\left(\frac{1}{2}\, bn^2 + i\, n\, \phi\right), \qquad \mathrm{Re}\ (b) < 0\ ,$$

(4) yields the usual cnoidal wave solution that is familiar from KdV theory

(e.g., see Sarpkaya & Isaacson, 1981). Figure 4 shows a typical cnoidal wave solution of (1); it bears a clear resemblance to the wave shown in Figure 1. Moreover, cnoidal waves from KdV-theory have been validated in extensive experimental tests as accurate models of one-dimensional waves in shallow water.

Cnoidal waves are periodic and nonlinear, but they are not two-dimensional. To achieve the desired model, we must generalize the cnoidal wave solutions of the KP equation to two dimensions.

The required generalization comes by taking $N = 2$ in (4), and using certain results of Dubrovin (1981). This family of exact solutions are called KP solutions of genus 2; they have eight free parameters. There is a sense in which they represent an oblique and nonlinear superposition of two trains of cnoidal waves. In one limit, these solutions resemble a slightly modulated cnoidal wave, as in Figure 5a. In another limit, they become periodic generalizations of the 2-soliton solution, as in Figure 5b. However, they also represent waves outside of either limit, as in Figure 5c.

These figures demonstrate that it is possible to construct a variety of wave forms from the KP solutions of genus 2. By itself, this possibility does not qualify this family of solutions as a physical model of water waves. We must also give an explicit algorithm to specify every free parameter of the solution in terms of measured physical quantities. One such algorithm was given by Segur, Finkel & Philander (1983). A different algorithm, based on a different kind of data, is given by Segur & Finkel (1984). Once this algorithm is given, the model can be tested experimentally, and we invite interested experimentalists to do so.

This work was supported in part by the Army Research Office, by the Office of Naval Research and by NSF Grant MCS 08814(A01). We are grateful to Terry Toedtemeier for permission to use Figure 2.

References

Ablowitz, M. J. & H. Segur, 1979, J. Fluid Mech., vol. 92, pp 691-715

Dubrovin, B. A., 1981, Russ. Math. Surveys, vol. 36, pp 11-92

Kadomtsev, B. B. & V. I. Petviashvili, 1970,) Sov. Phys. Doklady, vol. 15, pp 539-541

Krichever, I. M., 1976, Funct. Anal. Appl., vol. 10, pp 144-1146

Sarpkaya, T. & M. Isaacson, 1981, Mechanics of Wave Forces on Offshore Structures, Van Nostrand Reinhold, NY

Satsuma, J., 1976, J. Phys. Soc.Japan, vol. 40, pp 286-290

Segur, H., A. Finkel & H. Philander, 1983, "Integrable Models of Shallow Water Waves", in Nonlinear Phenomena, Lecture Notes in Physics #189, ed. by K. B. Wolf, Springer-Verlag, 1983

Segur, H., & A. Finkel, "An Analytical Model of Periodic Waves in Shallow Water", to be published

Stoker, J. J., 1957, Water Waves, Interscience, NY



Official photograph, U. S. Army Air Corps

TINY FISHING BOATS RIDING A PACIFIC SWELL LOOK LIKE FLIES ON A CORRUGATED TIN ROOF

As they near shallow water close to the coast of Panama, huge deep-sea waves, relics of a recent storm, are transformed into waves that have crests, but little or no troughs. A light breeze is blowing diagonally across the larger waves to produce a cross-chop. Three Army bombers, escorted by a training ship, are proceeding from Albrook Field, Canal Zone, to David, Panama.

Figure 1

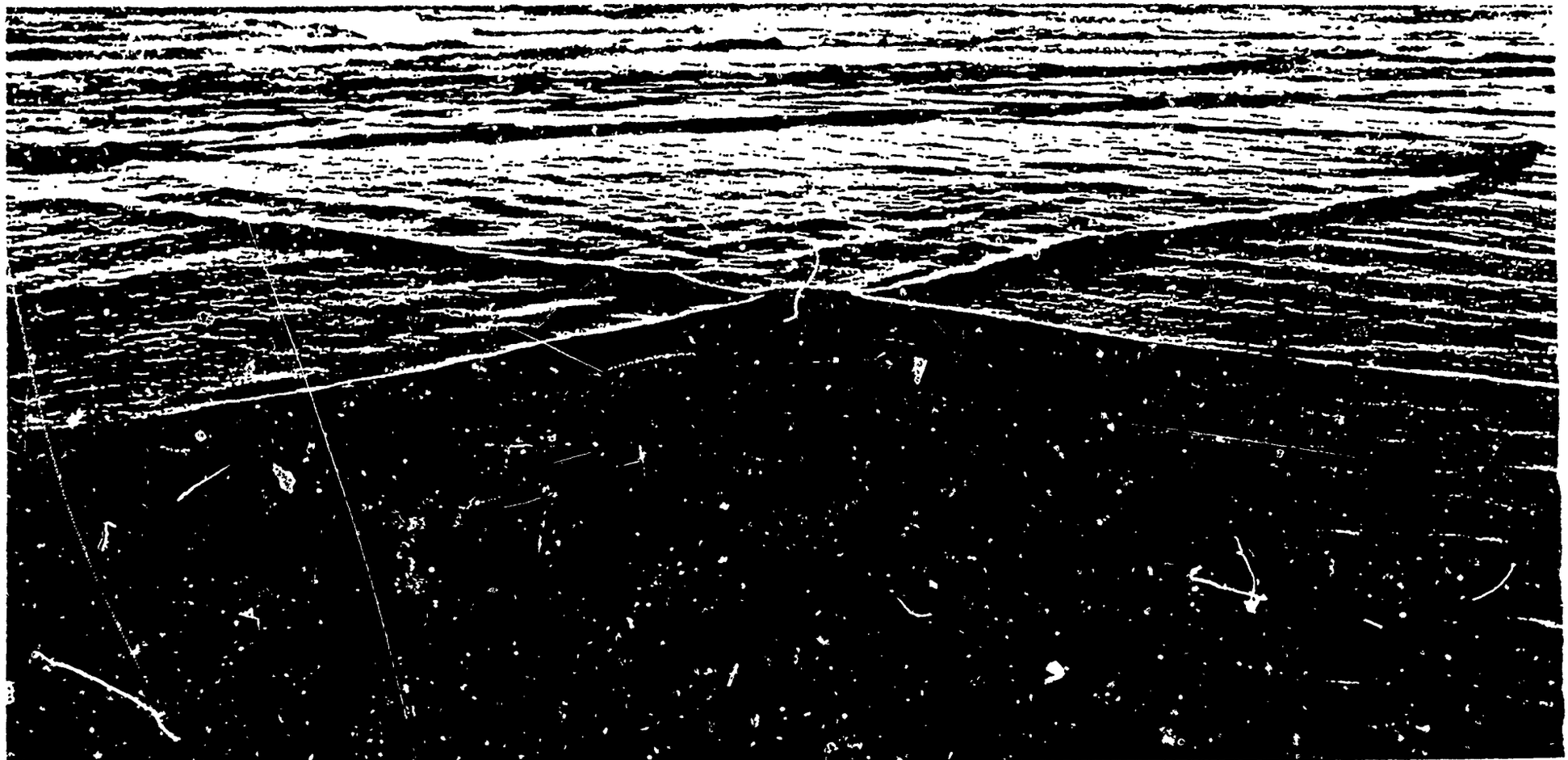


Figure 2

Oblique interaction of two waves in shallow water. (Photograph courtesy of T. Toedtemeier).

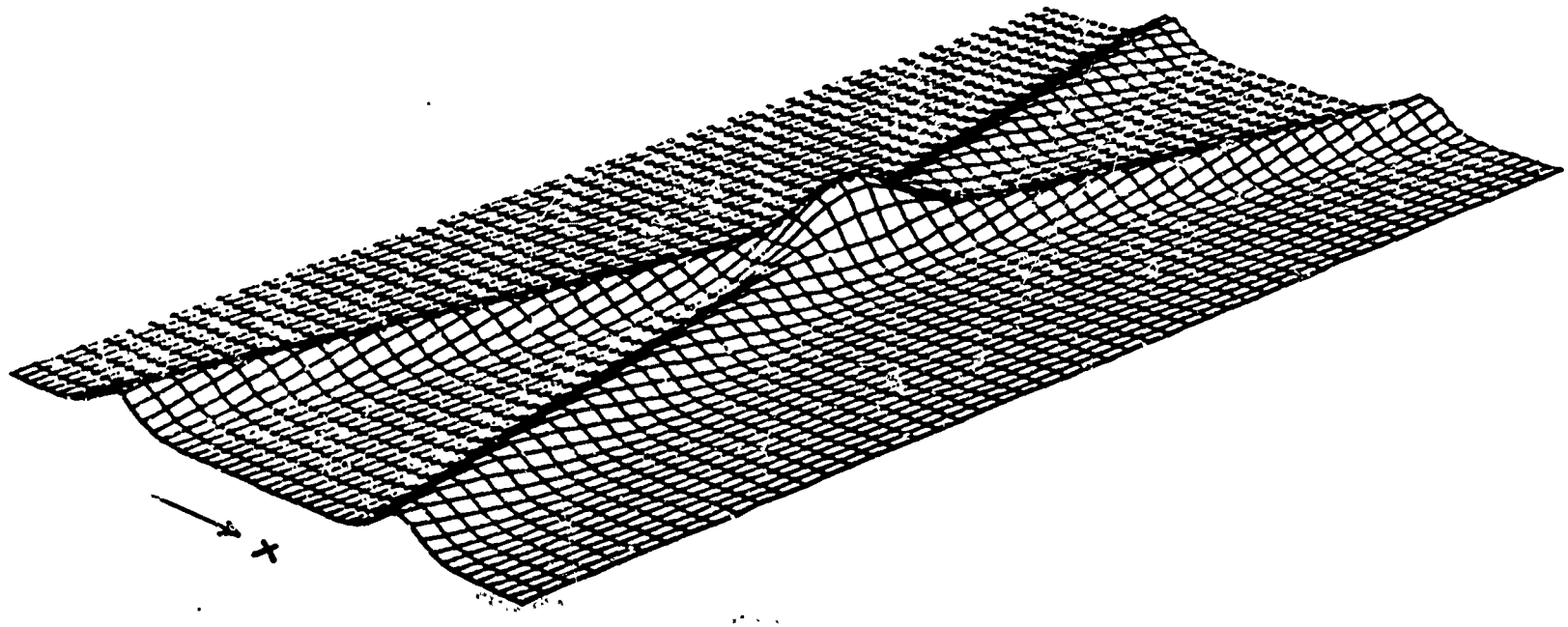



Figure 3

Two soliton solutions of the KP equation. In (3), $\kappa_1 = \kappa_2 = 1$, $\rho_1 = 4 = -\rho_2$, exp (A) = 16/15.

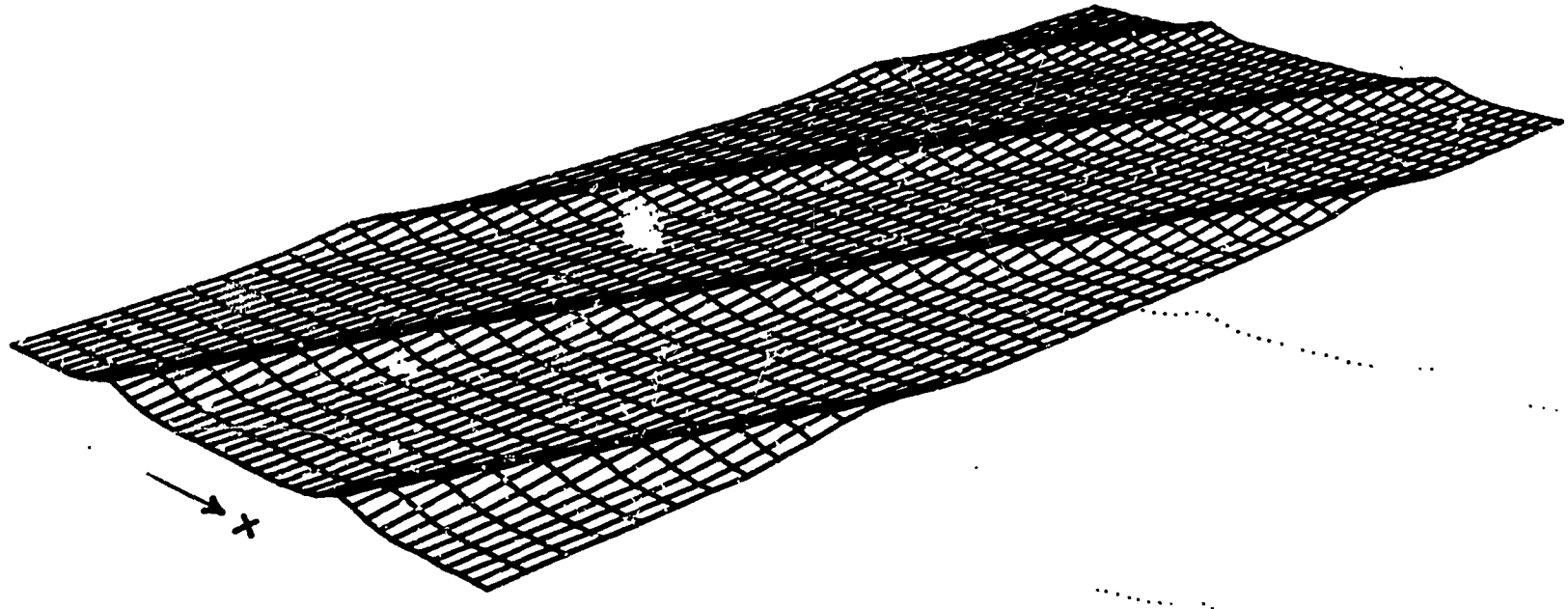



Figure 4

Cnoidal wave solution of the KP equation. In (b), $b = -3$, $\mu = 0.5$, $\rho = -0.43$.

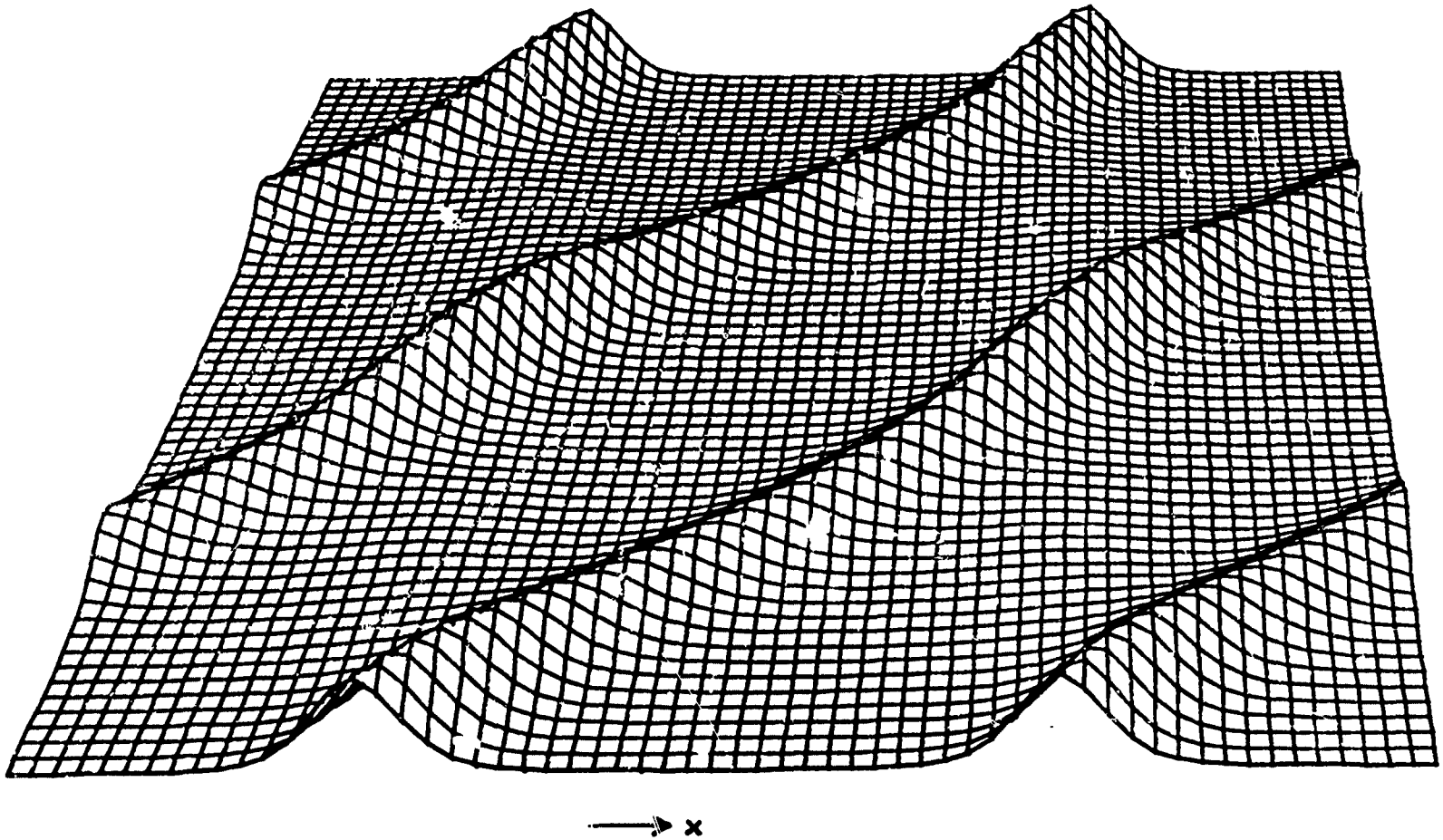



Figure 5a

An exact KP solution of genus 2 when one wave is dominant, so the weak waves simply modulate the strong one.

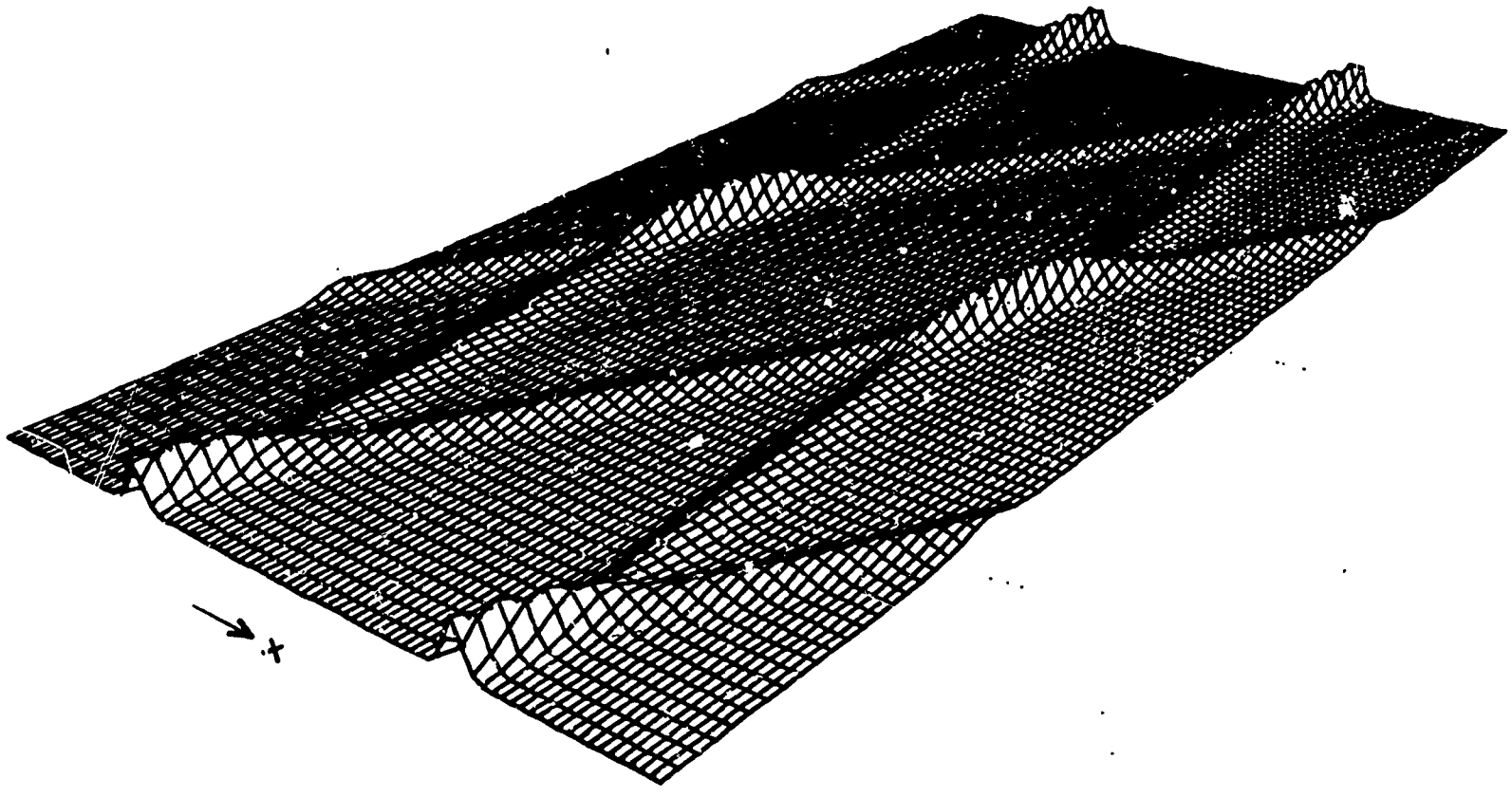



Figure 5b

An exact KP solution of genus 2 that is a periodic generalization of a 2-soliton solution.

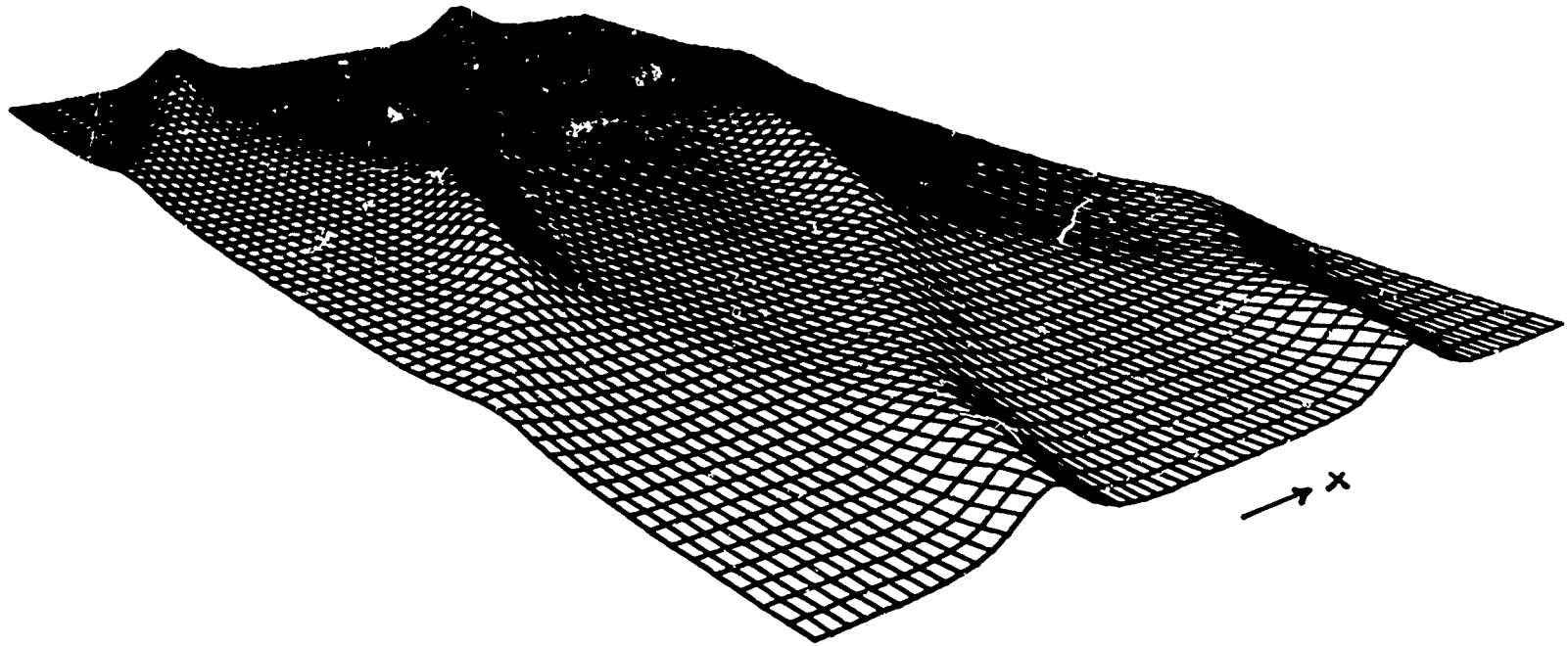



Figure 5c

An exact KP solution of genus 2 showing a "typical" pattern of periodic waves of finite amplitude in shallow water.

# ON AN IMPROVED T-MATRIX APPROACH TO STUDY THE SCALAR SCATTERING RESPONSE OF DOUBLY PERIODIC SURFACES

A. Lakhtakia, V.K. Varadan and V.V. Varadan
Wave Propagation Laboratory
Pennsylvania State University
University Park, PA 16802

## ABSTRACT

In investigating the scattering response of a periodic surface, the use of incomplete or inappropriate basis functions for representing the field(s) induced on the scattering surface has given rise to what is now called the Rayleigh Hypothesis (RH). Here we use normalised Fourier bases for this purpose and develop a T-matrix which completely characterises the scalar scattering response of such a surface. The Rayleigh limits are effectively bypassed, and the obtained solutions are seen to obey unitarity as well as reciprocity constraints. We also show that the measurement of the scattered field can lead to two different interpretations of the nature of the scattering surface in inverse shape problems.

## INTRODUCTION

The scattering of waves --- be they acoustic, electromagnetic, or elastic, --- by periodic surfaces has been the subject of much investigation ever since Rayleigh studied the scattering response of sinusoidal reflection gratings [1]. He expanded the incident and the scattered fields in terms of relevant incoming and outgoing planewaves, repsectively, and these decompositions are used to this day in such problems. However, he expressed the field(s) generated on the periodic surface in terms of outgoing plane-waves alone, a premise, --- now called the Rayleigh Hypothesis (RH), --- which involves an incomplete basis set, and can, therefore, be used for shallow corrugations. In a classic paper, Millar [2] has shown that for 2-D scalar problems involving a surface $S^2$: $x_3(x_2,x_1) = h \cos(2\pi x_1/L)$ the RH is applicable for $h/L \leq 0.072$. For the corresponding 3-D problems, we believe that Goodman's estimate [3] of $h/L \leq 0.0504$ is correct, the surface $S^3$: $x_3(x_2,x_1) = h[\cos(2\pi x_2/L) + \cos(2\pi x_1/L)]$.

Since then several efforts have been made to bypass the above-mentioned Rayleigh limits on the maximum gradient of periodic surfaces. Most of these methods fall into two categories. Methods of the first kind involve the solution of an integral equation (IE) [4,5]; while the second type are essentially matrix procedures [3, 5-12]. Though the IE methods have been very successful in dealing with highly corrugated $S^2$, their use for $S^3$ is extremely cumbersome because of tedious computations. Hence, matrix methods offer the only choice for 3-D problems. In this connection, DeSanto [6] has formulated coupled integral equations which are converted into matrix equations using relevant expansions for the fields of interest. A more elegant approach is due to Waterman [12] who used the 'extinction' theorem to formulate a T-matrix which characterizes the scalar scattering response of periodic $S^2$. This method, known as the T-matrix procedure, involves an understanding of the scattering problem from first principles using the Huyghen's and the Love's equivalence principles. Recently, this approach has also been extended to elastic scattering problems as well [8,9].

Nevertheless, the expansion of the surface field(s) in terms of only the incoming planewave bases for the T-matrix approach has proved to be a stumbling block in its application for highly corrugated surfaces. Such an expansion is as incomplete as the one used by Rayleigh; consequently, this method has suffered from the same limitations. Recently, however, using a hybrid T-matrix - point-matching technique, wherein the surface field(s) is expressed in terms of both incoming and outgoing planewave bases, the applicability of the method has been increased to higher corrugations than previously possible. We have used this hybrid technique for scalar [13] as well as elastic [9] scattering problems involving $S^2$ surfaces.

Specifically for scalar problems, Fourier bases have been used for representing the surface field(s) in the T-matrix framework [3] and with success as evinced by the data published in [14]. On the other hand, using these same Fourier bases for computing elastic responses by Chuang and Johnson [8] has not lifted the T-matrix approach from within the Rayleigh limit

on the maximum surface slope. However, the use of normalised Fourier bases of [6,7] has been more promising as shown by our work on the scalar response of $S^2$ [15].

In this paper we present a T-matrix formalism for computing a stable and accurate T-matrix which characterises the acoustic responses of hard and soft periodic $S^3$ surfaces. Normalised Fourier bases will be used to express the surface field; and the presented approach will also be valid for electromagnetic problems, where the relevant fields will have to be decomposed into TM-to and TE-to $x_3$ fields. The use of these bases for elastic problems is still under investigation and shall not be discussed here. From our results we shall show that the presented T-matrix method is useful for scalar problems involving surface slopes about 3 to 4 times the Rayleigh limits. We shall also discuss a non-uniqueness in the inverse shape problem when the field scattered by the periodic surface has been determined experimentally.

## THEORY

Let $Ox_1x_2x_3$ denote a 3-D Cartesian co-ordinate system. The surface $S^3$ is given by $x_3 = F(x_1,x_2)$, where F is assumed to be a single-valued, differentiable, periodic function with periodicities $L_1$ and $L_2$. This surface, in the mean, should be the flat plane $x_3 = 0$.

The region V above the surface $\{x_3 > F(x_1,x_2)\}$ is occupied by a non-viscous compressible fluid and an incident planewave

$$\psi_i(\underline{x}) = \psi_0 \exp(i\underline{k}_0 \cdot \underline{x}) \tag{1}$$

is incident on $S^3$ with a temporal variation $\exp(-i\omega t)$. The surface can be either acoustically soft (case S) or hard (case H), and the corresponding boundary conditions on the total field apply. The notation is as follows:

$$\left.\begin{aligned} &k = \omega/c \qquad \underline{k}_0 = k(\alpha_0\underline{u}_1 + \beta_0\underline{u}_2 - \gamma_{00}\underline{u}_3) \\ &\alpha_0 = \sin\theta_0\cos\phi_0 \qquad \beta_0 = \sin\theta_0\sin\phi_0 \\ &\gamma_{00} = \cos\theta_0 \qquad \underline{x} = x_1\underline{u}_1 + x_2\underline{u}_2 + x_3\underline{u}_3 \\ &\psi_0 = \text{constant amplitude.} \end{aligned}\right\} \tag{2}$$

The relevant boundary condition are:

$$\text{Case S:} \quad \psi(\underline{x})\big|_{S^3} = 0 \tag{3}$$

$$\text{Case H:} \quad \underline{\nu} \cdot \nabla\psi(\underline{x})\big|_{S^3} = 0 \tag{4}$$

$$\underline{\nu} = \text{unit vector normal to } S^3 \text{ into the fluid.} \tag{5}$$

The application of the Huyghens' principle, and the use of the free space

Green's function $G(\underline{x}';\underline{x}) = \exp(ik|\underline{x}'-\underline{x}|)/4\pi|\underline{x}'-\underline{x}|$, leads to [12]:

$$H_v(\underline{x}')\psi(\underline{x}') = \psi_i(\underline{x}') + \int_0^{L_1} dx_1 \int_0^{L_2} dx_2 \bullet$$

$$\bullet\{kG(\underline{x}';\underline{x})\ V(\underline{x}) - U(\underline{x})\ \zeta(\underline{x})\ \underline{\nu}\ .\ \nabla G(\underline{x}';\underline{x})\}, \tag{6}$$

with $H_v(\underline{x}') = 1$ if $\underline{x}' \in V$, and 0 otherwise; and

$$\left.\begin{aligned} V(\underline{x}) &= k^{-1}\ \zeta(\underline{x})\ \underline{\nu}\ .\ \nabla\psi(\underline{x})\big|_{S^3}\ , \\ U(\underline{x}) &= \psi(\underline{x})\big|_{S^3}\ , \\ \zeta(\underline{x}) &= \{1 + (\dot{F}_1)^2 + (\dot{F}_2)^2\}^{\frac{1}{2}},\quad \dot{F}_n = \partial F/\partial x_n . \end{aligned}\right\} \tag{7}$$

Equation (6) for $\underline{x}' \notin V$ is the 'extinction theorem' [12], $\psi = \psi_i + \psi_s$.

The free space Green's function can be expanded as

$$G(\underline{x}';\underline{x}) = (i/2kL_1L_2) \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} (1/\gamma_{pq}) \bullet$$

$$\bullet\ \exp\{ik[\alpha_p(x_1'-x_1) + \beta_q(x_2'-x_2) + \gamma_{pq}|x_3'-x_3|]\}, \tag{8}$$

with

$$\left.\begin{aligned} \alpha_p &= \alpha_0 + 2p\pi/kL_1,\ p = 0, \pm1, \pm2, \ldots. \\ \beta_q &= \beta_0 + 2q\pi/kL_2,\ q = 0, \pm1, \pm2, \ldots. \\ \gamma_{pq} &= [1 - \alpha_p^2 - \beta_q^2]^{\frac{1}{2}},\ \mathrm{Re}(\gamma_{pq}) \geq 0,\ \mathrm{Im}(\gamma_{pq}) \geq 0. \end{aligned}\right\} \tag{9}$$

At this juncture we also define two groups of wave vectors

$$\underline{k}^{\pm}_{pq} = k(\alpha_p \underline{u}_1 + \beta_q \underline{u}_2 \pm \gamma_{pq}\underline{u}_3) \tag{11}$$

with whose help we define the incident and the scattered fields as

$$\psi_{\substack{s\\i}}(\underline{x}) = \sum_{p=-\infty}^{\infty} \sum_{q=-\infty}^{\infty} a^{\pm}_{pq} \exp(i\underline{k}^{\pm}_{pq} \cdot \underline{x}) \tag{12}$$

with $a^{-}_{pq} = \psi_0\delta_{p0}\delta_{q0}$ and unknown $a^{+}_{pq}$ to be determined.

Let us first consider the case S. The boundary condition (3) would then apply and the Equations (6) would accordingly be modified. On substituting the expansions (12) in the modified (6) we obtain a set of equations:

$$a^{\mp}_{pq} = -\int_0^{L_2} dx_1 \int_0^{L_2} dx_2 \qquad (1/2ikL_1L_2\gamma_{pq}) \cdot$$

$$\cdot \underline{\chi} \cdot \{\exp(-i\underline{k}^{\pm}_{pq} \cdot \underline{x})\ \nabla_+ \psi(\underline{x})\}\ , \tag{13}$$

the vector

$$\underline{\chi} = (-\dot{F}_1, -\dot{F}_2, 1).$$

In order to solve the problem all we need now is the surface field representation, which we assume to be [6,15]

$$\underline{\chi} \cdot \nabla_+ \psi(\underline{x}) = 2 \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \alpha_{nm}\, \underline{\chi} \cdot \nabla_+ \{\exp(i\underline{k}^{*}_{nm} \cdot \underline{x})\};\ \underline{x} \in S^3 \tag{14}$$

with the wave vectors

$$\underline{k}^{*}_{nm} = k\ (\alpha_n, \beta_m, -\gamma_{00}); \tag{15}$$

finally, substitution of a truncated (14) in (13) leads to the matrix equations

$$a^{+} = Q^{+}_{d} \cdot (Q^{-}_{d})^{-1}\, a^{-} \tag{16}$$

where the matrices $Q^{\pm}_{d}$ are

$$(Q^{\pm}_{d})_{pq,nm} = \pm \int_0^{L_2} dx_1 \int_0^{L_2} dx_2 [(\gamma_{00} + \alpha_n \dot{F}_1 + \beta_m \dot{F}_2)/\gamma_{pq} L_1 L_2] \cdot$$

$$\cdot \exp[-i(\underline{k}^{\pm}_{pq} - \underline{k}^{*}_{nm}) \cdot \underline{x}]. \tag{17}$$

At this point we remark that replacing the $\underline{k}^{*}_{nm}$ by either of the $\underline{k}^{\pm}_{nm}$ vectors would completely debilitate the T-matrix procedure and subject it to the Rayleigh limits in its ability to handle deeply corrugated surfaces.

Likewise, for the case H the boundary condition (4) is substituted in (6) as also the assumed surface field expansion

$$\psi_+(\underline{x}) = 2 \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \alpha_{nm} \exp(i\underline{k}^{\pm}_{nm} \cdot \underline{x});\ \underline{x} \in S^3 \tag{18}$$

which yields the solution

$$a^{+} = Q_n^{+} \cdot (Q_n^{-})^{-1} a^{-}, \tag{19}$$

where

$$(Q_n^{\pm})_{pq;nm} = \pm \int_0^{L_1} dx_1 \int_0^{L_2} dx_2 [(\bar{+}\gamma_{pq} + \alpha_p \dot{F}_1 + \beta_q \dot{F}_2)/\gamma_{pq} L_1 L_2] \cdot$$

$$\cdot \exp[-i(\underline{k}_{pq}^{\pm} - \underline{k}_{nm}^{*}) \cdot \underline{x}]. \tag{20}$$

Defining the energy carried by the (pq)th mode of the scattered field as

$$P_{pq} = (\gamma_{pq}/\gamma_{00}) |a_{pq}^{+}|^2 / |\psi_0|^2, \tag{21}$$

provided, of course that $\gamma_{pq}$ is positive real, the conservation of energy relation is obtained by

$$E = \sum_p \sum_q P_{pq} = 1. \tag{22}$$

## NUMERICAL RESULTS

The system of equations (16) and (19) were programmed on a DEC vax 11/730 minicomputer. The inversion of matrices involved was carried out using a LU decomposition technique [16] via an IMSL subroutine LEQT1C, our numerical procedure being implemented in double precision arithmetic. The Q matrices, themselves, were computed using a two-dimensional Gauss-Legendre quadrature scheme [17], although for special cases of boundary profiles these matrices can be evaluated in closed forms. Convergence of the solution was checked by ensuring that the scattered field coefficients converged to within 0.5%. An additional check was also provided by (22) which had to be satisfied to be within ±0.005 of unity.

The general theory presented in the previous section holds for both $S^2$ and $S^3$ surfaces having periodic boundary profiles. However, here we consider surfaces described by $x_3(x_2, x_1) = f(x_2) + f(x_1)$, $f(x) = h \cos(2\pi x/L)$; for $S^2$, $f(x_2)$ is set to zero. The boundary conditions prevailing on the surface can be either Dirichlet or Neumann.

Consider, first, Fig.1 where we have plotted the scattered powers for a $S^2$, and have compared our calculations with those of Holford [5]. As is clear from this figure, the improved T-matrix scheme is applicable for much higher values of the parameter h/L than the Rayleigh limit of 0.072.

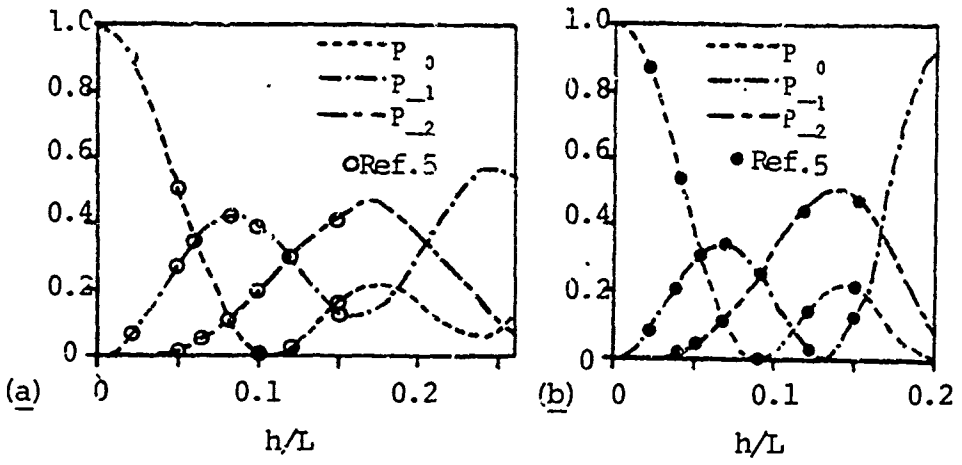



Fig. 1. Reflected mode power $P_0$ computed using the presented approach for a sinusoidal $S^2$, when a planewave is incident at $\theta_0 = 15°$, $\phi_0 = 0°$; $kL = 4\pi$. Ref. 5 are the IE results. (a) Dirichlet b.c., (b) Neumann b.c.

Similarly, in Table I we consider a doubly sinusoidal $S^3$ surface, for which case scattering is observed in 9 separate directions. Again, note that h/L = 0.15, which is roughly three times higher than Goodman's conjecture of 0.0504 for the Rayleigh limit. In these calculations, as in others made by us, we have been careful to tolerate only a 0.5% error in the check for the conservation of energy, and this seems to serve adequately as a check on the convergence of the scattered field coefficients as well. Reciprocity of the scattering solution has also been confirmed as is shown by comparing the data in

Table I. Scattering of a normally incident planewave from $S^3$; h = 0.426, L = 2.84, k = 3.5. Each entry represents $P_{pq}$ of (21).

| ±p, ±q | Dirichlet b.c. | Neumann b.c. |
|---|---|---|
| 0,0 | 0.11534 | 0.52718 |
| 1,0<br>0,1 | 0.03792 | 0.04348 |
| 1,1 | 0.18407 | 0.07574 |
| E | 1.0033 | 1.0041 |

Tables II and III. First, from Table II we see that the scattered power in the direction $\theta = 62.2461°$, $\phi = 0°$ is 0.13240 for the case S and 0.21042 for the case H, when a planewave is incident normally on $S^3$. Next, in Table III, the exciting planewave is incident at $\theta_0 = 62.2461°$, $\phi_0 = 0°$. For this latter excitation, the scattered power in the normal direction was computed to be 0.13231 for the Dirichlet b.c. and 0.21063 for the Neumann b.c., thus demonstrating the satisfaction of the reciprocity constraints.

Table II. Same as Table I except h = 0.284, L = 2.84, and k = 2.5.

| ±p, ±q | Case S | Case H |
|---|---|---|
| 0,0 | 0.47167 | 0.15755 |
| 1,0<br>0,1 | 0.13240 | 0.21042 |
| E | 1.0012 | 0.9992 |

Table III. Same as Table II except the incident wave is incident from $\theta_0 = 62.2461°$, $\phi_0 = 0°$.

| p, q | Case S | Case H |
|---|---|---|
| 0,0 | 0.83265 | 0.43256 |
| -1,0 | 0.13231 | 0.21063 |
| -1,±1 | 0.01476 | 0.11957 |
| -2,0 | 0.00818 | 0.11768 |
| E | 1.0027 | 1.0000 |

## A NON-UNIQUENESS OF THE INVERSE SHAPE PROBLEM

As has been seen in the preceding sections, a periodic surface scatters an incident planewave in discrete well-defined directions. Some of these directions, for which $\gamma_{pq}$ is real, have scattered planewaves which go upto $z = \infty$. Others, for which, $\gamma_{pq}$ is imaginary, represent evanescent planewaves. In the far zone, the reflection coefficients $a^+_{pq}$ can be measured for the propagating planewaves; hence the reflected field can be obtained from measurements as

$$\psi_s(x_3,x_1) = \sum_p a^+_{p0} \exp\{ikr\cos(\theta_p-\theta)\};\quad kx_3 \text{ large} \tag{23}$$

where we have considered, for the sake of brevity, only the 2-D problem; $\theta_p = \arctan(\alpha_p,\gamma_{p0})$; $\theta = \arctan(x_1, x_3)$; and r is the radial distance from the origin to the field point. The summation holds only for the propagating plane-waves. However, in most situations, the scattering surface is finite of total expanse B. Hence, if the surface $S^2$ were to be illuminated by a finite-aperture field

$$\psi^i(\underline{x}) = \begin{cases} \exp(ik_0\cdot\underline{x}), & \underline{x}\in S^2,\ -B/2 \le x \le B/2 \\ 0, & \underline{x}\in S^2,\ |x| > B/2 \end{cases} \tag{24}$$

then, for a sufficiently usual case when the Rayleigh-Wood anomalies are absent, the scattered field has been given by Jordan and Lang [7] to be

$$\psi_s(\underline{x}) = kB\ (2\pi kr)^{-\frac{1}{2}} \exp[i(kr-\pi/4)] \sum_p a^+_{p0}\ \mathrm{sinc}[kB(\sin\theta_p-\sin\theta)/2]. \tag{25}$$

For (25) to hold, kB must be large; and only the propagating plane waves need be accounted for.

Consider, next, a flat surface of the same expanse B which is illuminated also by the field (24). This flat surface has a periodic reflectivity profile $\rho$ of period L, the reflectivity function being dependent on the frequency. The scattered field field can be easily set down as

$$\psi^1(x_3, x_1) = \int_{-\frac{1}{2}B}^{\frac{1}{2}B} dx'_1\ \rho(x'_1)\exp(ikx'_1\sin\theta_0)\cdot \cdot \exp[ik\{x_3^2 + (x_1-x'_1)^2\}^{\frac{1}{2}}]\ \{x_3^2 + (x_1-x'_1)^2\}^{-\frac{1}{2}} \tag{26}$$

where $\theta_0 = \arcsin(\alpha_0)$. Because of the periodic nature of $\rho$, this can be reduced to

$$\rho(x) = \sum_p \rho_p \exp(ip2\pi x/L) \tag{27a}$$

$$\psi^1(x_3,x_1) = \sum_p \rho_p \sum_{\ell=-N}^{N} \exp(-ik\ell L\sin\theta_0) \int_{-\frac{1}{2}L}^{\frac{1}{2}L} dx'_1\ \{x_3^2 + (x_1-x'_1+\ell L)^2\}^{-\frac{1}{2}}\cdot \cdot\exp(ikx'_1\sin\theta_p)\exp[ik\{x_3^2 + (x_1-x'_1+\ell L)^2\}^{\frac{1}{2}}], \tag{27b}$$

and which, by approximating,

$$[x_3^2+ (x_1 - x'_1+\ell L)^2]^{\frac{1}{2}} = [x_3^2 + (x_1+\ell L)^2]^{\frac{1}{2}} - - x'_1\ (x_1+\ell L)\ [x_3^2+ (x_1+\ell L)^2]^{-\frac{1}{2}}, \tag{27c}$$

further reduces to

$$\psi^1(x_3,x_1) = \sum_p \rho_p \sum_{\ell=-N}^{N} \exp(-ik\ell L\sin\theta_0)\exp ik\{x_3^2+ (x_1+\ell L)^2\}^{\frac{1}{2}}\ r^{-\frac{1}{2}}\cdot \cdot\mathrm{sinc}\{(kL/2)(\sin\theta_p - (x_1+\ell L)[x_3^2+(x_1+\ell L)^2]^{-\frac{1}{2}})\} \tag{28}$$

with $\mathrm{sinc}(z) = \sin(z)/z$, and the ratio $B/L = 2N+1$ is considered integral. The factor $\sqrt{r}$ is introduced since the measurements are made in the far zone. Furthermore, by realizing that $x_1 = x_3 \tan\theta$, and on focussing our attention on the argument of the sinc function, we observe that this function reduces to

$$\mathrm{sinc}\{(kL/2)\ (\sin\theta_p - \sin\theta)\}, \tag{29a}$$

while the second exponential in (28) becomes

$$\exp[ik\ell L(\sin\theta - \sin\theta_0)]. \tag{29b}$$

Since these two are Fresnel-type approximations, $\ell$ must not assume high enough values so as to render them invalid. Therefore, the somewhat restrictive assumptions that the ratio $B/L \gtrsim 10$ while $NL \ll x_1$ are necessary. However, the product kB can be arbitrarily large. In effect, thus, this is also a high - frequency analysis.

Noting, however, that

$$\sum_{\ell=-N}^{N} \exp(i\ell\xi) = \mathrm{sinc}[(N+\tfrac{1}{2})\xi](2N+1)/\mathrm{sinc}(\tfrac{1}{2}\xi) \tag{30}$$

further simplifies (28) to

$$\psi^1(x_3,x_1) = (2N+1)\ (-)^{2N+1}\, r^{-\frac{1}{2}} \exp(ikr) \cdot \cdot \sum_p o_p\ \mathrm{sinc}\{kB(\sin\theta_p - \sin\theta)/2\} \tag{31}$$

after some manipulation of the various sinc functions involved.

Formally, the scattered field $\psi^1$ is indistinguishable from $\psi_s$ of (25) Furthermore, the Fourier components $\rho_p$ of the reflectivity $\rho$ may be obtained through a least-squares estimation procedure applied to repeated measurements of the scattered field for different angles of incidence. Thus, experimental measurements of the far scattered field for the purpose of determining the surface profile can lead to two different interpretations:

(a) the surface is periodically undulating, and

(b) the surface is flat with a periodic reflectivity profile.

## CONCLUSIONS

We have described a scalar T-matrix formalism which is applicable for highly corrugated periodic surfaces and have shown that the solutions obtained obey unitarity as well as reciprocity constraints. We have also shown that the measurement of the scattered field (which exists only in discrete well-defined directions) can give rise to two different interpretations of the nature of the scattering surface. Further work on extending the presented approach for bimaterial interfaces as well as for elastic scattering problems is in progress.

## REFERENCES


1. Lord Rayleigh, *The Theory of Sound*, Dover, New York (1945).
2. R.F. Millar, *Rad. Sci.*, *8*, 785 (1973).
3. F.O. Goodman, *J. Chem. Phys.*, *66*, 976 (1977).
4. J.T. Fokkema and P.M. van den Berg, *J. Acoust. Soc. Am.*, *62*, 1095 (1977).
5. R.L. Holford, *ibid*, *70*, 1116 (1981).
6. J.A. DeSanto, *ibid*, *57*, 1195 (1975).
7. A.K. Jordan and R.H. Lang, *Rad. Sci.*, *14*, 1077 (1979).
8. S.L. Chuang and R.R. Johnson, *J. Acoust. Soc. Am.*, *71*, 1368 (1982).
9. A. Lakhtakia, V.K. Varadan, V.V. Varadan and D.J.N. Wall, 'The T-matrix approach for scattering by a traction-free periodic rough surface,' *J. Acoust. Soc. Am.*, in press (1984).
10. A. Wirgin, *J. Acoust. Soc. Am.*, *75*, 345 (1984).
11. R.I. Masel, R.P. Merrill and W.H. Miller, *Phys. Rev. B*, *12*, 5545 (1975).
12. P.C. Waterman, *J. Acoust. Soc. Am.*, *57*, 791 (1975).
13. A. Lakhtakia, V.K. Varadan and V.V. Varadan, 'On the acoustic response of a deeply corrugated periodic surface - A hybrid T-matrix approach,' *J. Acoust. Soc. Am.*, (submitted) (1984).
14. S.L. Chuang and J.A. Kong, *Proc. IEEE*, *69*, 1132 (1981).
15. A. Lakhtakia, V.K. Varadan and V.V. Varadan, 'The T-matrix approach for EM scattering by perfectly conducting periodic surfaces,' *Proc. IEEE*, in press (1984).
16. G. Forsythe and C.B. Moler, *Computer Solution of Linear Algebraic Problems*, Prentice-Hall, New Jersey (1967).
17. M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York (1965).

# SOLITARY, PERIODIC AND CHAOTIC WAVES IN THIN FILMS

S. P. Lin and O. Suryadevara
Mechanical and Industrial Engineering Department
Clarkson University
Potsdam, NY 13676

ABSTRACT. A three-dimensional phase space analysis is carried out for a parabolic fourth order partial differential equation. This equation describes the time evolution of a class of nonlinear waves observed in viscous liquid films, diffusion flames and certain chemical oscillations. Solitary waves, almost periodic waves, as well as locally periodic but globally chaotic waves are found numerically.

I. INTRODUCTION. Lin [1] found that the supercritically stable monochromatic waves [2-4] in a liquid film flowing down an inclined plane were stable only with respect to side-band disturbances of small band-widths. With respect to disturbances of arbitrary band-widths, the waves were unstable due to modal interactions [5]. For the nonlinear wave evolution with strong modal interaction, he obtained the following evolution equation:

$$\alpha^{n-2}\frac{\partial f}{\partial T}+\frac{8}{15}(R-R_1)\alpha^{-1}\frac{\partial^2 f}{\partial x^2}+\frac{2}{3}\alpha W\frac{\partial^4 f}{\partial x^4}+E\frac{\partial^3 f}{\partial x^3}+F\frac{\partial^5 f}{\partial x^5}+4f\frac{\partial f}{\partial x}=0 \qquad (1)$$

where $\alpha \equiv (2\pi h_o/L)$, $h_o$ and $L$ being respectively the average film thickness and the wave length, is a small parameter, $n$ is a constant to be chosen for proper time scale of evolution, $f$, $T$ and $x$ are suitably normalized wave amplitude, time and distance in the direction of wave propagation respectively, and $E$, $F$ and $R_1$ are parameters depending on the Reynolds number R, the Weber number W, and the angle $\beta$ between the incline and the horizontal, i.e.

$$E=\frac{32}{63}R^2+2-\frac{40}{63}R\cot\beta\,,\qquad F=\frac{40}{63}\alpha^2 RW+\frac{16}{3}\alpha^2 RW\,,$$

$$R_1=\frac{5}{4}(\cot\beta-\alpha^2 W)\,,\qquad R=\frac{g h_o^3 \sin\beta}{2\nu^2}\,,\qquad W=\frac{c}{\rho g h_o^2 \sin\beta}$$

in which $g$ is the gravitational acceleration, $\nu$ the kinematic viscosity and $\sigma$ is the surface tension. Eq. (1) is an appropriate model equation in the flow parameter range such that

$$R - R_1 = O(\alpha) \ , \qquad \alpha W = O(1).$$

The appropriate time scale in this range is $n=2$. For a much larger surface tension such that $\alpha^2 W = O(1)$, we put $H = \alpha f$ into Eq. (1) and obtain

$$H_T + \frac{8}{15}(R-R_1)H_{xx} + \frac{2}{3}\alpha^2 W H_{xxxx} + 4HH_x = 0 \qquad (2)$$

where terms of $O(\alpha)$ have been neglected and $n=1$ for the appropriate time scale. Eq. (2) can be suitably normalized to read

$$G_\tau + G_\zeta + G_{\zeta\zeta} + G_{\zeta\zeta\zeta\zeta} = 0 \ , \qquad (3)$$

where G is the wave amplitude, and $\tau$ and $\zeta$ are respectively the time and space variables. Eq. (3) is a model equation expected to be asymptotically valid in the range $R-R_1 = O(1)$, $\alpha^2 W = O(1)$. Eq. (3) has also been independently obtained by Atherton and Homsy [6,12], Nepomnyashchii [7], and Sivashinski and Michelson [8] in the context of film waves. Eq. (3) was also found to describe the nonlinear evolution of diffusion flame instability [9], certain chemical oscillation [10] and the Rayleigh-Taylor instability in a top-heavy two-layered film [11]. Thus, despite its limited range of applicability, it does have some general dynamic properties. The common feature the different problems governed by Eq. (3) have is that all of these problems contain a nonlinear driving force, a diffusion process, and a restoring force, but no dispersion effect. The dispersion effect is represented by the third derivative term in Eq. (1). We point out that both equations (1) and (3) were derived in a reference frame moving with the Kinematic wave velocity which is three times the average fluid velocity in the film.

Numerical solutions of Eq. (3) were obtained by Atherton and Homsy [12] with a modified algorithm used by Tappert [13] for the solution of the Kortweg-de Vries

equation. They found that the consequence of nonlinear evolution was a stationary periodic wave which is insensitive to the initial data. The initial data used by them included large amplitude sine waves. In addition to periodic waves Tsvelodub *[14]* and others *[15,16,18]* also found stationary soliton solutions of Eq. *(3)*. However, Sivashinsky and Michelson *[8]* and Kuramoto *[10]* reported recently that the numerical solution of Eq. *(3)* with large amplitude sine waves as the initial data resulted in chaotic waves. There seems to be a puzzling inconsistency among existing works. The purpose of this work is to resolve this puzzle and to gain a deeper understanding of the physics of the nonlinear evolution of a class of problem modeled by Eq. *(3)*.

First we seek a stationary solution of Eq. *(3)*, and cast it into a dynamical system. Then we determine the properties of its fixed points. Guided by these properties we obtain the integral curves and study their structure by portraying them on the Poincare sections and phase planes. From these phase portraits, we seek the conditions which may lead to periodic, solitary or chaotic wave patterns.

II. WAVES IN PHASE SPACE. Consider stationary waves travelling at a constant speed $m$ in the $(\zeta-\tau)$-space. By use of the Gallilean transformation $\xi = \Omega - M\tau$, Eq. (3) can be written as

$$-MG' + GG' + G'' + G'''' = 0, \tag{4}$$

where primes denote differention with respect to $\xi$. The first integration of (4) gives

$$-(M-G/2)G + G' + G''' = c ,$$

where c is the integration constant. Let $G = F-d$, $m = M+d$, and $c=(d^2/2)+Md$ in this equation, we have

$$-(m-F/2)F + F' + F''' = 0, \tag{5}$$

where d is an arbitrary constant datum from which the amplitude of F is measured.

Eq. *(5)* can be cast in the dynamical system

$$\begin{aligned} F' &= F1 \\ F1' &= F2 \\ F2' &= (m-F/2)F-F1 . \end{aligned} \tag{6}$$

The integration of Eq. (6) with respect to $\xi$ gives a set of integral curves which may be displayed as trajectories in the phase space $(F, F1, F2)$. Any trajectory of Eq. (6) will come to a "rest" at a fixed points where $F' = F1' - F2' = 0$. The system (6) has only two fixed points $S1 = (F, F1, F2) = (0, 0, 0)$ and $S2 = (F, F2, F2) = (2m, 0, 0)$. It follows from Eq. (5) or equivalently from Eq. (6) that the eigenvalues, $\lambda$, at these fixed points are the roots of

$$\lambda^3 + \lambda + P = 0 \ , \qquad (7)$$

where $P = -m$ and $m$ respectively at the fixed points $S1$ and $S2$. Since $Q = [(P^2/4) + (1/27)] > 0$ for $P = \pm m$, Eq. (7) has one real and two complex conjugate roots. These roots are

$$\lambda_1 = a^{1/3} + b^{1/3} \ , \qquad \lambda_2 = wa^{1/3} + w^2 b^{1/3} \ , \qquad \lambda_3 = w^2 a^{1/3} + wb^{1/3}$$

where

$$a = -\frac{P}{2} + Q^{1/2} \ , \qquad b = \frac{P}{2} - Q^{1/2} \ , \qquad w = \frac{1}{2}(-1 + i\sqrt{3}) \ .$$

Consequently, at the fixed point $S1$, the real root $\lambda_1$ is of the same sign as $m$, and the sign of the real parts of the complex roots $\lambda_2$ and $\lambda_3$ is opposite to that of $m$. For a stationary wave which travels faster than the Kinematic wave speed, $m>0$ and thus the fixed point $S1$ has a one dimensional unstable manifold and a two-dimensional stable spiral manifold. On the other hand if $m<0$, $S1$ has a one-dimensional stable manifold and a two-dimensional unstable spiral manifold. Similarly if $m>0$, the fixed point $S2$ has a one-dimensional stable manifold and a two-dimensional unstable spiral manifold. If $m<0$, the dimensions of the two manifolds are interchanged at $S2$. When $m=0$, the two fixed points coalesce to form a center. The three characteristic roots at this center are $0$ and $\pm i$. As the shape of a trajectory depends on the initial values of $(F, F1, F2)$

as well as the location of the fixed point, which is fixed solely by $m$, the wave form corresponding to this trajectory will also heavily depend on the particular values of $m$. The fourth order Runge-Kutta method has been used to integrate Eq. *(6)*. The numerical results are given in the next section.

III. NUMERICAL RESULTS. The projection of the integral curve of Eq. *(6)* on the *F-F1* plan is plotted in Fig. *1a*, for the case of $m = .12161$, $F(0) = .00125$, $F1(0) = .01$ $F2(0) = .004$. The trajectory follows initially the one-dimensional unstable manifold near *S1 (0, 0, 0)* before it is attracted into the one-dimensional stable manifold near *S2 (2m, 0, 0)*. However, the transition is marked by several successive kinks, indicating the intersections of the one-dimensional unstable manifolds with the two-dimensional stable spiral manifold near *S1* and with the two-dimensional unstable manifold near *S2*. The trajectory then spirals cutward from *S2* along its two-dimensional unstable manifold before entering into the two-dimensional stable manifold of *S1*. Then, it comes to a point very close to the initial point *(.00125, .01, .004)* when $\xi = 175$. Further integration produced an extension of the trajectory which almost duplicates its predecessor from $\xi=0$ to *175*. The corresponding wave profile is plotted in Fig. *1b*. The intersection of the trajectory with the plane $F2 = 0$, i.e. a Poincare section, is given in Fig. *1c* for the first eight successive wave trains. This figure reveals clearly the nature of an almost periodic wave. Slightly different initial points lead to slightly altered trajectories. It appears that the pseudo periodic waves are structurally stable with respect to small disturbances. Among these trajectories, for a given $m$, there appears to be only one exactly periodic wave. Thus, it is infinitely more likely to find almost periodic waves than the exactly periodic one. Almost periodic waves have also been found for larger values of $m$. Fig. *2* gives such an example for $m=4$. Note that the larger amplitude almost periodic waves tend to appear more chaotic over a distance much larger than

the typical local peak to peak distance but remain locally almost periodic. Similar waves have also been found for negative values of $m$.

Fig. 3a gives the soliton with $m = 1.21599$. The corresponding homoclinic trajectory is projected on the $F$-$F1$ plane in Fig. 3b. This soliton is identical to that given in Fig. 1 of Tsvelodub except that his amplitude $H$ was scaled differently from our $F$. Tsvelodub obtained the soliton by representing the solution as a Fourier integral and by solving the resulting integral equation interatively. This soliton is characterized by a large elevated peak following a small amplitude wave train ahead of it. The second soliton with a deep trough trailed by small amplitude oscillations travelling at $m = -1.21599$ also exists, since Eq. (6) remains invariant upon changing the sign of $\xi$, $F$ and $m$.

The projection of a trajectory onto the $F$-$F1$ plane is given in Fig. 4a for the case of $m=0$. The corresponding wave profile is given in Fig. 4b. The intersection of the trajectory with the plane $F2 = 0$, i.e. the Poincare section, is given by Fig. 4c. It is seen from this figure that small finite amplitude periodic waves which travel exactly at the Kinematic waves speed do exist but they are neutrally unstable.

IV. DISCUSSION. The phase space study of Eq. (6) showed that the two types of solitons found by Tsvelodub [14] and Shkadov [15] do exist. It is very unlikely that these solitons are endowed with the stability properties of the $K$-$dV$ [16] solitons. Stationary periodic wave solutions [14,17] of Eq. (3) also exist. However, these pure waves are infinitely fewer in numbers than other forms. These forms include those which are almost periodic locally over a few "wave length" but appear chaotic over a much larger distance. Our findings that most of the stationary waves appear locally periodic but globally random may explain why Atherton and Homsy were able to obtain a purely periodic stationary wave of relatively short wave from Eq. (3) with an initial sinusiodal wave of a

small length, i.e. $F = .1 \sin\xi$. Our results may also explain why the initial sinusoidal wave of a much larger length i.e. $F = \sin(\pi\xi/100)$ used by Sivashinsky and Michelson resulted in locally periodic but globally chaotic waves. Our results are also consistent with the finding of Pumir *[19]* et al. that Eq. *(3)*, when discretized, possesses positive characteristic exponents only when the wave lengths are sufficiently large. The apparent inconsistencies in the existing works are now explained. It is almost clear that the ultimate forms of the equilibrated waves are very sensitive to the initial conditions. However, the precise initial conditions which must be met in order to produce specific types of stationary waves of Eq. *(3)* remain unknown.

It should be pointed out that not all of the predicted chaotic structure can be expected to be observable in experiments, especially those with large amplitude. Some of the chaotic structure may simply be the consequence of the neglect of the physical factors assumed to be insignificant in Eq. *(3)*. For example, the neglected three dimensional stabilizing surface tension effects *[20]* may alter the appearance of chaos dramatically. Only qualitative comparisons between theories and experiments are mentioned in the cited works.

## REFERENCES


1. S. P. Lin, J. Fluid Mech. 63, 417 (1974).
2. S. P. Lin, J. Fluid Mech. 36, 113 (1969).
3. B. Gjevik, Phys. Fluids, 13, 1918 (1970).
4. C. Nakaya, Phys. Fluids 18, 1407 (1975).
5. T. B. Benjamin and J. E. Feir, J. Fluid Mech. 31, 209 (1967).
6. R. W. Atherton, "Studies of the Hydrodynamics of a Viscous Liquid Film Flowing Down an Inclined Plane," Engineer's Thesis, Stanford University (1972).
7. A. A. Nepomnyashchii, Izv. Akad. Nauk SSSR, Mekh. Zhidk. Gaza, No. 3 (1974).
8. G. I. Sivashinsky, and D. M. Michelson. Prog. Theor. Phys. 63, 2112 (1980).
9. G. I. Sivashinsky, Acta Astronatica 4 1177 (1977).
10. Y. Kuramoto, Supplement, Prog. Theor. Phys, 64, 346 (1978).
11. A. J. Babchin, A. L. Frenkel, B. G. Levitch and G. I. Sivashinsky, Phys. Fluids, 26, 3159 (1983).
12. R. W. Atherton and G. M. Homsy, private communication.
13. F. Tappert, "Lectures in Applied Mathematics" 15 (A. Newell ed.) American Math. Soc.
14. O. Yu Tsvelodub, Izvestiya Akademii Nauk SSSR, Mekhnika Zhidk Gaza, No. 4, 142 (1980). See also Thermophysical Investigations, Novosibirsk (1977).
15. V. Ya. Shkadov, Izv. Akad. Nauk SSSR, Mekh. Zhidk. Gaza, No. 1 (1977).
16. N. J. Zabusky and M. D. Kruskal, Phys. Review Letters, 15, 140 (1965).
17. V. E. Nakoryakov, B. G. Pokusaev and S. V. Alekseenko, Inzh. Fiz. Zh., 30, No. 5 (1976).
18. H. Tougou, J. Phys. Soc. Japan, 50, 1017 (1981).
19. A. Pumir, P. Manneville, Y. Pomeau, J. Fluid Mech., (1984).
20. S. P. Lin and M. V. G. Krishna, Phys. Fluids 20, 2005 (1977).

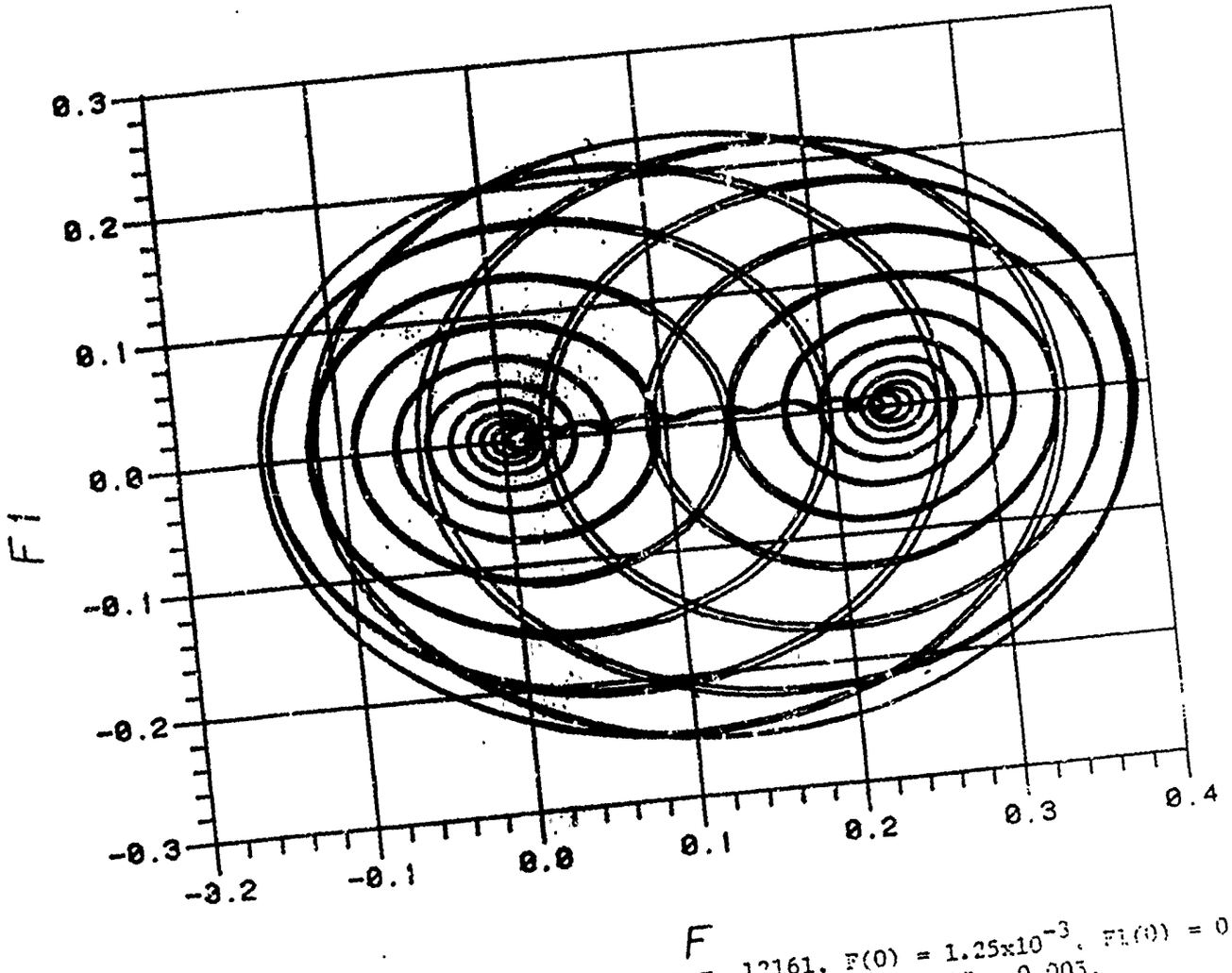



FIG. 1. Almost periodic waves. m = .12161, F(0) = $1.25 \times 10^{-3}$, F1(0) = 0.01, F2(0) = 0.004. Numerical integration step $\Delta\xi$ = 0.003.

FIG. 1a. Trajectory projection.

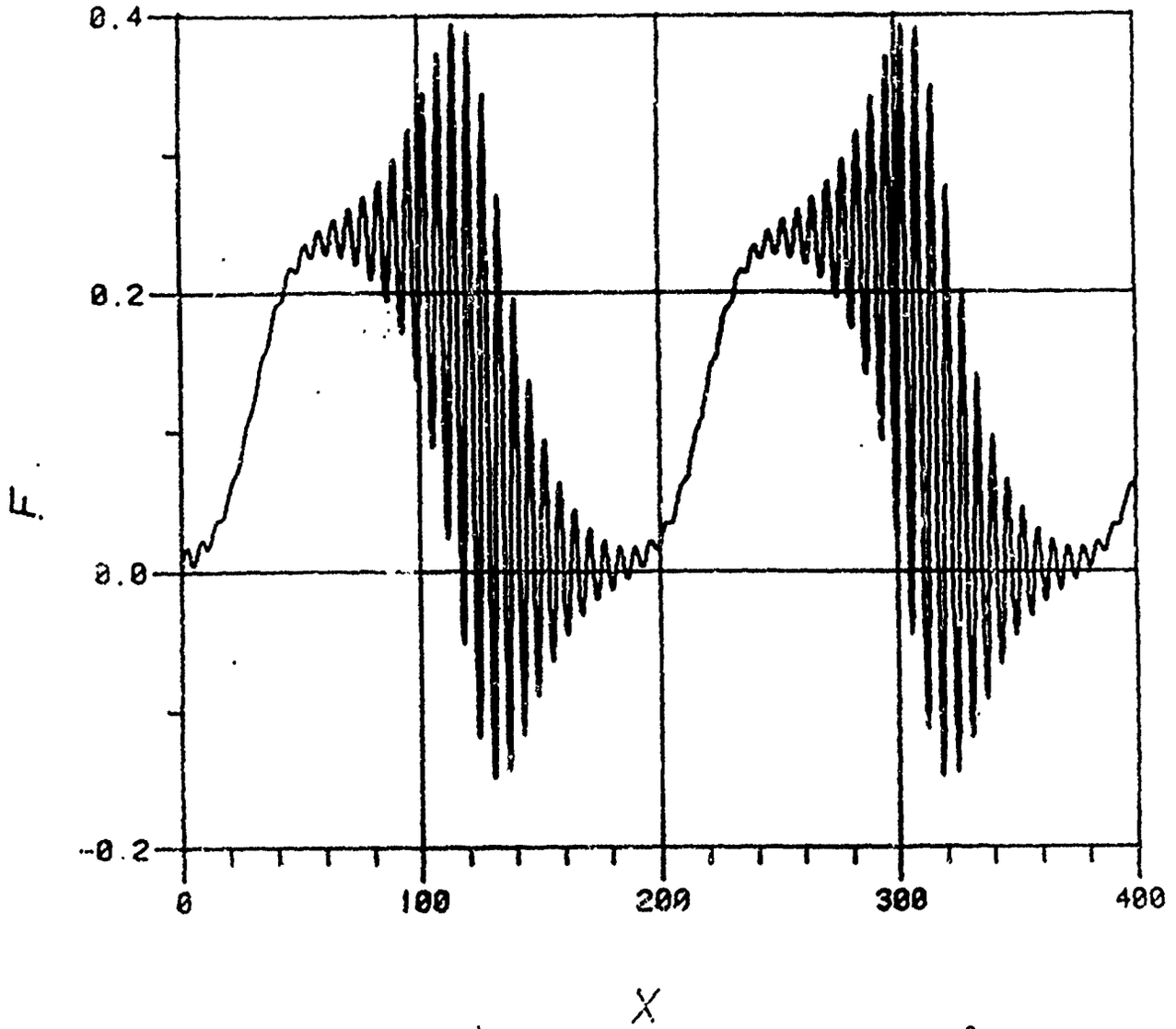



FIG. 1. Almost periodic waves. m = .12161, F(0) = $1.25 \times 10^{-3}$, F1(0) = 0.0, F2(0) = 0.004. Numerical integration step $\Delta\xi$ = 0.005.

FIG. 1b. Wave profile.

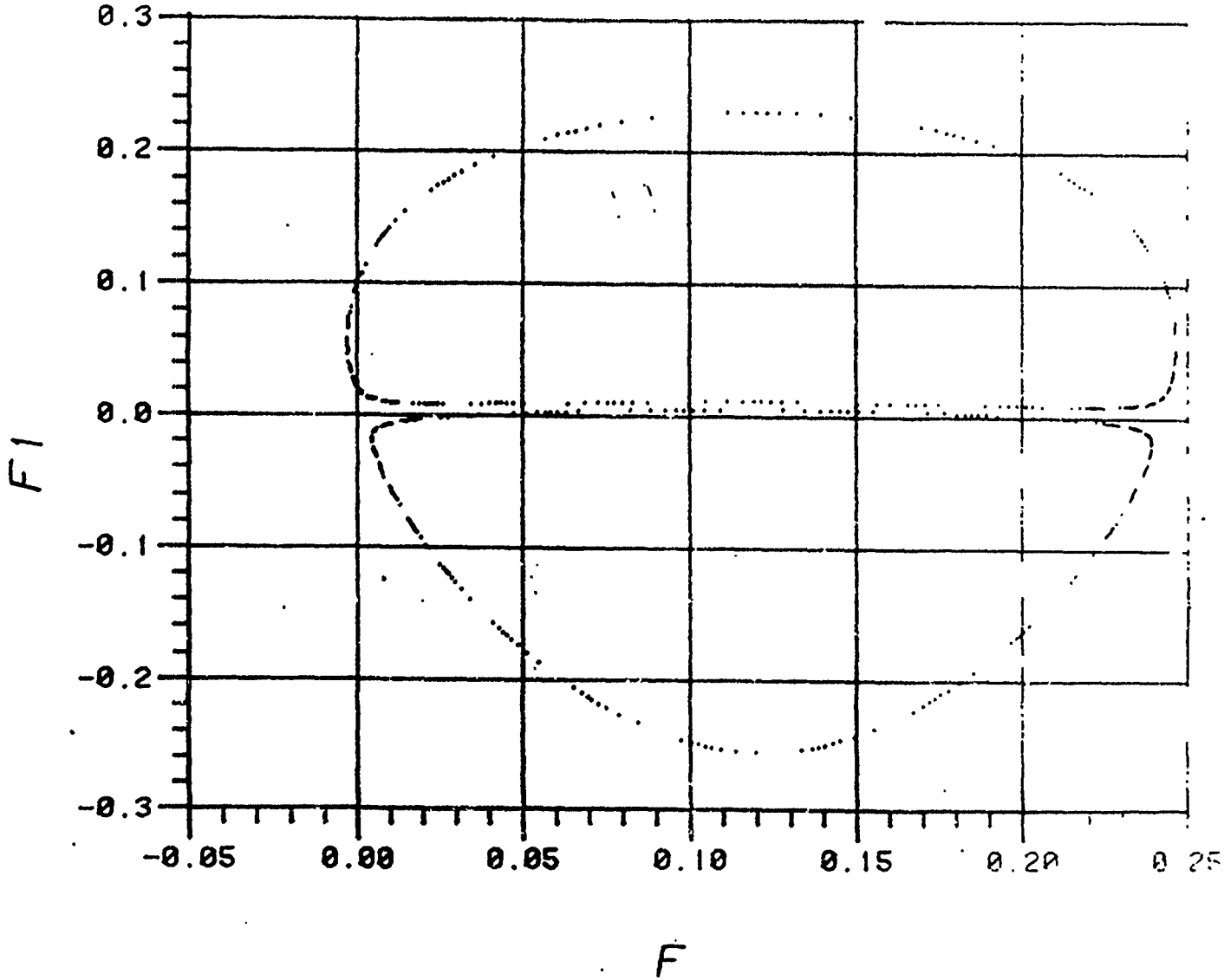



FIG. 1. Almost periodic waves. m = .12161, F(0) = $1.25 \times 10^{-3}$, F1(0) = 0.01 F2(0) = 0.004. Numerical integration step $\Delta\xi$ = 0.005.

FIG. 1c. Poincare section on F2 = 0.

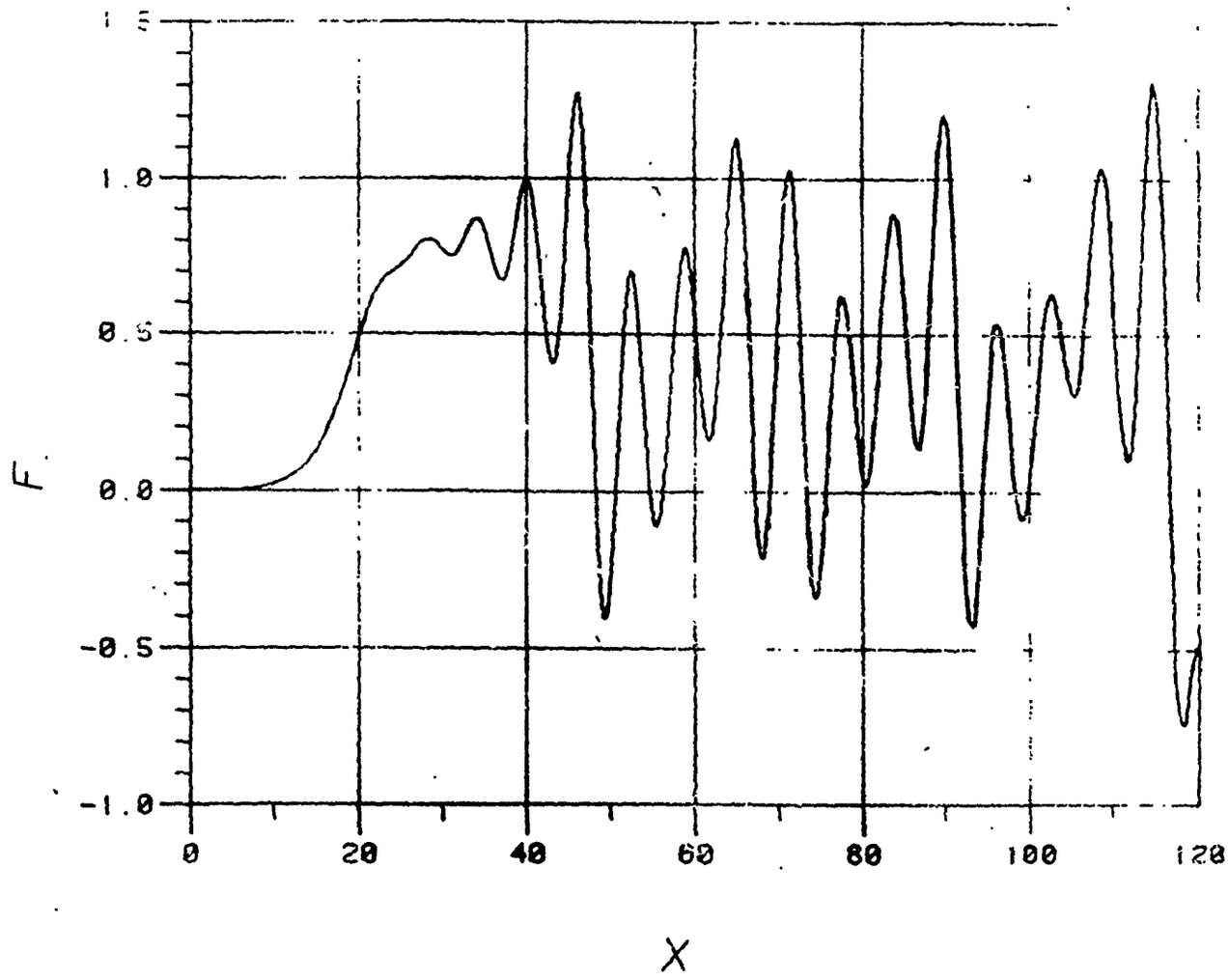



FIG. 2. Locally almost periodic globally chaotic wave.

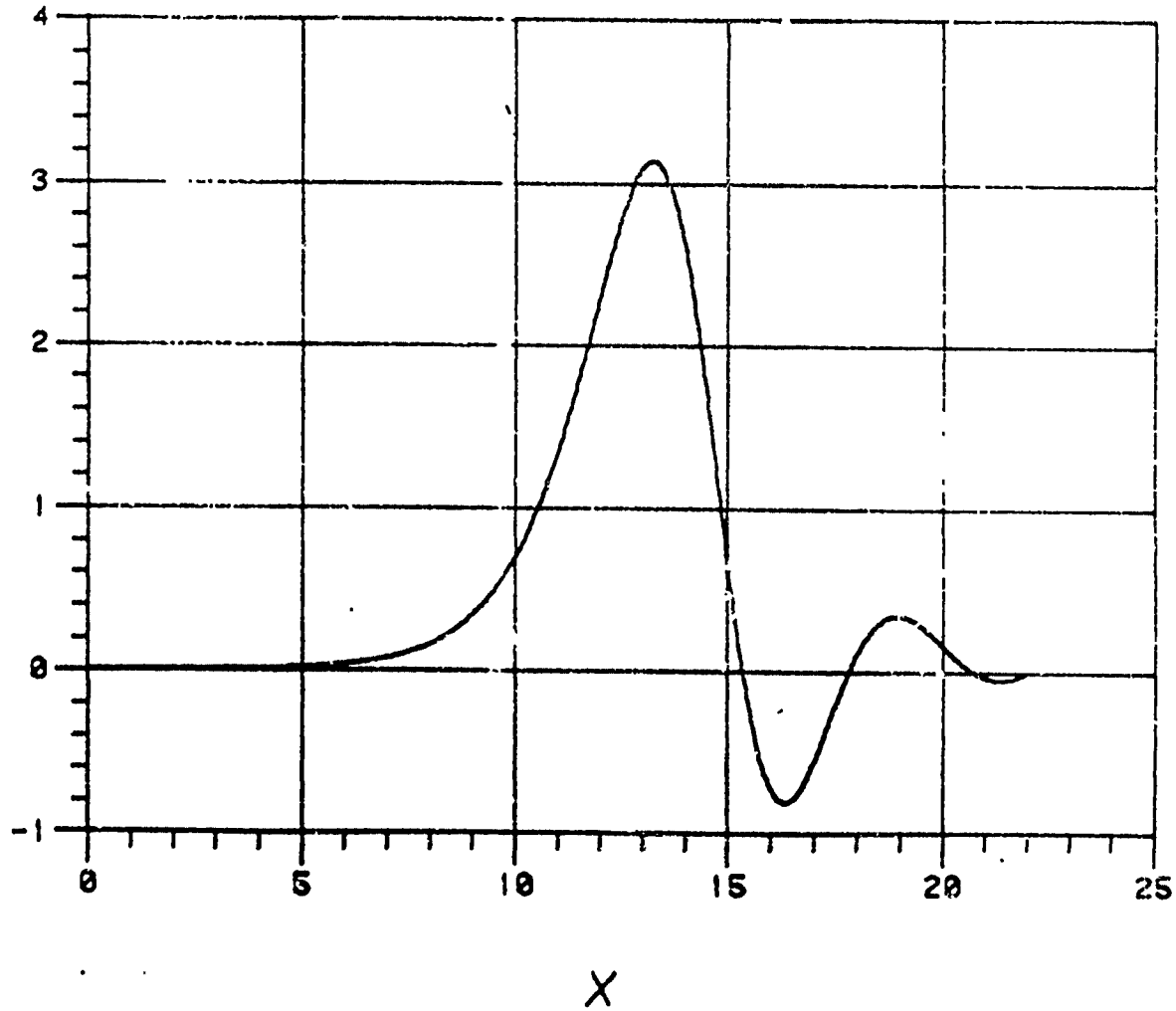



FIG. 3. Solitary wave. m = 1.21599, F(0) = 0, F1(0) = 0, F2(0) = 0.001, $\Delta\xi$ = 0.001.

FIG. 3a. Wave profile.

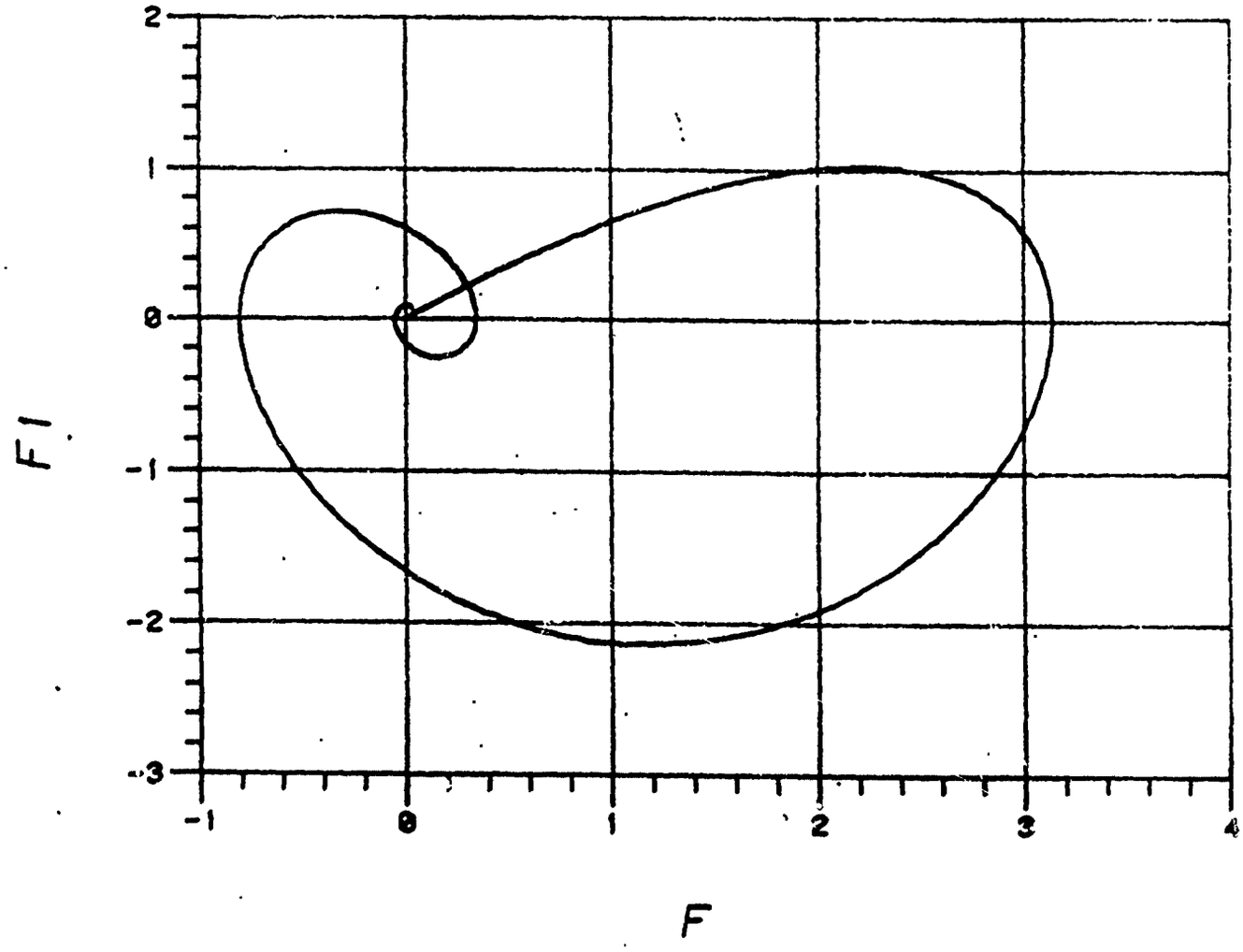



FIG. 3. Solitary wave. m = 1.21599, F(0) = 0, F1(0) = 0, F2(0) = 0.001, $\Delta\xi$ = 0.001.

FIG. 3b. Trajectory projection on F-F1 plane.

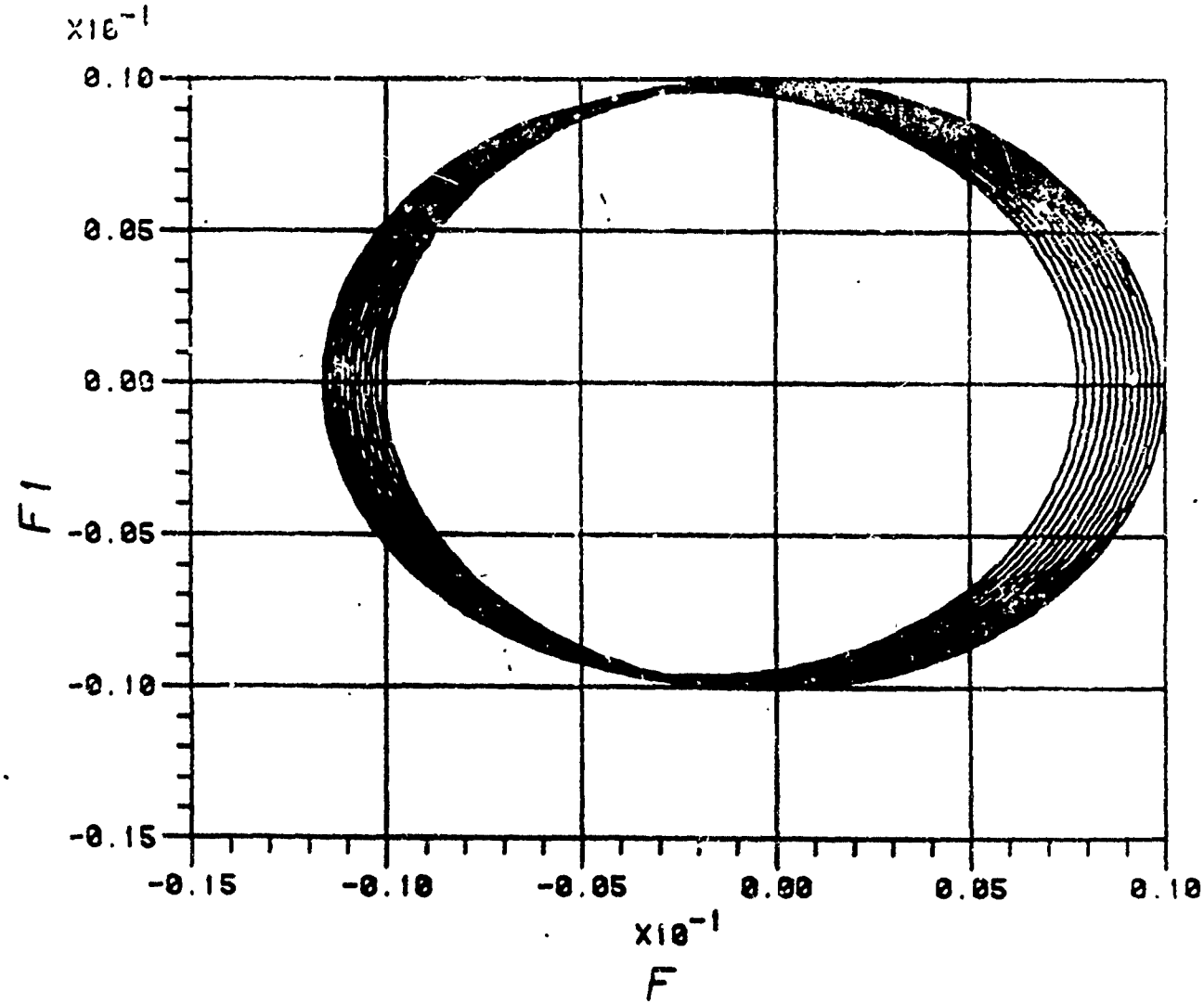



FIG. 4. Neutrally unstable periodic waves.

FIG. 4a. Trajectory projection

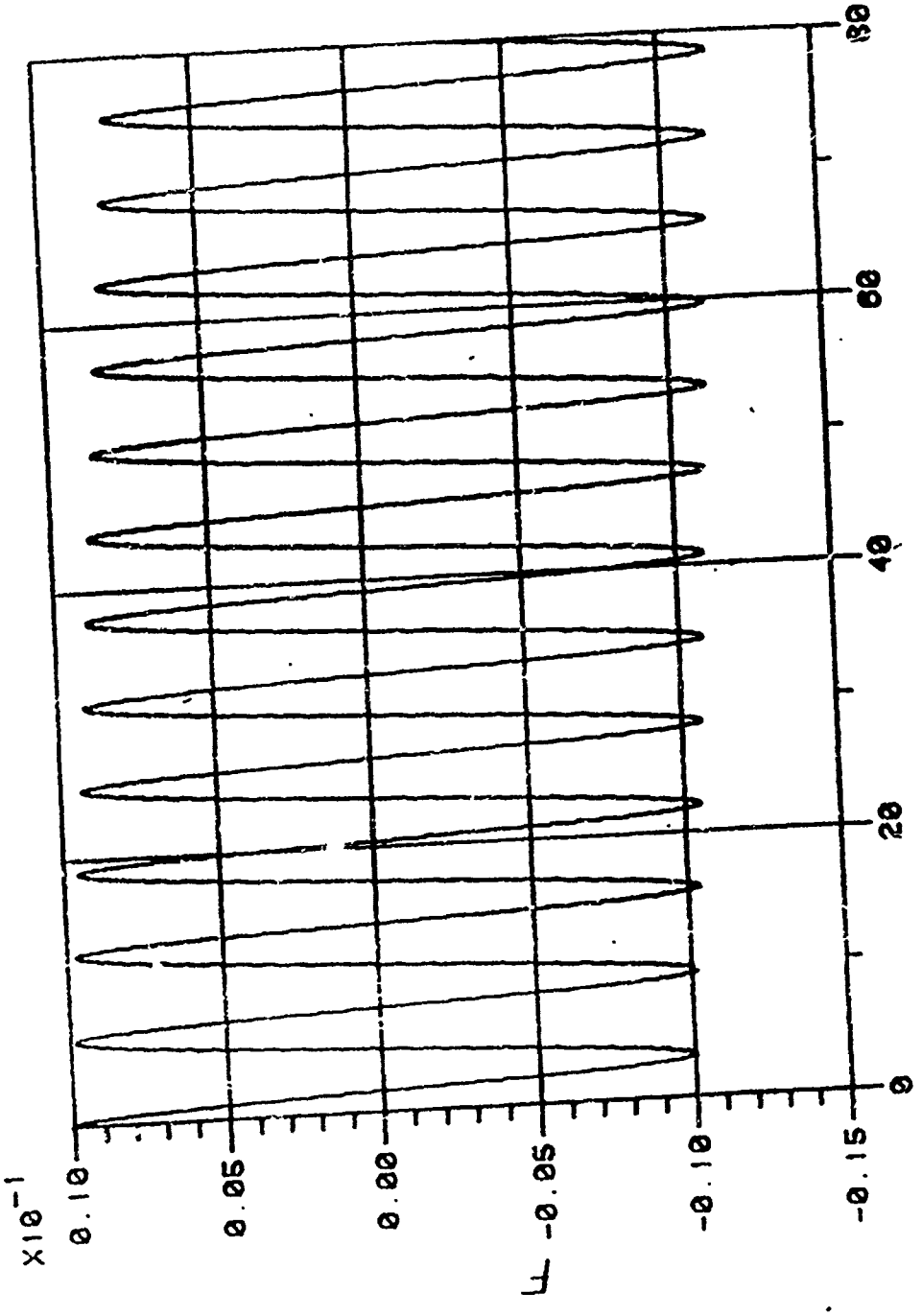



FIG. 4. Neutrally unstable periodic waves.

FIG. 4b. Wave profile.

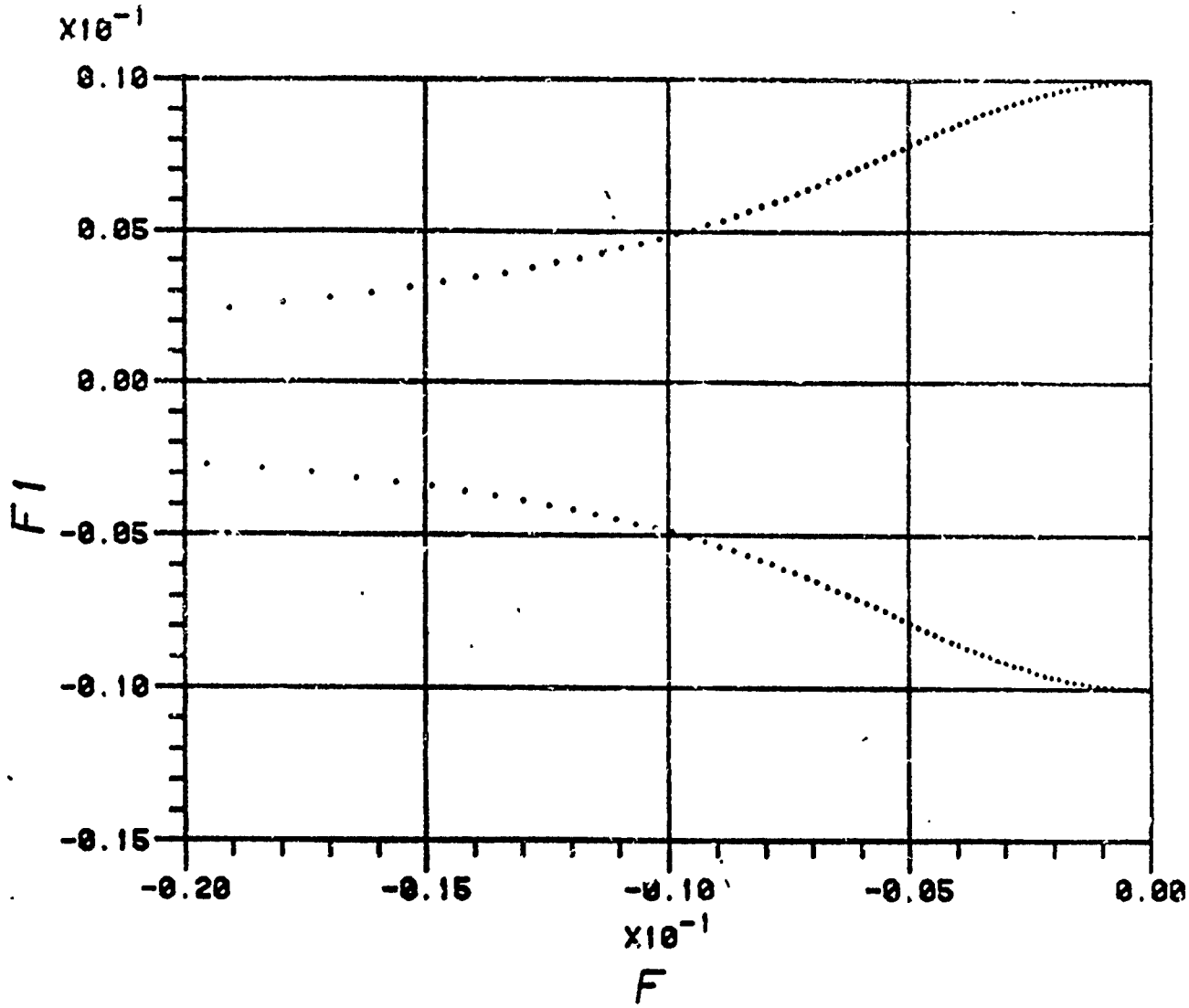



FIG. 4. Neutrally unstable periodic waves.

FIG. 4c. Poincare section on F2 = 0.

# DISCONTINUOUS DEPENDENCE OF SOLUTION ON BOUNDARY CONDITIONS FOR LARGE AMPLITUDE SHOCK WAVES[1]

T. C. T. Ting
Department of Civil Engineering,
Mechanics and Metallurgy
University of Illinois at Chicago
P.O. Box 4348
Chicago, Illinois 60680

ABSTRACT. Plane waves of finite amplitude in a simple elastic solid which occupies the half space $x \geq 0$ are studied. For isotropic materials there are two plane polarized simple waves as well as two shock waves and one circularly polarized simple wave which can be regarded as a shock wave. Using the wave curves and shock curves in the stress space, one can see clearly what combination of simple waves and/or shock waves is needed to satisfy the initial and boundary conditions. For hyperelastic materials, the wave curves associated with different wave speed are orthogonal to each other. Examples are presented for second order hyperelastic materials. In one example we show that the solution requires as many as four simple waves. In another, when the amplitude of the applied shock is large enough, the solution, though does not blow up with time, does not depend continously on the boundary conditions. Mathematically, this may create difficulties for a numerical solution of the problem. Physically, this means that if the applied load at the boundary is not properly controlled, any slight deviation in the applied load would result in a finite jump in the material response.

1. INTRODUCTION. Plane finite amplitude waves in simple elastic solids have been studied by many investigators [1-7]. The deformation gradient was invariably used as one of the dependent variables. In [4] Bland showed that there existed three plane simple waves and three associated plane shock waves in isotropic hyperelastic materials; two of the simple waves and two of the shock waves are plane polarized with respect to the deformation-gradient space and the remaining simple wave and shock wave are circularly polarized. Davison [5] obtained centered wave simple solutions to some initial and boundary conditions by a semi-inverse approach, i.e., one assumed a certain combination of simple waves and/or shock waves to see what deformation gradients should be prescribed as the initial and boundary conditions.

One of the main purposes of this paper is to present a means of determining the correct combination of simple waves and/or shock waves to satisfy the prescribed initial and boundary conditions. Since deformation gradients are seldom prescribed as the initial and boundary conditions, we use stress as the dependent variable. We introduce the 'stress paths' for simple waves and 'stress paths' for shock waves to find the solution. The idea of using the stress paths for simple waves in solving the problem was employed in the study of elastic-plastic wave propagation [8,9]. However, stress paths

---

[1]This is based on the work in collaboration with Yongchi Li. [12]



PREVIOUS PAGE IS BLANK

for shock wave were straight lines because the shock wave studied in [8,9] was linear and involved only one stress component. The stress paths are in fact the projection of wave curves [10] on the stress-space.

2. BASIC EQUATIONS FOR PLANE WAVES AND THEORY OF SIMPLE WAVES. In a fixed rectangular coordinate system, let $x_i$ and $X_i$ be, respectively, the position of a particle at time $t$ and at $t = 0$ which is taken as the undeformed state. The material occupies the half space $X_1 \geq 0$. Assuming that the plane wave is propagating in the $X_1$-direction, we have

$$x_i = X_i + u_i(X,t) \tag{1}$$

where $X = X_1$ and $u_i$ is the displacement. The deformation gradient $F$ then has the expression

$$F_{ij} = \frac{\partial x_i}{\partial X_j} = \begin{bmatrix} 1+p_1 & 0 & 0 \\ p_2 & 1 & 0 \\ p_2 & 0 & 1 \end{bmatrix} \tag{2}$$

$$p_i = \partial u_i / \partial X \tag{3}$$

Let $\underset{\sim}{T}$ be the Piola-Kirchhoff stress tensor of the first kind. For simple elastic materials $\underset{\sim}{T}$ is a given function of $\underset{\sim}{F}$

$$\underset{\sim}{T} = \underset{\sim}{T}(\underset{\sim}{F}) \tag{4}$$

Since $\underset{\sim}{F}$ is a function of $X$ and $t$, so is $\underset{\sim}{T}$. The equations of motion and the continuity of displacement can then be written as

$$\frac{\partial s_i}{\partial X} - \rho_o \frac{\partial v_i}{\partial t} = 0 \tag{5}$$

$$\frac{\partial v_i}{\partial X} - \frac{\partial p_i}{\partial t} = 0 \tag{6}$$

where $\rho_o$ is the mass density in the undeformed state and

$$v_i = \frac{\partial u_i}{\partial t} \tag{7}$$

$$s_i = T_{i1} = \sigma_{i1} \tag{8}$$

In Eq. (8) $\sigma_{ij}$ is the Cauchy stress which is related to $T_{ij}$ by [11]

$$\underset{\sim}{\sigma} = J^{-1} \underset{\sim}{T} \underset{\sim}{F}^T \tag{9}$$

where $J = ||\underset{\sim}{F}||$ and the superscript $T$ stands for the transpose. The second equality of Eq. (8) follows from the special form of $\underset{\sim}{F}$ given in Eq. (2).

Noticing that $s_i$ is a function of $p_1$, $p_2$ and $p_3$

$$s_i = s_i(p_1, p_2, p_3) \tag{10}$$

and assuming that its inverse as well as the derivatives

$$G_{ij} = \partial p_i / \partial s_j \tag{11}$$

exist, we write Eqs. (5,6) in matrix notations as

$$\left.\begin{aligned} \underset{\sim}{s}_{,X} - \rho_o \underset{\sim}{v}_{,t} &= 0 \\ \underset{\sim}{v}_{,X} - \underset{\sim}{G}\underset{\sim}{s}_{,t} &= 0 \end{aligned}\right\} \tag{12}$$

where a comma stands for the partial differentiation. Introducing the following matrices

$$\underset{\sim}{A} = \begin{bmatrix} \rho_o \underset{\sim}{I} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{G} \end{bmatrix}, \quad \underset{\sim}{B} = \begin{bmatrix} \underset{\sim}{0} & -\underset{\sim}{I} \\ -\underset{\sim}{I} & \underset{\sim}{0} \end{bmatrix}, \quad \underset{\sim}{w} = \begin{bmatrix} \underset{\sim}{v} \\ \underset{\sim}{s} \end{bmatrix}. \tag{13}$$

where $\underset{\sim}{I}$ is the unit matrix, Eq. (12) has the form

$$\underset{\sim}{A}\underset{\sim}{w}_{,t} + \underset{\sim}{B}\underset{\sim}{w}_{,X} = \underset{\sim}{0} \tag{14}$$

The formulation here differs from that of [1-7] in that the dependent variables are $\underset{\sim}{s}$ and $\underset{\sim}{v}$, not $\underset{\sim}{p}$ and $\underset{\sim}{v}$. Notice that $\underset{\sim}{B}$ is a constant matrix while $\underset{\sim}{A}$ depends on $\underset{\sim}{s}$ only.

3. SECOND ORDER ISOTROPIC HYPERELASTIC MATERIALS. For illustrative purposes we will consider isotropic hyperelastic materials of second order, i.e., the deformation gradients are functions of $s_1$, $s_2$, $s_3$ of order up to two. Hence

$$\varepsilon = a\sigma + \frac{b}{2}\sigma^2 + \frac{e}{2}\tau^2 \tag{15a}$$

$$\gamma = d + e\sigma \tag{15b}$$

where a, b, d and e are material constants and

$$\left.\begin{aligned} \sigma &= s_1 \ , \quad \tau^2 = s_2^2 + s_3^2 \\ \varepsilon &= p_1 \ , \quad \gamma^2 = p_2^2 + p_3^2 \end{aligned}\right\} \tag{16}$$

For linear materials b = e = 0 and hence a and d can be identified with Lame constants $\lambda$ and $\mu$ by

$$\begin{aligned} d &= 1/\mu \\ a &= 1(\lambda + 2\mu) \end{aligned} \tag{17}$$

Since Lame constants are positive, we have

$$d > 2s > 0 \tag{18}$$

Eqs. (14) yield three wave speeds $c_i$, $(i = 1,2,3)$. We assume that

$$c_1 \geq c_2 \geq c_3 \tag{19}$$

Simple waves associated with $c_1$ and $c_3$ will be called, respectively, the 'fast' and 'slow' simple waves. The simple wave associated with $c_2$ degenerates into a shockwave.

Detail derivations of equations for simple waves can be found in [12]. The stress paths for the fast and slow simple waves lie on the $(\sigma,\tau)$ radial planes. They are orthogonal to each other. The stress paths for $c_2$ - wave is a circle $\tau$ = constant on a $\sigma$ = constant plane. For the second order hyperelastic materials given by Eqs. (15), the stress paths for simple waves and shock waves depend on one non-dimentional parameter k:

$$k = 1 - \frac{b}{e} \tag{20}$$

There are four cases depending on the values of k. The stress paths for simple waves for case 3 $(0 \leq k \leq 1)$ are shown in Fig. 1. Solid lines are stress paths for fast simple waves and dotted line are for slow simple waves. The arrows indicate the direction along which the wave speed is decreasing. Thus, if the stress state at $t = 0$ is at point B while the boundary condition specified at $X = 0$ is the point A, the admissable stress path is BC*A. The wave pattern in the $(X,t)$ plane is shown in Fig. 2. We see that we have four simple waves in this example. This is due to the existence of the point $\sigma^*$ in Fig. 1 where the fast and slow wave speeds are identical. This is a singular point which always provides some unexpected phenomena [13].

4. PLANE SHOCK WAVES. If we denote the jump of a quantity f across a shock wave by

$$[f] = f^- - f^+ \tag{21}$$

where $f^-$ and $f^+$ are the values of f behind and ahead of the shock wave, the jump conditions representing the conservation of momentum and the continuity of displacements can be written as [11]

$$\left.\begin{aligned} [s_i] + \rho_0 V[v_i] &= 0 \\ [v_i] + \rho_0 V[p_i] &= 0 \end{aligned}\right\} \tag{22}$$

By eliminating $[v_i]$ in Eqs. (22) and in making use of (15), one obtains equations for $(\sigma^-,\tau^-)$ when $(\sigma^+,\tau^+)$ and V are known. For a fixed $(\sigma^+,\tau^+)$, $(\sigma^-,\tau^-)$ traces a stress path for shock wave as V varies. It can be shown that $V \to c$ as $(\sigma^+,\tau^+) \to (\sigma^-,\tau^-)$. We will denote by

$V_1$ (or $V_3$) the shock wave speed associated with points on the stress path for shock wave which reduces to $c_1$ (or $c_3$) when $(\sigma^-,\tau^-)$ approaches $(\sigma^+,\tau^+)$.

It should be noted that not every point on the stress path for shock wave is an admissible shock. According to Lax [14], a shock wave is stable if

$$c_i^- \geq V \geq c_i^+ , \qquad (i = 1,2,3) \tag{23a}$$

Otherwise, the shock wave may develop into a simple wave. We may also write Eq. (23a) as

$$c_i(\sigma^-,\tau^-) \geq V_i(\sigma^-,\tau^- ; \sigma^+,\tau^+) \geq c_i(\sigma^+,\tau^+) \tag{23b}$$

5. SOLUTION WHICH INVOLVES SHOCK WAVES. For illustrative purposes we consider the following initial and boundary conditions

$$(\sigma,\tau) = (0,0) \quad , \quad \text{at } t = 0 \tag{24}$$

$$= (\sigma^a,\tau^a) \quad , \quad \text{at } X = 0$$

Depending on the location of $(\sigma^a,\tau^a)$ in the $(\sigma,\tau)$ plane, the solution may consist of a different combination of simple waves and/or shock waves. We again use case 3 as an example. With the initial value being zero and the boundary value prescribed as $(\sigma^a,\tau^a)$, the solution depends on the position of $(\sigma^a,\tau^a)$ in the $(\sigma,\tau)$ plane, Fig. 3 (see also Fig. 1). If $(\sigma^a,\tau^a)$ is the point $A_1$ in region I bounded by the positive $\sigma$-axis and OP which is the stress path for slow simple wave from the origin, we have the solution which consists of two simple waves as shown in Fig. 4. If $(\sigma^a,\tau^a)$ is outside of region I, there is no solution which consists of simple waves only. In this case we consider a shock wave $V_1$ with $(\sigma^+,\tau^+) = (0,0)$. The stress path for shock wave $V_1$ is the negative $\sigma$-axis $0\hat{\sigma}m_2$ and the curve $\hat{\sigma}Q$ which is one branch of a hyperbola. The other branch of the hyperbola violates the stability condition and hence is not admissible. $\hat{\sigma}$ is the stress at which

$$V_1(0,0 ; \hat{\sigma},0) = c_3(\hat{\sigma},0) \tag{25}$$

We use double solid lines (or double dashed lines) to denote the stress path for admissible shock wave $V_1$ (or $V_3$). The arrows on the path indicates the direction from $(\sigma^+,\tau^+)$ to $(\sigma^-,\tau^-)$. Thus if $(\sigma^a,\tau^a)$ is the point $A_2$ in region II bounded by $Q\hat{\sigma}OP$, the solution consists of a shock wave $V_1$ which carries the stress from point 0 to M followed by a slow simple wave $c_3$ which carries the stress to $A_2$. If $(\sigma^a,\tau^a)$ is the point $A_3$ in region III, Fig 3, which is bounded by $\hat{\sigma}m_2$ and $\hat{\sigma}Q$ the solution consists of shock waves $V_1$ and $V_3$, Fig. 4.

Notice that there are actually more than three types of solutions as shown in Fig. 4. For instance, if $(\sigma^a,\tau^a)$ is on the positive $\sigma$-axis (or on OP), the solution consists of one simple wave $c_1$ (or $c_3$) only. Similarly, if $(\sigma^a,\tau^a)$ is on the negative $\sigma$-axis or on $\hat{\sigma}Q$, the solution consists of one shock wave $V_1$ only.

6. DISCONTINUOUS DEPENDENCE OF SOLUTION ON BOUNDARY CONDITIONS. The stress path $MA_4$ for shock wave $V_3$ in Fig. 3 intersects the $\sigma$-axis at point $A_4$. If $(\sigma^a,\tau^a)$ is at point $A_3$, we have the solution in which a $V_1$ shock wave carries the stress from point 0 to M followed by a $V_3$ shock wave which carries the stress from M to $A_3$. On the other hand, if $(\sigma^a,\tau^a)$ is at point $A_4$, we have the solution in which only one shock wave $V_1$ carries the stress from point 0 to $A_4$. As point $A_3$ approaches $A_4$, it can be shown that

$$V_1(0;M) = V_3(M;A_4) = V_1(0;A_4) \tag{26}$$

The constant region m, Fig. 4, between the two shocks diminishes to zero and the two shocks coalesce into one. Therefore, in the limit as $A_3$ approaches $A_4$ the two solutions are identical. However, as long as $A_3$ is not on the $\sigma$-axis, we have a constant stress region m in which the shear stress $\tau^m$ is finite; although the constant region may be small for small time t.

Thus for a half-space which is initially stress free, an impact at the boundary $X = 0$ with $\sigma^a < \hat{\sigma}$ and $\tau^a = 0$ produce one shock wave which is a longitudinal shock because no shear stress is generated. If $\tau^a$ is nonzero but very small, one has two shock waves (neither of them is longitudinal) with a finite shear stress between the two shocks although the region between the two shock waves may be small for small time. In experiments this implies that a slight misalignment of the longitudinal impact at $X = 0$ can produce quite different response at $X \neq 0$.

## REFERENCES


[1] Chu, Boa-Teh, 'Finite Amplitude Waves in Incompressible Perfectly Elastic Materials,' Journal of the Mechanics and Physics of Solids, Vol. 12, 45-57 (1964).

[2] Bland, D. R., 'On Shock Waves in Hyperelastic Media,' IUTAM Symposium on Second-Order Effects in Elasticity, Plasticity and Fluid Dynamics, 93-108 (1964).

[3] Bland, D. R., 'Dilatational Waves and Shocks in Large Displacement Isentropic Dynamic Elasticity,' Journal of the Mechanics and Physics of Solids, Vol. 12, 245-267 (1964).

[4] Bland, D. R., 'Plane Isentropic Large Displacement Simple Waves in a Compressible Elastic Solids,' Z. Angew, Math. Phys., Vol. 16, 752-769 (1965).

[5] Davison, L., 'Propagation of Plane Waves of Finite Amplitude in Elastic Solids,' Journal of the Mechanics and Physics of Solids, Vol. 14, 249-270 (1966).

[6] Collins, W. D., 'One-Dimensional Non-Linear Wave Propagation in Incompressible Elastic Materials,' Q. J. Mech. Appl. Math., Vol. 19, 2[illegible]-328 (1966).

[7] Howard, I. C., 'Finite Simple Waves in a Compressible Transversely Isotropic Elastic Solid,' Q. J. Mech. Appl. Math., Vol. 19, 329-341 (1966).

[8] Clifton, R. J., 'An Analysis of Combined Longitudinal and Torsional Plastic Waves in a Thin-Walled Tube,' Proc. 5th U.S. Nat. Congress Appl. Mech., ASME, New York, 465-480 (1966).

[9] Ting, T. C. T. and Nan, N., 'Plane Waves Due to Combined Compressive and Shear Stresses in a Half-Space,' J. Appl. Mech., Vol. 36, 189-197 (1968).

[10] Dafermos, C. M., 'Hyperbolic Systems of Conservation Laws,' Brown University, Appl. Math. Report, LCDS #83-5. (1983)

[11] Truesdell, C. and Noll, W., 'The Non-Linear Field Theories of Mechanics,' Handbuch der Physik, Vol. III/3, Springer-Verlag (1965).

[12] Li, Yongchi and Ting, T. C. T., 'Plane Waves in Simple Elastic Solids and Discontinuous Dependence of Solution on Boundary Conditions,' Int. J. Solids Structures, Vol. 19, 989-1008. (1983)

[13] Ting, T. C. T., 'On Wave Propagation Problems in Which $c_f = c_s = c_2$ Occurs,' Quarterly of Applied Mathematics, Vol. 31, No. 3, 275-286 (1973).

[14] Lax, P. D., 'Hyperbolic Systems of Conservation Laws II,' Communications on Pure and Applied Mathematics, Vol. 10, 537-566 (1957).

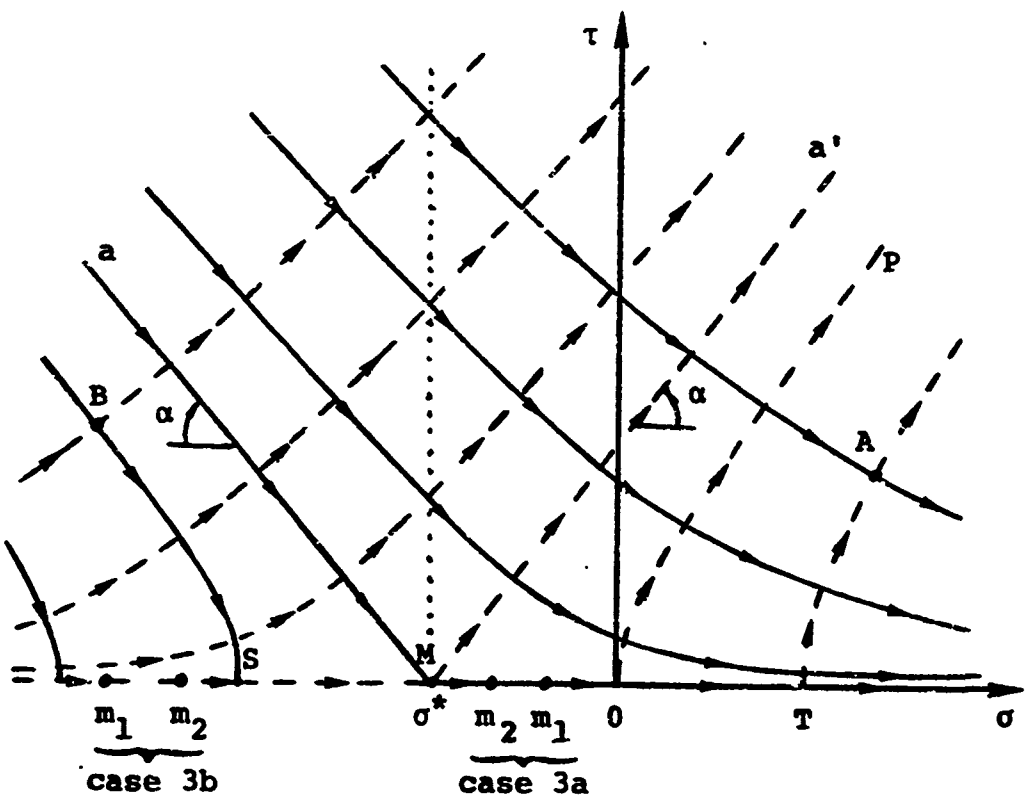



Fig. 1 Case 3a: $e > 0$, $0 < k \leq 1 - a/d$,
Case 3b: $e > 0$, $1 - a/d < k \leq 1$,
($m_1 = m_2 = \sigma^*$ when $k = 1 - a/d$, $m_1 \to -\infty$ as $k \to 1$,
$1 < \tan \alpha = (1+k)^{\frac{1}{2}} \leq \sqrt{2}$ )

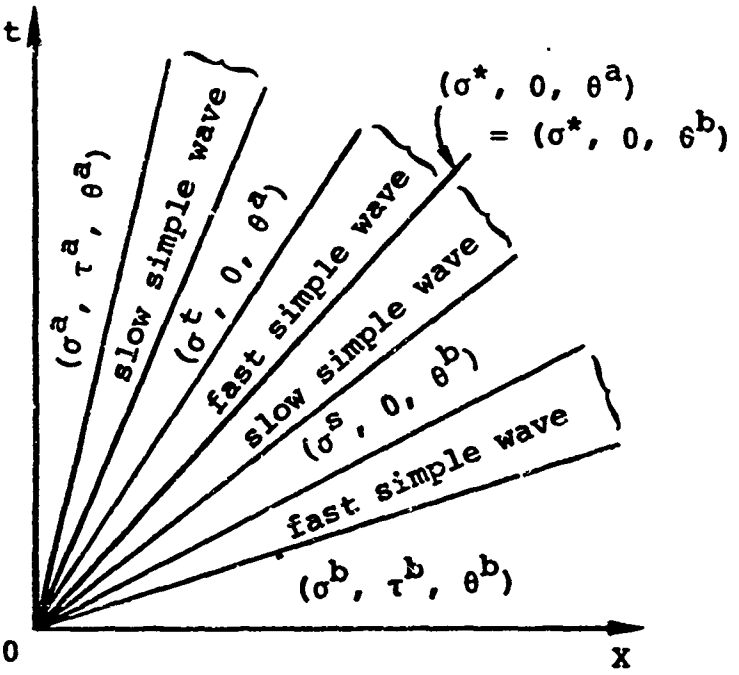



Fig. 2 The solution for $(\sigma^b, \tau^b) = B$ and $(\sigma^a, \tau^a) = A$ in Fig. 1

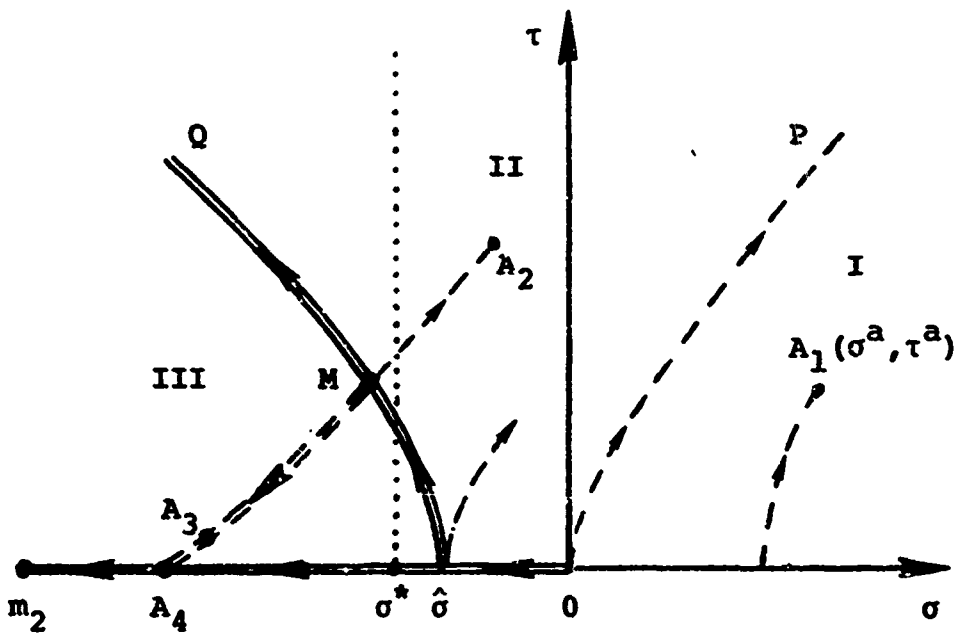
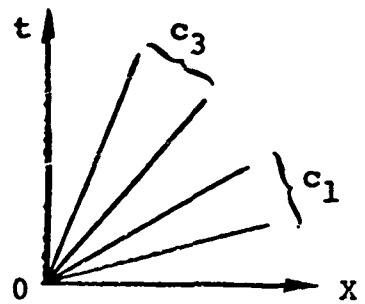
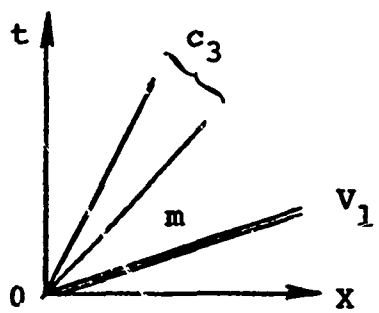


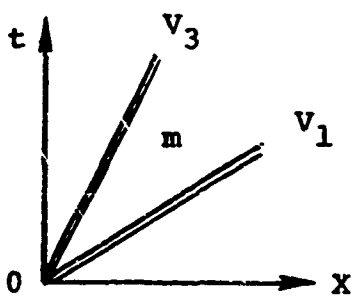

Fig. 3 Case 3b: Initial condition $(\sigma, \tau) = (0, 0)$, boundary condition $(\sigma^a, \tau^a)$ arbitrary.




Region I




Region II




Region III

Fig. 4 Case 3b: Solution in the (X,t) plane for $(\sigma^a, \tau^a)$ in the region shown in Fig. 3.

# FRONT TRACKING AND TWO DIMENSIONAL RIEMANN PROBLEMS: A CONFERENCE REPORT

*James Glimm* [1, 3, 4, 6]
*Christian Klingenberg* [1, 4]
*Oliver McBryan* [1, 3, 5, 7]
*Bradley Plohr* [1, 3]
*David Sharp* [2, 8]
*Sara Yaniv* [1, 3, 4]

*ABSTRACT*

A substantial improvement in resolution has been achieved for the computation of jump discontinuities in gas dynamics using the method of front tracking. The essential feature of this method is that a lower dimensional grid is fitted to and follows the discontinuous waves. At the intersection points of these discontinuities, two-dimensional Riemann problems occur. In this paper we study such two-dimensional Riemann problems from both numerical and theoretical points of view. Specifically included is a numerical solution for the Mach reflection, a general classification scheme for two-dimensional elementary waves, and a discussion of problems and conjectures in this area.

## 1. Introduction

Many phenomena in nature are modeled by nonlinear hyperbolic systems of conservation laws:

$$u_t + \nabla \cdot \vec{f}(u) = 0 . \tag{1.1}$$

The example considered here is the system of Euler equations for a compressible, inviscid, polytropic gas. The equation (1.1) represents an idealization. Its solutions are the limits, as viscosity parameters tend to zero, of the solutions of more complete equations such as the Navier-Stokes equations. The solutions of interest for the system (1.1) are frequently

1. Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, N.Y. 10012.
2. Los Alamos National Laboratory, Los Alamos, N.M. 87545.
3. Supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Dept. of Energy, Contract DE-A02-76ER03077.
4. Supported in part by the Army Research Office, Contract No. DAAG29-83-K-007.
5. Supported in Part by the National Science Foundation, Grant No. MCS-8207965.
6. Supported in part by the National Science Foundation, Grant No. MCS-8243730.
7. Alfred P. Sloan Foundation Fellow.
8. Supported by U.S. Department of Energy.



PREVIOUS PAGE IS BLANK

found to be piecewise smooth. For the Euler equations in one space dimension the jump discontinuities between the smooth pieces are contact discontinuities and shock waves. In two space dimensions these same wave modes give rise to surface singularities of codimension one. The Rankine-Hugoniot conditions, as derived from the integral form of the Euler equations, hold across these jumps.

When solving the system (1.1) numerically, the discontinuities that may occur in its solution may be resolved on coarser grids by the method of front tracking than by conventional finite difference methods. For two space dimensions, front tracking may be described as follows. A one-dimensional grid is placed onto the discontinuity. Its evolution in time is given by a two step procedure, using first the Rankine-Hugoniot relations to propagate the front normally and then using tangential equations to propagate surface waves. This approach works away from the points where the discontinuity curves meet. At such intersection points the geometry does not in general allow an operator splitting into normal and tangential directions, so the evolution of intersection points must be determined as the solution of a two-dimensional Riemann problem. To solve two-dimensional Riemann problems it is crucial to classify the coherent waves, which are defined to be dynamically stable intersection points of one-dimensional coherent waves. The region between the fronts is treated as an initial/boundary-value problem and is solved using (almost) standard finite difference methods. The front and interior schemes are connected in a strip $O(\Delta x)$ in width about the front. For a detailed description see [1].

The front tracking method appears to allow an increase of linear resolution by a factor of three or better, i.e. an improvement in the number of space-time computational grid units by a factor of 27 or better. The method has been tested on various problems. In Sec. 2 and 3 we compare the results of our numerical calculation to experimental results for two specific test problems. An example of how the motion of a two-dimensional coherent wave is determined numerically is given in Sec. 3. In Sec. 4 we give a classification of the two-dimensional coherent waves for compressible gas dynamics and indicate its derivation. In Sec. 5 we conclude with a discussion of outstanding questions related to Riemann problems.

## 2. Regular Reflection of Shock Waves

The front tracking scheme for two-dimensional gas dynamics has been tested on several problems that admit solution by other means. These problems are: an expanding or contracting circular shock followed by a contact discontinuity, the steady-state supersonic flow past a wedge, the Kelvin-Helmholtz instability, regular reflection of a shock wave, and Mach reflection of a shock wave. For details on the first five see [1].

In this section the numerical solution for nonsteady regular reflection of a shock wave is compared with experimental results [3]. The experiment consists of a planar shock (I) moving down a shock tube and impinging on a wedge with a sufficiently large

angle. When the incident shock strikes the wedge corner a reflected shock (R) is formed, which extends from the reflection point to the shock tube wall, where it forms a bow wave in front of the wedge, as shown in Fig. 2.1. We will refer to the region enclosed by the reflected shock as the "bubble". As with Riemann problems in general, the solution is self-similar, i.e $u(t,x) = u(\sigma t,\sigma x)$ for every $\sigma>0$. In the computation presented here, the Mach number of the incident shock is 2.05 and the angle of the wedge is 63.4°.

The numerical calculation was initialized just after the reflected shock had formed and enclosed only a small region about one quarter of a mesh interval in height. Data at the two ends of the reflected shock were obtained using a shock polar analysis. At the wedge corner, the two velocity components vanish. The remaining solution components at the corner and the full solution in the interior are determined by interpolation. One arbitrary parameter is used in the initialization, which is the oblateness of the bubble, defined as the ratio of the distance of the reflection point from the corner to the distance of the bow shock to the corner (the ratio of the lengths of the segments BC and AB in Fig. 2.1). The initial oblateness was taken from experimental data, but it can be determined approximately by a preliminary calculation because it is constrained to lie in a bounded interval by theoretical considerations. In fact the computational results are quite insensitive to its value. The initialization algorithm can be regarded as an approximate solution of the two-dimensional Riemann problem in a special case.

In Fig. 2.2 the contours of constant density and constant entropy that were obtained numerically are shown. In Fig. 2.3 the density distribution along the wall obtained from the calculation is superimposed on the experimental data.

### 3. Mach Reflection of Shock Waves

The intersection point of discontinuity curves will be called a node. In a neighborhood of a node the curves are approximated by straight lines separating wedge shaped regions. In analogy with the one-dimensional Riemann problem, we define a two-dimensional Riemann problem to be an initial value problem for a two-dimensional conservation law having data that is either a constant state or a simple rarefaction wave in each of a finite number of wedges. Such problems have been studied for scalar conservation laws [8,5,6], but only special solutions are known for systems of conservation laws. As with the solution of a one-dimensional Riemann problem, the solution of a two-dimensional Riemann problem will evolve into a more complicated configuration containing several elementary waves. Thus in front tracking we must solve a subcase of the full Riemann problem: determining the velocity and states associated with the one specific elementary wave (node) being tracked. We shall report in this section on a method for tracking the Mach node. Its propagation is a fully two-dimensional problem that cannot be be reduced to one-dimensional problems by spatial operator splitting. In fact successive solutions to one-dimensional problems still play a key role in solution of the Mach node, but their composition is governed by the geometry of the waves entering the node and not by an orthogonal set of coordinate axes.

Consider a planar shock moving down a shock tube and incident on a wedge with a small angle (see Fig. 3.1). In contrast with the regular reflection case we obtain a Mach reflection. The point where the incident (I) and the reflected shock (R) meet (the "Mach node") has lifted off the wall and is connected to the wall by a nearly straight shock called the Mach stem (M). Behind the Mach node a contact discontinuity (C) is formed between the reflected shock and the Mach stem.

The corresponding two-dimensional Riemann problem is shown in Fig. 3.2. We move to a frame where the node is at rest and denote the states as indicated. The contact discontinuity has a jump in density and tangential velocity across it. Each state is given by the two components of velocity $\vec{q}$, the density $\rho$, and the pressure $p$. Given the state in one sector, the Rankine-Hugoniot conditions determine a one parameter family of states that can occur across a shock or contact discontinuity in a neighboring sector. These conditions may be written as follows [2, pp. 301-302 and 329]:

$$\frac{p_i - p_j}{\rho_j} = \vec{q}_j \cdot (\vec{q}_j - \vec{q}_i)$$

for

$$(i,j) \in \{(0,1), (1,0), (1,2), (2,1), (0,3), (3,0)\}, \tag{3.1}$$

$$\frac{p_i}{p_j} = \frac{\mu^2 - \frac{\rho_i}{\rho_j}}{\mu^2 \frac{\rho_i}{\rho_j} - 1}$$

for

$$(i,j) \in \{(1,0), (3,0), (2,1)\}, \tag{3.2}$$

$$\vec{q}_2 \times \vec{q}_3 = 0, \tag{3.3}$$

and

$$p_2 = p_3. \tag{3.4}$$

Here $p$ is the pressure, $\vec{q}$ is the fluid velocity, and $\mu^2 = \frac{\gamma - 1}{\gamma + 1}$, where $\gamma$ is the polytropic gas constant. Relations (3.1) and (3.2) are the Rankine-Hugoniot conditions for a shock, while relations (3.3) and (3.4) express the existence of a contact discontinuity.

There are eleven equations in sixteen unknowns. From the point of view of an experimentalist, the initial conditions in a shock tube experiment, viz. four parameters specifying the density and pressure scales and the strength and orientation of the incident shock, are not sufficient to determine the solution. For this reason it could be stated that there is a missing equation for the Mach interaction. However, this missing equation is only an absence of an analytic or closed form solution to give the node trajectory on the basis of equations (3.1)–(3.4), and does not indicate an incompleteness of the Cauchy

problem for the Euler equations. From a mathematical point of view there is no missing equation, since the solution from the previous time step provides complete Cauchy data.

The front tracking problem is to obtain a complete Riemann problem solution for given Cauchy data and to select the Mach node out of that solution. As formulated this problem is too difficult. Hence we proceed with equations (3.1)−(3.4). With sixteen state variables and eleven equations at the node, we see that the Mach node lies in a five-dimensional manifold within the space of states $u_0$, $u_1$, $u_2$, and $u_3$. In lieu of solving a full two-dimensional Riemann problem to select a point in this manifold, five of the above sixteen parameters are selected to specify the node. Using physical intuition, we selected five parameters from the complete set of Cauchy data [4]. Based on numerical evidence, we believe this method does a satisfactory job of picking out the Mach node from the waves emanating from the complete solution of a two-dimensional Riemann problem that is close to a Mach node.

We compared the numerical solution for the nonsteady Mach reflection with experiments in [3]. In the experiment an incident shock with a Mach number 2.03 impinges on a wedge with angle 27°. The calculations were initialized just after the Mach configuration has appeared. The reflected shock and the Mach stem then enclose a region of about one mesh interval in height. After the bubble enclosed a region several mesh intervals in height the solution had settled down to its self-similar form as seen in the experiment. In Fig. 3.3 the constant density contours are shown. In Fig. 3.4 the wall density distribution obtained in our calculation is superimposed on the experimental data. However the calculation is preliminary in two respects. The present form of the algorithm for the propagation of the Mach triple point seems to be stable only when the initial oblateness is chosen near the experimentally determined value. Moreover the algorithm for the propagation of the point of intersection of the contact discontinuity with the wall is discernably unstable: the fluid velocity at this point is very sensitive to the pressure upstream, so the end of the contact tends to curl up. The causes of these instabilities have not yet been determined.

## 4. The Classification of Two Dimensional Elementary Waves

In this section we classify the elementary waves for two-dimensional gas dynamics. Front tracking employs a normal and a tangential operator splitting at jump surfaces and a solution of two-dimensional Riemann problems at the point singularities formed by the intersection of jump surfaces ("nodes"). An example of such a two-dimensional Riemann problem was described in the previous section. For two-dimensional compressible gas dynamics there are only a small number of such nodes.

We make some general assumptions that idealize the problem, but which we believe apply to a generic set of possible point singularities formed by the meeting of jump surfaces and centered rarefaction waves. Then we refine these general assumptions into a precise mathematical formulation, and using the latter, derive a classification scheme for

the allowed point singularities.

Excluded from this classification scheme are point singularities formed by centered waves (implosions) and points in a neighborhood of which the solution is not piecewise smooth.

**Definition 4.1** A *pressure wave* is a shock wave or a centered rarefaction wave. A *wave* is either a pressure wave or a contact discontinuity. A *node* is the point singularity formed by the intersection of waves. A rarefaction wave centered at a node is called an *incoming (forward facing) rarefaction wave* if its straight line $C^+$ or $C^-$ characteristics, in the frame in which the node is stationary, point towards the node. A shock wave emanating from a node is said to be an *incoming shock wave* if, in the stationary frame, it turns the flow towards the node. Similarly we define an *outgoing rarefaction wave* and an *outgoing shock wave*. We observe that every pressure wave at a node is either incoming or outgoing.

**Assumption 4.2.** We assume our solution $u$ to be an *elementary wave*, which, in general terms, satisfies the following:

4.2.1 $u$ is a stationary solution of the Euler equations for a polytropic gas: $u_t = 0$ and $\nabla \cdot \vec{f}(u) = 0$;

4.2.2 $u$ has the form

$$u = u_j \text{ for } \theta_{j-1} < \theta < \theta_j, \quad j = 1, \ldots, n$$

where $\theta_n = \theta_0 + 2\pi$ and each $u_j$ is constant or a centered rarefaction wave;

4.2.3 the only jumps allowed in $u$ are shock waves and contact discontinuities;

4.2.4 $u$ is generic;

4.2.5 $u$ is an entropy increasing solution, with

$$u = \lim_{\nu \to 0} \tilde{u},$$

where $\tilde{u}$ is a solution of the Navier-Stokes equations with viscosity $\nu$.

First consider all possible elementary waves containing only contact discontinuities. If one of the sectors has a 180° opening, there can be a nonzero discontinuity in the tangential velocity across its boundary. All other contact discontinuities contain only temperature jumps. It is possible for any number of them to occur, and at any set of angles.

From now on assume that the elementary wave contains at least one pressure wave. Assumptions 4.2.4 and 4.2.5 are not written in mathematical terms, so we formulate the ideas that they express in a manner that we can use in our analysis.

4.2.4a No incoming rarefaction waves are allowed. There can be at most two incoming pressure waves, which are necessarily shock waves. If the flow on the ahead side (the side with the lower pressure) of an incoming shock wave is adjacent to a contact discontinuity that is within 90° of the incoming shock,

only this one incoming shock wave is allowed.

4.2.5a No sectors of zero velocity bounded by contact discontinuities ("embedded wedges") are allowed.

Observe that the elementary waves that satisfy 4.2.1 through 4.2.4 but fail to satisfy 4.2.5a can be interpreted as solutions of an initial/boundary-value problem if the contact discontinuities bounding the embedded wedge are replaced with reflecting walls.

In [4] we have determined that an elementary wave containing a pressure wave is restricted by the following. There is a unique streamline through the node. There is either one or two incoming shock waves. There is no more than one outgoing pressure wave on each side of the streamline leaving the node, only one of which may be a rarefaction. The streamline is a contact discontinuity if there are two outgoing pressure waves. This then leaves us with only a small number of possible elementary waves, which we shall determine presently.

**Theorem 4.3.** Under Assumptions 4.2 an elementary wave containing at least one pressure wave is one of the following types, as specified in detail below: cross, overtake, Mach, diffraction, and transmission.

What follows is an explanation of the types together with figures. For a proof of the theorem see [4].

**Diffraction.** The diffraction of a shock impinging on a contact discontinuity, causing a reflected and a transmitted shock is a possible solution, see Fig 4.1.a. We show in Fig. 4.1.b how the solution may be constructed by drawing the appropriate shock polars in the $\theta, p$ plane.

**Diffraction.** The configuration in Fig. 4.2.a is an elementary wave. It is as in Fig. 4.1.a, but with a rarefaction wave in place of a reflected shock. The solution is constructed using shock polars, as indicated in Fig. 4.2.b.

**Transmission.** A shock impinging on a contact discontinuity and causing a transmitted shock but no reflected wave (see Fig. 4.3.a) is a possible elementary wave; see the corresponding shock polar in Fig. 4.3.b.

**Mach node.** A direct Mach reflection, where the incident shock breaks into two shocks, the reflected shock and the Mach stem, is a possible elementary wave; see Fig 4.4.a. The solution is found using shock polars as in Fig 4.4.b.

**Overtake.** It is possible to have two incoming shocks overtake each other and give rise to two outgoing shocks separated by a contact discontinuity (see Fig. 4.5.a). The solution may be constructed using shock polars as in Fig. 4.5.b. The special case where one of the two incoming shocks has zero strength coincides with the Mach node case.

**Overtake.** One shock may overtake the other, resulting in a reflected rarefaction and a transmitted shock (see Fig 4.6.a). The solution is found using shock polars as in Fig. 4.6.b. Note that for the same parameters of the two incident shocks, both this case and the previous case are possible.

**Cross or Mach node.** Two incident shocks colliding to form two reflected shocks separated by a streamline (see Fig. 4.7.a) is a possible elementary wave. The solution is found using shock polars as in Fig. 4.7.b. The special case of one incident shock having zero strength gives the direct Mach node. A single shock wave in the outgoing sector defines the inverted Mach node; this interaction is a limit of the cross case above, where the shock between region 2 and 3 in Fig. 4.7.a reduces to zero.

## 5. Some Problems and Conjectures Concerning Riemann Problems

In this section we drop the restriction to two dimensions and to gas dynamics, but we retain the terminology of Sec. 4. Recall that in steady supersonic two-dimensional gas dynamics, where the direction of the flow defines a timelike direction, the equations can be reduced to the form of a one-dimensional time-dependent system of conservation laws. Then the two-dimensional elementary waves viewed in the stationary frame are Riemann problems for a distinct but related one-dimensional system. Similarly Riemann problem solutions in $n-1$ space dimensions are qualitatively similar to elementary waves in $n$ dimensions.

We list some problems of general interest in this area.

1. The possible $n$-dimensional elementary waves for a system of conservation laws could be classified. The elementary waves in two-dimensional polytropic gas dynamics were classified in the previous section.

2. Let the incoming wave operator be the solution operator bringing two or more elementary waves to a single point and thereby defining a Riemann problem. The incoming wave operator also acts on single elementary waves by mapping to the configuration at a time of bifurcation, or dynamic instability; this also defines a Riemann problem. The range of this operator is limited to the possible mergers or bifurcations of the elementary waves found in the classification above. Can this range be categorized?

3. The outgoing wave operator gives the possible elementary waves that may occur in the solution of the Riemann problems in the range of the incoming wave operator. We pose the question of existence of solutions for this restricted set of data. Are solutions piecewise smooth, so that there is a finite number of outgoing elementary waves? The answer depends on the order of the system, the dimension of space, and the convexity or number of inflection points in the flux function, as examples [6] show and analogies [7] suggest.

4. A logical scattering matrix $S$ is a map from sets of incoming wave types to sets of outgoing wave types as labeled by the solutions to problem 1. It decides which types of incoming waves produce which types of outgoing waves. In the language of quantum mechanics, the problem here is to classify the possible $S$ matrix graphs. Let us consider this problem from the point of view of two-dimensional gas dynamics. We restrict attention to Riemann data contained in the range of the incoming wave operator as defined in 2 above. Under such restriction, the waves will be said to be in incoming

order. The allowed nodes of Sec. 4 provide interchange of wave order to an outgoing order. In general, however, the interchange of wave order produces three outgoing waves from two incoming waves and need not reduce the total number of wave pairs that fail to be in outgoing order. On this basis, we expect that even simple incoming configurations could produce complicated outgoing wave interactions. It is possible that the complication (e.g. the number of nodes), while not bounded *a priori*, is still finite. In fact the wave interactions typically decrease the Mach number of the flow, and may give rise to a subsonic region, inside of which no pressure waves can occur. Related to this possibility is the occurrence of nodes with only two outgoing waves. Such nodes allow the interchange of wave order with a reduction in the total number of pairs that are out of order.

5. Uniqueness is an open problem. Well known problems of nonuniqueness are not understood on a fundamental level. For example, consider an incident shock hitting a wedge, resulting in either a regular reflection or Mach reflection. There are regions where both solutions are possible. By introducing additional physical effects such as viscosity, with a resulting boundary layer, or surface roughness on a certain length scale, this overlapping region of nonuniqueness might disappear.

6. Extended or nonlocal Riemann problems may be considered, where the restriction of constancy in sectors between the waves is replaced by linear or higher order data. This has been implemented for one dimension in the normal propagation of the front [1].

7. Lower order terms in the equations and new waves in the Riemann solution may be caused by geometrical effects and by external sources.

8. The geometry in the large defined on the state space by the flux function needs to be understood. For gas dynamics the qualitative behavior of solutions may be studied by considering the acoustic waves of the linearized problem. But this is not the case for all hyperbolic conservation laws, and new families of waves may be possible in the large. The topology defined by the critical points of the flux function $\vec{f}$ in (1.1) is important here. A critical point is a point in state space where the gradient $A = \nabla \vec{f}$ has coinciding eigenvalues. At the critical points equation (1.1) is no longer strictly hyperbolic; moreover $A$ can fail to have a complete set of eigenvectors. Such a loss of strict hyperbolicity is not pathological in applications, and the mathematical consequences of this fact have not been developed. An extension of this phenomena are the problems for which the equations in different regions of state space change type. The applications are of general interest: oil reservoir simulation, transonic flow in gas dynamics, chemically reacting flows, and nonlinear elasticity.

**Acknowledgements**

We thank Jonathan Goodman for helpful discussions.

**References**

1. I-L. Chern, J. Glimm, O. McBryan, B. Plohr, and S. Yaniv, "Front Tracking for Gas Dynamics," Submitted to J. Comp. Phys., 1984.
2. R. Courant and K. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience, New York, 1948.
3. R. Deschambault and I. Glass, "An Update On Non-Stationary Oblique Shock-Wave Reflections: Actual Isopycnics and Numerical Experiments," *J. Fluid Mech.*, vol. 131, pp. 27-57, 1983.
4. J. Glimm, C Klingenberg, O. McBryan, B. Plohr, D. Sharp, and S. Yaniv, "Front Tracking and Two Dimensional Riemann Problems," *Adv. Appl. Math.*, 1984. To appear.
5. J. Guckenheimer, "Shocks and Rarefactions in Two Space Dimensions," *Arch. Rational Mech. Anal.*, vol. 59, pp. 281-291, 1975.
6. B. Lindquist, "The Two Dimensional Scalar Riemann Problem," New York University Preprint, 1984.
7. J. Rauch and M. Reed, "Jump Discontinuities of Semilinear, Strictly Hyperbolic Systems in Two Variables," *Comm. Math. Phys.*, vol. 81, pp. 203-227, 1981.
8. D. Wagner, "The Riemann Problem in Two Space Dimensions for a Single Conservation Law," *J. Math. Anal.*, vol. 14, pp. 534-559, SIAM J. Math. Anal., 1983.

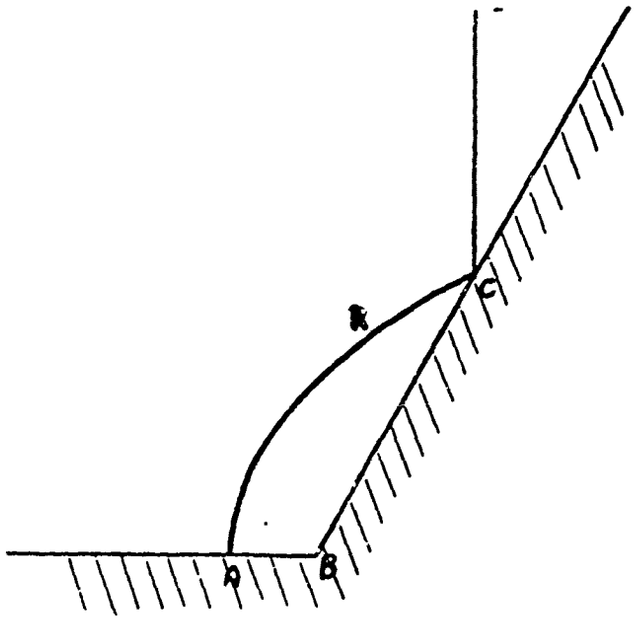



**Fig. 2.1.** Regular reflection of a shock wave by a wedge. A vertical shock *I* has struck a 63° wedge from the left, causing a reflected shock *R*, which forms a bowshock in front of the wedge.

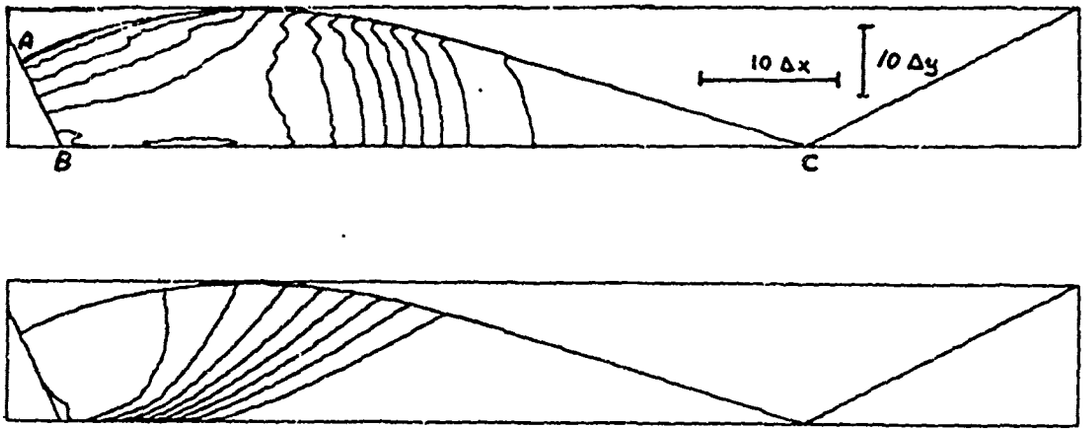



**Fig. 2.2.** The numerical simulation of a regular reflection, where the incident shock has Mach number 2.05 and the wedge angle is 63.4°. The calculation was performed on a 80 by 20 grid. The top picture shows the lines of constant density inside the bubble formed by the reflected shock. The bottom picture shows the lines of constant entropy. They should coincide with the integral curves for the self-similar velocity field. Theoretical arguments in the text suggest that these integral curves all terminate at the corner.

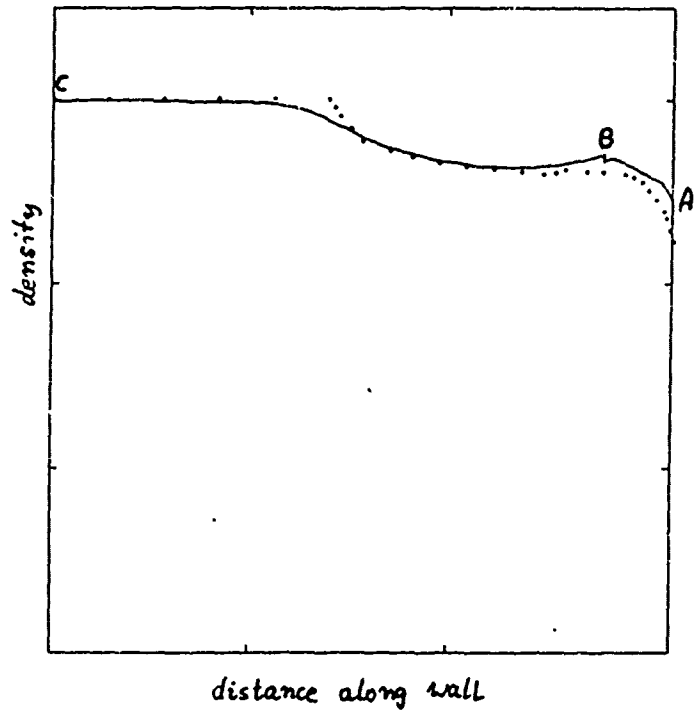



**Fig. 2.3.** The computed (solid line) density distribution along the wall for the regular reflection run compared to the data obtained in the experiments of Deschambault and Glass (dots).

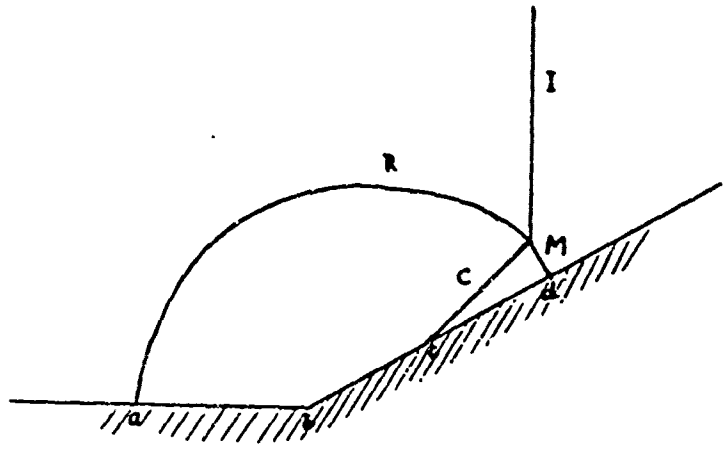



**Fig. 3.1.** Mach reflection of a shock wave by a wedge. A vertical shock *I* has struck a 27° wedge from the left. The point where the incident shock *I* and reflected shock *R* meet is connected to the wall by a shock called a Mach stem (*M*). Behind the triple point, where the three shocks meet, a contact discontinuity *C* is formed between the reflected shock and the Mach stem.

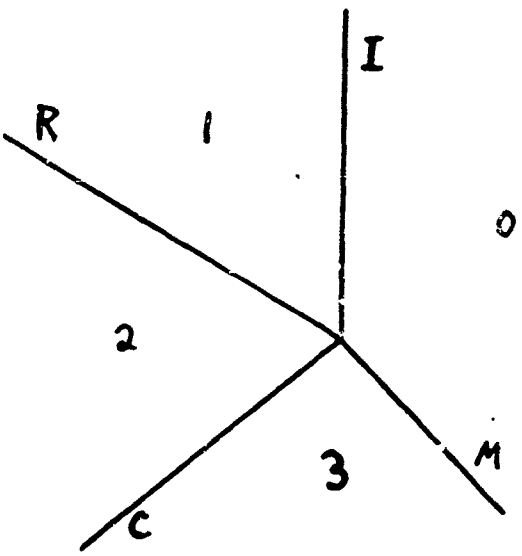



**Fig. 3.2.** The Riemann problem corresponding to the triple point obtained in a Mach reflection. We assume the shocks and contact discontinuities are straight lines (labeled as in Fig. 3.1) with constant states 0 through 3 in the wedges.

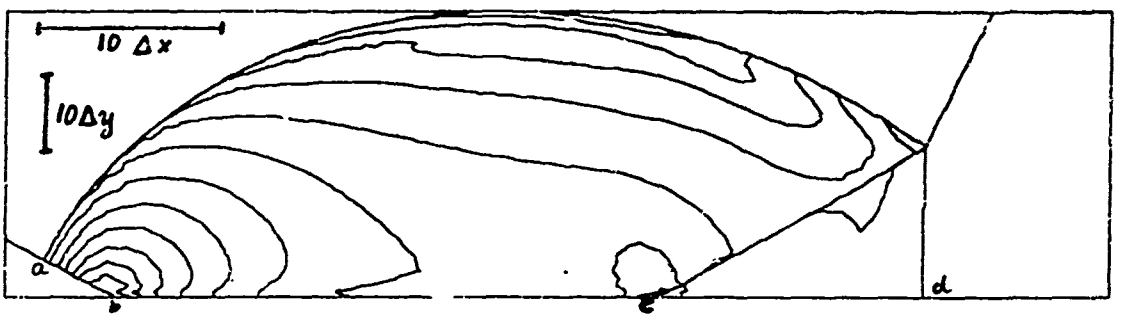



**Fig. 3.3.** The numerical simulation of a Mach reflection, where the incident shock has Mach number 2.03 and the wedge angle is 27°. Inside the bubble formed by the reflected shock we show the calculated lines of constant density. The calculations were performed on a 60 by 40 grid.

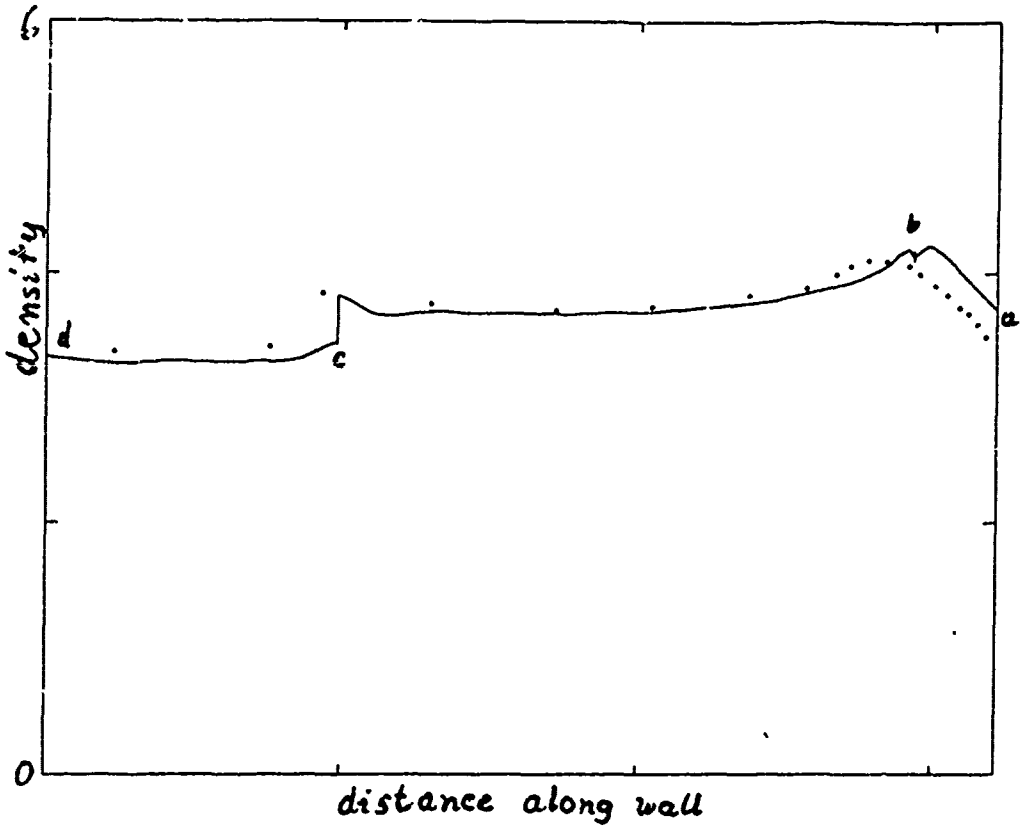



**Fig. 3.4.** The density distribution along the wall of the Mach reflection run (solid line) shown superimposed on the data found experimentally by Deschambault and Glass (dots).

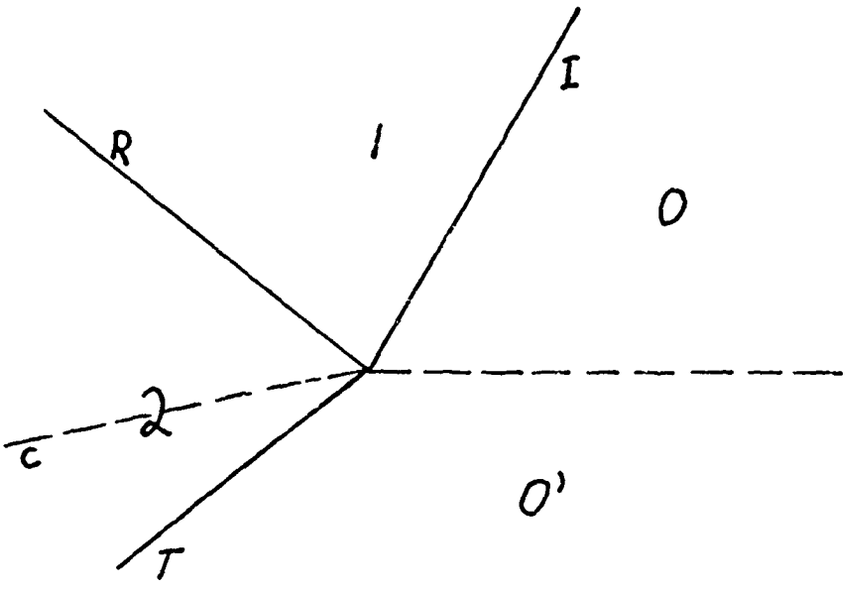



**Fig. 4.1.a. Diffraction. A shock *I* can diffract through a contact discontinuity *C* to cause a reflected shock *R* and a transmitted shock *T*.**

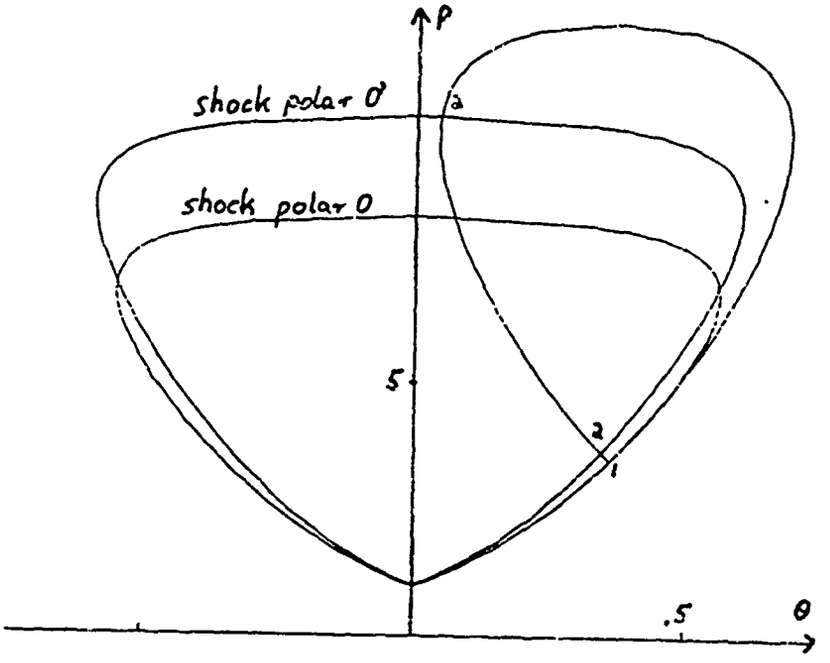



**Fig. 4.1.b. The shock polars corresponding to Fig. 4.1.a. The Mach number of state 0 is 2.7 and that of state 0' is 3. The shock strength of *I* is 3.**

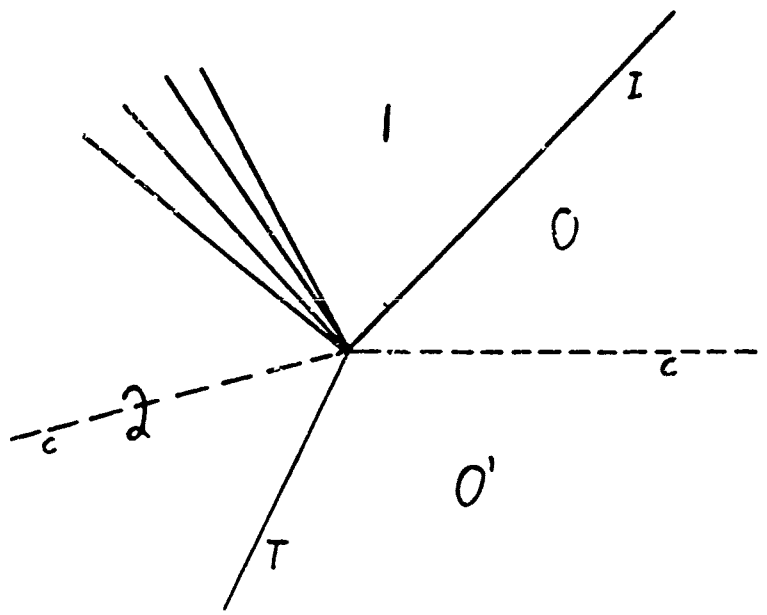



**Fig. 4.2.a. Diffraction.** The diffraction of a shock $I$ by a contact discontinuity $C$ can cause a reflected rarefaction wave $R$ and a transmitted shock $T$.

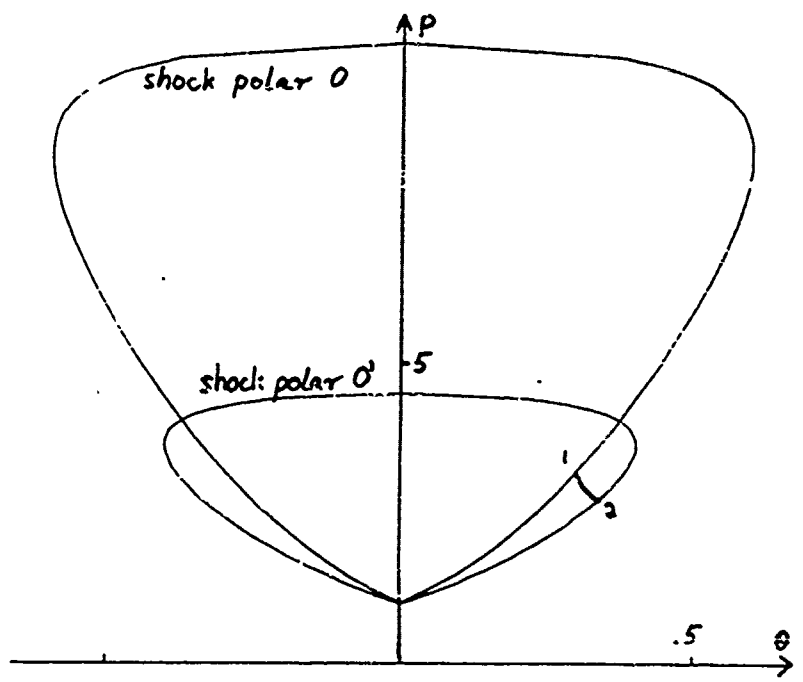



**Fig. 4.2.b.** The shock polars corresponding to Fig. 4.2.a. The Mach number of state 0 is 3 and that of state 0' is 2. The image of a $\Gamma-$characteristic in this $p$ , $\theta$ plane is drawn, connecting the states $p = 3$ on one shock polar to $p = 2.8$ on the other.

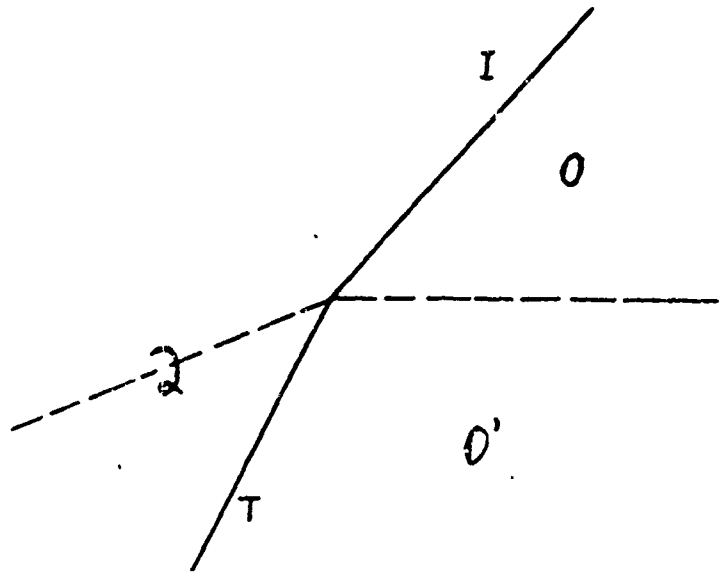



**Fig. 4.3.a. Transmission. A shock *I* incident on a contact discontinuity *C* causes a transmitted shock *T* but no reflected wave is possible.**

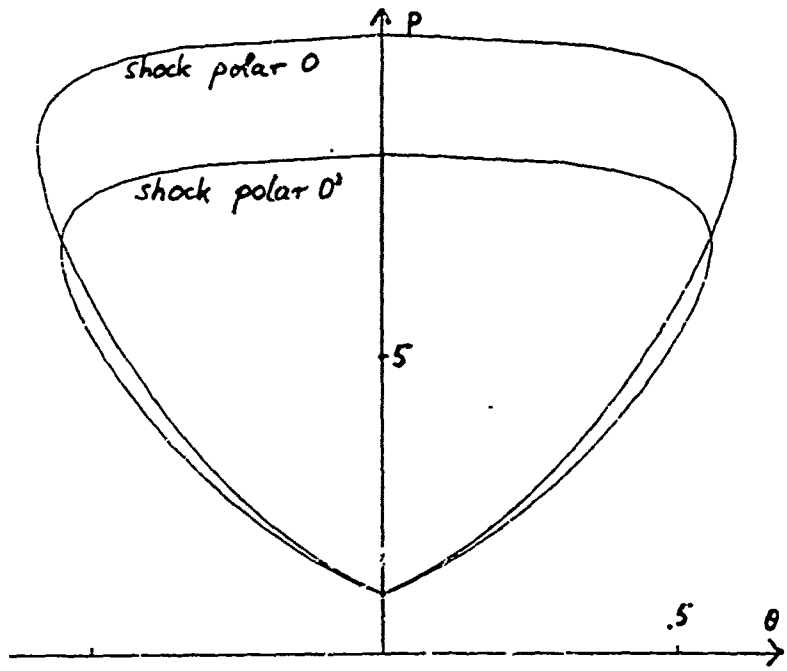



**Fig. 4.3.b. The shock polars corresponding to Fig. 4.3.a are shown. The Mach number of state 0 is 3 and that of state 0' is 2.7.**

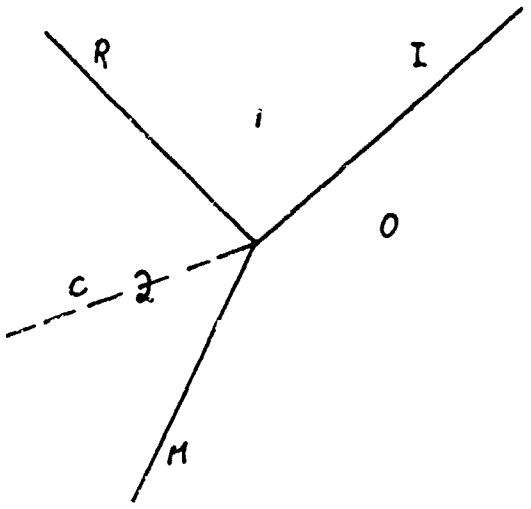



**Fig. 4.4.a.** Mach node. Direct Mach reflection is shown, with the incident shock *I* breaking into a reflected shock *R* and a Mach stem *M* separated by a contact discontinuity *C*.

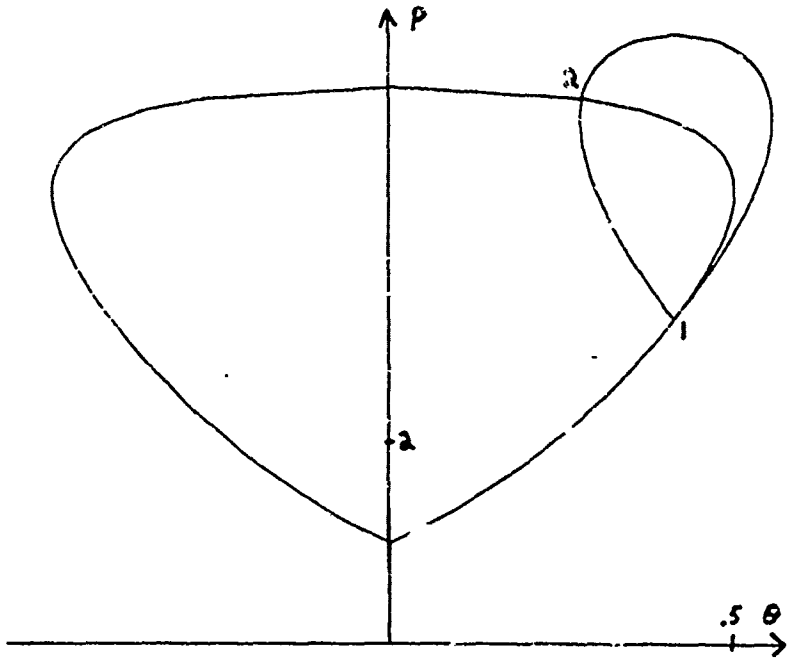



**Fig. 4.4.b.** The shock polars corresponding to Fig. 4.4a. The Mach number of state 0 is 2.2 and the shock strength of *I* is 3.2.

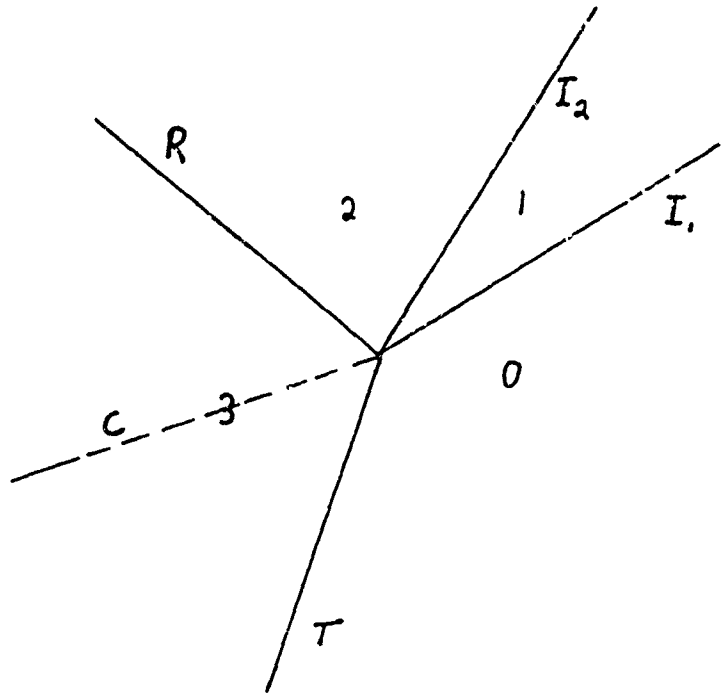



**Fig. 4.5.a.** Overtake. It is possible to have one incoming shock $I_1$ overtake another $I_2$ to cause a reflected shock $R$ and a transmitted shock $T$ with a contact discontinuity $C$ between them.

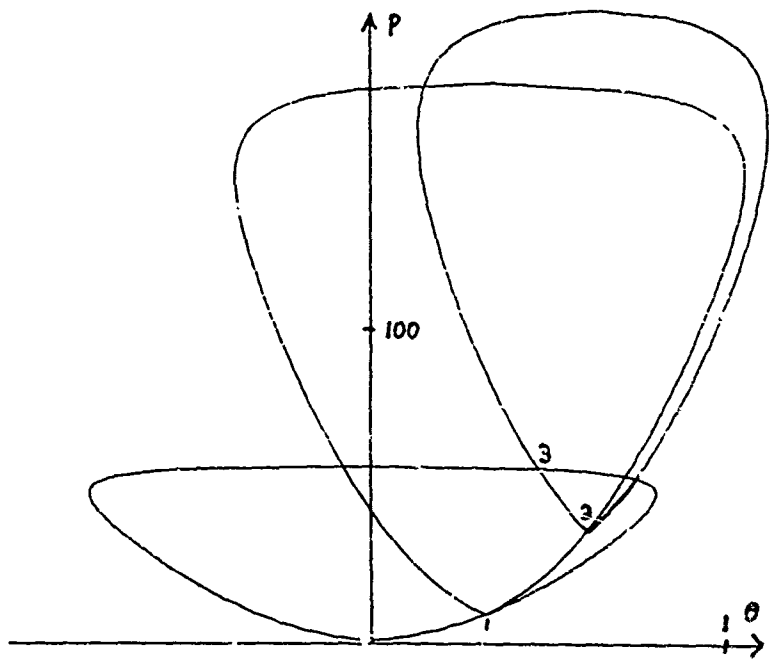



**Fig. 4.5.b.** The shock polars corresponding to Fig. 4.5.a. The Mach number of state 0 is 7, the shock strength of the incident shock $I_1$ is 9.5, and the shock strength of $I_2$ is 3.9.

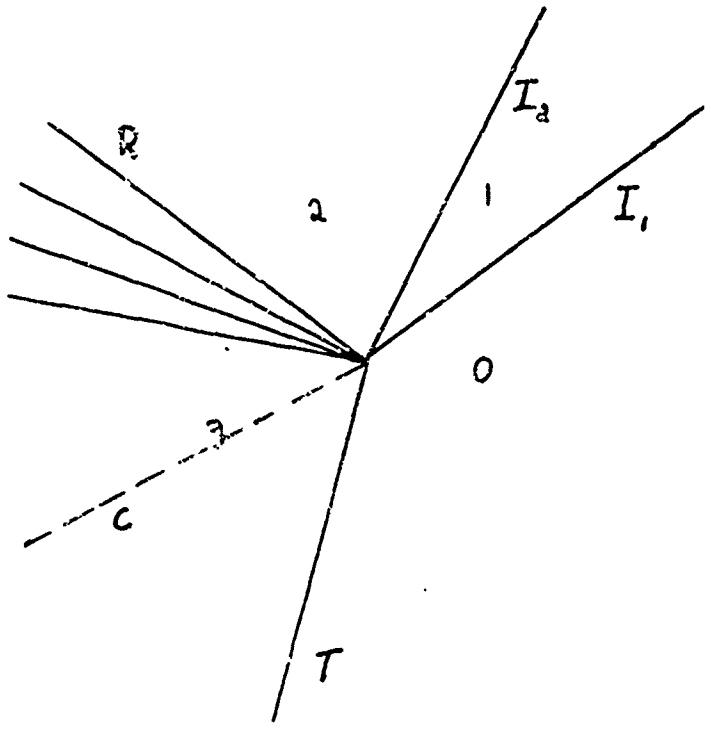



**Fig. 4.6.a.** Overtake. It is possible for one incoming shock $I_1$ to overtake another $I_2$ t- cause a reflected rarefaction wave and a transmitted shock wave separated by a contact discontinuity.

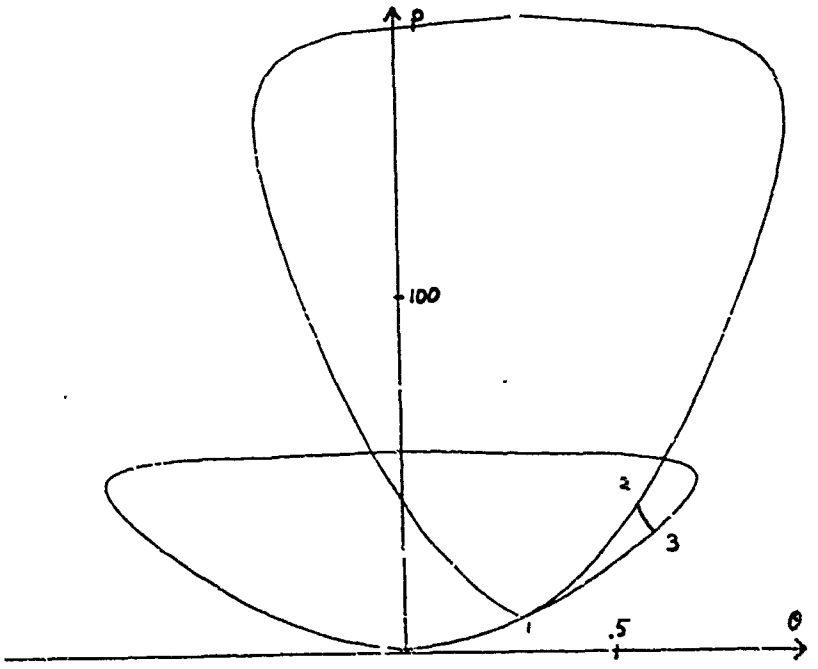



**Fig. 4.6.b.** The shock polars corresponding to Fig. 4.6.a. The Mach number of state 0 is 7, the shock strength of the incident shock $I_1$ is 9.5, and the shock strength of $I_2$ is 3.9.

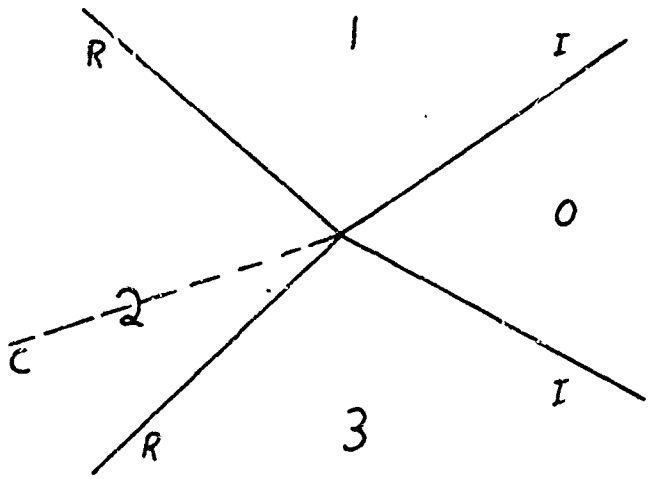



**Fig. 4.7.a.** Cross. Two may shocks collide and cause two reflected shocks separated by a contact discontinuity.

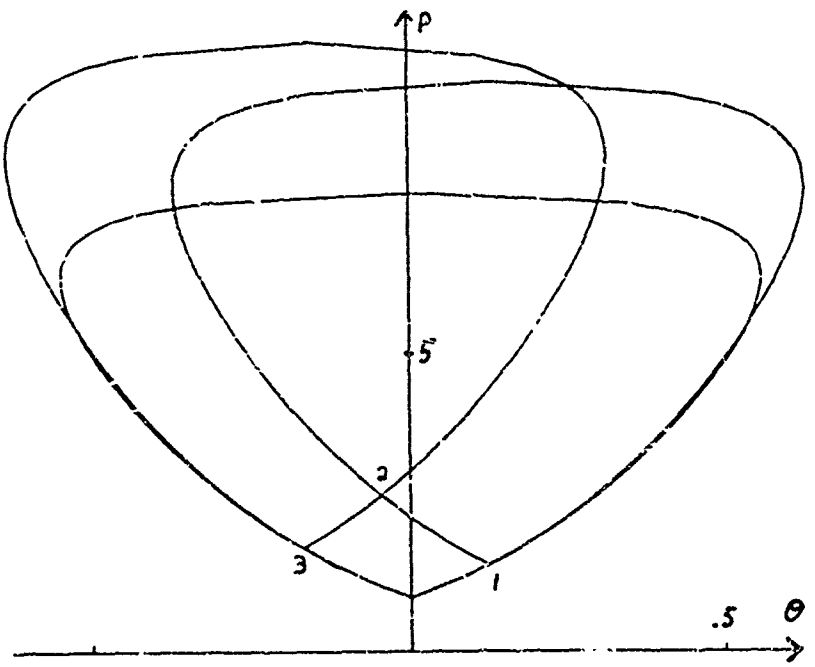



**Fig. 4.7.b.** The shock polars corresponding to Fig. 4.7.a. The Mach number of state 0 is 2.7 and the shock strengths of the incident shocks are 1.6 and 1.9.

# MIGRATION OF THE GAS GLOBE FROM UNDERWATER EXPLOSIONS: THE EFFECTS OF DRAG AND RADIATIVE ENERGY LOSS

K.C. Heaton
Weapon Systems Section, Armaments Division
Defence Research Establishment Valcartier
P.O. Box 8800, Courcelette
Quebec, G0A 1R0

ABSTRACT. In this work, the problem of the motion of the bubble from an underwater explosion is considered. Previous derivations of the equations of motion are summarized. It is shown how a Lagrangian method can be utilised to obtain the equations of motion, and how generalised dissipative forces can be formally incorporated.

The generalised dissipative force for energy loss by radiation of sound is obtained from first principles, and found to be approximately proportional to the cube of the rate of change of the bubble radius with respect to time.

The algorithm by which the equations of motion were solved numerically is briefly discussed, and some results of the computations presented.

It is found that the approximate formula for energy loss by radiation of sound by a bubble in the absence of vertical motion (derived by Herring) gives results similar to those obtained using the more exact formulation developed in this work. It is also found that the rate of energy loss is affected slightly by processes analogous to radiation reaction in electromagnetic theory.

Results are also presented for the cases for which energy is lost by drag as well as by radiation of sound, and for the case for which the effects of free and rigid surfaces on the bubble motion, but not energy loss, are considered.

I. INTRODUCTION. One of the more prominent phenomena associated with underwater detonations, after the initial shock produced by the explosion itself, is the formation of a bubble of gaseous explosion products. This bubble rises toward the surface of the water, responding as it does so to changes in the external pressure distribution with radial oscillatory motion during the course of which it loses some of its energy through the emission of sonic pulses. Although the bubble may have been initially spherical, the effect of its upward motion distorts it into a non-spherical shape, which becomes most pronounced in the neighbourhood of the minimum bubble radius. The alteration in the bubble's shape further affects both the pulsations and the upward translational motion of the bubble. By means of finite element techniques, the equations of motion of the bubble can be solved, taking into account the effects of the changing shape of the bubble, although the amount of computing time required by this method limits its utility. Finite element methods have a further disadvantage in that physical insights into the systems considered are rather more difficult to come by than might otherwise have been the case.

Herring (1942) and others (eg. Taylor (1942) and Shiffman and Friedman (1944)) have treated the problem of the motion of the bubble by considering it to be a perfect sphere throughout its entire motion. This treatment yields values for the periods of radial pulsations of the bubble which are in good agreement with experimental data, but predicts a much more rapid movement toward the surface than that which is actually observed. This arises because the largest upward velocities of the bubble occur at those times when the bubble is near its minimum radius; it is precisely then that the largest departures from sphericity have been observed. Hicks (1972) added a drag term to the equations of motion, chosing the drag coefficient such that the predicted distance travelled by the bubble was in agreement with that which had been observed.

In this paper, previous formulations of the problem of the motion of an underwater bubble are discussed. The ways in which the effects of drag, as well as those of the energy radiated by a bubble during its motion, can be incorporated into its equations of motion are considered. A more rigorous derivation of the dissipation function for the bubble is presented. The equations of motion incorporating the effects of gravity and loss of energy by drag and the radiation of sound by the bubble, are derived using a Lagrangian formalism. The algorithm by which various forms of the equations of motion were solved is briefly discussed. Some computational results are presented, along with suggestions for the further extension of this work.

## II. EQUATIONS OF MOTION FOR A SPHERICAL BUBBLE.

II.1 Review of Previous Work. Taylor (1942) derived equations describing the motion of a spherical bubble of gas undergoing both radial pulsations and translational motion toward the water surface. These are:

$$2\pi\rho a^3\left(\frac{da}{dt}\right)^2 + \frac{\pi\rho a^3 U^2}{3} + \frac{4\pi\rho a^3 gz}{3} = Y - E(a), \qquad [1]$$

$$U = -\frac{dz}{dt} = \frac{2g}{a^3}\int_0^t a^3 dt \qquad [2]$$

where a is the radius of the bubble as a function of the time t, U its upward velocity, z the position of the bubble below the pressure datum (i.e. below the zero pressure level), E(a) the internal energy of the gas comprising the bubble, $\rho$ the density of the water, g the gravitational acceleration, and Y the total energy of the bubble. In Taylor's formulation, there was no mechanism included for energy loss, and hence Y was taken to be a constant. For TNT explosions, it has been found (Herring 1942) that approximately 50% of the total explosion energy is retained by the bubble; in that case:

$$Y = (1.85 \times 10^{10})M \qquad [3]$$

where Y is measured in ergs, and M, the original mass of the explosive charge, is given in grams. If one assumes that the gaseous explosion products obey the ideal gas law, then the internal pressure, P, is given by:

$$P = k(\rho_g)^{\gamma} \qquad [4]$$

where $\rho_g$ is the density of the explosion products and $\gamma$ the ratio of specific heats. Assuming that the entire mass, M, of the explosive has been converted to gas:

$$\rho_g = \frac{M}{\frac{4\pi a^3}{3}}, \qquad [5]$$

and hence:

$$E(a) = \int_a^{\infty} P dV \qquad [6]$$

$$= \frac{kM^{\gamma} a^{-3(\gamma-1)}}{(\gamma-1)\left(\frac{4\pi}{3}\right)^{\gamma-1}}$$

where $dV = 4\pi a^2 da$. Taylor, using the work of Jones, set:

$$k = 7.83 \times 10^9 \quad ,$$
$$\gamma = 1.25 \qquad [7]$$

for TNT where, in eq. [6], E(a) is measured in ergs and $\rho_g$ in gm/cm$^3$.

Herring (1942) derived an equation of motion for the bubble, neglecting the effects of gravity and translational motion U, but incorporating the effects of the loss of energy by the radiation of sound. This is:

$$a\frac{d^2a}{dt^2} + \frac{3}{2}\left(\frac{da}{dt}\right)^2 - \frac{1}{ac}\frac{d}{dt}\left[a^2\left(\frac{da}{dt}\right)^2\right] = \frac{a}{\rho c}\frac{dPa}{dt}\left(1 - \frac{1}{c}\frac{da}{dt}\right) - \int_a^{\infty}\frac{dP}{\rho} \qquad [8]$$

where Pa is the gas pressure in the buble when its radius is a, c is the speed of sound in the water, P the hydrostatic pressure, and all other variables are as defined previously. Herring has shown that the energy loss, which is contained in the term $\frac{a}{\rho c}\frac{dPa}{dt}\left(1-\frac{1}{c}\frac{da}{dt}\right)$, is significant only during the times at which the bubble is near its minimum radius. Using the values of k and $\gamma$ from eq. [7] for TNT, one finds that the loss of energy, $\Delta Y$, over one cycle of radial pulsation of the bubble, is approximately:

$$\frac{\Delta Y}{Y} = 1.87\frac{P_{rad}^{\frac{1}{2}}}{c\rho^{\frac{1}{2}}} \qquad [9]$$

where $P_{rad}$ is the gas pressure at the minimum radius of the bubble during that cycle.

Finally, the presence of the water surface and the ocean bottom has an effect on the bubble motion. Using the theory of images, Shiffman and Friedman (1944) have modified Taylor's equations of motion for the bubble, thusly:

$$2\pi\rho a^3(1 + f_o)\left(\frac{da}{dt}\right)^2 - 4\pi\rho a^3 f_1 U \frac{da}{dt} + \frac{\pi}{3}\rho a^3(1 + 3f_2)U^2$$
$$+ \frac{4\pi}{3}\rho a^3 gz = Y - E(a) \quad , \qquad [10]$$

$$\frac{d}{dt}\left[\frac{1}{3}a^3(1 + 3f_2)U - 2a^3 f_1 \frac{da}{dt}\right] = -\frac{a^4}{2b^2}\left[\frac{df_o}{d\alpha}\left(\frac{da}{dt}\right)^2\right.$$
$$\left. + \frac{1}{2}\frac{df_2}{d\alpha}U^2 - 2\frac{df_1}{d\alpha}U\frac{da}{dt}\right] + \frac{2}{3}ga^3 \qquad [11]$$

where $f_o$, $f_1$, $f_2$ are infinite series, b is the height of the centre of the bubble above the ocean floor, and $\alpha$ is given by:

$$\alpha = \frac{a}{2b} \qquad [12]$$

(i.e. the ratio of the bubble radius to twice the distance from the bubble centre to the ocean floor). The leading terms of the series $f_o$, $f_1$, $f_2$ are:

$$f_o = \alpha + \frac{1}{2}\alpha^4,$$
$$f_1 = \frac{1}{2}\alpha^2 + \frac{1}{2}\alpha^5, \qquad [13]$$
$$f_2 = \frac{1}{2}\alpha^3 + \frac{1}{2}\alpha^6 \ .$$

The preceeding suggests an approximate method for determining the amount of energy lost through radiation of sound by a spherical bubble undergoing both translational motion and radial pulsation. In this procedure, one ignores the loss of energy until just after the bubble has passed its minimum radius, then subtracts the energy radiated away, using eq. [9], and proceeds with the integration until the next minimum is reached. In other words, one integrates Taylor's eqs. [1-2], keeping Y constant, with its value being given by eq. [3]. After the bubble has gone through its minimum radius, one replaces Y in eq. [1] by Y', where:

$$Y' = Y\left(1 - \frac{\Delta Y}{Y}\right) \qquad [14]$$

and $\Delta Y$ is given by eq. [9]. One then repeats this procedure whenever the bubble radius passes through a local minimum, using the new value of Y, Y', in eq. [9].

Hicks (1972) incorporated a drag force, $F_D$, into the equations of motion, where $F_D$ is given by:

$$F_D = \frac{1}{2} C_D \pi \rho a^2 U^2 \qquad [15]$$

and the value of the drag coefficient, $C_D$, was chosen to be $C_D \approx 2.25$ in order to bring the distance travelled upward by the bubble at its first maximum into agreement with that actually observed for 500 lbs of TNT detonated 150 ft below the surface.

By differentiating [2], he obtained the rate of change of momentum with respect to time, and equated that to $F_D$. The momentum equation then would then become:

$$\frac{d}{dt}(a^3 U) = 2a^3 g - \frac{3}{4} C_D a^2 U^2 \ . \qquad [16]$$

The energy dissipated by drag forces is given by:

$$\frac{dE}{dt} = \frac{1}{2} C_D \pi \rho a^2 U^3 \ . \qquad [17]$$

The rate of energy loss, -dE/dt, was then equated to the time derivative of [1] to obtain the energy equation:

$$\frac{d}{dt}\left(2\pi\rho a^3 \left(\frac{da}{dt}\right)^2 + \frac{\pi}{3}\rho a^3 U^2 + \frac{4\pi}{3}\rho a^3 gz + E(a)\right)$$

$$= -\frac{1}{2} C_D \pi \rho a^2 U^3 \ . \qquad [18]$$

It should be noted that eqs. [16] and [18] are not precisely identical to those of Hicks, since they neglect his correction for the effect of the water surface.

II.2 Equations of Motion including Energy Loss by Drag and Radiation of Sound. The method described in the previous section for determining the energy loss through radiation of sound by a bubble has a number of drawbacks, not the least of which is that the subtraction of energy from eq. [1] at the minimum radius can adversely affect the convergence of numerical solutions to the system of differential eqs. [1-2]. As well, the energy loss expression, eq. [9], was expressly derived under the assumption that any effects of gravity and translational motion on the energy loss would be negligible; one really has no a priori reason for believing this to be the case.

For these reasons, there seems to be a need for a treatment which produces a more rigorous formulation of the equations of motion of a bubble in the presence of gravity, incorporating the effects of the radiation of energy by sound, and which can be easily extended to include other forms of energy loss, such as drag. Such a treatment is presented in this section.

The kinetic energy, T, of the water surounding a sphere of radius a moving with an upward translational velocity U, is given by:

$$T = 2\pi\rho a^3\left(\frac{da}{dt}\right)^2 + \frac{\pi}{3}\rho a^3 U^2 \qquad [19]$$

(Taylor 1942, Cole 1948). The potential energy, V, is given by:

$$V = E(a) + V_p \qquad [20]$$

where E(a) is, as before, the internal energy of the bubble, eq. [6], and $V_p$, the energy associated with the hydrostatic pressure around the bubble:

$$V_p = \frac{4\pi}{3}\rho a^3 gz \qquad [21]$$

The Lagrangian of the flow around the bubble, L, is then:

$$\begin{aligned} L &= T - V \\ &= 2\pi\rho a^3\left(\frac{da}{dt}\right)^2 + \frac{\pi}{3}\rho a^3 U^2 - \frac{4\pi}{3}\rho a^3 gz \\ &\quad - \frac{kM^\gamma a^{-3\gamma+3}}{(\gamma-1)\left(\frac{4\pi}{3}\right)^{\gamma-1}} \end{aligned} \qquad [22]$$

and the Hamiltonian, H:

$$H = T + V$$

$$= 2\pi\rho a^3 \left(\frac{da}{dt}\right)^2 + \frac{\pi}{3}\rho a^3 U^2 + \frac{4\pi}{3}\rho a^3 gz + \frac{kM^\gamma a^{-3\gamma+3}}{(\gamma-1)\left(\frac{4\pi}{3}\right)^{\gamma-1}} \qquad [23]$$

The equations of motion of the bubble are therefore given by:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{a}}\right) - \frac{\partial L}{\partial a} = Q_a,$$

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{z}}\right) - \frac{\partial L}{\partial z} = Q_z \qquad [24]$$

where $\dot{a} = \frac{da}{dt}$, $\dot{z} = \frac{dz}{dt} = -U$, and the $Q_i$ are the non-conservative generalized forces.

Evaluating [24], one obtains, as equations of motion for a bubble:

$$4\pi\rho a^3 \frac{d^2a}{dt^2} + 6\pi\rho a^2\left(\frac{da}{dt}\right)^2 - \pi\rho a^2 U^2 + 4\pi\rho a^2 gz - \frac{3kM^\gamma}{\left(\frac{4\pi}{3}\right)^{\gamma-1}} a^{-3\gamma+2} = Q_a, \qquad [25]$$

$$\frac{d}{dt}(a^3 U) = 2a^3 g - \frac{Q_z}{\left(\frac{2\pi}{3}\rho\right)} \qquad [26]$$

As might have been expected, when $Q_a = Q_z = 0$, i.e. when there is no change in the total energy of the system, eqs. [25] and [26] can be integrated, and yield Taylor's eqs. [1] and [2].

In order to complete the derivation of the equations of motion of the bubble, it is necessary to determine $Q_a$ and $Q_z$.

Now, the velocity potential, $\phi$, is given by:

$$\phi = \phi_1 + \phi_2 \qquad [27]$$

where

$$\phi_1 = \frac{a^2}{r}\frac{da}{dt} = \frac{f(t)}{r} \quad , \qquad [28]$$

$$\phi_2 = \frac{1}{2}\frac{a^3}{r^2} U \cos\Theta = \vec{\nabla}\cdot\left(A(t)\frac{\cos\Theta}{r}\hat{n}\right) \qquad [29]$$

where r and $\Theta$ are the radial and angular co-ordinates, respectively, shown in Fig. 1 and $\hat{n}$ is the unit vector normal to the bubble surface (e.g. Cole 1948). Since the bubble is explicitly assumed to remain spherical, $\hat{n} \equiv \hat{r}$, where $\hat{r}$ is the unit radial vector.

Following a development similar to that of Landau and Lifshitz (1966), in the wave zone:

$$\phi = \frac{f(t')}{r} + \vec{\nabla}\cdot\left(A(t')\frac{\cos\Theta\,\hat{r}}{r}\right) \qquad [30]$$

where

$$t' = t - \frac{r}{c} \qquad [31]$$

and c is, as before, the velocity of sound in water. The velocity, $\vec{V}$, of the water in the wave zone must therefore be given by:

$$\vec{V} = \vec{\nabla}\phi$$

$$\approx \left(\frac{1}{cr}\frac{\partial f(t)}{\partial t} - \frac{\cos\Theta}{rc^2}\frac{\partial^2 A(t)}{\partial t^2}\right)\hat{r} + \cdots \qquad [32]$$

where terms of higher negative order in r have been neglected.

The total energy emitted as sonic radiation per unit time, $\frac{dE}{dt}$, is then:

$$\frac{dE}{dt} = \rho c \int\int_{(S)} (\vec{V} \cdot \vec{V})\, dS$$

$$= \frac{4\pi\rho}{c} \left(\frac{\partial f(t)}{\partial t}\right)^2 + \frac{4\pi\rho}{3c^3} \left(\frac{\partial^2 A(t)}{\partial t^2}\right)^2 \qquad [33]$$

where the surface integral has been taken over a sphere of radius r. To a good approximation, the term in eq. [33] proportional to $\frac{1}{c^3}$ can be neglected for low translational velocities, since it will be 2 orders of magnitude smaller than that proportional to 1/c.

Now,

$$\frac{\partial f(t)}{\partial t} = a^2 \frac{d^2 a}{dt^2} + 2a \left(\frac{da}{dt}\right)^2 \qquad [34]$$

and hence:

$$\frac{dE}{dt} = \frac{4\pi\rho}{c} \left(4a^2 \left(\frac{da}{dt}\right)^4 + 4a^3 \left(\frac{da}{dt}\right)^2 \frac{d^2 a}{dt^2} + a^4 \left(\frac{d^2 a}{dt^2}\right)^2\right) \qquad [35]$$

Now, in view of the foregoing, it would seem reasonable to conclude that, in the absence of drag forces, the dissipative forces depend solely upon the rate of change of the radius of the bubble, and hence that:

$$Q_z = 0 \qquad [36]$$

In that event, it is possible to write that:

$$\dot{a} Q_a = -\frac{dE}{dt} \qquad [37]$$

and hence:

$$Q_a = -4\pi\rho \frac{a^2}{c} \dot{Q}_a - \dot{a} \frac{\partial Q_a}{\partial \dot{a}} \qquad [38]$$

where

$$\bar{Q}_a = \left[16\left(\frac{da}{dt}\right)^3 + 8a\left(\frac{da}{dt}\right)\frac{d^2a}{dt^2} + \left(4a\left(\frac{da}{dt}\right)^2\right.\right.$$

$$\left.\left. + 2a^2\frac{d^2a}{dt^2}\right)\frac{\partial}{\partial\dot{a}}\frac{d^2a}{dt^2}\right] \qquad [39]$$

By substituting eqs. [36] - [39] into eqs. [25] and [26], one obtains:

$$a\frac{d^2a}{dt^2} + \frac{3}{2}\left(\frac{da}{dt}\right)^2 - \frac{U^2}{4} + gz - \frac{kM^\gamma a^{-3\gamma}}{\left(\frac{4\pi}{3}\right)^\gamma \rho} \qquad [40]$$

$$= \frac{Q_a}{\left(4\pi\rho a^2\right)}$$

$$\frac{dU}{dt} = 2g - \frac{3U}{a}\frac{da}{dt} \qquad [41]$$

Differentiating eq. [40] with respect to $\dot{a}$, one obtains:

$$\frac{\partial}{\partial\dot{a}}\left(\frac{d^2a}{dt^2}\right) = -\frac{3}{a}\dot{a} + \frac{1}{4\pi\rho a^3}\frac{\partial Q_a}{\partial\dot{a}} \qquad [42]$$

Substituting eq. [42] into eqs. [38] and [39] one finds:

$$Q_a = -4\pi\rho\frac{a^2}{c}\left[4\left(\frac{da}{dt}\right)^3\right] - 8\pi\rho\frac{a^3}{c}\left(\frac{da}{dt}\right)\frac{d^2a}{dt^2} - \left[4a\left(\frac{da}{dt}\right)^2\right.$$

$$\left. + 2a^2\frac{d^2a}{dt^2}\right]\frac{1}{ac}\frac{\partial Q_a}{\partial\dot{a}} - \left(\frac{da}{dt}\right)\frac{\partial Q_a}{\partial\dot{a}} \qquad [43]$$

The evaluation of the last three terms in eq. [43] still presents some difficulty. However, by equating the rate of change of the Hamiltonian

H of the bubble, given by eq. [23], with respect to time, to -dE/dt, thusly:

$$\frac{dH}{dt} = -\frac{dE}{dt} \qquad [44]$$

one obtains an alternative form of eq. [40], the equation of motion. A detailed comparison of terms shows that:

$$Q_a = -4\pi\rho\frac{a^2}{c}\left[4\left(\frac{da}{dt}\right)^3 + 4a\left(\frac{da}{dt}\right)\frac{d^2a}{dt^2} + a^2\left(\frac{d^2a}{dt^2}\right)^2/\left(\frac{da}{dt}\right)\right] \qquad [45]$$

By comparing eqs. [45] to [43], it follows that:

$$-8\pi\rho\frac{a^3}{c}\left(\frac{da}{dt}\right)\frac{d^2a}{dt^2} = -\left[4\left(\frac{da}{dt}\right)^2 + 2\frac{d^2a}{dt^2}\right]\frac{1}{c}\frac{\partial Q_a}{\partial \dot{a}} + \frac{4\pi\rho}{c}a^4\left(\left(\frac{d^2a}{dt^2}\right)^2/\left(\frac{da}{dt}\right)\right) - \left(\frac{da}{dt}\right)\frac{\partial Q}{\partial \dot{a}} \qquad [46]$$

The terms on the left hand side of eq. [46] are recognizable as being analogous to radiation reaction terms in electromagnetic theory, and hence are ignorable in a first approximation. By extension, it follows that $8\pi\rho\frac{a^3}{c}\left(\frac{da}{dt}\right)\frac{d^2a}{dt^2}$ is also ignorable, and hence:

$$Q_a = -4\pi\rho\frac{a^2}{c}\left[4\left(\frac{da}{dt}\right)^3\right] \qquad [47]$$

If greater accuracy is desired, these neglected terms can be evaluated, most conveniently by means of [45], and added to the generalised force $Q_a$.

If, following Hicks (1972), one introduces a drag force, $F_D$, of the form of eq. [15], one can show that $Q_z$ in [24] is given by:

$$Q_z = \frac{1}{2}C_D\pi\rho a^2U^2 \qquad [48]$$

Finally, the equations of motion of a bubble incorporating energy loss by radiation can, to a good approximation, be written as:

$$a\frac{d^2a}{dt^2} + \frac{3}{2}\left(\frac{da}{dt}\right)^2 - \frac{U^2}{4} + gz - \frac{kH^\gamma a^{-3\gamma}}{\left(\frac{4\pi}{3}\right)^\gamma \rho} = -\frac{1}{c}\,Q_a' \qquad [49]$$

$$\frac{dU}{dt} = 2g - 3\frac{U}{a}\left(\frac{da}{dt}\right) - \frac{3}{4}C_D\frac{U^2}{a} \qquad [50]$$

where $Q_a'$ is given by:

$$Q_a' = 4\left(\frac{da}{dt}\right)^3 \qquad [51]$$

or, if second order terms are included, by:

$$Q_a' = 4\left(\frac{da}{dt}\right)^3 + 4a\left(\frac{da}{dt}\right)\frac{d^2a}{dt^2} + \left(a^2\left(\frac{d^2a}{dt^2}\right)/\left(\frac{da}{dt}\right)\right) . \qquad [52]$$

In this section, we have presented 2 sets of equations of motion which had been previously derived by various authors to describe the motion of a spherical bubble under different conditions. We have extended these to incorporate more rigorous approximations to the effects of energy loss by the radiation of sound as well an estimate of the influence of drag upon the motion.

To sum up: eqs. [1-2] are the equations of motion for a spherical bubble undergoing radial pulsations and translational motion, without energy loss. The introduction of eq. [9] into this system at the time at which the bubble's radius has reached a local minimum provides a crude mechanism for incorporating energy loss by sonic radiation. Equations [10-11] apply when the effects of surfaces such as the sea bed and surface significantly affect the motion of the bubble, and do not incorporate energy loss. Finally, eqs. [49-52] are the equations of motion of a spherical bubble undergoing radial pulsations and translational motion in an infinite medium and in which the effects of energy loss from drag and radiation of sound have taken account of in a reasonably rigourous fashion.

III. NUMERICAL METHODS OF SOLUTION. Before one attempts numerical solutions of any of these sets of equations, it is useful to make the equations of motion non-dimensional. Thus, the substitution of:

$$\begin{aligned} a &= a^* L, \\ b &= b^* \\ z &= z^* L, \\ t &= t^* L, \\ U &= U^* L/T, \end{aligned} \qquad [53]$$

into the equations of motion used, where

$$\begin{aligned} L &= \left(\frac{Y}{g\rho}\right)^{1/4}, \\ T &= \sqrt{\frac{L}{g}} \end{aligned} \qquad [54]$$

yields a dimensionless form of the equations of motion. As before, Y is given by eq. [3] and g and $\rho$ are, respectively, the gravitational acceleration and the density of water. These particular scaling factors in eq. [2] were originally used by Taylor (1942).

Since all of the equations of motion have the unfortunate property of singularity at the origin, it is necessary to begin the integration with a series solution. Taylor (1942) suggested initial values for the dimensionless variables of:

$$\begin{aligned} a^* &= \left(\frac{t^*}{1.0025}\right)^{2/5}, \\ U^* &= \left(\frac{10}{11}\right) t^* \quad , \\ z^* &= z_o - \left(\frac{5}{11}\right) (t^*)^2 \end{aligned} \qquad [55]$$

for values of $t^*$ near zero.

A similar problem arises if one desires to use the more accurate approximation to $Q'_a$, eq. [52], in the equations of motion. Since da/dt becomes 0 at several points during the motion of the bubble, the

last term in eq. [53] is formally undefined at those points. However, if one uses the expression for the total energy, eq. [1], one can write, in terms of the dimensionless variables:

$$\left(\frac{da^*}{dt^*}\right)^2 = \left(\frac{Y(t)}{Y} - \frac{E^*(a^*)}{Y}\right)\frac{1}{2\pi(a^*)^3} - \frac{(U^*)^2}{6} - \frac{2}{3}z^* \qquad [56]$$

where

$$E^*(a^*) = \frac{\frac{1}{2}M^{\gamma}(a^*)^{-3(\gamma-1)}}{(\gamma-1)\left(\frac{4\pi}{3}\right)^{\gamma-1}L^{3(\gamma-1)}} \quad , \qquad [57]$$

Y(t) is the total energy at time t, and all other variables are as previously defined.

If one defines:

$$\alpha = \left(\frac{Y(t) - E^*(a^*)}{Y}\right)\frac{1}{2\pi(a^*)^3}$$

$$\beta = \frac{(U^*)^2}{6} + \frac{2}{3}z^* \qquad [58]$$

one can write:

$$\left(\frac{da^*}{dt^*}\right)^{-1} = \pm\frac{1}{\alpha^{\frac{1}{2}}}\left(1 + \frac{1}{2}\frac{\beta}{\alpha} + \frac{3}{8}\left(\frac{\beta}{\alpha}\right)^2 + \cdots\right) \qquad [59]$$

where the positive value is taken while the bubble is expanding, and the negative while contracting. Equation [59] may be used to evaluate the expression for $Q'_a$, eq. [52], if one wishes to include the final two terms. When such was done in this work, the value of Y(t) was approximated at each step of the integration by the substitution of the current estimates for da/dt, U and z into eq. [1]. An estimate for $\frac{d^2a}{dt^2}$, was obtained by substituting the current estimates for da/dt, U and z into eq. [49] with $Q'_a$ set to 0. The series, eq. [59], can be terminated at any point, and the truncation error determined by reperforming the integration with the inclusion of the next higher power of $\beta/\alpha$. In this work, the last term included in the series was $(\beta/\alpha)^4$.

The actual integrations were carried out using a 4 point Runge-Kutta algorithm incorporating automatic error controls.

The initial parameters which must be supplied to the programme are listed in Fig. 2.

IV. NUMERICAL RESULTS AND ANALYSIS. Taylor (1942) solved eqs. [1] and [2], neglecting the internal energy of the gas (i.e. setting E(a) = 0). His results were therefore independent of the mass of the explosive charge, and hence were amenable to scaling, a convenience which was achieved at the expense of some accuracy, especially near the minimum radius. Figures 3-5 show a comparison of the radius, velocity, and height above the original explosion for a charge of TNT of mass 2.1136 kg. detonated at a depth of 6.1 metres below the surface, calculated using Taylor's equations, with and without the internal energy having been neglected. As can be seen, the internal energy does have a significant effect on the behaviour of the bubble. Accordingly, in all subsequent calculations, it should be understood that the effect of the internal energy has been taken into account.

Figures 6-8 compare the results obtained by solving Taylor's equation of motion [1]-[2], and the ones derived in this paper eqs. [49]-[50] with the non-conservative generalized forces, $Q_a$ and $Q_z$, having been set to zero for the bubble produced by an explosive charge of 2.1136 kg. detonated at a depth of 34.81 metres below the ocean surface. Since one would expect the integration of the two sets of equations to yield identical results under these conditions, Figs. 6-8 serve as a test of both the theoretical derivation and the numerical algorithm. As can be seen, the results of the two sets of computations are virtually identical. It should be noted that the solution of the set of equations of motion [49]-[50] require the provision of an initial value for da/dt. The programme permits one to do this by one of two ways: either by differentiating the expression for a in eq. [55] (SKIP = 0), or by substituting the values for a, U, and z obtained from eq. [55] into eq. [1] and solving for da/dt at the initial time (SKIP = 1). This latter procedure seems to be the more accurate, since it amounts to specifying the initial total energy to be equal to that given by eq. [3]. Figures 9-11 show the dependence of the calculated values for the radius, translational velocity, and height above the explosion, using eqs. [49] and [50], (with $Q_a = Q_z = 0$) on the method chosen to obtain the initial value of da/dt. Although the general behaviour of the two graphs is qualitatively the same, there is substantial detailed disagreement, indicating that the algorithm is sensitive to the initial choice for da/dt.

Figures 12-15 show the effects of energy loss by the radiation of sound, only, (i.e. $Q_z = 0$) on the motion of a bubble produced by an explosive charge of the mass and at the depth. The energy loss was calculated by using the more accurate expression for $Q_a'$, eq. [52],

including the radiation reaction terms. The effect of omitting these terms (i.e. using the value for $Q'_a$, given by eq. [51] is shown in Figs. 16-19. Evidently, the characteristics of the motion of are substantially affected by the radiation of energy. Specifically, the maximum radius of the bubble is decreased and the minimum radius is increased in comparison with those which one would obtain by neglecting energy loss. The radiation of energy also decreases the calculated periods of oscillation slightly, as shown in Fig. 20.

In the particular example illustrated here, whether the energy loss is calculated using eq. [9] or the formalism developed in section II.2 seems to have little effect on the characteristics of the motion for this case, at least early in the bubble motion, although this difference is more striking when the approximate radiation reaction terms are included. However, as shown in Figs. 15 and 19, the distribution of the energy loss over time depends quite strongly on which formula for energy loss is used, and whether the radiation reaction terms are included.

Figures 21-24 show the effect of including drag in the equations of motion, with and without the addition of radiative loss, for the bubble produced by 227.27 kg. of TNT at a depth of 45.73 metres below the surface. It has been observed that a bubble from such an explosion rises approximately 3.35 metres from the location of the explosion in the time taken to reach its first maximum. Hicks found that where drag is the only source of dissipation, a drag coefficient of $C_D = 2.25$ was necessary to reproduce this behaviour. In this work, it was found that a coefficient of $C_D = 1.85$ best matched the observed rise to the first minimum, in the absence of radiation of sound. When radiative dissipation was considered, a drag coefficient of $C_D = 1.6$ seemed best to fit the observed motion, and it is this value which was used in Figs. 21-24.

Finally, Figs. 25-27 show the changes in the motion of the bubble resulting from the inclusion of the effects of the sea bed, located at a depth of 8.94 metres below the explosion, and the surface, but not those of energy loss, for 2.1136 kg. of TNT, detonated at 34.81 metres below the ocean surface. It can be seen that, for this case, the effects of the radiation of sound produce more significant changes in the bubble's motion than do those of the ocean floor and surface, although this may not always be true. It should also be noted that the precise characteristics of the motion seem to quite sensitive to the initial values of the variables. For example, if one uses the initial values for as given eq. [3] at a time $t^* = .001$, integrates eq. [1] and [2] (i.e. the equations of motion ignoring bottom and surface effects) to a time $t^* = .01$, and uses the values obtained as starting values at

$t^* = .01$ in the integration of eqs. [10] and [11], (i.e. the equations of motion incorporating surface and bottom effects), one obtains Figs. 28-30, which exhibit some differences from Figs. 25-27, notably in the values obtained for the upward velocity of the bubble near the first minimum.

V. CONCLUDING REMARKS. In this work, we have derived a more precise expression for the energy loss experienced by a spherical bubble through the radiation of sound, and shown that the dissipative force from this cause is proportional to the cube of the rate of change of the bubble radius, at least as a first approximation. We have also shown how the effects of drag may be combined with those of radiation of sound.

We have presented some results of an algorithm which solves the various forms of the equations of motion. We suggest that the form of the equations of motion derived in section II.2 of this paper are more amenable to numerical solution than those derived by Taylor (1942), especially when one wishes to incorporate the effects of energy loss, and would hence recommend their employment.

From the results of our computations, it would seem that one should not neglect the effects of internal energy and energy loss from radation and drag in the calculation of the bubble's behaviour, but that the effects of free and rigid surfaces on the motion are usually of lesser significance.

Finally, since the non-sphericity of the bubble near its minimum radius is of considerable importance to the motion, the equations of motion should be extended to cover this case. We suggest that an approach using the Lagrangian formalism might have some utility in this endeavour.

VI. BIBLIOGRAPHY. Cole, R.H. 1948, Underwater Explosions Princetown University Press, Princeton

Herring, C. 1942, in Underwater Explosion Research, Vol. II, 35 Office of Naval Research, Dept. of the Navy, Washington, D.C., 1950, UNCLASSIFIED

Hicks, A.N. 1972, The Theory of Explosion Induced Whipping Ship Motions, Report no. NCRE/R579, Naval Construction Research Establishment, St. Leonard's Hill, Dunfermline, Fife UNCLASSIFIED

Holt, R.A. 1977, Annual Review of Fluid Mechanics, 9, 187 Pao Alto, California

Landau, L.D. and Lifshitz, E.M. 1966, Fluid Mechanics, Addison-Wesley Inc., Don Mills, Ontario

Shiffman, M. and Friedman, B. 1944, in Underwater Explosion Research, Vol II, 245, Office of Naval Research, Dept. of the Navy, Washington, D.C., 1950, UNCLASSIFIED

Taylor, Sir G.I. 1942, in The Scientific Papers of Sir Geoffrey Ingram Taylor, Vol III, 320, Cambridge University Press, Cambridge 1963

Figure 1

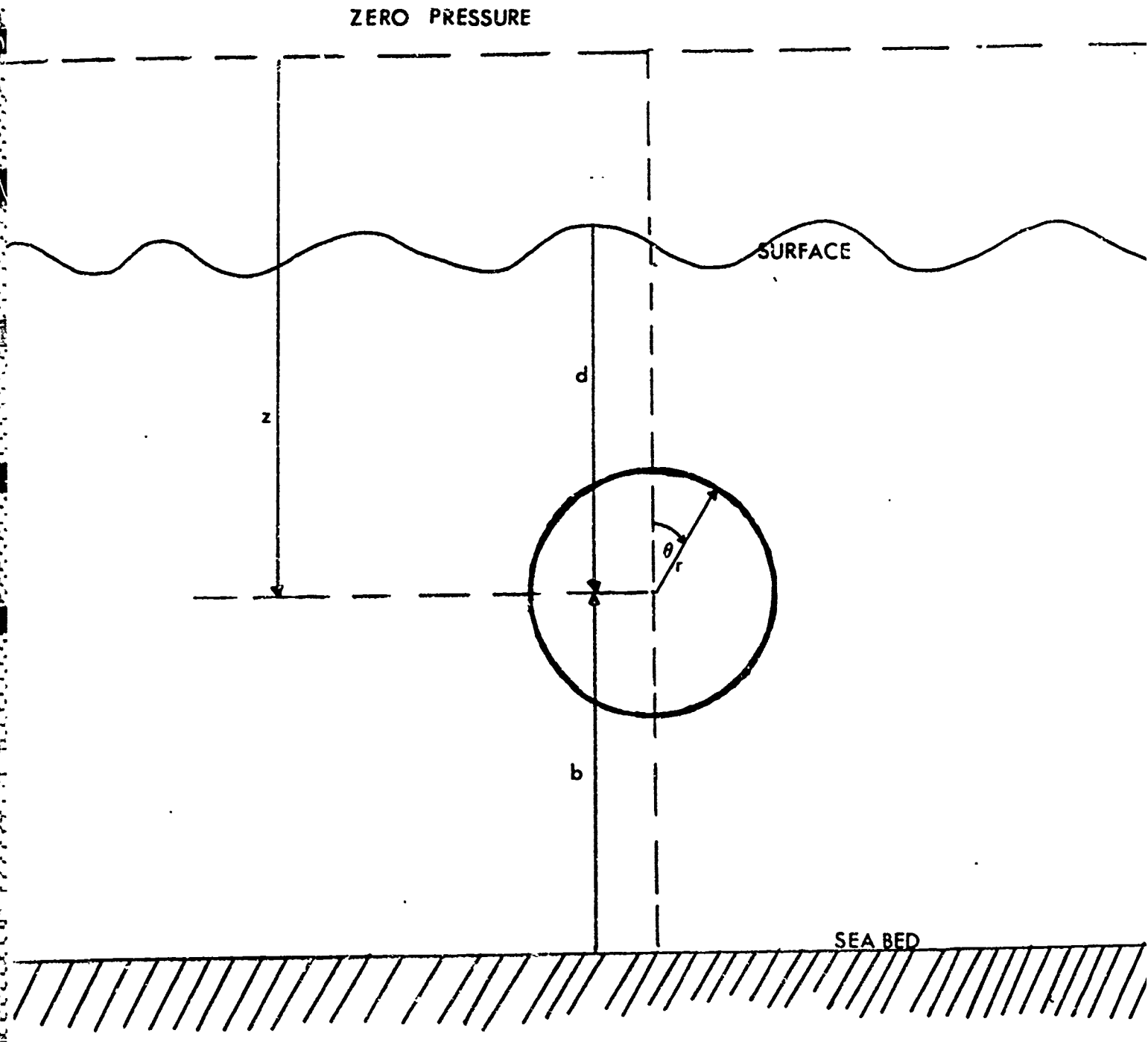


Figure 2

```
771.200                          INPUT PARAMETERS
771.300
771.400  MA = MASS OF EXPLOSIVE IN GRAMS
771.500  Z0 = SCALED DEPTH OF EXPLOSION BELOW ZERO PRESSURE LEVEL
771.600  B0 = SCALED DISTANCE OF BOTTOM FROM LOCATION OF EXPLOSION
771.700  TMIN = INITIAL TIME AT WHICH INTEGRATION IS STARTED
771.800  TINC = VALUE OF INCREMENT TO TMIN
771.900  NINC = NUMBER OF INCREMENTS TO TMIN
771.910  CRIT = MINIMUM CONVERGENCE CRITERION FOR RUNGE-KUTTA ROUTINE
771.920  BOUN = MAXIMUM CONVERGENCE CRITERION FOR RUNGE-KUTTA ROUTINE
771.930
771.940
772.000                          INPUT OPTIONS
773.000
774.000
775.000  INK = 1    SOLVES  TAYLOR'S EQUATIONS OF MOTION (2.1)-(2.2) FOR BUBBLE
777.000  INK = 3    USES METHOD OF IMAGES TO INCLUDE EFFECT OF SURFACES AND
777.100             SOLVES (2.10)-(2.11)
779.000  INK = 4    SOLVES EQUATIONS OF MOTION (2.44)- (2.45) DERIVE' IN THIS
780.000             PAPER
781.000
782.000  E = 0      IGNORES EFFECTS OF INTERNAL ENERGY
783.000  E = 1      INCLUDES EFFECTS OF INTERNAL ENERGY
784.000
785.000  IS = 0     IGNORES RADIATION OF ENERGY
786.000  IS = 1     INCLUDES RADIATION OF ENERGY
787.000             WHEN INK = 1, USES (2.9) TO OBTAIN ENERGY LOSS
788.000             WHEN INK = 4, USES (2.43) TO OBTAIN ENERGY LOSS
788.100
789.000
790.000  SKIP = 0   CALCULATES INITIAL da/dt FROM DERIVATIVE OF INITIAL
790.100             VALUE OF a
791.000  SKIP = 1   CALCULATES INITIAL da/dt FROM TOTAL ENERGY
```

Figure 3

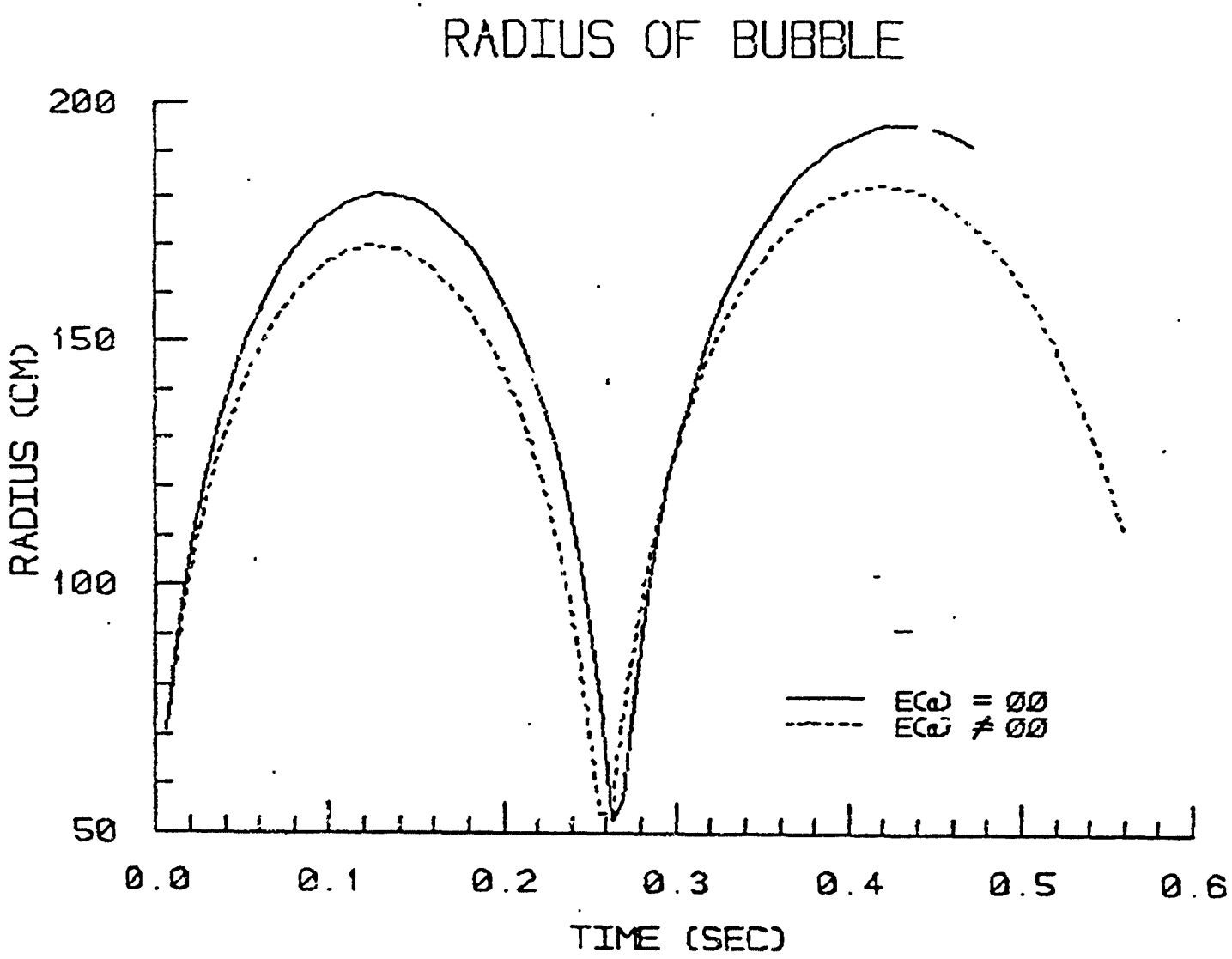
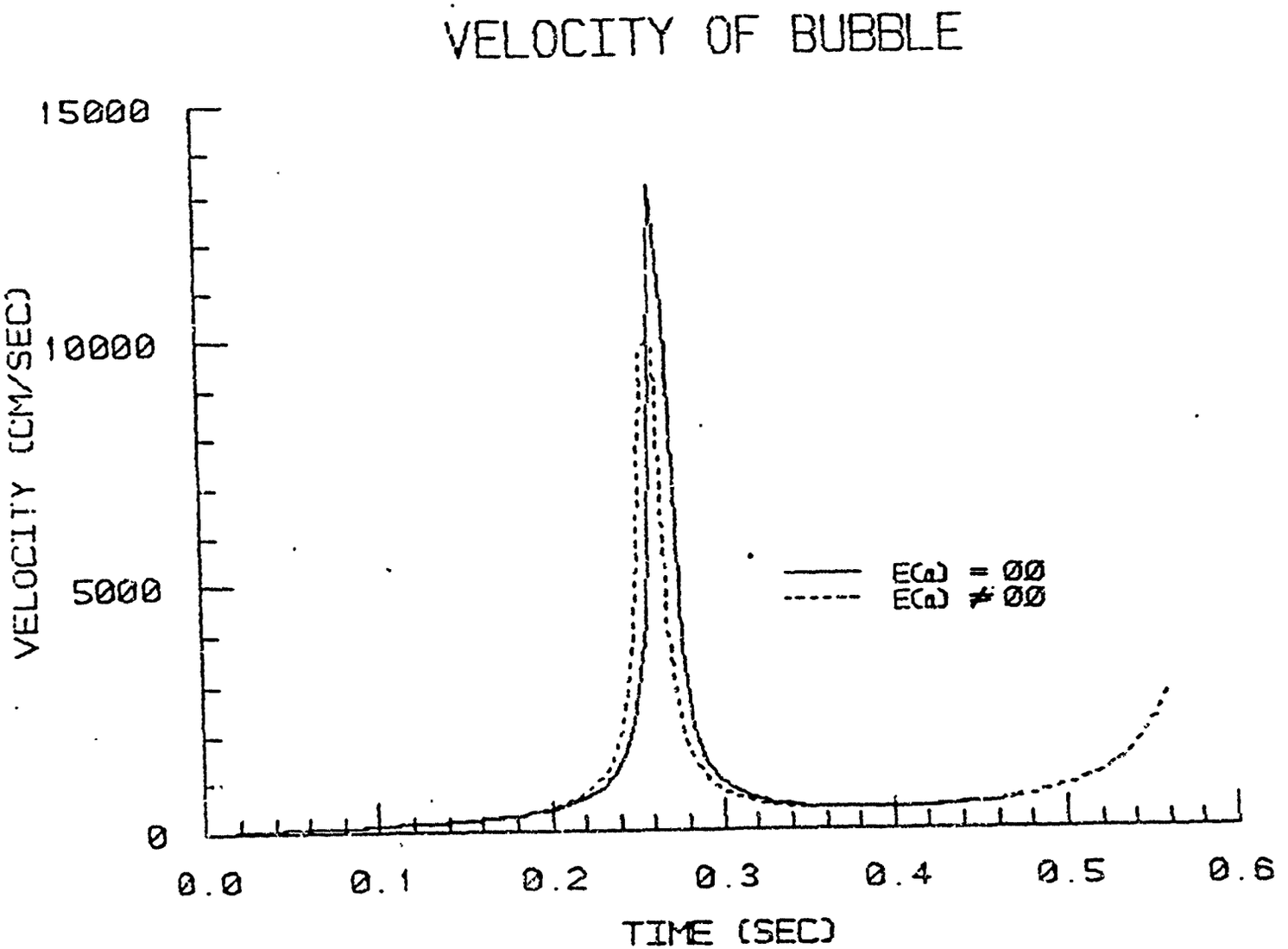

RADIUS OF BUBBLE
200
150
100
50
RADIUS (CM)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TIME (SEC)
E(a) = ØØ
E(a) ≠ ØØ

Figure 4


VELOCITY OF BUBBLE
15000
10000
5000
0
VELOCITY (CM/SEC)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TIME (SEC)
E(a) = ØØ
E(a) ≠ ØØ

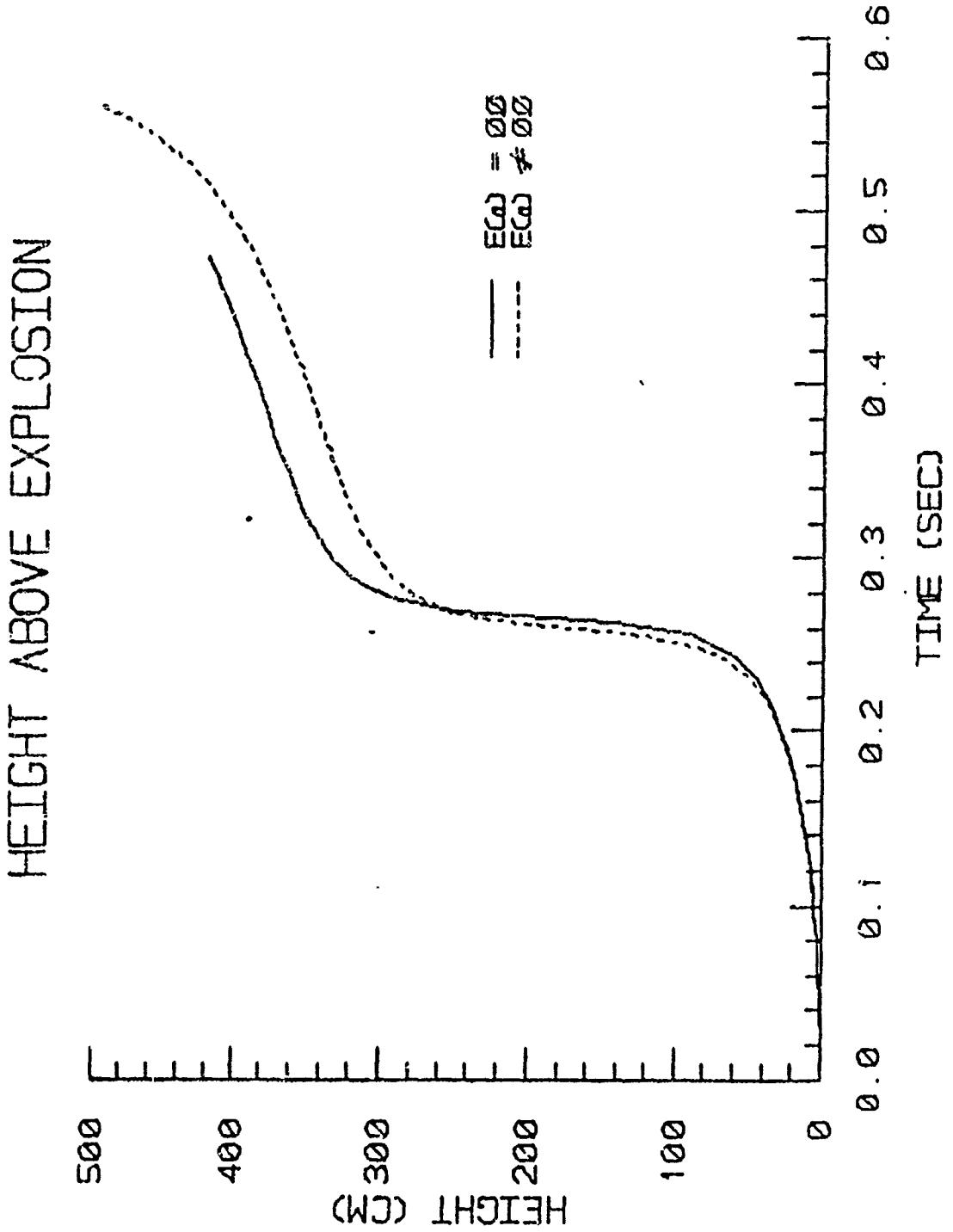

HEIGHT ABOVE EXPLOSION
E(ω) = ØØ
E(ω) ≠ ØØ
HEIGHT (CM)
500
400
300
200
100
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
TIME (SEC)

Figure 5

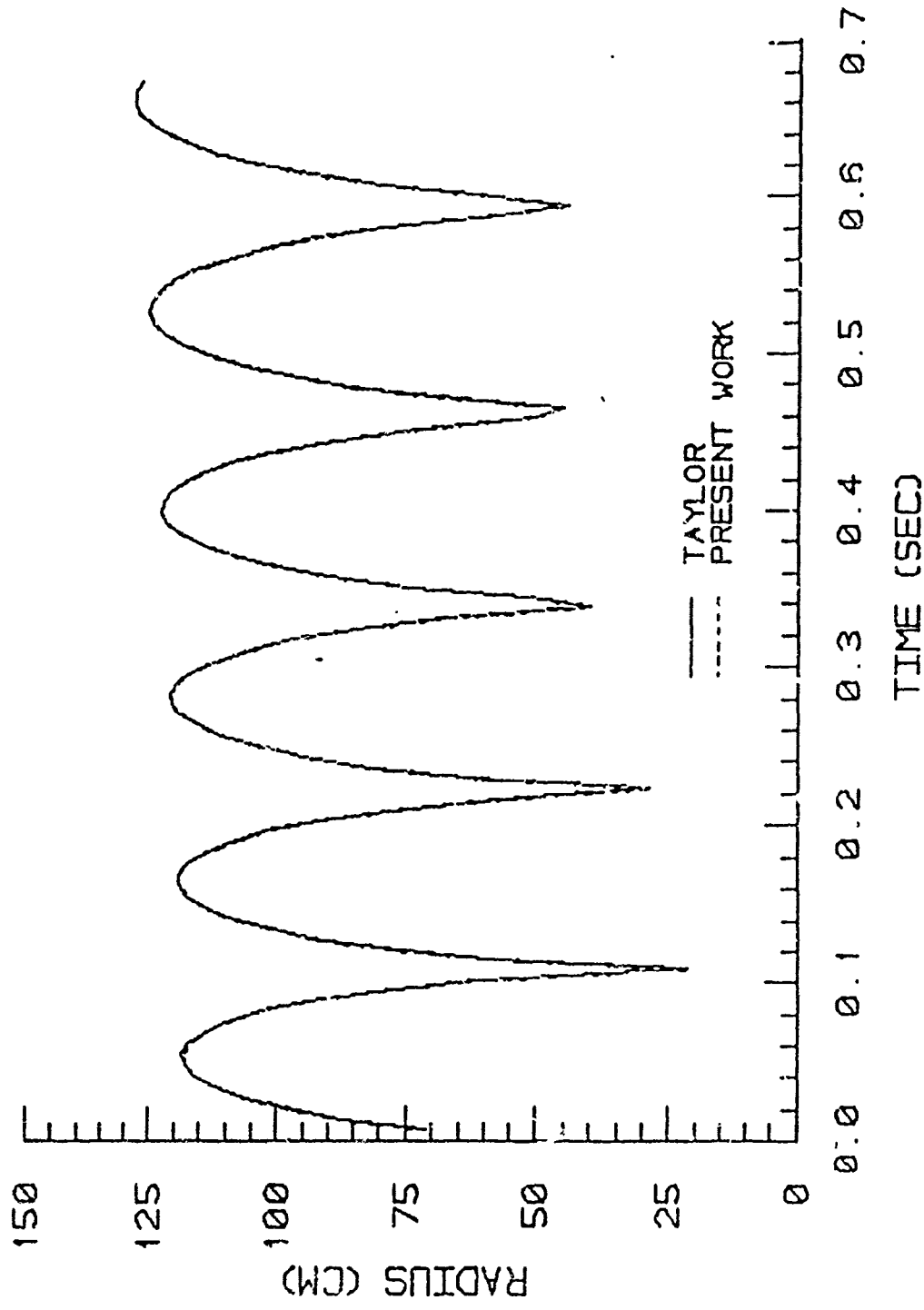

RADIUS OF BUBBLE
RADIUS (CM)
150
125
100
75
50
25
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)
TAYLOR
PRESENT WORK

Figure 6

Figure 7

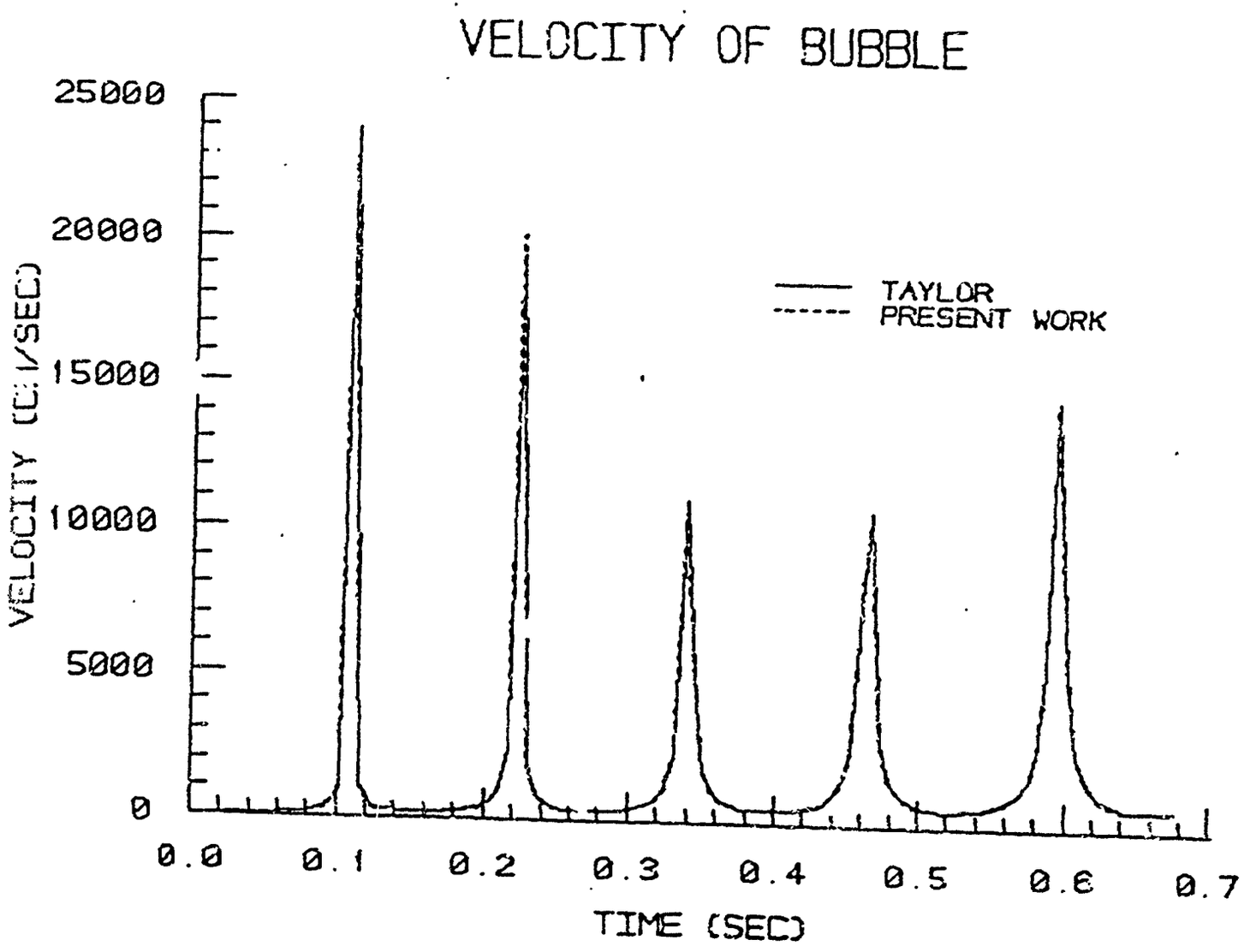
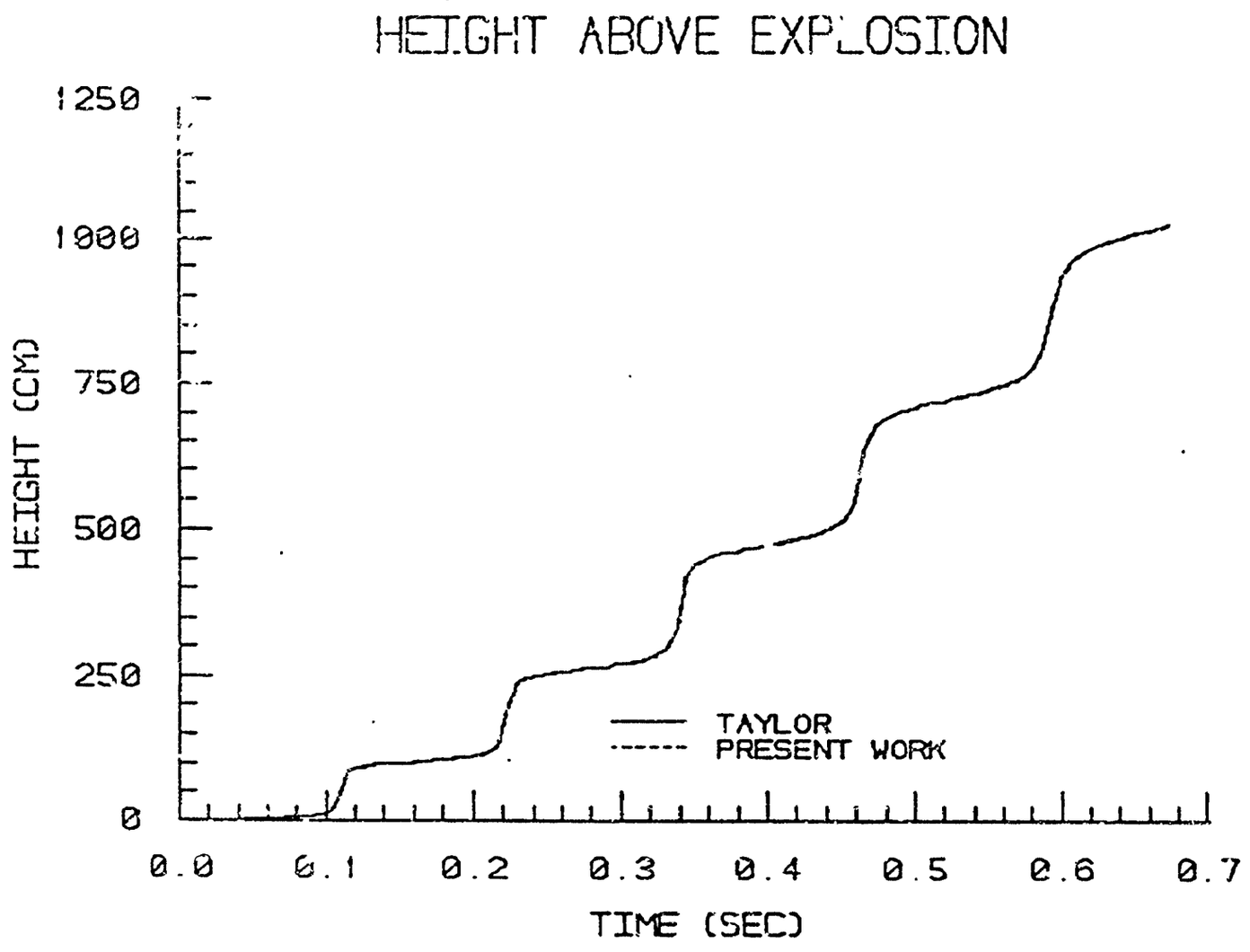

VELOCITY OF BUBBLE
25000
20000
15000
10000
5000
0
VELOCITY (CM/SEC)
TAYLOR
PRESENT WORK
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 8


HEIGHT ABOVE EXPLOSION
1250
1000
750
500
250
0
HEIGHT (CM)
TAYLOR
PRESENT WORK
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

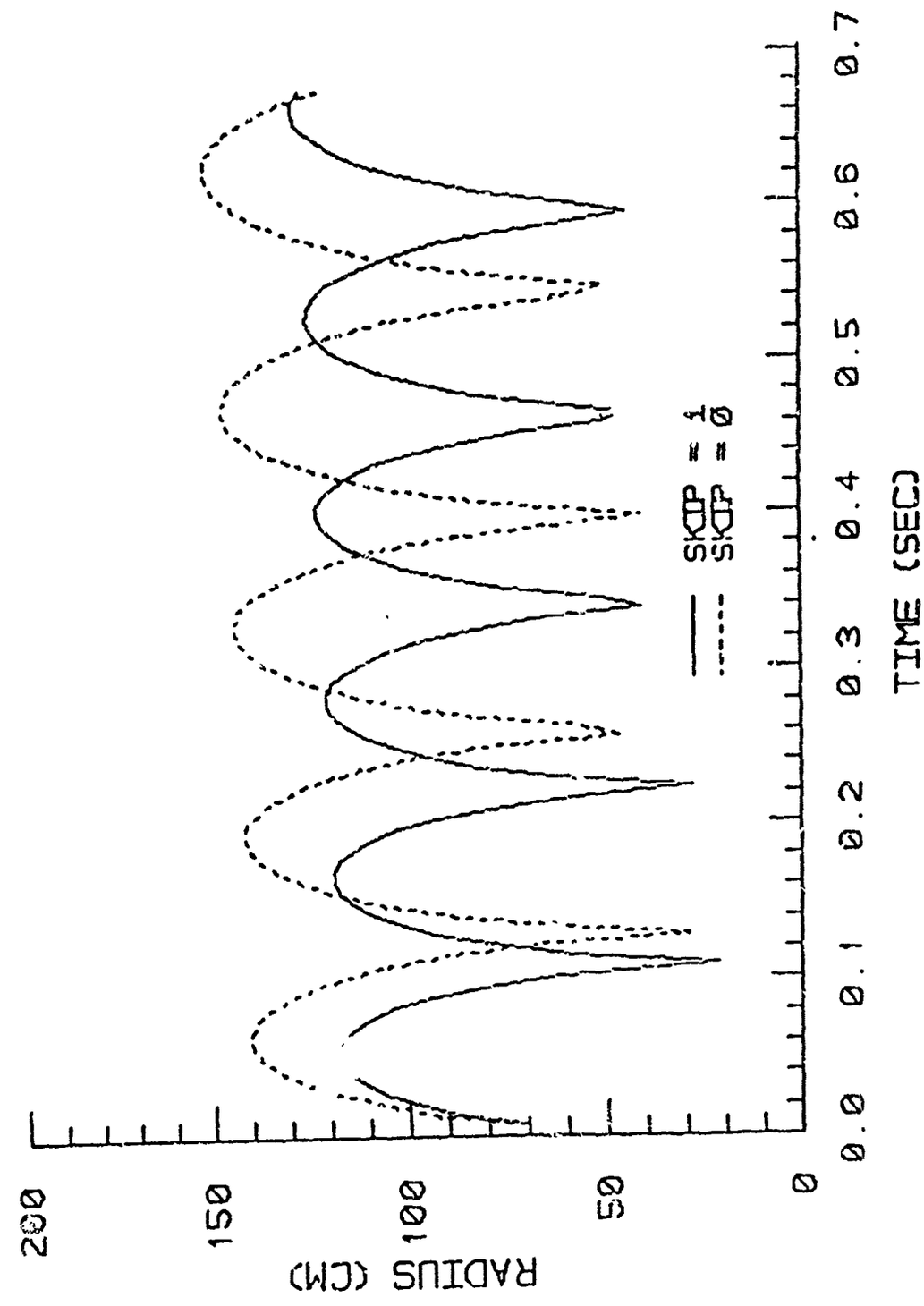

RADIUS OF BUBBLE
RADIUS (CM)
200
150
100
50
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)
SKIP = 1
SKIP = 0

Figure 9

Figure 10

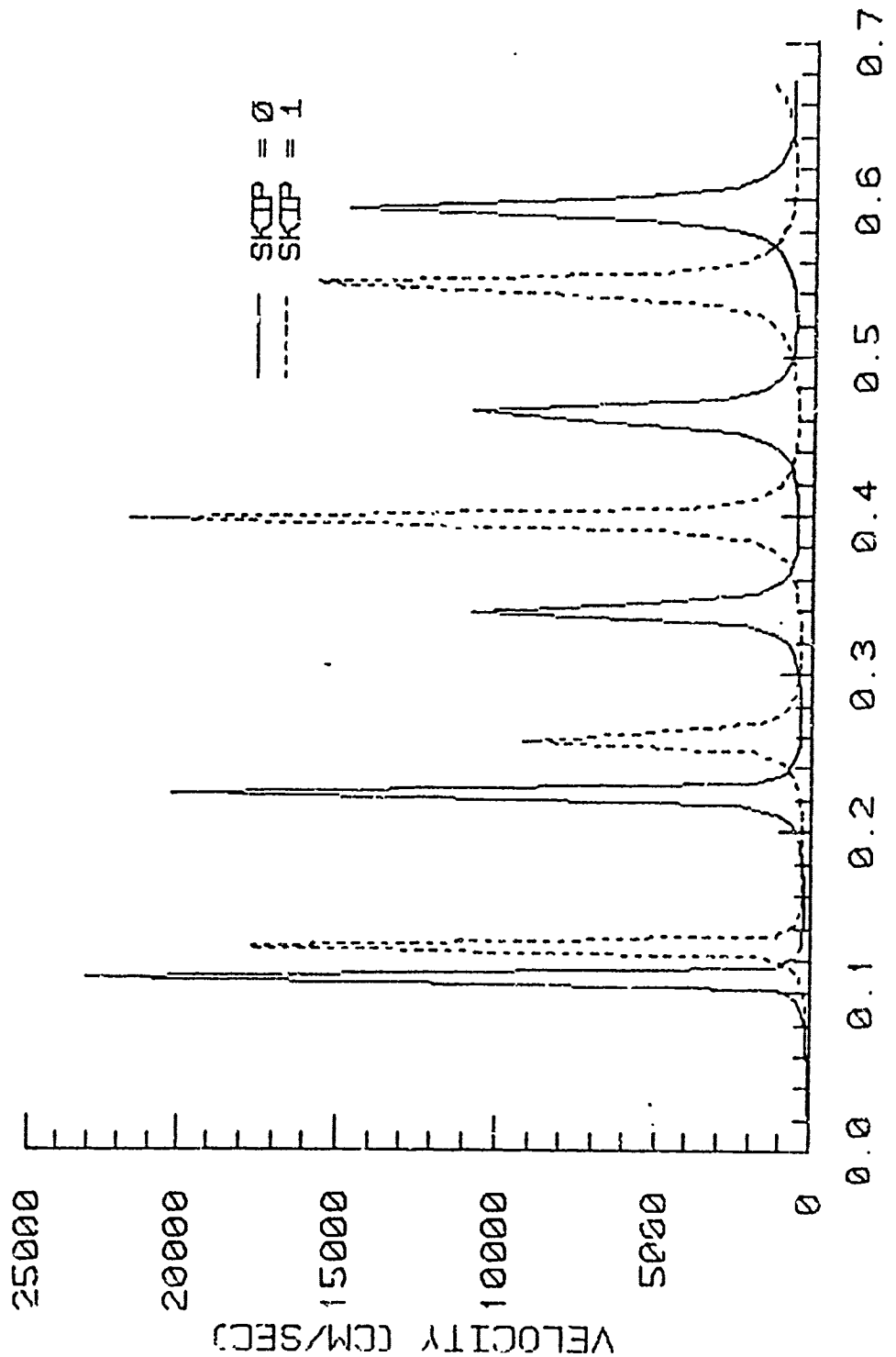
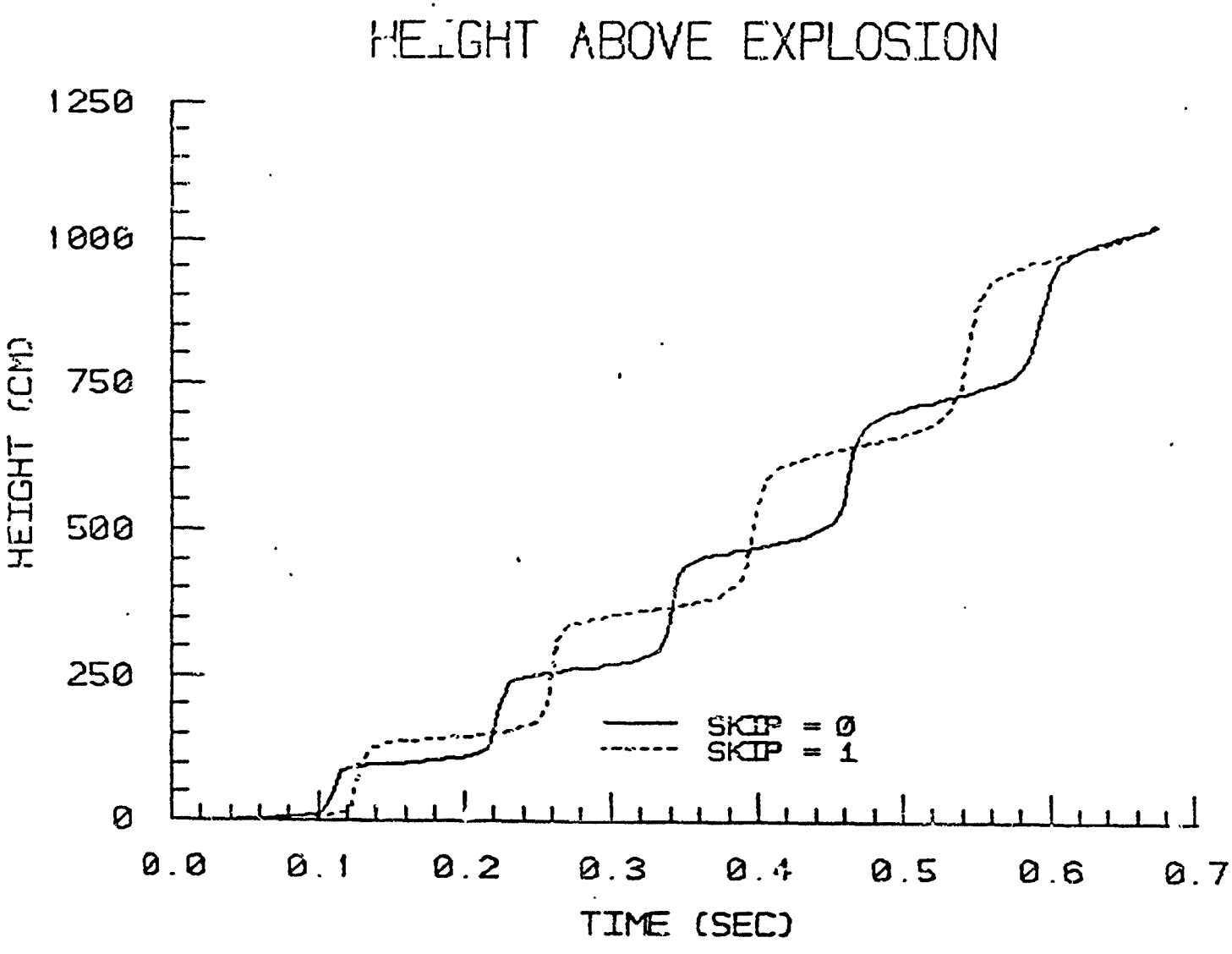

VELOCITY OF BUBBLE
SKIP = 0
SKIP = 1
VELOCITY (CM/SEC)
25000
20000
15000
10000
5000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 11


HEIGHT ABOVE EXPLOSION
1250
1000
750
500
250
0
HEIGHT (CM)
SKIP = 0
SKIP = 1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

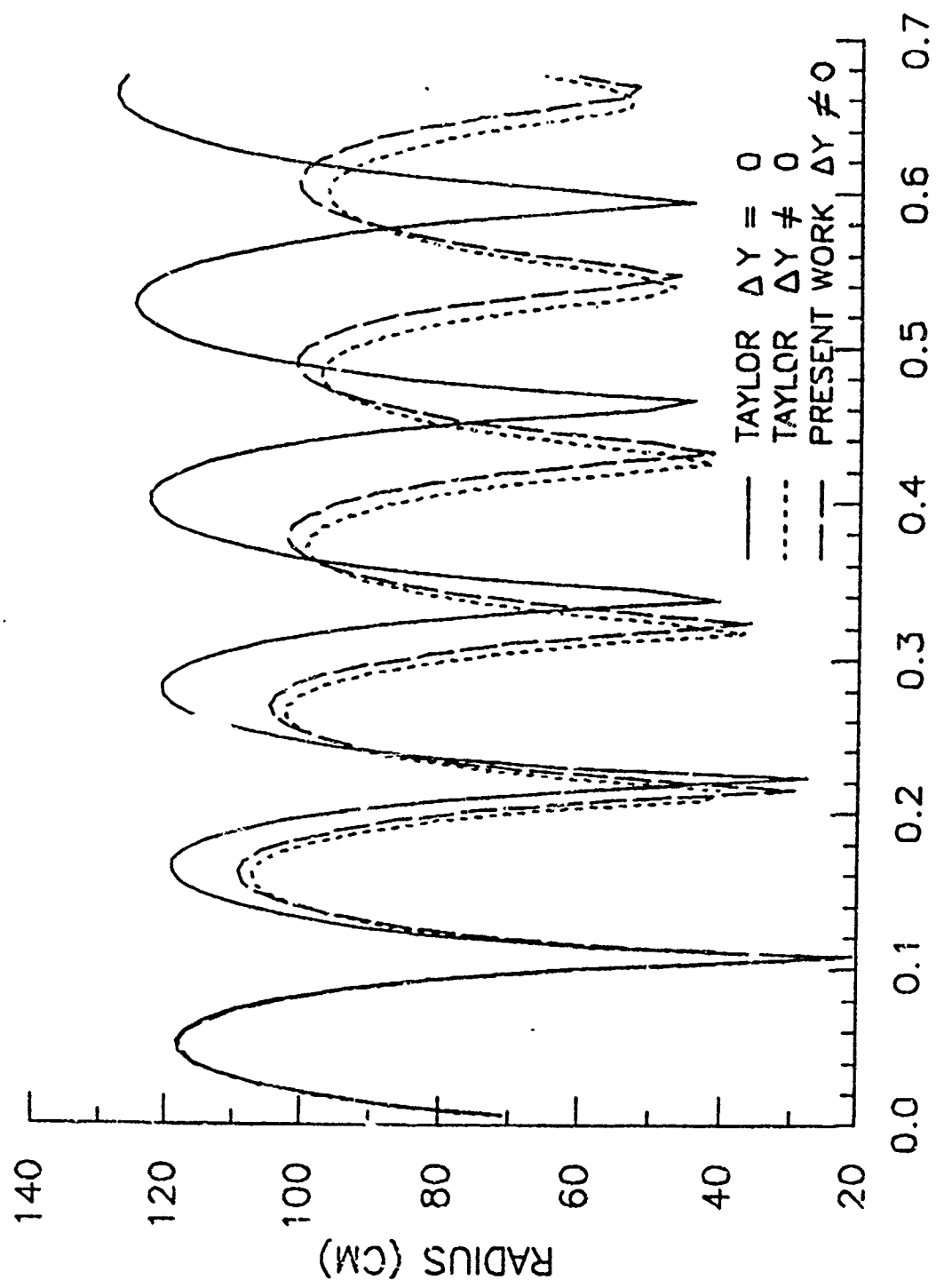

RADIUS OF BUBBLE
RADIUS (CM)
140
120
100
80
60
40
20
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)
TAYLOR ΔY = 0
TAYLOR ΔY ≠ 0
PRESENT WORK ΔY ≠0

Figure 12

Figure 13

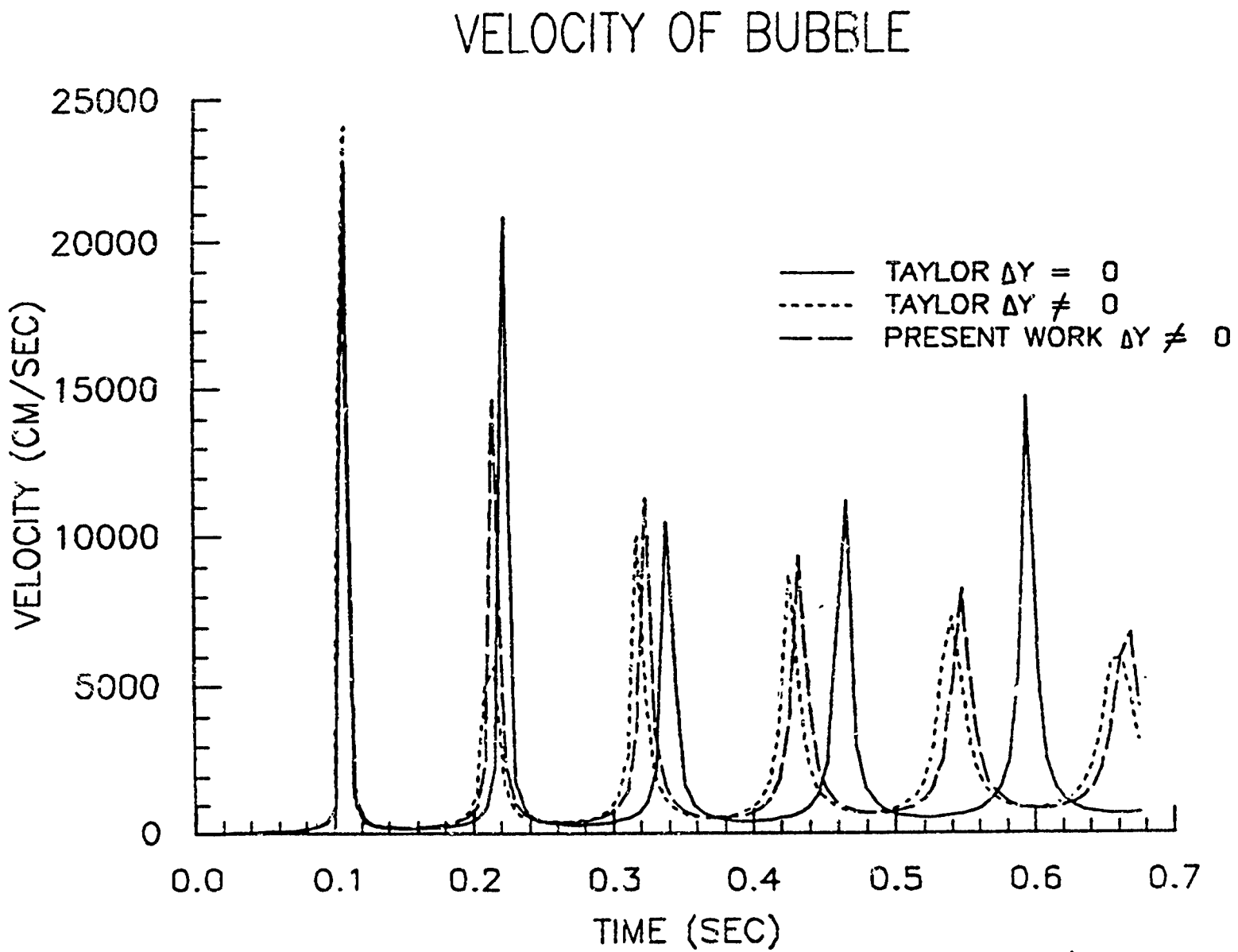

VELOCITY OF BUBBLE
25000
20000
15000
10000
5000
0
VELOCITY (CM/SEC)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)
TAYLOR ΔY = 0
TAYLOR ΔY ≠ 0
PRESENT WORK ΔY ≠ 0

Figure 14

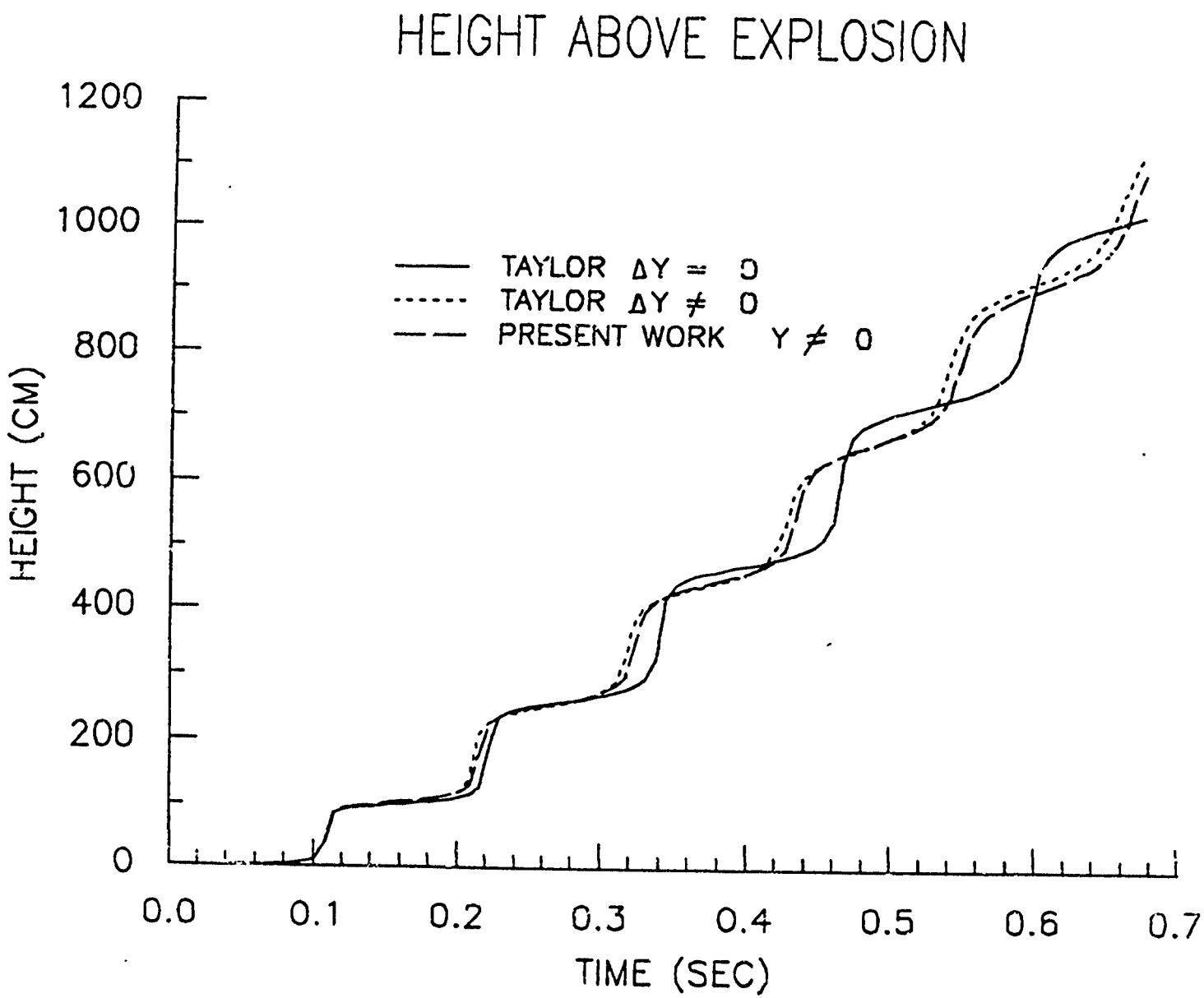

HEIGHT ABOVE EXPLOSION
1200
1000
800
600
400
200
0
HEIGHT (CM)
TAYLOR ΔY = 0
TAYLOR ΔY ≠ 0
PRESENT WORK Y ≠ 0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 15

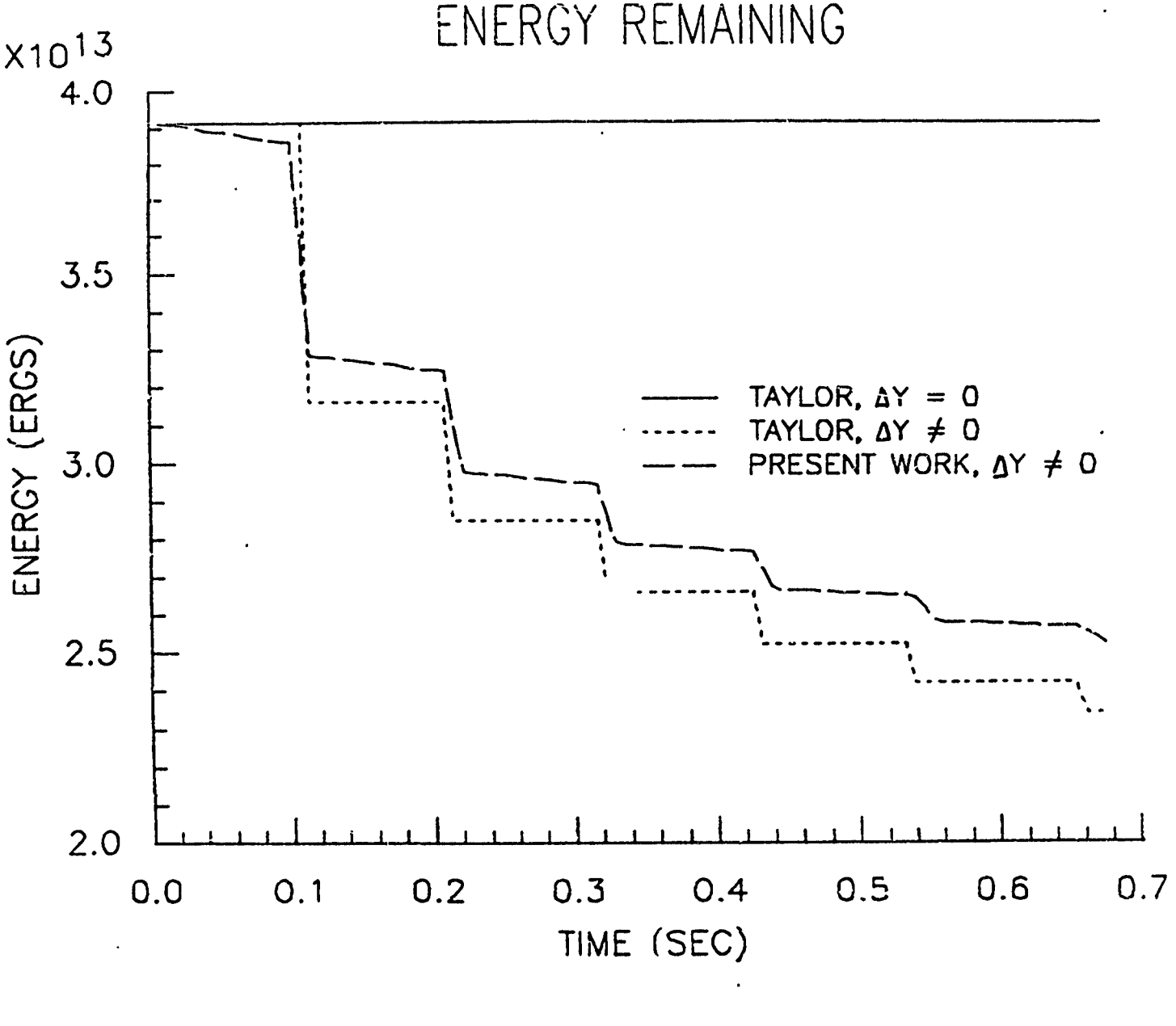

ENERGY REMAINING
X10^13
4.0
3.5
3.0
2.5
2.0
ENERGY (ERGS)
TAYLOR, ΔY = 0
TAYLOR, ΔY ≠ 0
PRESENT WORK, ΔY ≠ 0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 16

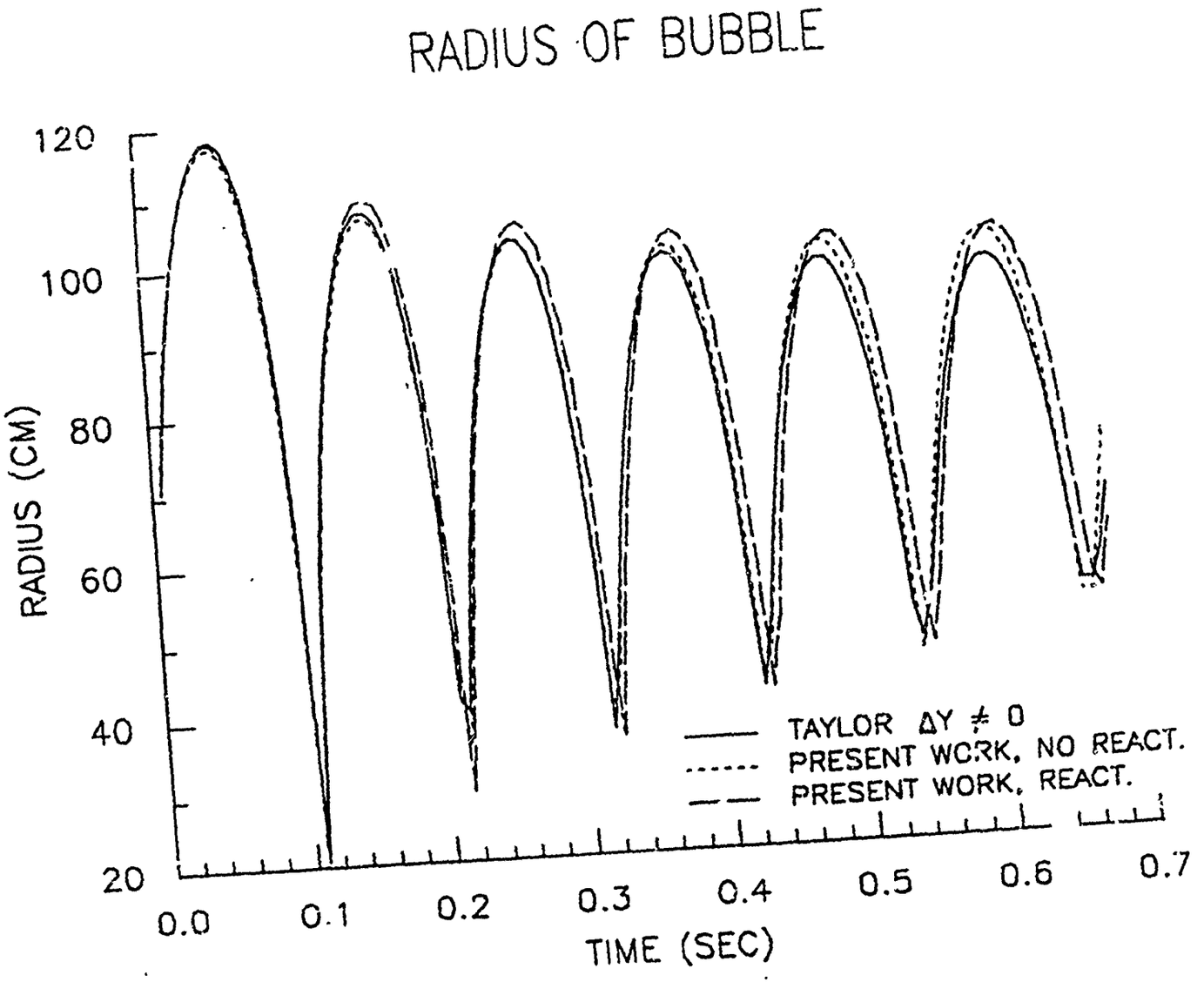
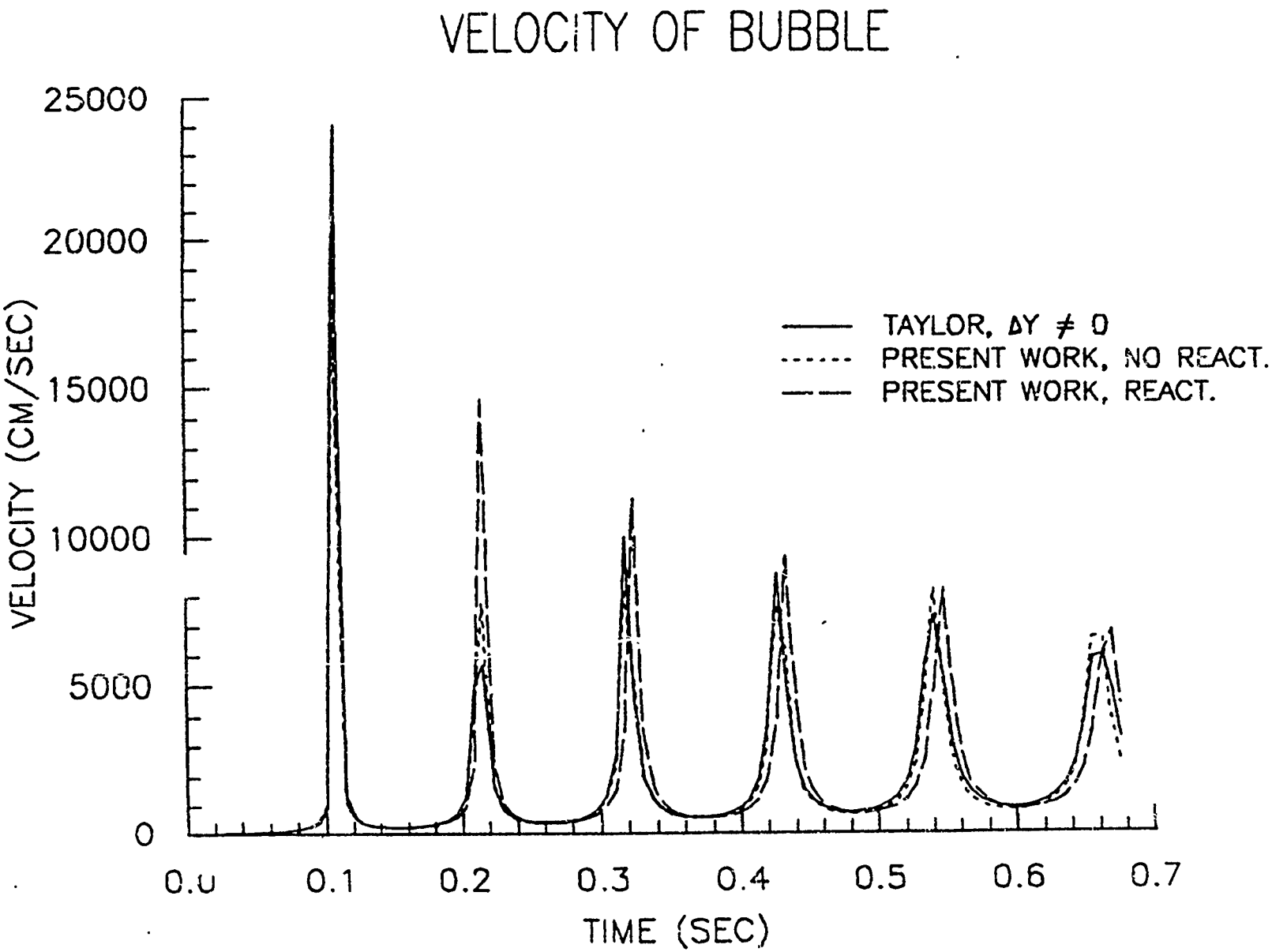

RADIUS OF BUBBLE
RADIUS (CM)
120
100
80
60
40
20
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)
TAYLOR ΔY ≠ 0
PRESENT WORK, NO REACT.
PRESENT WORK, REACT.

Figure 17




570

Figure 18

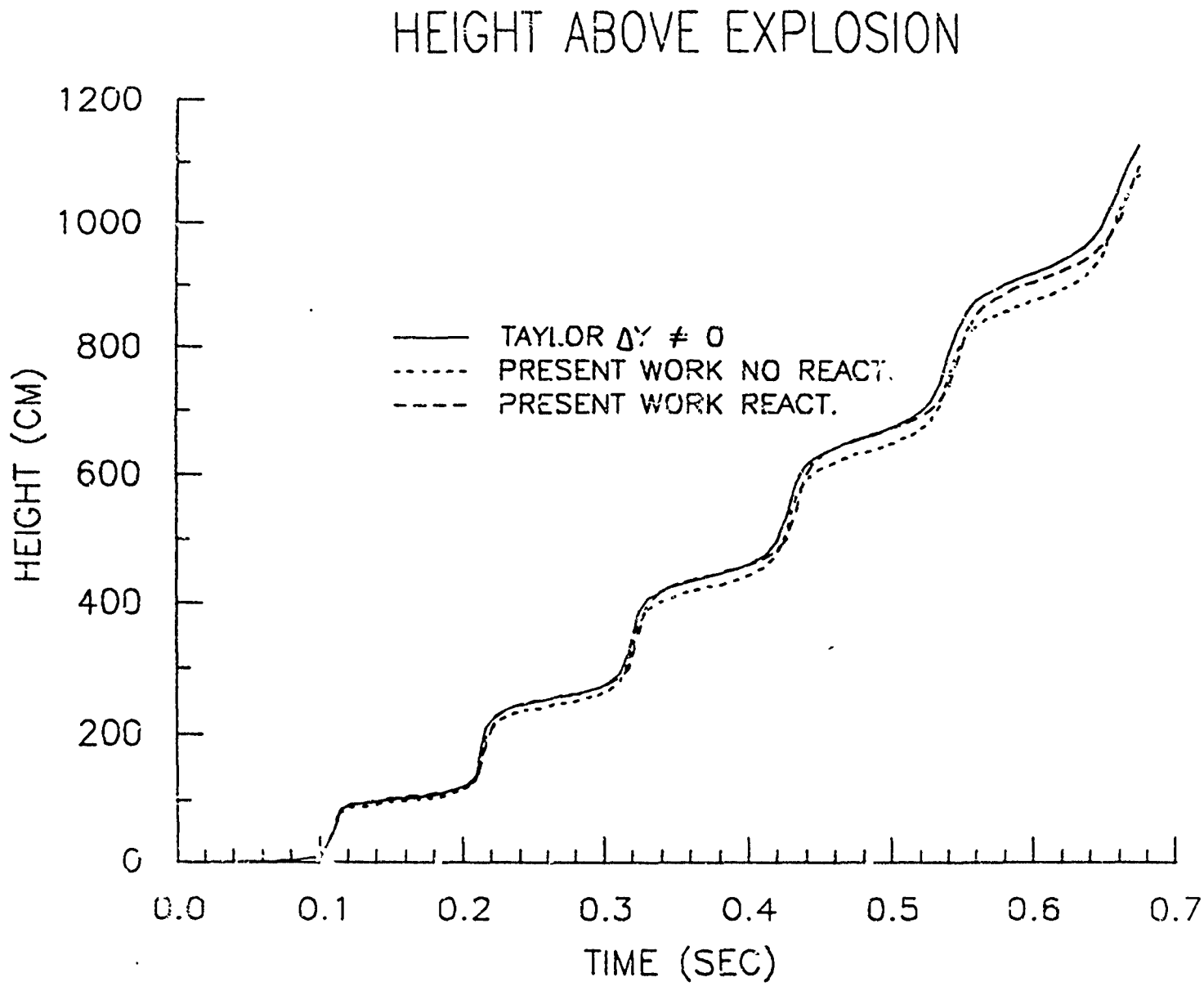

HEIGHT ABOVE EXPLOSION
1200
1000
800
600
400
200
0
HEIGHT (CM)
TAYLOR ΔY ≠ 0
PRESENT WORK NO REACT.
PRESENT WORK REACT.
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 19

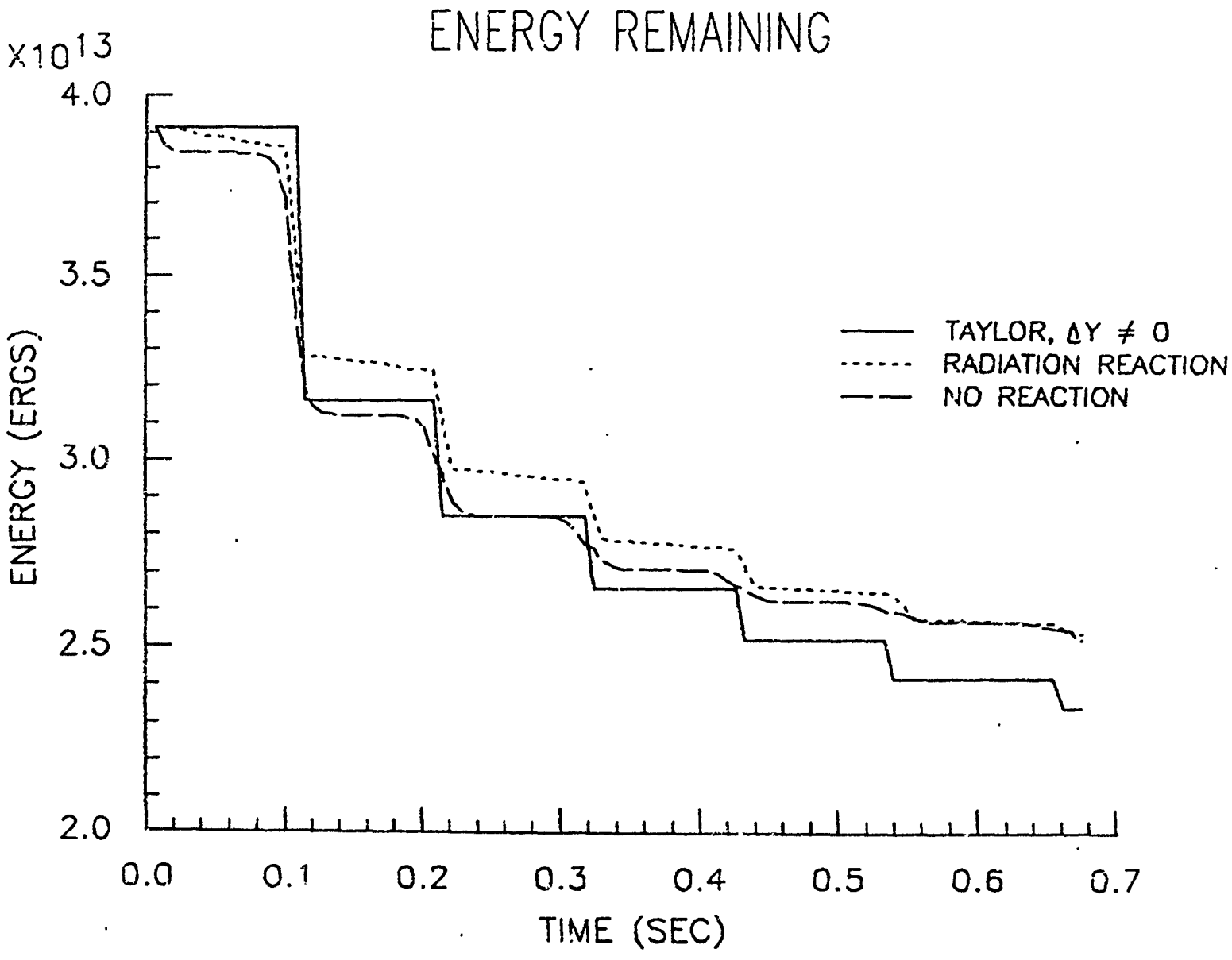

ENERGY REMAINING
X10^13
4.0
3.5
3.0
2.5
2.0
ENERGY (ERGS)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)
TAYLOR, ΔY ≠ 0
RADIATION REACTION
NO REACTION

Figure 20

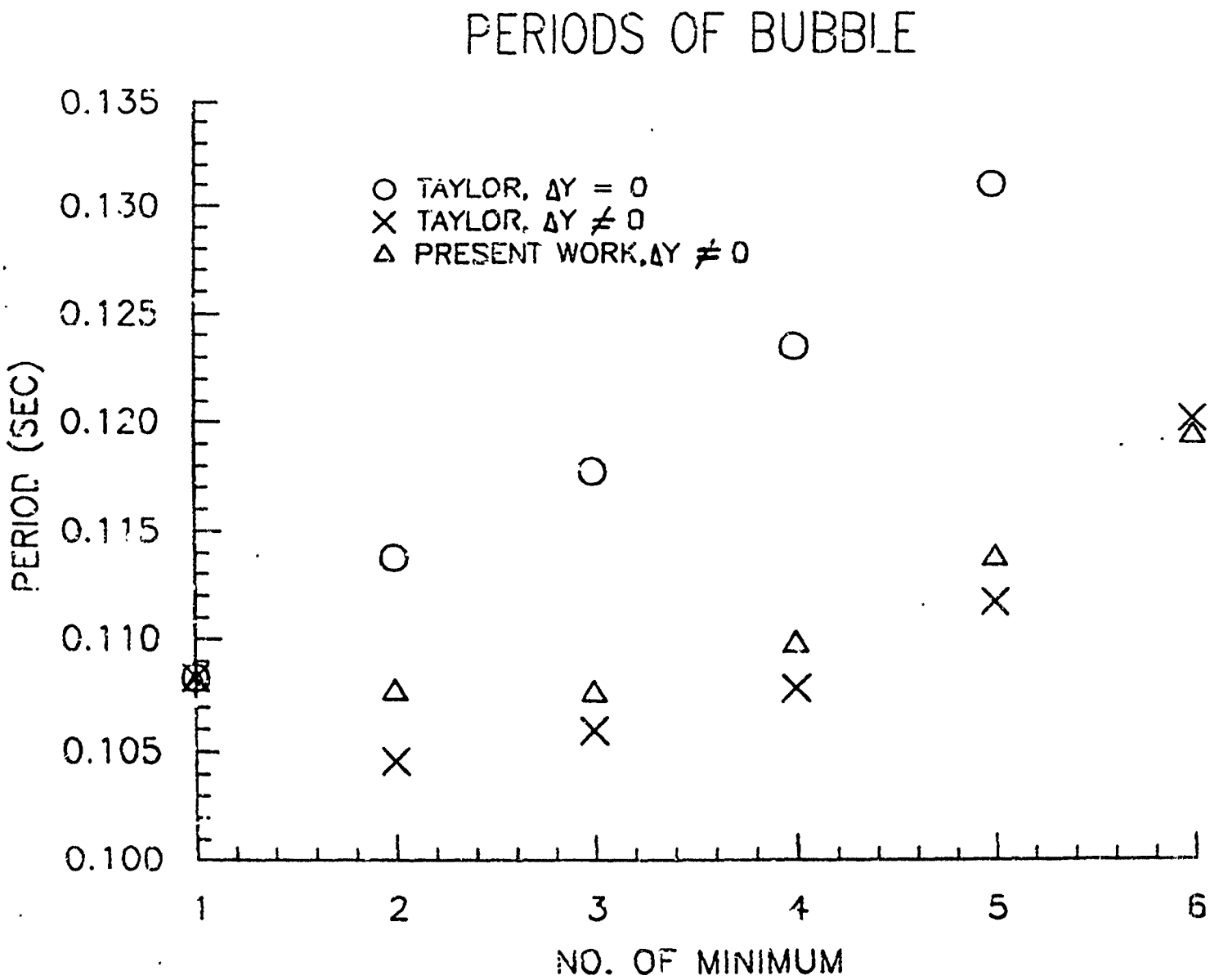

PERIODS OF BUBBLE
O TAYLOR, ΔY = 0
X TAYLOR, ΔY ≠ 0
Δ PRESENT WORK, ΔY ≠ 0
PERIOD (SEC)
0.135
0.130
0.125
0.120
0.115
0.110
0.105
0.100
1
2
3
4
5
6
NO. OF MINIMUM

Figure 21

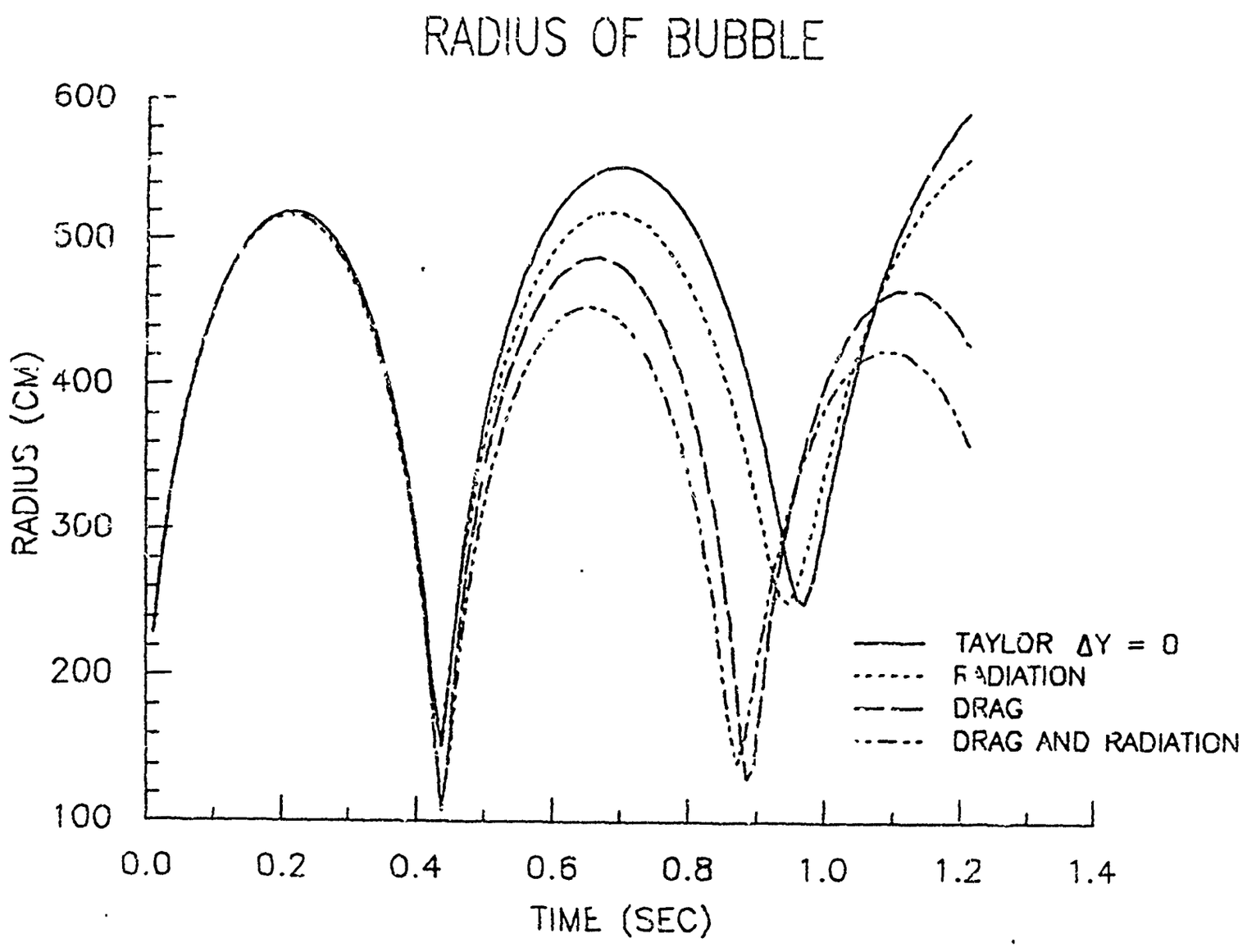

RADIUS OF BUBBLE
600
500
400
300
200
100
RADIUS (CM)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
TIME (SEC)
TAYLOR ΔY = 0
RADIATION
DRAG
DRAG AND RADIATION

Figure 22

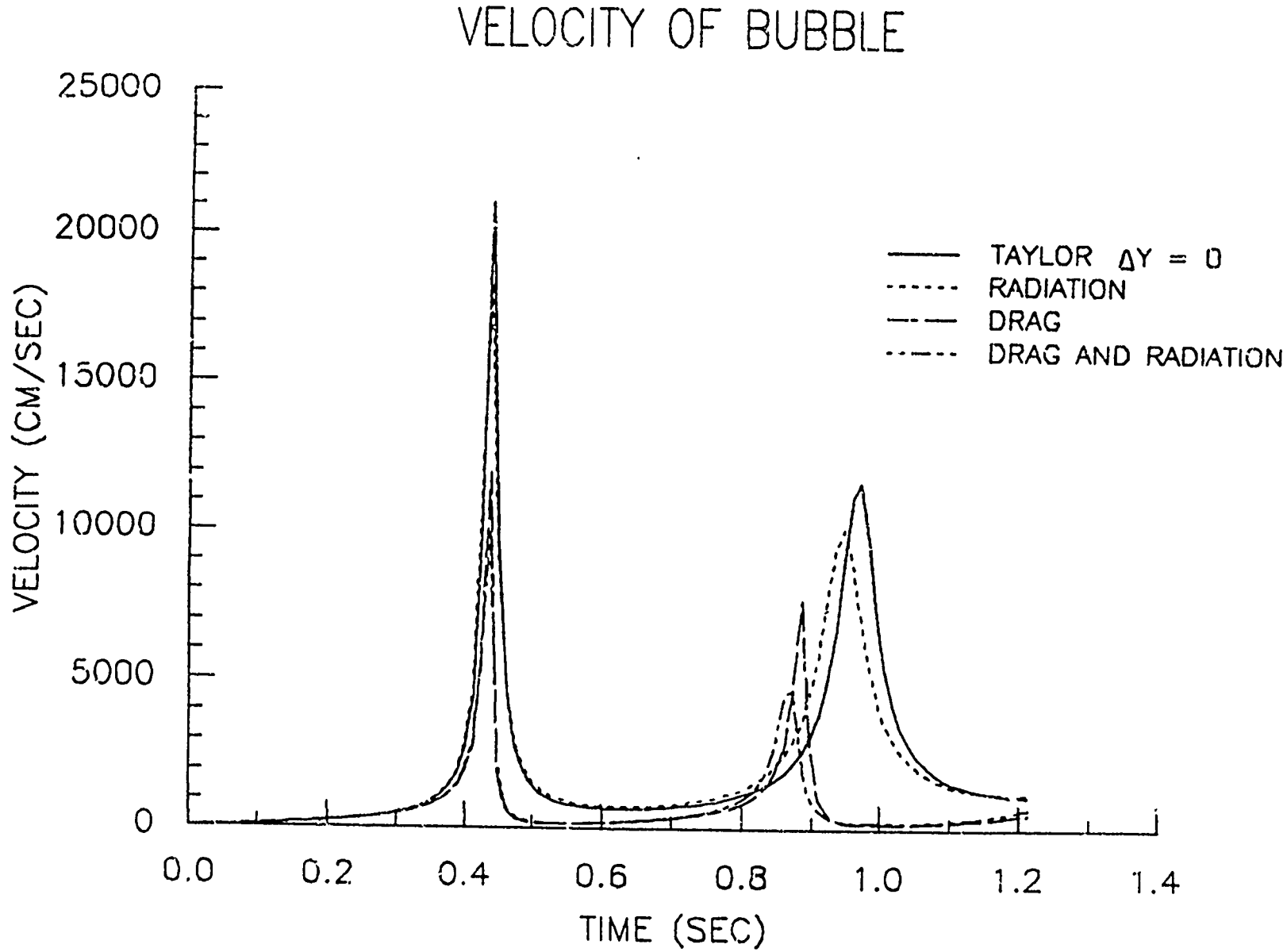

VELOCITY OF BUBBLE
25000
20000
15000
10000
5000
0
VELOCITY (CM/SEC)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
TIME (SEC)
TAYLOR ΔY = 0
RADIATION
DRAG
DRAG AND RADIATION

Figure 23

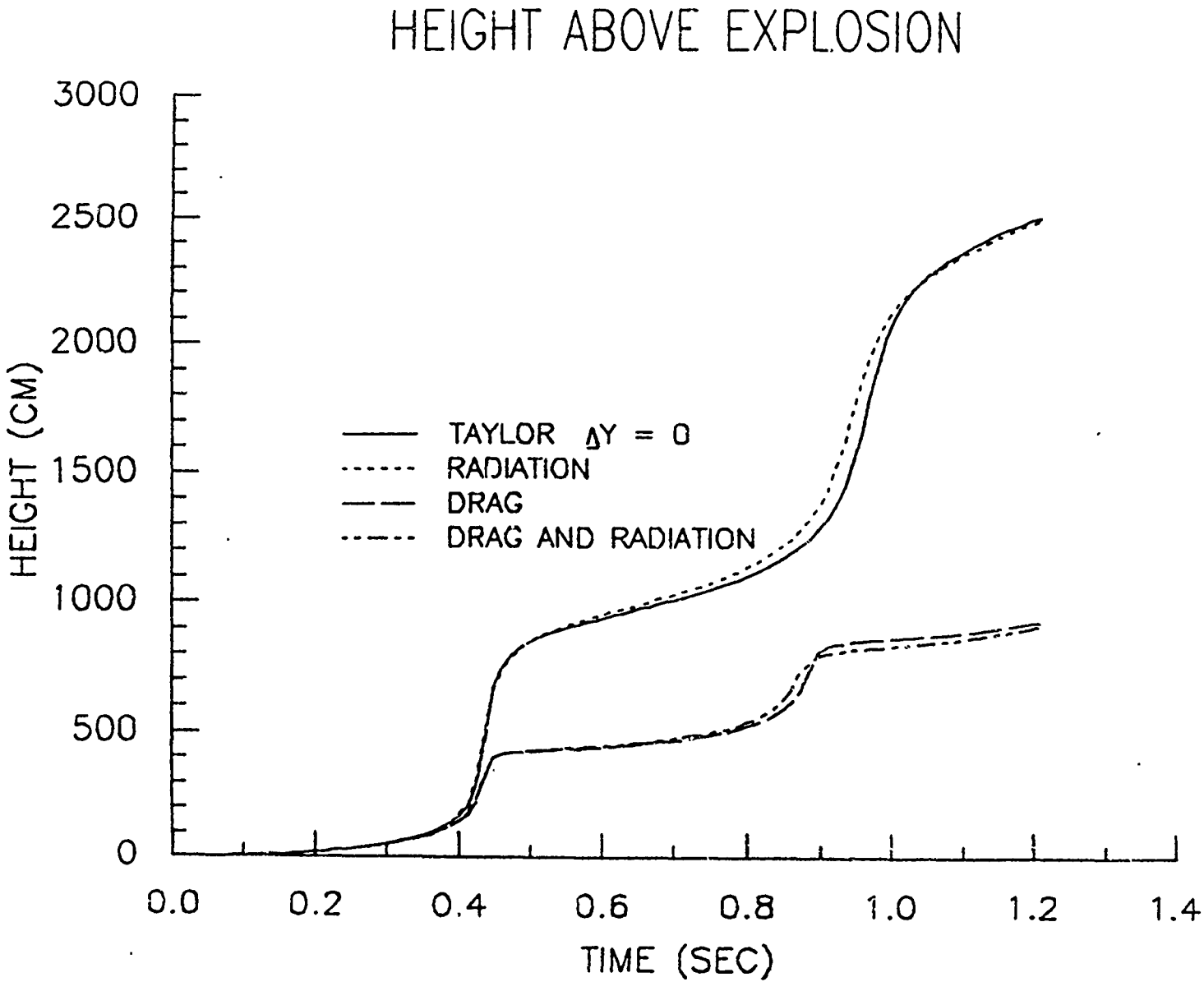
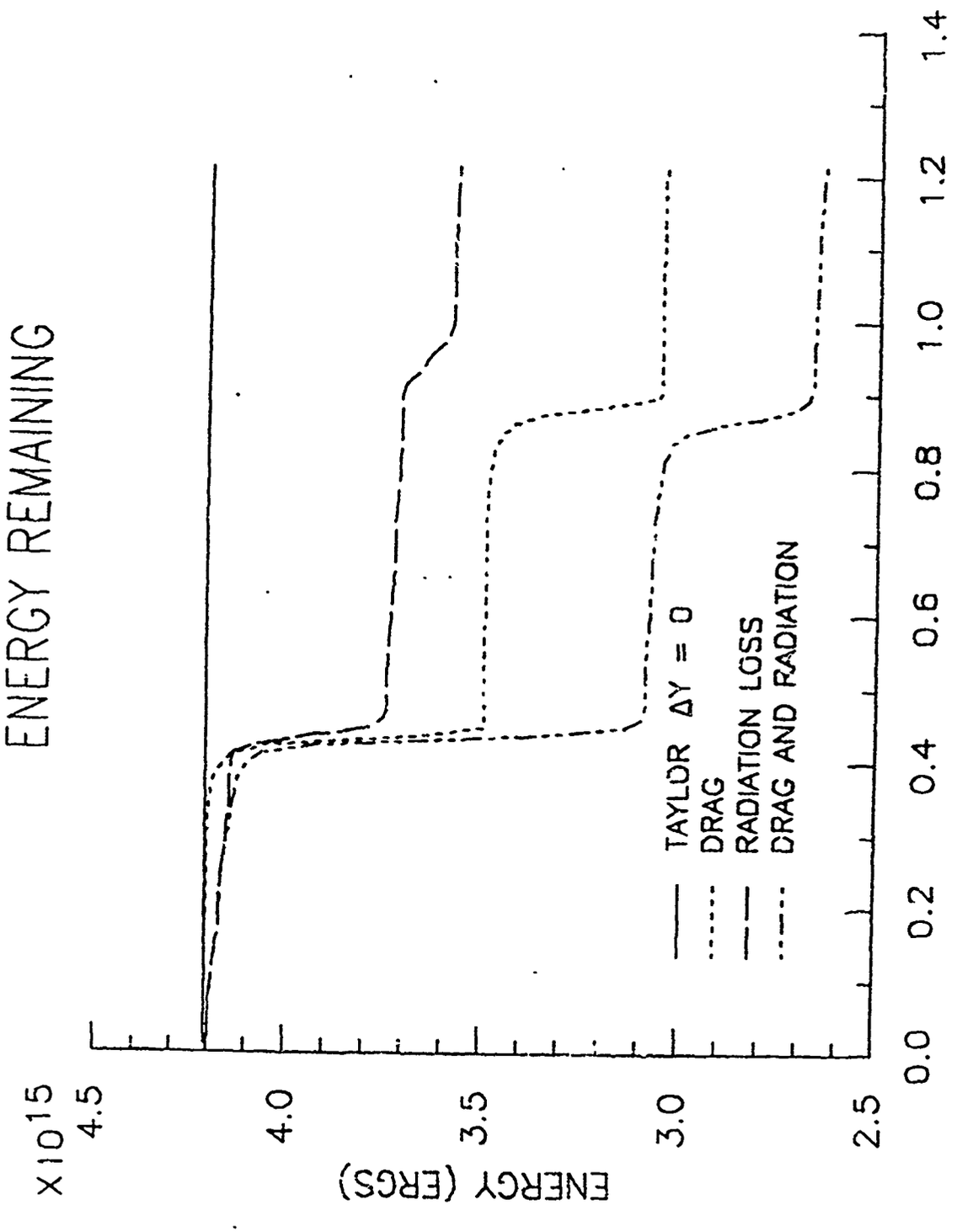

HEIGHT ABOVE EXPLOSION
HEIGHT (CM)
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
TIME (SEC)
TAYLOR ΔY = 0
RADIATION
DRAG
DRAG AND RADIATION

Figure 24


ENERGY REMAINING
X 10^15
4.5
4.0
3.5
3.0
2.5
ENERGY (ERGS)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
TIME (SEC)
TAYLOR ΔY = 0
DRAG
RADIATION LOSS
DRAG AND RADIATION

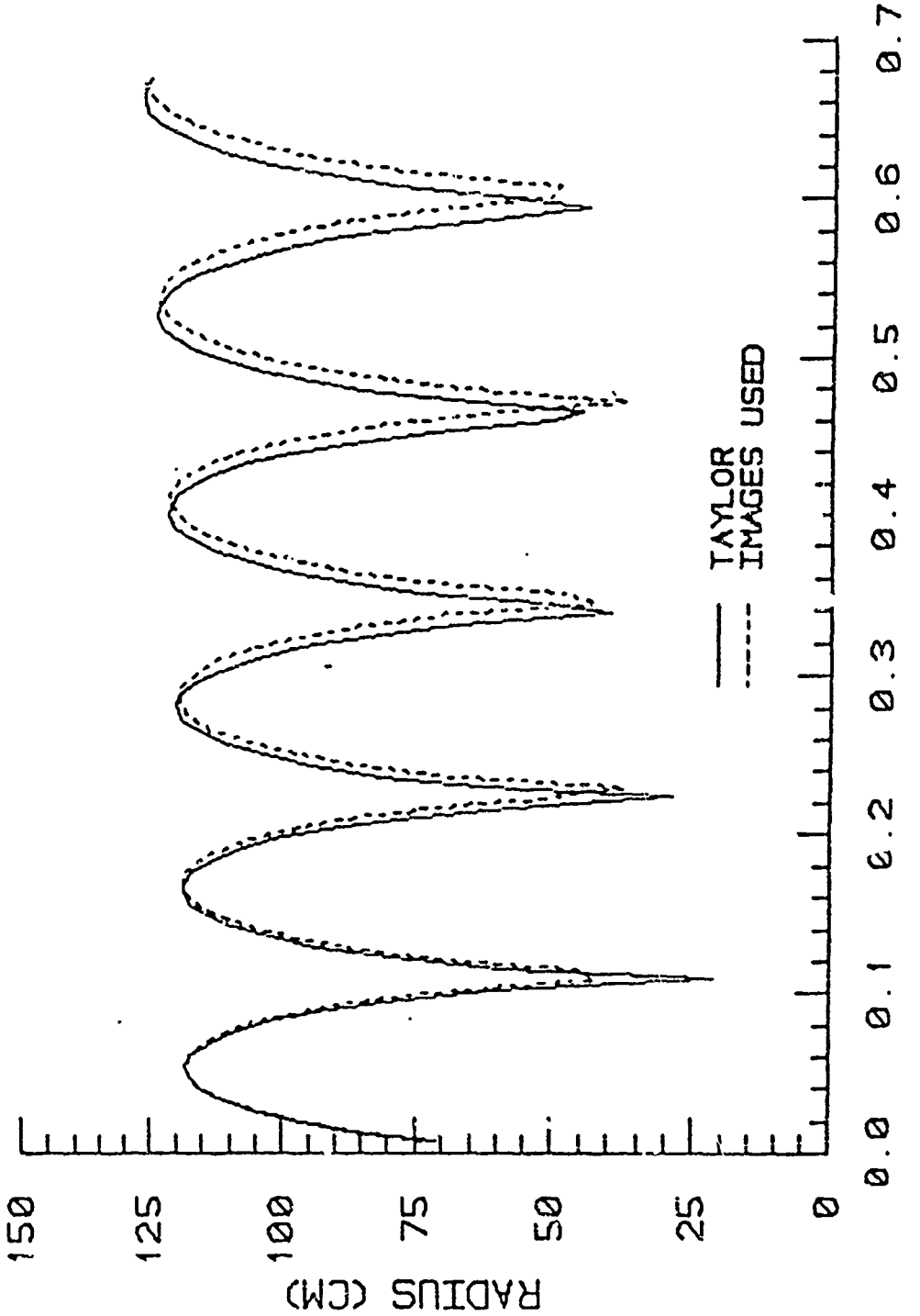

RADIUS OF BUBBLE
RADIUS (CM)
150
125
100
75
50
25
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)
TAYLOR
IMAGES USED

Figure 25

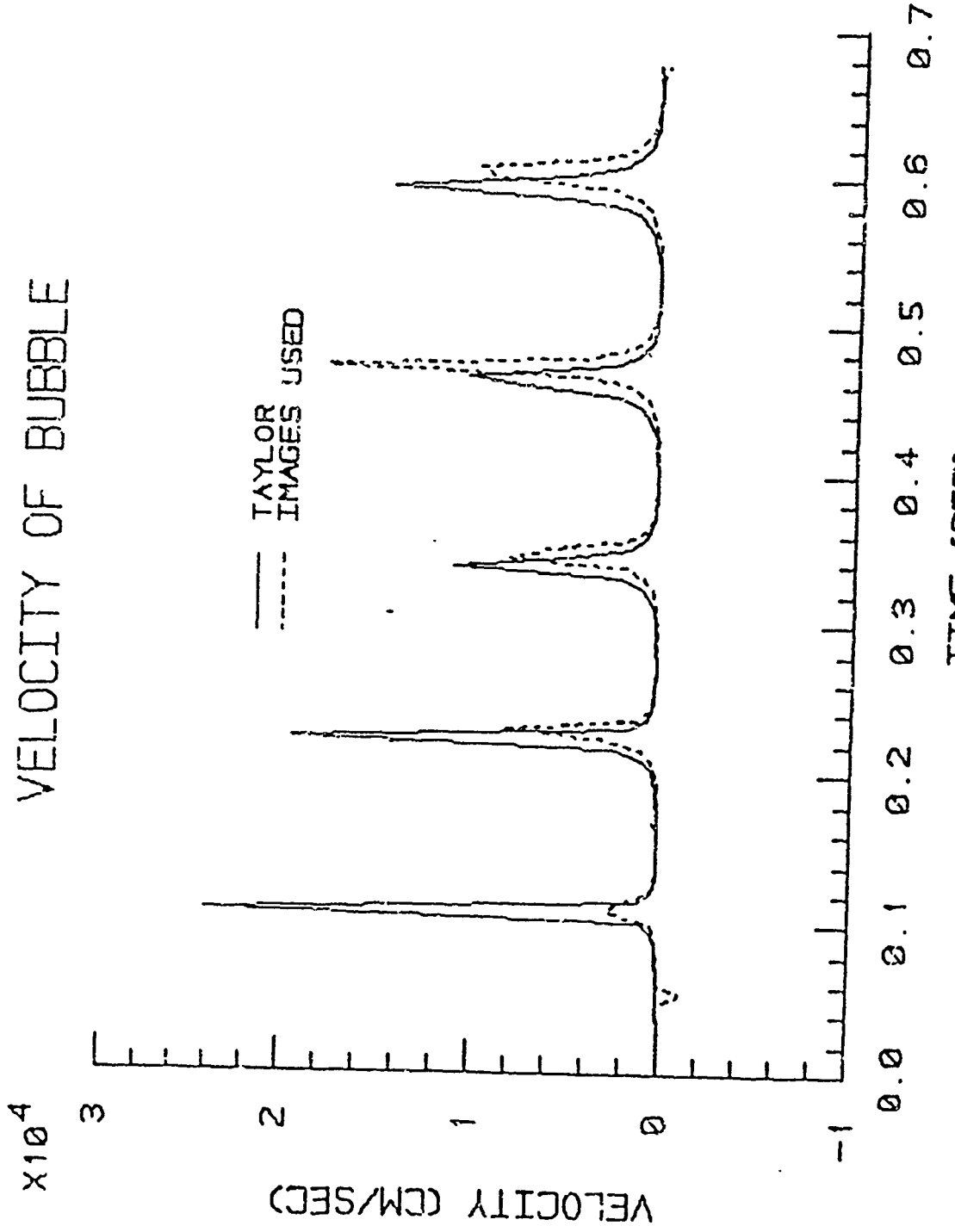

VELOCITY OF BUBBLE
TAYLOR
IMAGES USED
X10^4
3
2
1
0
-1
VELOCITY (CM/SEC)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 26

Figure 27

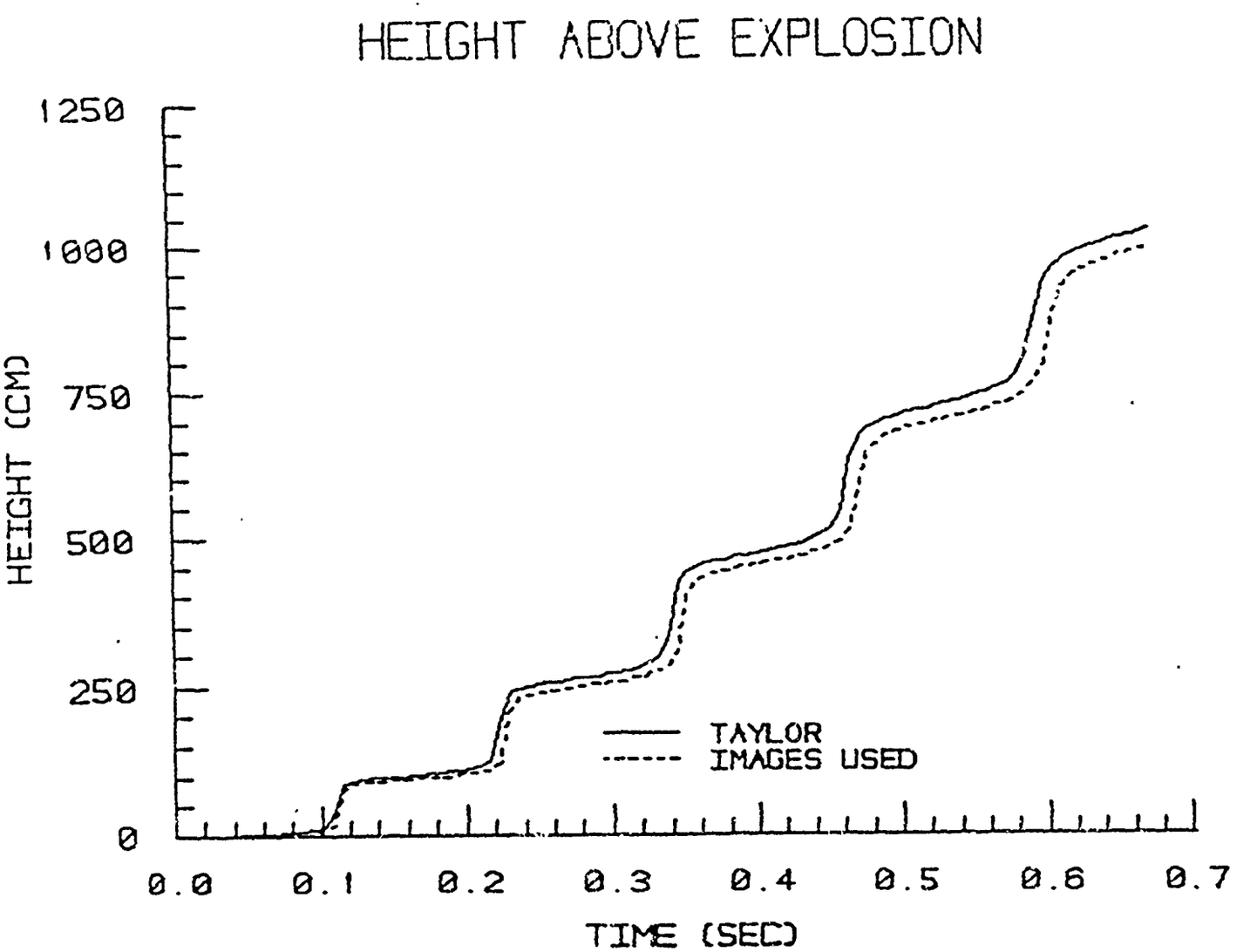
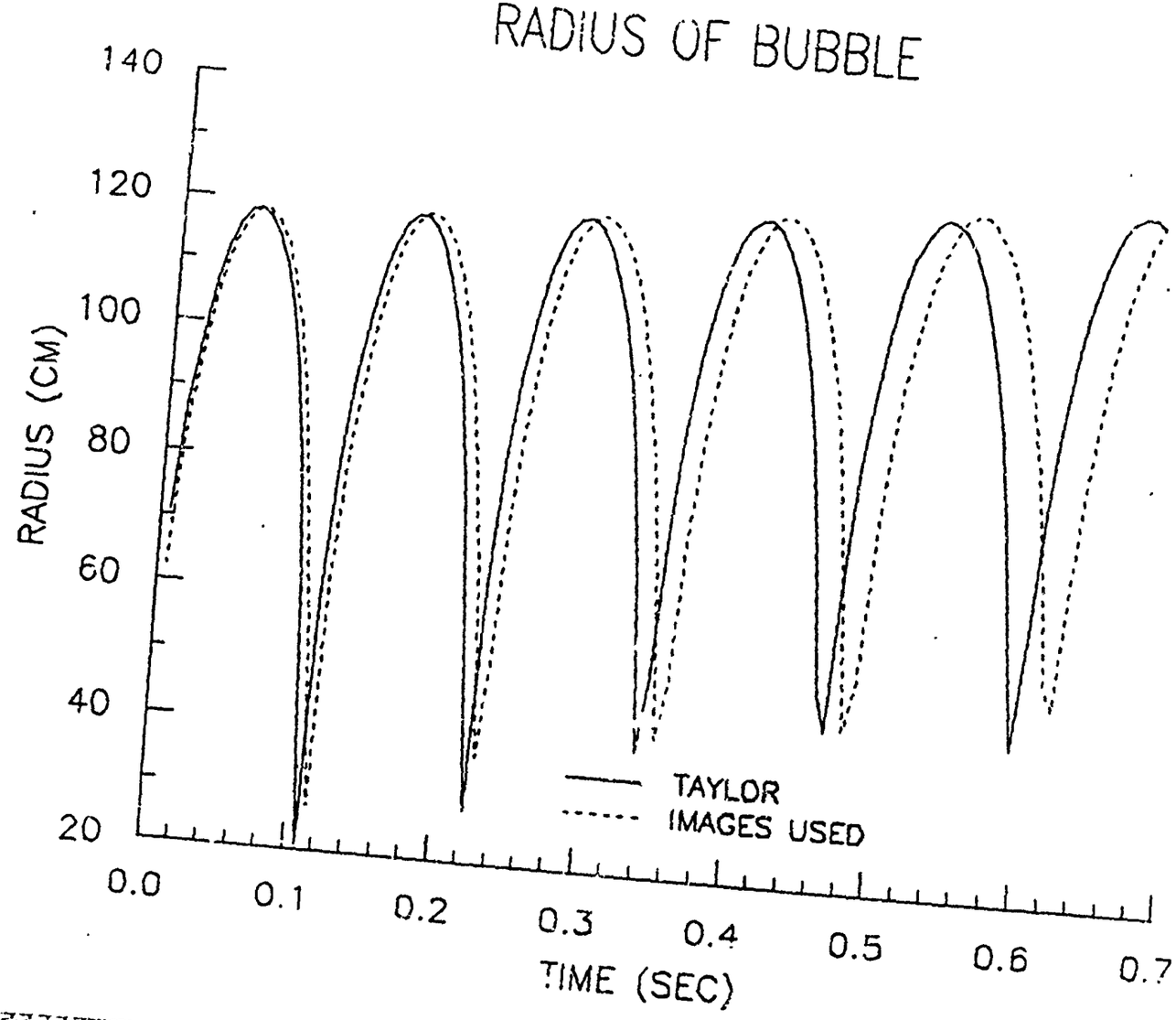

HEIGHT ABOVE EXPLOSION
1250
1000
750
500
250
0
HEIGHT (CM)
TAYLOR
IMAGES USED
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 28




581

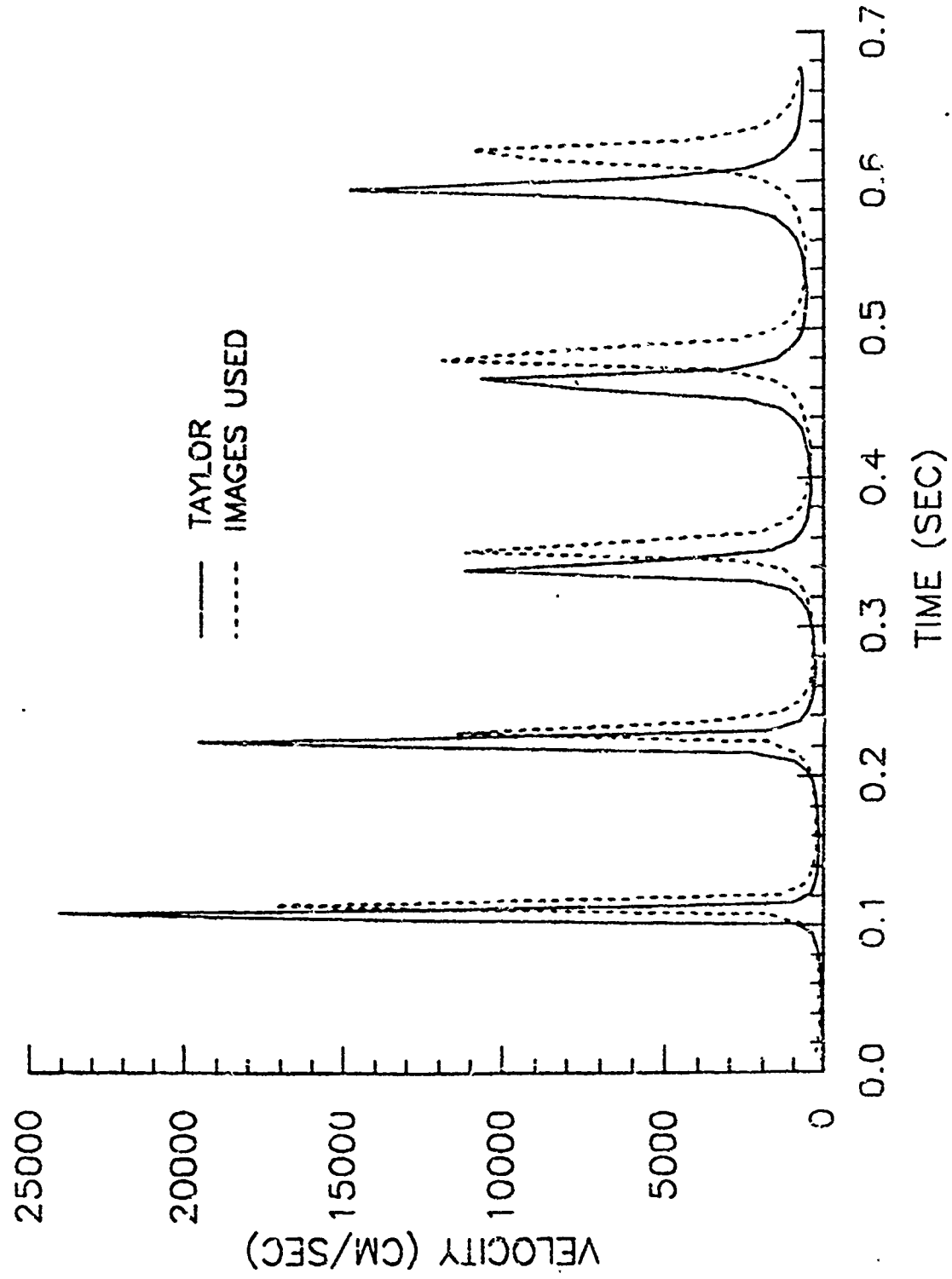
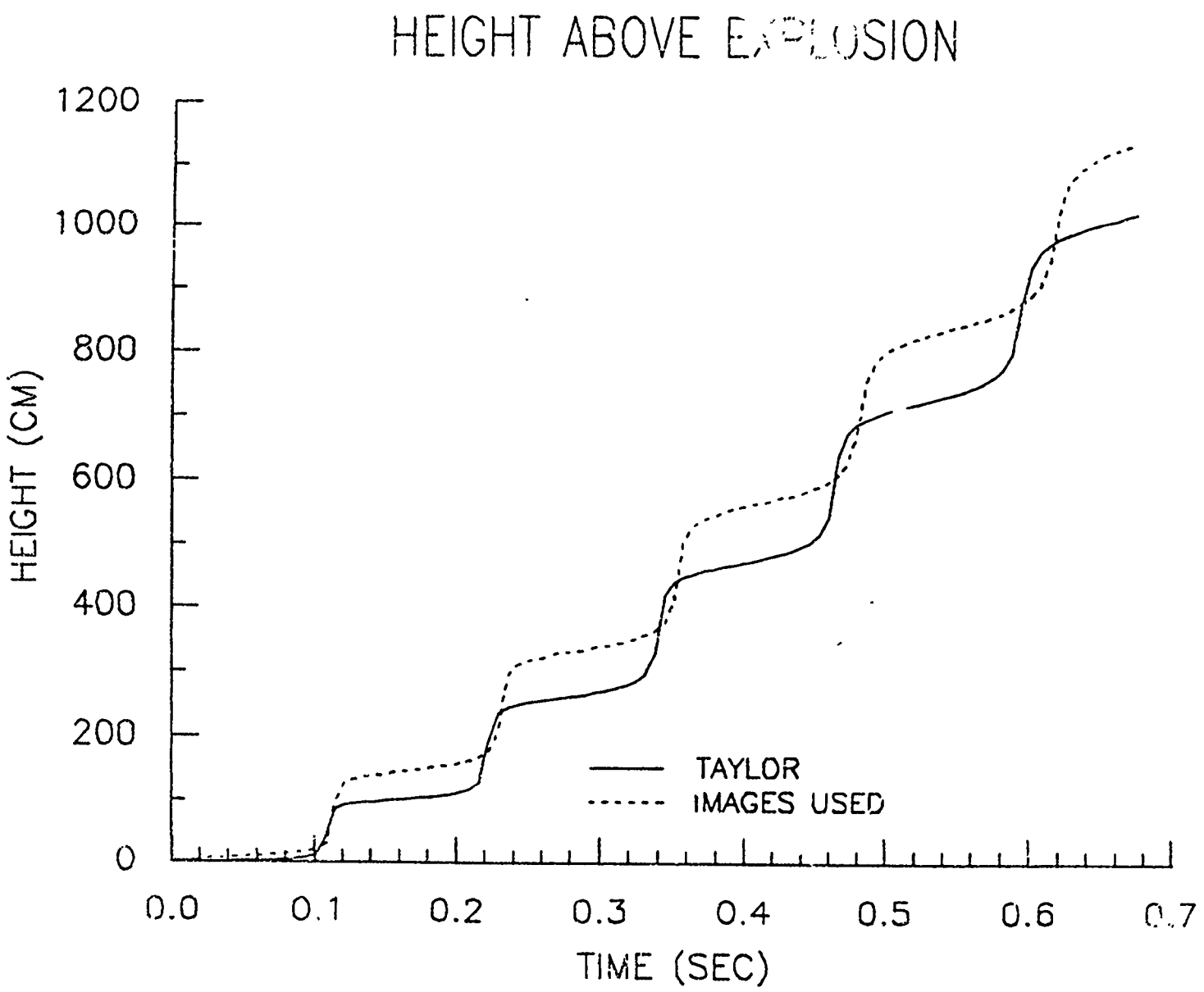

VELOCITY OF BUBBLE
TAYLOR
IMAGES USED
VELOCITY (CM/SEC)
25000
20000
15000
10000
5000
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

Figure 29

Figure 30


HEIGHT ABOVE EXPLOSION
1200
1000
800
600
400
200
0
HEIGHT (CM)
TAYLOR
IMAGES USED
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
TIME (SEC)

# A LOCAL REFINEMENT FINITE ELEMENT METHOD FOR TIME DEPENDENT PARTIAL DIFFERENTIAL EQUATIONS [1]

Joseph E. Flaherty
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12181

and

U. S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Calibre Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189-5000

and

Peter K. Moore
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12181

ABSTRACT. We discuss an adaptive local refinement finite element method for solving initial-boundary value problems for vector systems of partial differential equations in one space dimension and time. The method uses piecewise bilinear rectangular space-time finite elements. For each time step, grids are automatically added to regions where the local discretization error is estimated as being larger than a prescribed tolerance. We discuss several aspects of our algorithm, including the tree structure that is used to represent the finite element solution and grids, an error estimation technique, and initial and boundary conditions at coarse-fine mesh interfaces. We also present computational results for a simple linear hyperbolic problem, a problem involving Burgers' equation, and a model combustion problem.

1. INTRODUCTION. There is an ever increasing need to solve problems of greater complexity and a corresponding need for reliable and robust software tools to accurately and efficiently describe the phenomena. Adaptive techniques are good candidates for providing the computational methods and codes necessary to solve some of these difficult problems. Two popular adaptive techniques are: (i) moving mesh methods, where a grid of a fixed number of finite difference cells or finite elements is moved so as to follow and resolve local nonuniformities in the solution, and (ii) local refinement methods, where uniform fine grids are added to coarser grids in regions where the solution is not adequately resolved. A representative sample of both types of methods is contained in Babuska, Chandra, and

---

[1] The authors were partially supported by the U. S. Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant Number AFOSR 80-0192 and the U. S. Army Research Office under Contract Number DAAG29-82-K-0197.

PREVIOUS PAGE IS BLANK

Flaherty [2]. Recently, Adjerid and Flaherty [1] developed a finite element method that combines r esh moving and refinement.

Herein, we discuss a local refinement finite element procedure for finding numerical solutions of M-dimensional vector systems of partial differential equations having the form

$$Lu := u_t + f(x,t,u,u_x) - [D(x,t,u)u_x]_x = 0,$$

$$a < x < b,\ t > 0, \qquad (1.1)$$

subject to the initial conditions

$$u(x,0) = u^0(x)\ ,\ a \le x \le b\ , \qquad (1.2)$$

and appropriate boundary conditions so that the problem has a well posed solution.

We discretize (1.1,2) for a time step using a finite element-Galerkin procedure with piecewise bilinear approximations on a rectangular space-time net. At the end of each time step we estimate the local discretization error, add finer subgrids of space-time elements in regions of high error, and recursively solve the problem again in these regions. The process terminates when the error estimate on each grid is less than a prescribed tolerance. The original coarse space-time grid is then carried forward for the next time step and the strategy is repeated. Our algorithm is discussed further in Flaherty and Moore [9] and some of this discussion is repeated in Section 2.

Berger [3] used a similar local refinement procedure to solve one-and two-dimensional hyperbolic systems. She used explicit finite difference schemes to discretize the partial differential equations, while we use implicit finite element techniques since we are primarily interested in parabolic problems.

In addition to the discretization technique, the major numerical questions that must be answered as part of the development of a local refinement code are (i) the estimation of the discretization error and (ii) the appropriate initial and boundary conditions to apply at coarse-fine mesh interfaces. Of course, Computer Science questions, such as which language to use to describe and implement the various algorithms and what data structures to use to represent and store the grids and solutions must also be answered. Our work in all of these areas is still far from complete and herein we only discuss our progress and thoughts on error estimation techniques, data structures, and interface conditions (cf. Section 2). In Section 3, we present the results of three examples that illustrate our method and the discussion of Section 2, and in Section 4, we present some preliminary conclusions and future plans.

2. FINITE ELEMENT ALGORITHM. We discretize equation (1.1) on a strip $\alpha < x < \beta$, $p < t < q$ using a finite element-Galerkin method with a uniform grid of N rectangular elements of size $(\beta - \alpha)/N$ by $(q - p)$. We refer to this grid as $R(\alpha,\beta,p,q,N,f,s)$, where f and s are pointers to the father and son grids discussed later. Each grid uses records to store the appropriate information.

We generate the discrete system on $R(\alpha,\beta,p,q,N,f,s)$ in the usual manner; thus, we approximate u by $U(x,t)$ and select test functions $V(x,t)$, where U and V are elements of a space of $C^0$ bilinear polynomials with respect to the grid R. We then take the inner product of equation (1.1) and V, replace u by U, and integrate any diffusive terms by parts to obtain

$$\int_R [V^T U_t + V^T f(x,t,U,U_x) + V_x^T D(x,t,U)U_x]dxdt$$

$$-\int_p^q V^T D(x,t,U)U_x|_\alpha^\beta dt = 0 . \qquad (2.2)$$

Equation (2.2) must vanish for all bilinear functions V on the grid R. The integrals are approximated using a four point Gauss quadrature rule and the resulting nonlinear system is solved by Newton iteration (cf.,e.g., [7] for additional details). Appropriate initial and boundary conditions for (2.2) are discussed later in this section.

We describe our local refinement procedure for solving problem (1.1,2) for one time step $(t^0,t^1)$ on a coarse grid with $N^0$ elements, i.e, on $R(a,b,t^0,t^1,N^0,0,s)$ (where the pointer $f = 0$ signifies that this grid has no father). To solve this problem we simply call the procedure "locref" with the arguments $R(a,b,t^0,t^1,N^0,0,s)$, tol, tsub for each coarse grid time interval. A pseudo-PASCAL description of the procedure "locref" is shown in Figure 1.

```
procedure locref (R(α,β,p,q,N,f,s), tol, tsub)
   begin
      Solve the finite element equations (2.2) on R(α,β,p,q,N,f,s);
      Estimate the error on R(α,β,p,q,N,f,s);
      if error > tol then
         begin
            calculate where error > tol and return the son grids;
            for j := 1 to tsub do
               for i := 1 to number of sons do
                  begin
                     p[j] := p + (j-1)*(q-p)/tsub;
                     q[j] := p[j] + (q-p)/tsub;
                     locref (R(α[i],β[i],p[j],q[j],N[i],
                        R(α,β,p,q,N,f,s),s[i],tol,tsub)
                  end
         end
   end;
```

Figure 1. Algorithm for local refinement solution of (1.1,2) on $R(\alpha,\beta,p,q,N,f,s)$ with an error tolerance of tol and dividing the local time step by tsub each time the error test is not satisfied.

The recursive algorithm locref sets up a tree structure of grids with $R(a,b,t^0,t^1,N^0,0,s)$ being the root node and with the solution being

generated by a preorder traversal of the tree at each local time step. For example, if the root grid is refined to give two subgrids and the time step is halved, then the problem is solved on the first subgrid on its first time step, then on the second subgrid on the same time step, then this procedure is repeated for the second time step. The error is estimated by Richardson extrapolation, i.e., the space and time steps are halved and the problem is solved again on this new grid. The two solutions that are obtained at each original grid point are used to generate an error estimate. If this pointwise estimate exceeds the tolerance "tol", finer grids are added as leaf nodes to the tree. This procedure is similar to one used by Berger [3]; however, there are more economical error estimation strategies (cf., e.g., Bieterman and Babuska [5, 6]) which we are currently investigating.

In order to solve the finite element system (2.2) we need to supply initial and boundary conditions. On any grid with $p = 0$, $\alpha = a$, or $\beta = b$ these can be obtained from the initial condition (1.2) or prescribed boundary conditions. However, artificial initial and boundary conditions must be created at all other coarse-fine mesh interfaces. This is a difficult and crucial problem that is discussed for explicit finite difference methods by Berger [3, 4]; however, it is largely unanswered for finite element applications. Instabilities or incorrect solutions (cf. Example 1 of Section 3) can result if inappropriate conditions are specified.

For initial conditions, two strategies immediately come to mind: (i) saving all fine grid data for propagation in time or (ii) interpolating the best coarse grid data to finer grids. We consider a blend of the two strategies which consists of saving the fine grid data down to a given level $\lambda$ in the tree and subsequently interpolating for finer grids. Each grid in the first $\lambda$ levels either has a linked list of the initial data directly associated with it or uses an initial data list of an ancestor grid. To find the value of the solution at some new initial point, the coordinate of that point is sequentially compared to values in the linked list until an interval containing the point is found so that interpolation can be used. This is costly and we are investigating more efficient procedures that use the natural ordering that already exists. We used either piecewise linear interpolation or piecewise parabolic interpolation with shape preserving splines developed by McLaughlin [10]. For each grid in the first $\lambda$ levels of the tree, a linked list is created to store the initial data. We are studying several alternative ways of determining a proper value for $\lambda$.

At the present time, we prescribe internal Dirichlet boundary conditions by linearly interpolating from coarse to finer grids. A buffer zone of two elements is added to each end of regions of high error that do not intersect the boundaries $x = a$ and $b$. If two buffer zones overlap or are separated from one another by one element, the two grids are joined. Similarly, if the buffer is only one element away from either a or b, that element is added to the grid.

**3. NUMERICAL EXAMPLES.** An experimental code based on the algorithms in Section 2 has been written in FORTRAN-77. We are testing it on several examples, some of these follow and others are presented in [9]. All results were computed in double precision on an IBM 3081D computer.

Example 1. In order to illustrate the importance of adequately resolving initial conditions at each time step we solve the linear hyperbolic initial value problem

$$u_t + u_x = 0 ,$$

$$u(x,0) = u^0(x) = \begin{cases} (1/2)(\cos(20\pi(x-0.45)) - 1) , & \\ 0 , \text{ otherwise} & \end{cases} \qquad 0.35 < x < 0.75$$

We solve this problem for one coarse time step of $\Delta t = 0.05$, 10 elements on $0 < x < 1$, tol = 0.01. For small enough times the exact solution is $u^0(x-t)$. If initial conditions are interpolated from the coarse to the fine grid, the oscillations are missed and an incorrect solution is computed, possibly without a user realizing that there is anything wrong. However, saving initial values for the first 8 levels of the tree of grids calculates the correct solution to the prescribed accuracy. The incorrect and correct solutions are shown at $t = 0.05$ in Figure 2.

Example 2. We solve the following problem for Burgers' equation:

$$u_t + uu_x = du_{xx} , \quad 0 < x < 1 , \quad 0 < t < 1 ,$$

$$u(x,0) = \sin\pi x , \quad 0 < x < 1 ,$$

$$u(0,t) = u(1,t) = 0 , \quad t > 0 .$$

We choose $d = 0.00003$, a coarse grid of 10 elements and $\Delta t = 0.1$, and piecewise parabolic approximations for the initial conditions with $\lambda = 6$. It is well known, that the solution of this problem is a "pulse" that steepens as it travels to the right until it forms a shock layer at $x = 1$. After a time of $O(1/d)$ the pulse dissipates and the solution decays to zero. We solve this problem for tol = 0.01 and 0.001 and show the solutions at $t = 0.4$ in Figure 3. The solution with the cruder tolerance is exhibiting some oscillations that are within our bounds. These, however, are not visible when the finer tolerance is used to solve the problem.

Example 3. We solve the model combustion problem

$$u_t + u_x - 2e^u = u_{xx} , \quad 0 < x < 1 , \quad 0 < t < 1 ,$$

$$u(x,0) = 0 , \quad u(0,t)=0 , \quad u_x(1,t) = 0 .$$

The exponential nonlinearity is typical in combustion problems having Arrhenius chemical kinetics. However, in this case the solution develops a "hot spot" at $x = 1$ and becomes infinite when $t$ is approximately 0.85. We choose a coarse grid of 20 elements and $\Delta t = 0.05$, tol = 0.001, and $\lambda = 6$. In Figure 4 we show the computed solution $U(x,t)$ as a function of $x$ for $t = 0.05$, 0.6, and 0.8 and in Figure 5 we show the mesh that was used to solve the problem. We see that the mesh is initially concentrated in the region near $x = 0$ where the curvature of the solution is largest. As time progresses and the curvature diminishes, excessive refinement is not necessary. Finally, as the solution begins to "blow-up" our algorithm generates a fine mesh only in the region near $x = 1$.

4. DISCUSSION AND CONCLUSIONS. We have briefly described an adaptive local refinement algorithm for solving time dependent partial differential equations. Even though this is very much a working algorithm, and not a production code, we are very encouraged by the preliminary results. We are investigating several possible ways of improving the efficiency and robustness of our algorithm. These include adding higher order polynomial finite element approximations, adaptively changing the number of elements that are carried forward in the coarse grid at each coarse time step, how to select the appropriate buffer length, adaptively determining the optimal number of levels of initital conditions to keep at coarse-fine interfaces, and the best boundary conditions to apply at internal boundaries. We are encouraged by the performance of McLaughlin's [10] shape preserving parabolic splines; however, the entire area of interpolating from coarse to fine grids needs further study. We are also developing non-Dirichlet "natural" boundary conditions to use at coarse-fine mesh interfaces.

Finally, we are very interested in combining the moving mesh strategy of, e.g., [7, 8] with the present local refinement strategy and extending our methods to two and three dimensions.

## References


1. ADJERID, S. AND FLAHERTY, J. E., A Moving Finite Element Method for Time Dependent Partial Differential Equations with Error Estimation and Refinement, in preparation, 1984.

2. BABUSKA, I., CHANDRA, J., AND FLAHERTY, J. E. (Eds.), *Adaptive Computational Methods for Partial Differential Equations*, SIAM, Philadelphia, 1983.

3. BERGER, M. J., Adaptive mesh refinement for hyperbolic partial differential equations, Report No. STAN-CS-82-924, Department of Computer Science, Stanford University, 1982.

4. BERGER, M. J., Stability of interfaces with mesh refinement, Report No. 83-42, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, 1983.

5. BIETERMAN, M. AND BABUSKA, I., The finite elememt method for parabolic equations, I. a posteriori error estimation, *Numer. Math.*, Vol. 40, pp 339-371, 1982.

6. BIETERMAN, M. AND BABUSKA, I., The finite elememt method for parabolic equations, II. a posteriori error estimation and adaptive approach, *Numer. Math.*, Vol. 40, pp 373-406, 1982.

7. DAVIS, S. F. AND FLAHERTY, J. E., An adaptive finite element method for initial-boundary value problems for partial differential equations, *SIAM J. Sci. Stat. Comput.*, Vol. 3, pp. 6-27, 1982

8. FLAHERTY, J. E., COYLE, J. M., LUDWIG, R., AND DAVIS, S. F., Adaptive finite element methods for parabolic partial differential equations, *Adaptive Computational Methods for Partial Differential Equations*, Babuska, I., Chandra, J., and Flaherty, J. E. (Eds.), SIAM,

Philadelphia, 1983.

9. FLAHERTY, J. E. AND MOORE, P. K., An adaptive local refinement finite element method for parabolic partial differential equations, *Proc. of the International Conference on Accuracy Estimates and Adaptive Refinements in Finite Element Computations,* Technical University of Lisbon, Lisbon, Vol. 2, pp. 139-152, 1984.

10. McLAUGHLIN, H. W., Shape preserving planar interpolation: an algorithm, *IEEE Computer Graphics and Applics.,* Vol. 3, pp 58-67, 1983.

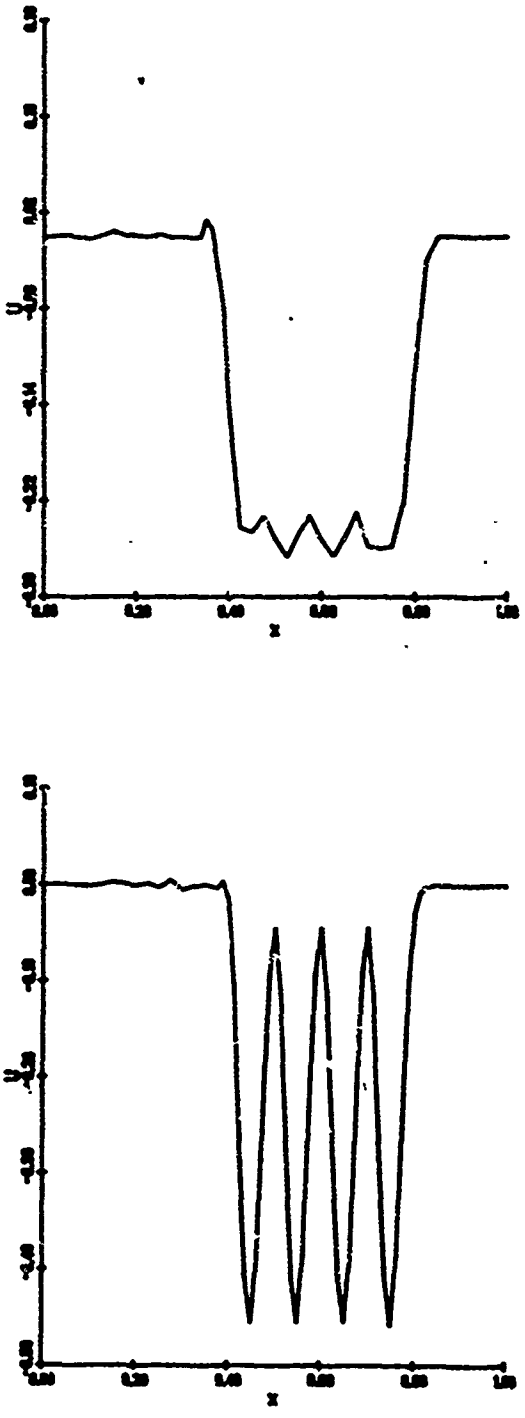


Figure 2. Solution of Example 1 at time t = 0.05 using interpolation from the coarse grid to the fine grid (top) and saving the initial values for the first 8 levels of the tree (bottom). The upper solution overlooks the oscillations and is incorrect.

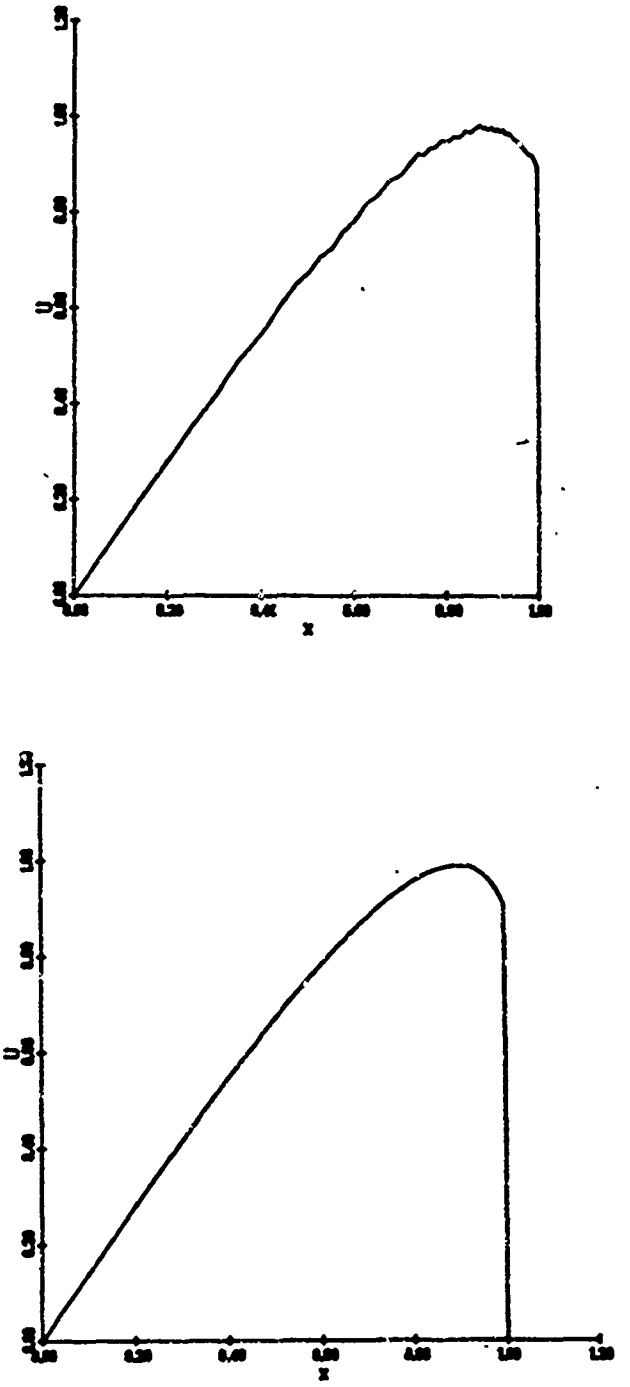


Figure 3. Solution of Example 2 at time t = 0.4 with toierances of 0.01 (top) and 0.001 (bottom).

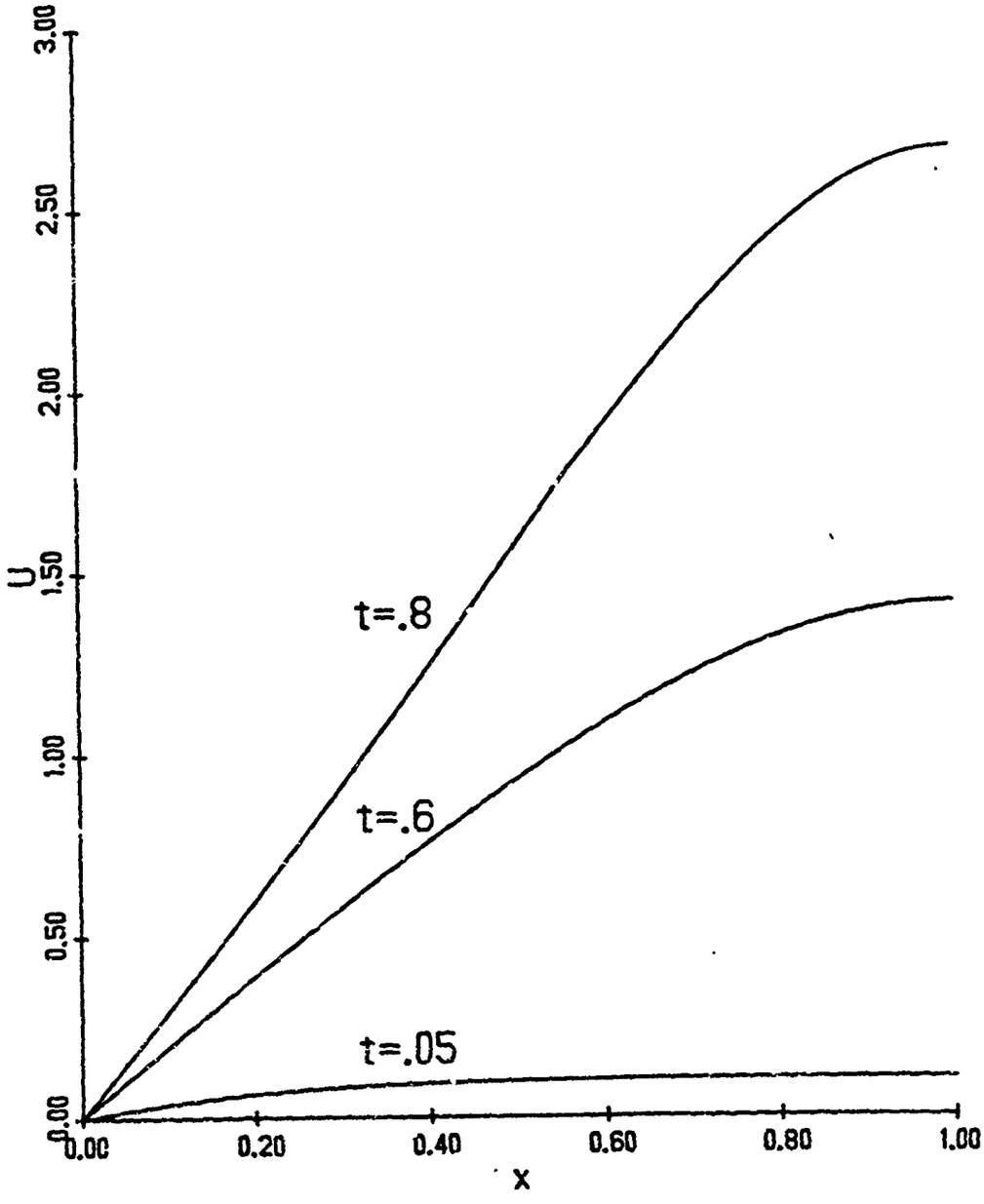



Figure 4. Solution of Example 3 at times t = 0.05, 0.6, and 0.8 with a tolerance of 0.001.

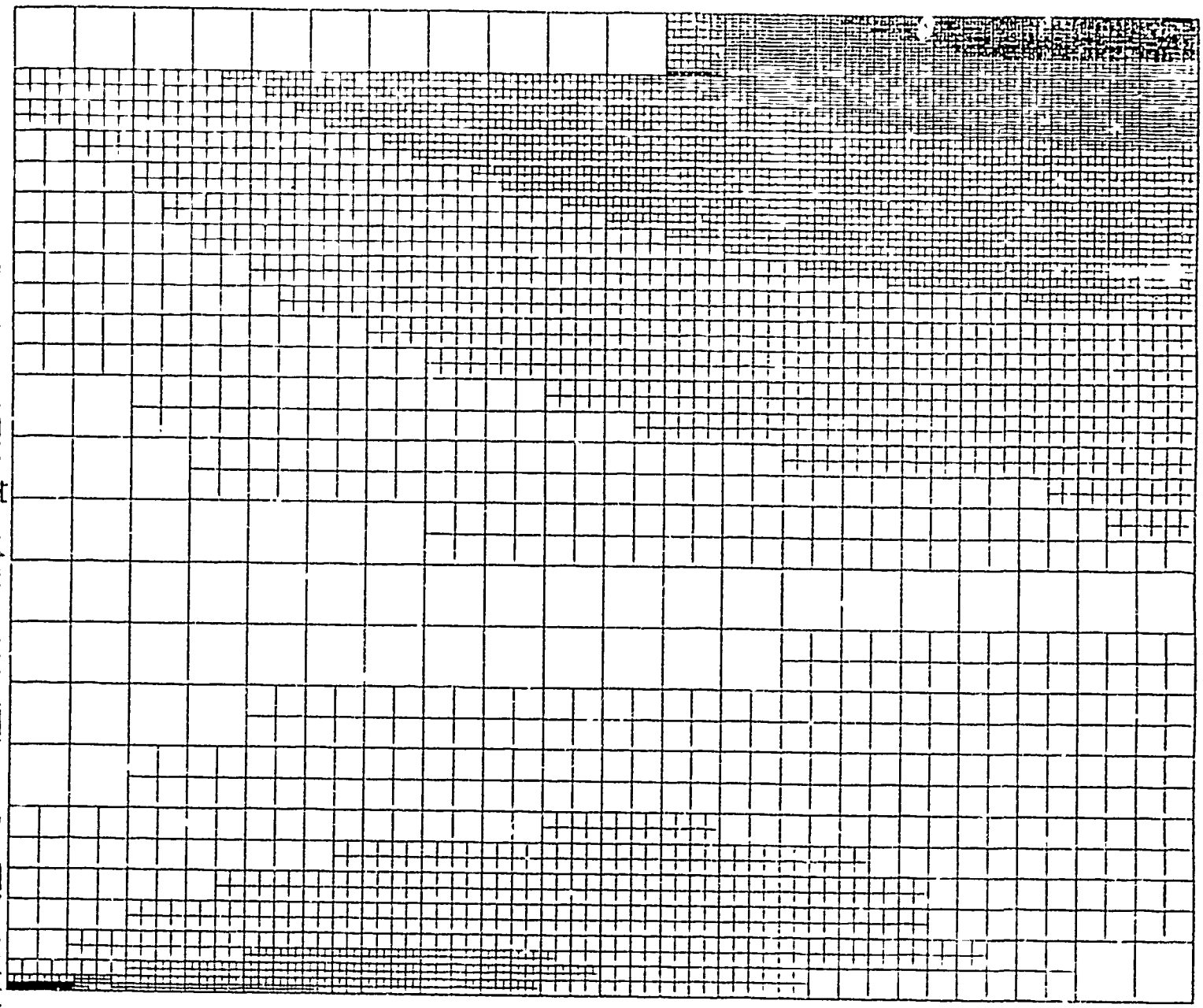



Figure 5. The grids generated in solving Example 3 for $0 < t < 0.8$. The initial coarse mesh has 20 elements with $\Delta t = 0.05$.

# Post-Buckling Analysis of an Elastica with with One and Many Critical Loads

Iradj G. Tadjbakhsh*

**Abstract**

Static stability of an elastic system composed of a flexible and a rigid link leads to a nonlinear eigenvalue problem with a single buckling load or infinite buckling loads when the flexible link is in tension or compression, respectively. Integration of equations of equilibrium leads to nonlinear singular integral equations which can be analyzed and solved numerically.

*Professor of Civil Engineering, Rensselaer Polytechnic Institute, Troy, New York

PREVIOUS PAGE IS BLANK

## Introduction

Among the problems arising from consideration of stability of elastic rods under axial loads is the class of nonlinear eigenvalue problems. This type of problems describe the buckling loads of the rod, bifurdation of the solutions at critical stability conditions and their post buckling behavior [1,2].

A particular case is the buckling of a straight column, fixed at the base and acted upon by a rigid crank moving up and down at an end along the column axis and hinged at the other end to the free tip of the column. This state of affairs produce alternately states of tension and compression in the column. For the upward motion of the crank, Fig. 1, the column is in tension and, yet, it will buckle under a single critical load. By contrast, the downward motion of the crank is characterized by the column being in compression and by infinitely many critical loads.

Integrations of the equations of equilibrium lead to singular nonlinear integral equations which can be solved numerically after appropriate limits are obtained. Results indicate that postbuckling behavior is characterized by gradual decrease in the load carrying capacity of the system as deflections grow. This is the type of behavior which is indicative of sensitivity to imperfections in column geometry or material [3].

Formulation and Solution

Referring to Fig. 1, an elastic rod BC in considered. The rod is fixed at C and alternately in tension or compression as the rigid rod AB is pushed upward or downward along the axis AC. For load P less than a critical value the elastica will be in an unbuckled state and lying along the x-axis. The critical value of P and the post-buckling behavior of the elastic rod will be different for the two states of loading.

Considering first the case of upward motion of hinge A, one may express the axial force T and the transverse shear N in the elastica by

$$\begin{aligned} T\cos\theta - N\sin\theta &= P \\ T\sin\theta + N\cos\theta &= P\tan\beta \end{aligned} \qquad (1)$$

Also noting that $N = -EI\,\theta''$, we obtain after eliminating T

$$EI\,\theta'' + P(\tan\beta\cos\theta - \sin\theta) = 0,\ 0 < s < \ell,\ (' = d/ds) \qquad (2)$$

The relation between the deflection y(s) and $\theta$ is

$$y = \int \sin\theta(s)\,ds \tag{3}$$

and furthermore

$$\tan\beta = y(\ell)/[d^2 - y^2(\ell)]^{1/2} \tag{4}$$

where d is the length of AB. The boundary conditions that accompany (2) are

$$\theta(0) = \theta'(\ell) = 0 \tag{5}$$

During the downward motion of A, again acted upon by a force P which is considered positive, (3) - (5) remain the same but (2) changes and becomes

$$EI\,\theta'' + P\,(\tan\beta\,\cos\theta + \sin\theta) = 0 \tag{6}$$

The linearized version of the problem defined by (2) - (5) has only a single eigenvalue and corresponding eigenfunction. By contrast when (6) is considered in place of (2) the problem possesses infinitely many eigenvalues and eigenfunctions. The post-buckling regime is characterized by a gradual decrease in the value of P as

the amplitude of deflection grows.

Let

$$\epsilon = \theta(l),\ \xi = s/l,\ r = l/d,\ \phi(\xi,\epsilon) = \theta/\epsilon$$

$$\delta(\epsilon) = y(l)/l,\ \alpha(\epsilon) = l\sqrt{P/EI} \tag{7}$$

Then (2) - (5) yield

$$\phi_{\xi\xi} + (\alpha^2/\epsilon)(\tan\beta\cos\epsilon\phi - \sin\epsilon\phi) = 0,\ 0 < \xi < 1 \tag{8}$$

$$\phi = 0;\ \xi = 0 \tag{9}$$

$$\phi_\xi = 0,\ \phi = 1;\ \xi = 1 \tag{10}$$

Additionally

$$\delta = \int_0^1 \sin\epsilon\phi\, d\xi \tag{11}$$

$$\tan\beta = r\delta/(1-r^2\delta^2)^{1/2} \tag{12}$$

On physical grounds one may expect that the load should be an even function of amplitude $\epsilon$ and that the deflection should be odd in

$\epsilon$. This expectation, also borne out by the equations, can be expressed as

$$\begin{aligned} \phi &= \phi_o(\xi) + \phi_1(\xi)\epsilon^2 + \phi_2(\xi)\epsilon^4 + \dots \\ \alpha &= \alpha_o + \alpha_2\epsilon^2 + \alpha_2\epsilon^4 + \dots \\ \delta &= \delta_o\epsilon + \delta_1\epsilon^3 + \delta_2\epsilon^5 + \dots \end{aligned} \tag{13}$$

In the limit as $\epsilon \to 0$, (8) - (10) yield

$$\phi_{o\xi\xi} + \alpha_o^2(\kappa\delta_o - \phi_o) = 0, \qquad 0 < \xi < 1 \tag{14}$$

$$\phi_o = 0; \qquad \xi = 0 \tag{15}$$

$$\phi_{o\xi} = 0, \ \phi_o = 1; \qquad \xi = 1 \tag{16}$$

while from (11)

$$\delta_o = \int_0^1 \phi_o \, d\xi \tag{17}$$

The solution of (14) is in terms of hyperbolic sine and cosine functions. As a consequence of satisfying (15) - (17) one obtains the relationship

$$\tanh \alpha_o = (1 - 1/r)\, \alpha_o \tag{18}$$

for the determination of the eigenvalue $\alpha_o$. Since $r > 1$, (18) has only one positive real root. In view of the scaling defined in (7) the eigenfunction is uniquely determined

$$\phi_o = r\delta_o\, (-\cosh \alpha_o \xi + \tanh \alpha_o \cdot \sinh \alpha_o \xi + 1) \tag{19}$$

where

$$\delta_o = -\cosh\alpha_o / r(1-\cosh\alpha_o) \tag{20}$$

For the downward stroke of the hinge A, (18) is replaced by

$$\tan \alpha_o = (1 + 1/r)\, \alpha_o \tag{22}$$

with infinitely many roots and corresponding eigenfunctions.

To obtain the nonlinear solution, we note that (8) has the first integral

$$1/2\, \phi_\xi^2 + (\alpha^2/\epsilon^2)\, (\tan\beta \sin\epsilon\phi + \cos\epsilon\phi) = C \tag{23}$$

Application of (10) yields

$$C = (\alpha^2/\epsilon^2)\,(\tan\beta\,\sin\epsilon + \cos\epsilon) \tag{24}$$

Using this and solving (23) for $\phi_\xi$, we have

$$d\phi/d\xi = (\sqrt{2}\alpha/\epsilon)\,Q\,(\phi,\epsilon,\beta) \tag{25}$$

where

$$Q = [\tan\beta\,(\sin\epsilon - \sin\epsilon\phi) + (\cos\epsilon - \cos\epsilon\phi)]^{1/2} \tag{26}$$

Separating variables in (25) and integrating results in

$$\int_0^1 Q^{-1}\,d\phi = \sqrt{2}\alpha\epsilon^{-1} \tag{27}$$

From (11) and (30)

$$\delta = (\epsilon/\sqrt{2}\alpha)\int_0^1 Q^{-1}\,\sin\epsilon\phi.d\phi \tag{28}$$

Eliminating $\alpha$ between (27) and (28) yields

$$\delta = \int_0^1 Q^{-1} \sin\epsilon\phi . d\phi / \int_0^1 Q^{-1} d\phi \qquad (29)$$

which determines $\delta$ as a function of $\beta$ and $\epsilon$. Substituting (29) into the right side of (12) produces an implicit equation determining $\beta$ as a function of $\epsilon$ for fixed r. Once $\beta(\epsilon)$ is found, (29) determines $\delta(\epsilon)$ and subsequently (28) will yield $\alpha(\epsilon)$.

In calculating the integrals that occur in (27) - (29) singularities are encountered for (i) $\phi = 1$, (ii) $\epsilon = 0$ and (iii) $\tan\beta = \infty$. Briefly we describe how in each case the singularities may be removed.

(i) Consider the integral in (27) and note that by using binomial expansion we have

$$1/Q = 1/Q^* + R \qquad (30)$$

where

$$Q^* = [\epsilon\ (\tan\beta\ \cos\epsilon - \sin\epsilon)(1-\phi)]^{1/2} \qquad (31)$$

and $R = [0\ (1-\phi)^{1/2}]$. Now $1/Q^*$ is integrable analytically and R vanishes at $\phi = 1$. The integration in (27) can proceed with the aid of decomposition indicated in (30). A further removable singularity

in integral of (27) occurs when $R|_{\phi=0}$ is considered for $\epsilon = 0$. Now

$$R|_{\phi=0} = (\tan\beta \sin\epsilon+\cos\epsilon-1)^{-1/2}-[(\tan\beta\cos\epsilon-\sin\epsilon),\epsilon]^{-1/2}$$

which becomes indefinite of the form $\infty - \infty$ as $\epsilon \to 0$. A limiting process shows that $\lim R|_{\phi=0} = 0$.

(ii) the limit of $\epsilon = 0$ reduces the nonlinear problem to the linearized problem described earlier. This is accomplished if we obtain limit of $\alpha$ from (27) which is

$$\alpha_o = (2)^{-1/2} \int_o^1 \{(1-\phi)[r\delta_o-(1+\phi)/2]\}^{-1/2} d\phi \qquad (32)$$

Similar limiting process when applied to (29) produces

$$\delta_o = (1/\sqrt{2}\alpha_o) \int_o^1 \phi \{(1-\phi)[r\delta_o-(1+\phi)/2]\}^{-1/2} d\phi \qquad (33)$$

Carrying out the integrations and solving (32) - (33) for $\alpha_o$ and $\delta_o$ the results in (18) and (20) are obtained.

(iii) This is the limit when the rigid rod AB is in a horizontal position in its up and down motion. Corresponding to this situation $\tan\beta = \infty$, $\alpha = 0$, $\delta = 1/r$ and $\epsilon$ tends to a limit which will

be denoted by $\hat{\epsilon}$. Since

$$\lim_{\tan\beta\to\infty} (\tan\beta)^{1/2}\, Q^{-1} = q(\hat{\epsilon},\phi) = (\sin\epsilon - \sin\epsilon\phi)^{1/2} \tag{34}$$

it follows from (29) that

$$r = \int_0^1 q(\hat{\epsilon},\phi)\,d\phi \Big/ \int_0^1 q(\hat{\epsilon}\phi)\sin\epsilon\phi.d\phi \tag{35}$$

This relation determines the inclination $\theta(t) = \hat{\epsilon}$ for the position $\tan\beta = \infty$.

Fig. 2 shows the result of numerical determinations giving $\alpha$ as a function of $\delta$ for various values of r. The results for $\alpha > 0$ pertain to the upward stroke of the hinge A and those for $\alpha < 0$ show the load-deflection relationships when A moves downward. For column in compression only the first buckling loads are shown. The complete picture includes points obtained by symmetric reflection about the $\alpha$-axis.

Acknowledgement

The author wishes to acknowledge the assistance of his graduate student C. Younis in checking manipulations and performing numerical calculations.

## References


1. A.E.H. Love, A Treastie or the Mathematical Theory of Elasticity, Dover, 1944, p. 402-414.

2. J.B. Keller and S. Antman Editors, Bifurcation Theory and Nonlinear Eigenvalue Problems, W.A. Benjamin, 1969.

3. B. Budiansky, "Theory of Buckling and Post-Buckling Behavior of Elastic Structures," Advances in Applied Mechanics, 14, (1974), 1-40.

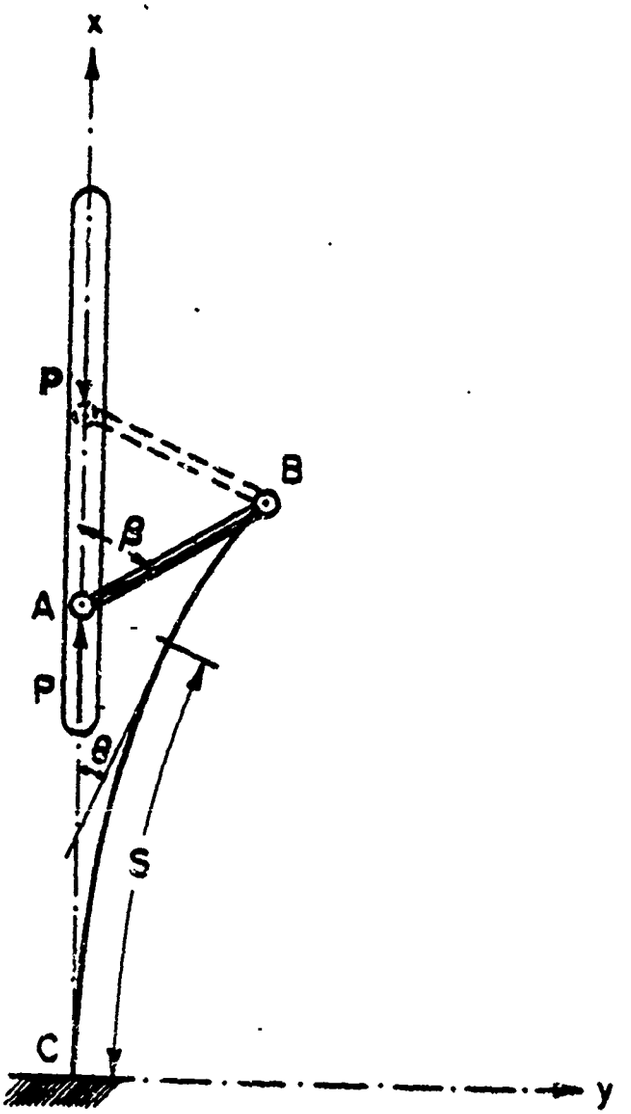



Figure 1

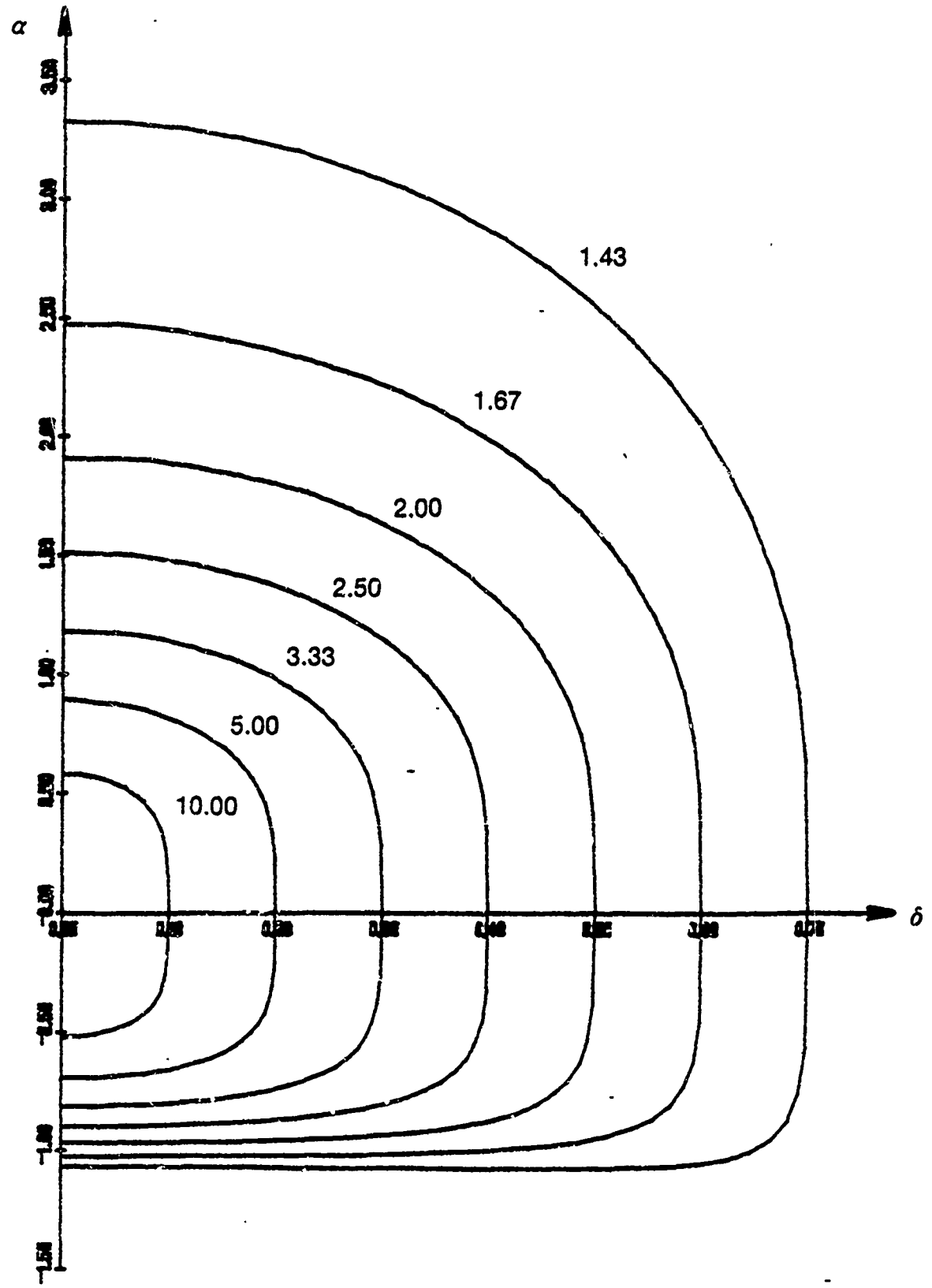



Figure 2

# A MESH MOVING TECHNIQUE FOR TIME DEPENDENT PARTIAL DIFFERENTIAL EQUATIONS IN TWO SPACE DIMENSIONS[1]

David C. Arney
Department of Mathematics
United States Military Academy
West Point, NY 10996
and
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12181

and
Joseph E. Flaherty
U. S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Calibre Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189-5000
and
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12181

ABSTRACT. We discuss an adaptive mesh moving technique that can be used with a finite difference or finite element scheme to solve initial-boundary value problems for vector systems of partial differential equations in two space dimensions and time. The mesh moving technique is based on an algebraic node movement function determined from the propagation of significant error regions. The algorithm is designed to be flexible, so that it can be used with many existing finite difference and finite element methods. To test the mesh moving algorithm, we implemented it in a system code with an initial mesh generator and a MacCormack finite volume scheme on quadralateral cells for hyperbolic vector systems. Results are presented for several computational examples. The moving mesh scheme reduces dispersion errors near shocks and wave fronts and thereby reduces the grid requirements necessary to compute accurate solutions while increasing computational efficiency.

1. INTRODUCTION. Mesh moving , an adaptive technique that has been used successfully to improve the accuracy of both finite element and finite difference schemes for a variety of time dependent problems in one space dimension (cf., e.g., [1,2,10,11,14,18,20,23]). The essential idea is to derive equations so that the mesh moves either

---

[1]The authors were partially supported by the U. S. Army Research Office under Contract Number DAAG29-82-K-0197 and the U. S. Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant Number AFOSR 80-0192.

to extremize some quantity, e.g., to minimize the discretization error, or to follow some local nonuniformity, e.g., a wave front. This generally reduces dispersive errors and allows the use of larger time steps while maintaining accuracy. For example, with a fixed mesh a wave front may move through a cell in one time step causing significant dispersive errors (cf. Figure 1a); whereas, a moving mesh with the same time step can follow the wave front and keep it within the same cell (cf. Figure 1b)

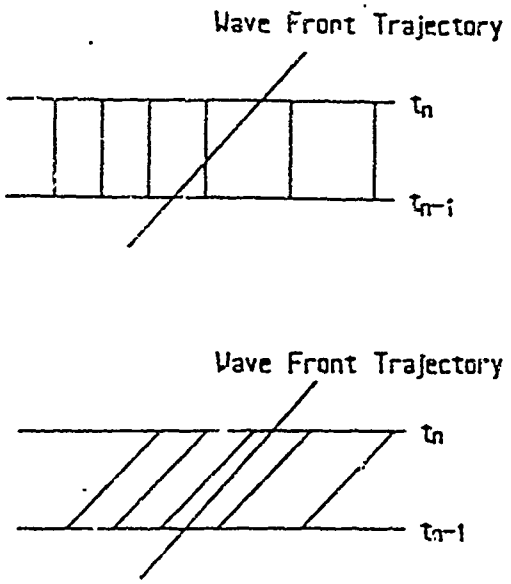



Figure 1. Wave Front Trajectory on a) Stationary Mesh (top) and b) Moving Mesh (bottom).

Mesh moving algorithms have often been related to the numerical integration scheme and/or the problem being solved. In one dimension, for example, Hyman [19] moves a mesh to minimize the time variation of the solution at the nodes. This scheme uses finite difference approximations for solving hyperbolic conservation laws. Davis and Flaherty [9] develop a finite element code for parabolic systems that moves the mesh so as to equidistribute the spatial component of the discretization error. Miller et al. [15,20,21] couple the node position equations into the finite element variational equations and minimize the residual in solving parabolic problems. Bell and Shubin [2] solve the Euler-Lagrange equations of an extremizing functional and use a finite difference scheme to solve hyperbolic conservation laws. All of these schemes have successfully demonstrated that mesh moving can reduce error and provide improvements in computational efficiency for one-dimensional problems.

With some modification the Hyman and Miller algorithms can be extended to higher dimensions; however, many other mesh moving techniques are not directly applicable to two- and three-dimensional problems. One difficulty is that equidistribution strategies fail to produce unique solutions. Brackbill and Saltzman [7,26] have overcome this problem by adding the constraints of mesh smoothness and orthogonality to a variational problem.

A sucessful mesh moving scheme for higher dimensional problems that is somewhat similar to the method presented here is the algorithm of Rai and Anderson [23,24,25]. Their algorithm is based on a gravitional principle and calculates the velocity of a node based on the difference between its error and the mean error. The displacement of one node with respect to another is inversely proportional to the distance between them. A summation over all nodes is necessary to determine each node's speed in a computational grid.

A different adaptive technique is local mesh refinement which consists of dividing or refining elements in regions where the solution is not adequately resolved. The advantage of this technique relative to mesh moving is that enough fine grids can be added to resolve the small scale structures of the solution and provide solutions to within user prescribed error tolerances. The local mesh refinement schemes of Berger [3,4], Flaherty and Moore [13], Gannon [16] and Bieterman and Babuska [5,6], have successfully satisfied user tolerances for different problems using finite element or finite difference schemes in either one or two dimensions.

The most promising algorithms appear to be those that combine both mesh moving and local mesh refinement. While neither adaptive technique or their combination is likely to be optimal for most problems, a combination can accurately solve for the solution in regions where it varies rapidly and devote little effort in regions where it varies slowly. It is our intention to consider such schemes; however, the computational procedures discussed here do not as yet contain local refinement.

The mesh moving scheme we have developed is simple, efficient, and independent of the numerical method being employed to discretize the partial differential equations. At each time step it uses the current node locations and the nodal values of a mesh movement indicator. We use local error estimates as mesh movement indicators, but other computable values such as solution gradients or curvature could be used. Nodes with "statistically significant error" (cf. Section 2) are grouped into rectangular error clusters. This clustering separates spatially distinct phenomena of the solution. As time evolves the clusters can move, change size, change orientation, collide, separate, reflect off boundaries, or pass through boundaries. At each time step new clusters can be created, and old ones can vanish. The clustering algorithms we use are briefly described in Section 2 and were developed by Berger [3,4] for a mesh refinement scheme for solving hyperbolic problems.

Mesh movement is determined by the node's relationship to its nearest error cluster. Movement is done in two steps, each in a direction along a principal axis of a cluster rectangle. The amount of movement in each direction is determined by a movement function which insures that the center of error of the cluster moves according to a differential equation suggested by Coyle et al. [8]. Additionally, the movement function smoothes mesh motion, reduces distortion, mesh tangling, or overlapping, and prevents nodes from moving outside the domain boundaries.

In Section 2 we discuss error clustering, movement of the center of mass of the error cluster, the node movement function, and the initial mesh generator used in the computational examples. In Section 3 we discuss the MacCormack finite volume scheme for hyperbolic equations and the error estimates used in the computational examples. The results of the computational examples are given in Section 4, and Section 5 contains a discussion of the results of the experiments and the status of our algorithm.

2. MESH MOVING SCHEME AND INITIAL MESH GENERATION. We suppose that an approximate solution of the partial differential equation and a pointwise error estimate have been calculated by some numerical technique at the current time step, we then flag "significantly high error nodes" as nodes with error greater than twice the mean nodal error and also greater than a user supplied error tolerance. If there are no significant error nodes, computation is performed on a stationary mesh. Next the nearest neighbor clustering algorithm of Berger [3,4] is used to cluster flagged error nodes. The nearest neighbor clusters have internodal distances in the cluster less than intercluster distances, which are the minimum distances between clusters. The formation of a cluster is done iteratively by starting with a node and including nodes in a cluster if the distance from the node to the cluster is less than a specified distance. When a node is determined to belong to two or more clusters, those clusters are merged.

Berger [3,4] shows that near minimum area rectangles that contain all the nodes within the cluster can be easily generated. The principal axes of such a rectangle are the major and minor axes of an enclosed ellipse with the same first and second moments as the clustered nodes. Thus, if $x_m$ and $y_m$ are the mean coordinates of the clustered nodes, then the axes of the rectangle are the eigenvectors of the symmetric (2x2) matrix

$$\begin{bmatrix} \Sigma\, x_i^2 - x_m^2 & \Sigma\, x_i\, y_i - x_m y_m \\ \Sigma\, x_i y_i - x_m\, y_m & \Sigma\, y_i^2 - y_m^2 \end{bmatrix} . \qquad (2.1)$$

For problems with significant error nodes located on a long curved line, the entire region will belong to one unacceptably large cluster. In order to prevent this inefficieny and provide better alignment with curved fronts, the rectangular clusters are checked for efficiency by determining the percentage of flagged nodes in th cluster to the total nodes in the cluster. If a 50 percent efficiency not achieved, the rectangle is iteratively bisected in the direction of the major axis. This is repeated until all clusters have a 50 percent efficiency or more. This nearest neighbor clustering separates spatially distinct phenomena as shown by the dotted clusters on the two dimensional mesh of Figure 2 and provides some linear alignment with long curved gradient fronts as shown by the clusters in Figure 3.

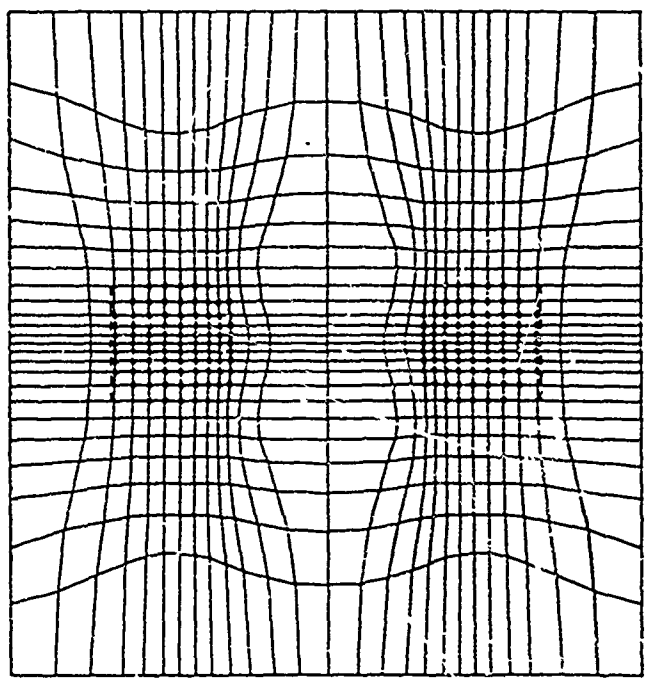


Figure 2. Two spatially distinct Clusters.

In order to determine proper node movement, as shown for a one dimensional problem in Figure 1b. the speed of propagation of the error clusters must be determined. Hyman [19] and Hyman and

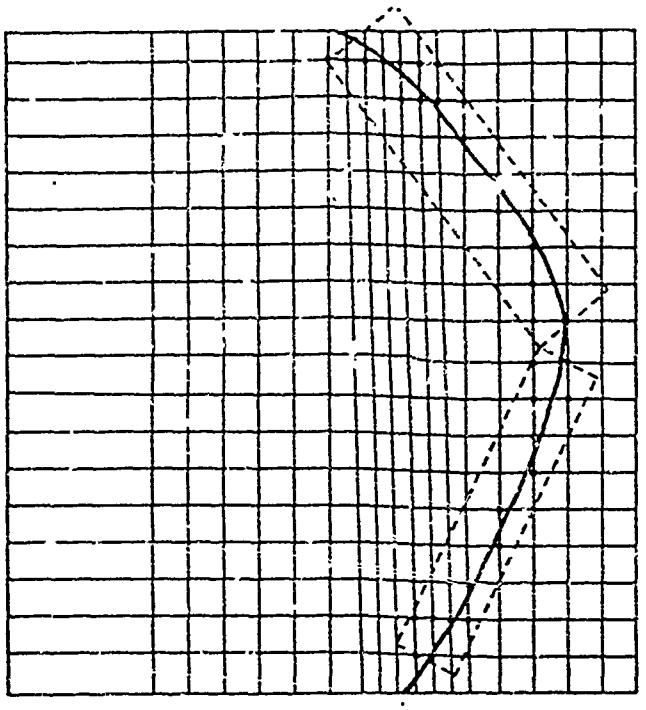


Figure 3. Clusters aligned with a Curved Front.

Harten [18] move nodes based on minimizing the time variation of the solution components. For hyperbolic systems this allows the mesh to move at a weighted average of the characteristic speeds at the node. Front tracking schemes also move the mesh so that isolated discontinuities are stationary in reference to the mesh. Since tracking error is possible for all time dependent problems, our approch is more general and is also an approximation to these schemes for hyperbolic problems, where error propagates in a characteristic direction. We assume that nodes in the same cluster have related solution characteristics, so that we can determine individual node movement from the propagation of the center of mass of the error cluster.

In the Hyman and Harten algorithm [18], when there is multiple wave interaction in a vector system, the best that can be done is to move the mesh with a weighted average of characteristic velocities. The same principle applies to our algorithm when multiple error clusters have merged, the mesh is still able to move as their combined error cluster moves, which is a form of weighted averaging. Comparisons between center of mass propagation and characteristic paths for an example problem are made in Section 4.

We attempted to move nodes based on a procedure that was based on extrapolating for the center of masses; however, this showed an unstable oscillatory effect. Indeed, Coyle et al. [8] showed that node movement based on extrapolation can be unstable. Using their suggestion, we stablize the movement by solving the differential

equation

$$\ddot{r} + \lambda \dot{r} = 0 , \qquad (2.2)$$

where r(t) is the position vector of the center of mass of an error cluster and $\dot{(\ )} := d(\ )/dt$. Equation (2.2) is conditionally stable, and when solved numerically with reasonable choices of $\lambda > 0$ the oscillations in the mesh movement were no longer present.

We solve (2.2) from, say, $t_n$ to $t_{n+1}$ and hence, determine $r(t_{n+1})$ and the vector $r(t_{n+1}) - r(t_n)$ which is projected in the two axial directions of the rectangular cluster to determine the maximum movement (MM) in each axial direction. Along the two axial directions the movement function is one dimensional. A profile of the movement function that we use is shown in Figure 4; however, the algorithm is designed to be used with any general one dimensional movement function. Slopes a and b depend on distance from the cluster to the domain boundaries and other clusters.

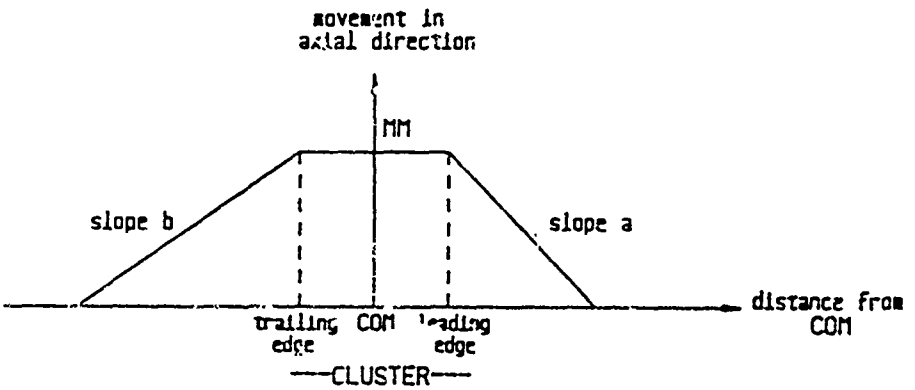



Figure 4. Profile of the Node Movement Function.

As shown in Figure 4, we let

$$MA_{inside} = \begin{cases} MM(1-ax) & \text{if nodes are x distance ahead of leading edge} \\ MM & \text{if nodes are inside the cluster projection} \\ MM(1-bx) & \text{if nodes are x distance behind trailing edge .} \end{cases} \qquad (2.3)$$

Equation (2.3) determines the movement distance for nodes inside the projection of the cluster on the axis, i.e., the shaded region of Figure 5. In order to provide smooth node movement throughout the domain, nodes outside this region move in a reduced amount as determined by

$$MA_{outside} = MA_{inside}[\ 1-(2z/DIAM)]\ , \qquad (2.4)$$

where z is the distance outside the cluster projection as shown in Figure 5 and DIAM is the diameter of the Domain.

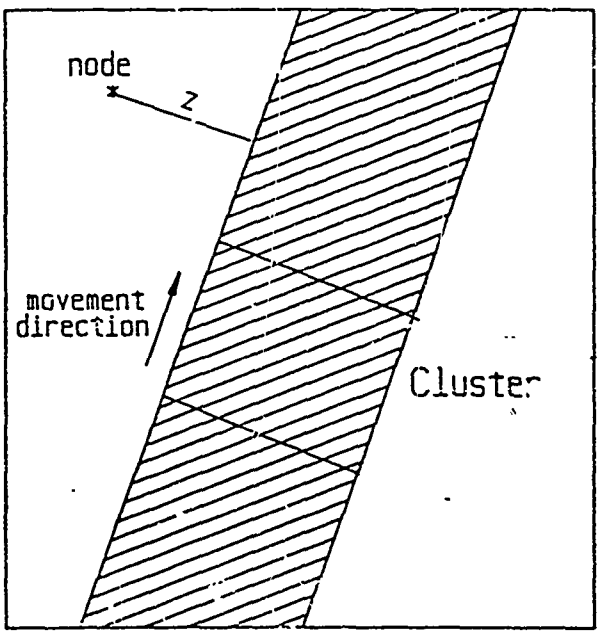



Figure 5. Nodes outside the Projection of the Error Cluster.

The generation of a proper initial mesh is critical to the success of the mesh moving scheme. Without refinement the mesh moving algorithm can not provide suitable error control unless the initial mesh spacing properly resolves initial data. An initial error measure appropriate for the finite volume scheme on quadralateral cells of Section 3 is the error in interpolating the prescribed initial condition $u_0(x,y)$ on each cell by a bilinear polynomial. The error on each cell

is determined as the difference between the value of the initial function and its bilinear interpolant at the center of each cell. Therefore, the initial mesh must be generated so that the condition

$$|1/4(u_0(x_i,y_i)+u_0(x_j,y_j)+u_0(x_k,y_k)+u_0(x_l,y_l))-u_0(\bar{x},\bar{y})| < TOL \qquad (2.5)$$

holds on each cell when using the vertex and center point labelling as shown for a general cell in Figure 6. TOL is a user supplied error tolerance. We used the following iterative scheme to satisfy condition (2.5) for the computational examples of Section 4:

1. Input domain boundaries and initial data function.

2. Generate a uniform mesh.

3. Compute cell error from the left hand side of (2.5).

4. Cluster high error nodes and move influenced nodes toward center of clusters according to (2.3) and (2.4).

5. Smooth the mesh by solving the Euler-Lagrange equations of Brackbill and Saltzman [7].

6. Recompute error on cells.

*Repeat*

7.1. Add a mesh row and column to divide cells with error greater than TOL

7.2. Smooth the mesh by the algorithm of Brackbill and Saltzman[7]

7.3. Recompute the error

*Until* the error tolerance condition (2.5) is satisfied.

Initial meshes generated with this algorithm are shown in Figures 2 and 8 for the initial condition functions (4.6) and (4.2) respectively.

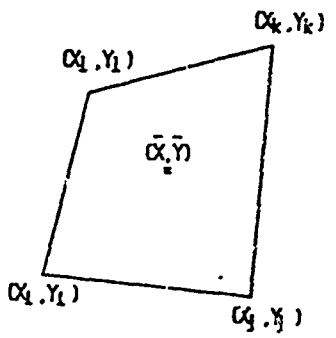



Figure 6. Cell Labelling for Equation (2.5).

3. MacCORMACK FINITE VOLUME SOLVER AND ERROR ESTIMATION. In order to test our mesh moving scheme, we used the explicit finite volume MacCormack scheme on nonuniform quadralateral grids for hyperbolic vector systems of the form

$$u_t + f_x(x,y,u,t) + g_y(x,y,u,t) = 0, \tag{3.1}$$

$$u(x,y,0) = u_0(x,y), \tag{3.2}$$

with appropriate well posed boundary conditions.

We index the nodes in a logically rectangular fashion where the time dependent node locations have cartesian coordinates $(x_{i,j}(t), y_{i,j}(t))$ and therefore, the time derivative of $u(x_{i,j}(t), y_{i,j}(t), t)$ is given by

$$\dot{u} = u_x \dot{x}_{i,j} + u_y \dot{y}_{i,j} + u_t \quad . \tag{3.3}$$

At time step n, we approximate

$$\dot{x}_{i,j}(t_n) \approx \Delta x_{i,j}(t_n)/\Delta t_n = (x_{i,j}^{n+1} - x_{i,j}^n)/\Delta t_n \tag{3.4}$$

$$\dot{y}_{i,j}(t_n) \approx \Delta y_{i,j}(t_n)/\Delta t_n = (y_{i,j}^{n+1} - y_{i,j}^n)/\Delta t_n \tag{3.5}$$

where $x_{i,j}^n = x_{i,j}(t_n)$, $y_{i,j}^n = y_{i,j}(t_n)$, and $\Delta t_n$ is the current time step.

The finite volume scheme is obtained by integrating Equation (3.1) over each cell, where the general cell(i,j) with center $(x_{i,j}, y_{i,j})$ is shown in Figure 7. The area integrals involving the spatial derivatives of f and g are converted to line integrals of f and g around the cell boundaries using Green's Theorem. The integral of the time derivative term over cell(i,j) is approximated by

$$\iint u_t dxdy \approx A_{i,j}^n u_t \tag{3.6}$$

where $A_{i,j}^n$ is the area of cell(i,j) at timestep n. The line integrals are approximated by using values of the solution at nodes on appropriate sides of the cell boundaries. The predictor step of the MacCormack scheme uses node values to the left and below the boundaries, while the corrector uses node values to the right and above the boundaries.

After substitution of (3.4), (3.5), (3.6), and the appropriate line integral approximations into Equation (3.1) and using

$$F(u_{i,j}^n) = f(u_{i,j}^n) + (\Delta x_{i,j}(t_n)/\Delta t_n) u_{i,j}^n \tag{3.7}$$

$$G(u_{i,j}^n) = g(u_{i,j}^n) + (\Delta y_{i,j}(t_n)/\Delta t_n) u_{i,j}^n \tag{3.8}$$

where $u_{i,j}^n$ is the calculated approximation to $u_{i,j}(t_n)$ to simplify the function evaluations, the predictor and corrector Equations (3.9) and (3.10) for the MacCormack finite volume scheme on a moving mesh are obtained.

The predictor step is

$$\bar{u}^{n+1}_{i,j} = u^n_{i,j} - \Delta t_n / A^n_{i,j} \{ F(u^n_{i,j}) (y_{i+1/2,j+1/2} - y_{i+1/2,j-1/2}) - F(u^n_{i-1,j}) (y_{i-1/2,j+1/2} - y_{i-1/2,j-1/2}) + F(u^n_{i,j}) (y_{i-1/2,j+1/2} - y_{i+1/2,j+1/2}) - F(u^n_{i,j-1}) (y_{i-1/2,j-1/2} - y_{i+1/2,j-1/2}) - G(u^n_{i,j}) (x_{i+1/2,j+1/2} - x_{i+1/2,j-1/2}) + G(u^n_{i-1,j}) (x_{i-1/2,j+1/2} - x_{i-1/2,j-1/2}) - G(u^n_{i,j}) (x_{i-1/2,j+1/2} - x_{i+1/2,j+1/2}) + G(u^n_{i,j-1}) (x_{i-1/2,j-1/2} - x_{i+1/2,j-1/2}) \}. \quad (3.9)$$

The corrector step is

$$u^{n+1}_{i,j} = 1/2 \{ u^n_{i,j} + \bar{u}^{n+1}_{i,j} - \Delta t_n / A^n_{i,j} [ F(\bar{u}^{n+1}_{i+1,j}) (y_{i+1/2,j+1/2} - y_{i+1/2,j-1/2}) - F(\bar{u}^{n+1}_{i,j}) (y_{i-1/2,j+1/2} - y_{i-1/2,j-1/2}) + F(\bar{u}^{n+1}_{i,j+1}) (y_{i-1/2,j+1/2} - y_{i+1/2,j+1/2}) - F(\bar{u}^{n+1}_{i,j}) (y_{i-1/2,j-1/2} - y_{i+1/2,j-1/2}) - G(\bar{u}^{n+1}_{i+1,j}) (x_{i+1/2,j+1/2} - x_{i+1/2,j-1/2}) + G(\bar{u}^{n+1}_{i,j}) (x_{i-1/2,j+1/2} - x_{i-1/2,j-1/2}) - G(\bar{u}^{n+1}_{i,j+1}) (x_{i-1/2,j+1/2} - x_{i+1/2,j+1/2}) + G(\bar{u}^{n+1}_{i,j}) (x_{i-1/2,j-1/2} - x_{i+1/2,j-1/2}) ] \}. \quad (3.10)$$

On a stationary rectangular mesh, this scheme reduces to the standard MacCormack finite difference scheme, which when the predictor and corrector are combined for a linear partial differential equation is the same as the Lax-Wendroff finite difference scheme.

Accurate error estimation is important to insure that user tolerances are achieved and to refine proper regions when doing local mesh refinement. However, mesh moving is not as sensitive to error estimation. As long as the error estimator shows the error propagation, proper error magnitudes are not necessary. Therefore, in the

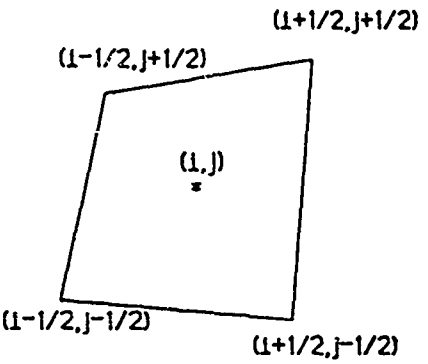



Figure 7. Labelling for general cell(i,j) for finite volume Equations (3.9) and (3.10).

computational examples of Section 4, we were able to use the difference between the predicted solution of Equation (3.9) and the corrected solution of Equation (3.10) as the error estimation or movement indicator. This error estimation is actually an estimation for the first order predicted solution, not the second order corrected solution, but does have the proper propagation characteristic.

A more accurate error estimation will be needed when local mesh refinement is implemented. Accurate error estimators that could be used are Richardson extrapolation for finite difference schemes [3] and hierarchical methods for finite element schemes [28].

4. COMPUTATIONAL EXAMPLES. The following linear hyperbolic equations were solved as tests of our mesh moving technique. We used the initial mesh generator of Section 2 and the MacCormack solver described in Section 3. For each problem the two-dimensional domain was a square with sides of length 2 centered at the origin.

Example 4.1 Consider the initial-boundary value problem

$$u_t - yu_x + xu_y = 0, \qquad t > 0 \tag{4.1}$$

$$u(x,y,0) = \begin{cases} 0, & \text{if } (x-1/2)^2 + 1.5\,y^2 \geq 1/16 \\ 1-16((x-1/2)^2+1.5y^2), & \text{otherwise} \end{cases} \tag{4.2}$$

$$u(1,y,t) = u(-1,y,t) = u(x,-1,t) = u(x,1,t) = 0\ . \tag{4.3}$$

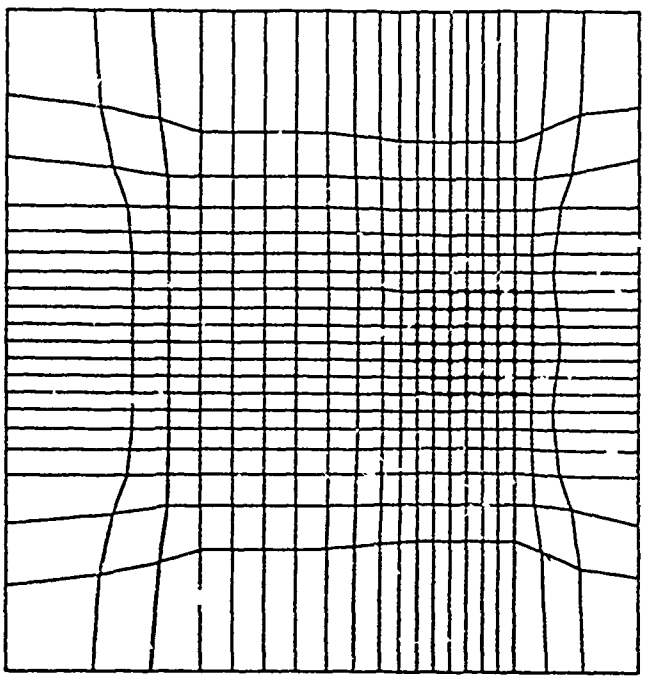


Figure 8. Initial Mesh for Example 4.1.

The exact solution of this problem is

$$u(x,y,t) = \begin{cases} 0, & \text{if } C < 0 \\ C, & \text{if } C \geq 0, \end{cases} \tag{4.4}$$

$$C = 1-16((x\cos t+y\sin t-1/2)^2 + 1.5(y\cos t-x\sin t)^2). \tag{4.5}$$

Equations (4.4) and (4.5) represent a moving elliptical cone rotating counterclockwise around the origin with period $2\pi$. It was proposed as a test problem by Gottlieb and Orszag [17] and we selected it because the rotational quality of the the error region is a good test of a mesh moving scheme.

The initial mesh generated for this problem is shown in Figure 8. This mesh has an initial interpolation error less than 0.08. Figure 9 shows the mesh at t = 1.6, and Figure 10 shows the mesh at t = 3.2. The nodes follow the moving cone to keep it within the refined region. The dashed lines on Figures 8, 9, and 10 are the error cluster rectangles at the appropriate time steps. Figures 11 and 12 show the contour and surface plot of the solution at t = 3.2. The dispersion error in the form of a wake behind the cone for the moving mesh solution is reduced significantly from the wake in the solution using the Lax-Wendroff finite difference scheme on a 20x20 uniform stationary mesh. For comparison the contour plot and surface plot of the Lax-Wendroff solution at the same total time t = 3.2 and same number

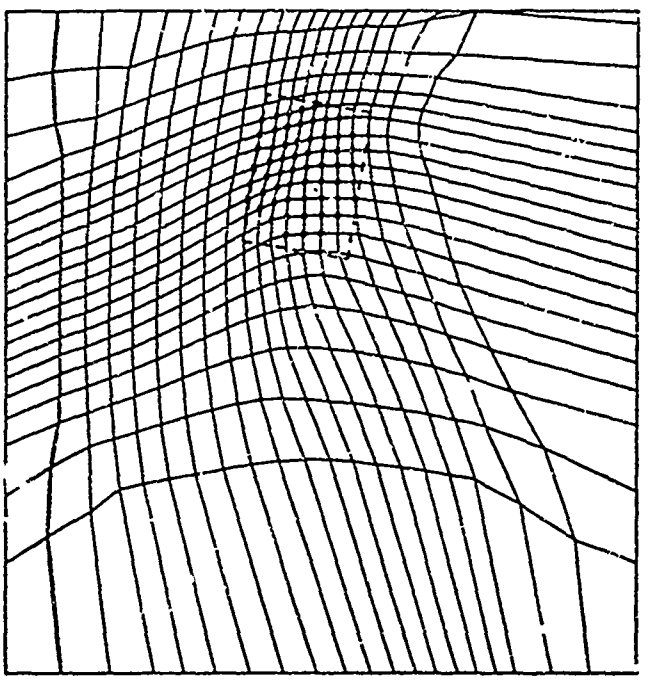


Figure 9. Mesh for Example 4.1 at t = 1.6.

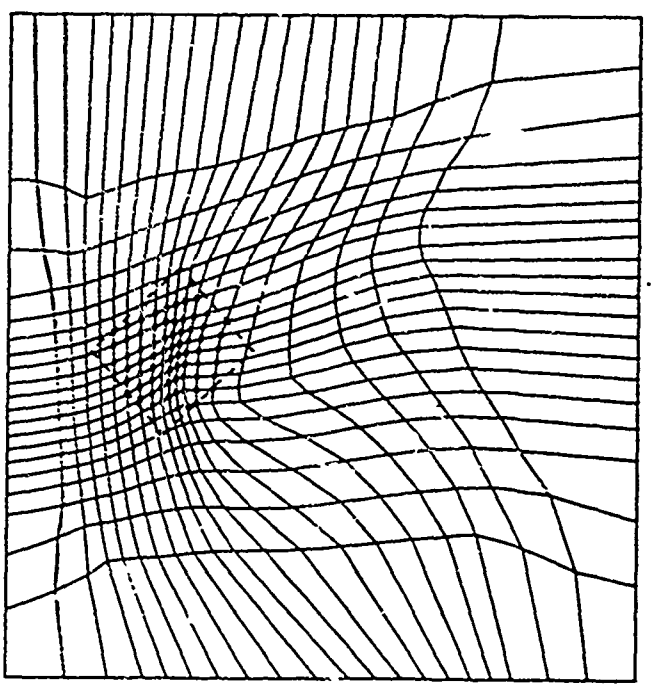


Figure 10. Mesh for Example 4.1 at t = 3.2.

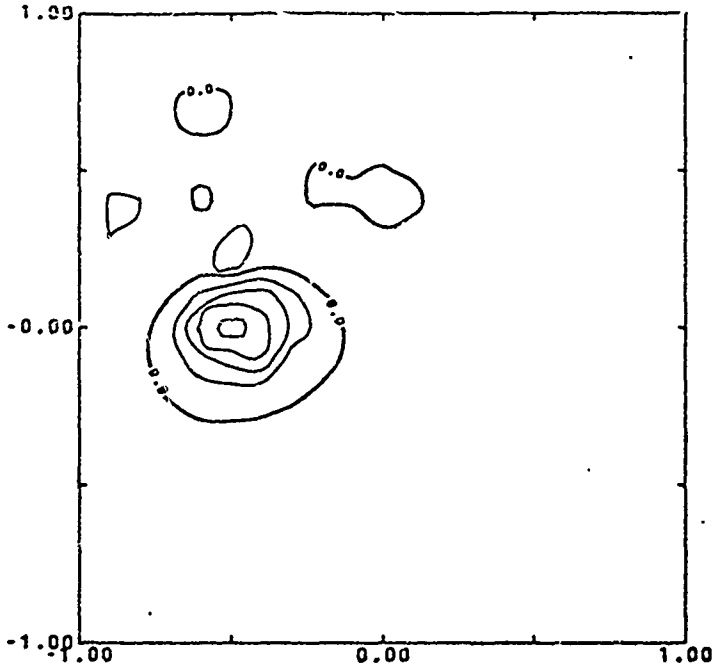



Figure 11. Contour Plot of Solution of Example 4.1 by MacCormack finite volume method on a moving mesh at t = 3.2

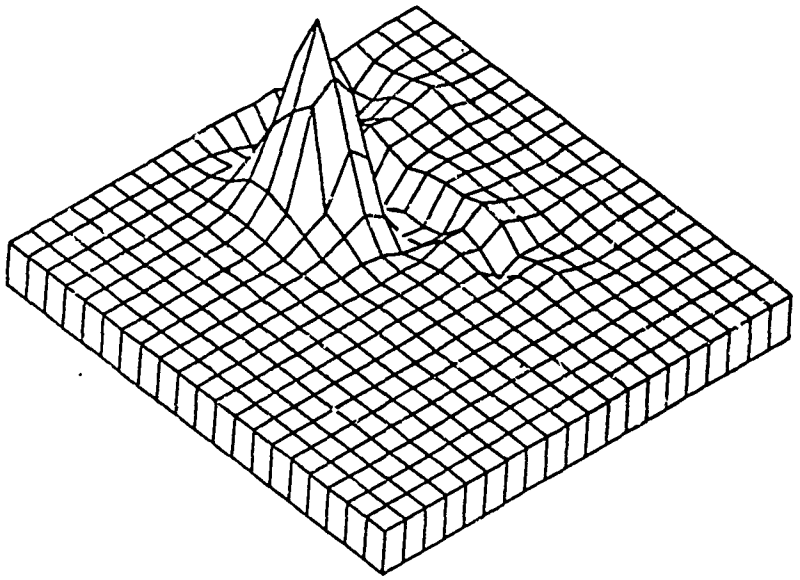


Figure 12. Surface Plot of Solution of Example 4.1 by MacCormack finite volume method on a moving mesh at t = 3.2

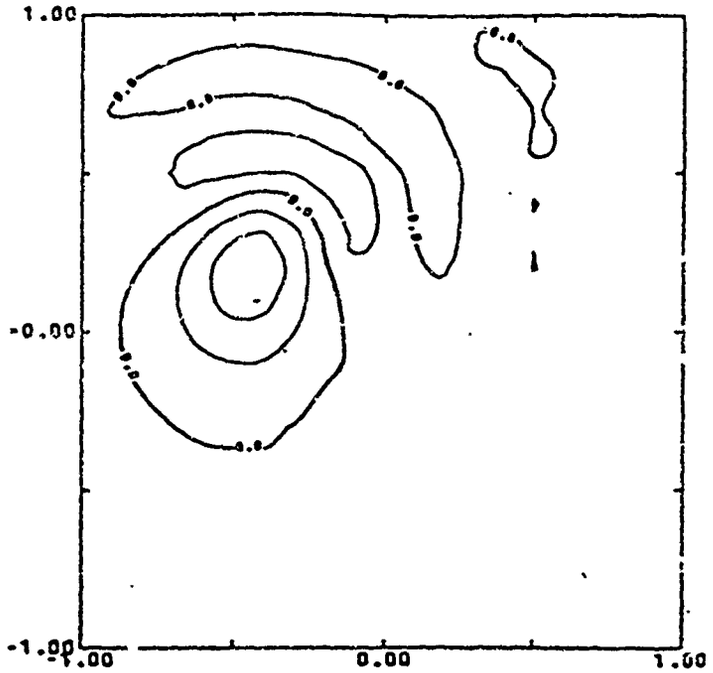



Figure 13. Contour Plot of Solution by the Lax-Wendroff scheme on a fixed 20x20 Mesh at t = 3.2.

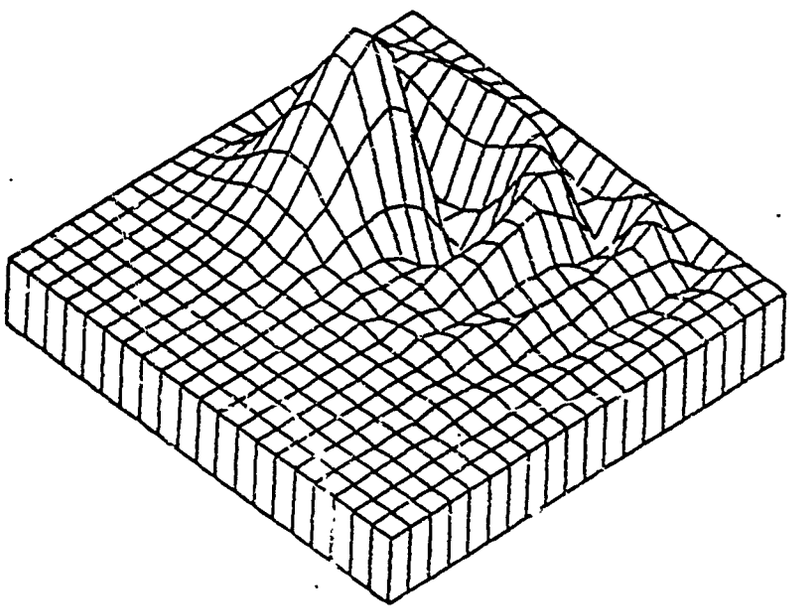


Figure 14. Surface Plot of Solution by the Lax-Wendroff scheme on a fixed 20x20 Mesh at t = 3.2.

of time steps are shown in Figures 13 and 14. Figure 15 compares the path of the center of mass propagation using Equation (2.2) and the real characteristic path of the peak of the cone. As expected for this scalar hyperbolic problem, the vectors for the movement of the center of error mass determined by Equation (2.2) closely approximate the characteristic vectors of the center of the cone with a maximum difference of 15 percent in length and direction.

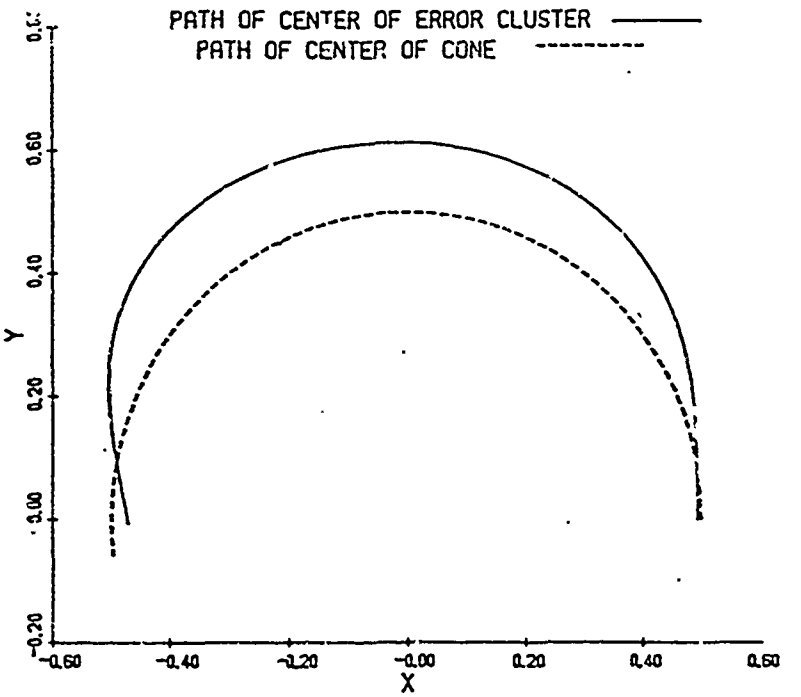



Figure 15. Comparison of characteristic path of center of cone and path of center of error mass as determined by Equation (2.2) for Example 4.1.

Example 4.2 This problem is a scalar double rotating cone problem with two symmetric cones rotating counterclockwise around the origin. The problem is given by Equations (4.1), (4.3), and new initial conditions provided by

$$u(x,y,0) = \begin{cases} 1-16((x-1/2)^2+1.5y^2), & \text{if } (x-1/2)^2+1.5y^2 \le 1/16 \\ 1-16((x+1/2)^2+1.5y^2), & \text{if } (x+1/2)^2+1.5y^2 \le 1/16 \\ 0, & \text{otherwise} \end{cases} \qquad (4.6)$$

Figure 16 shows the mesh at $t = 1.15$. Figure 16 shows the poor aspect ratio and mesh distortion caused by the rotation. The mesh tangles as the cones rotate further. When mesh tangling occurs a static rezone that creates a new mesh using an algorithm similar to the

one that generated the initial mesh must be employed. The data for the mesh can to be obtained by interpolation from the calculated solution at the nodes.

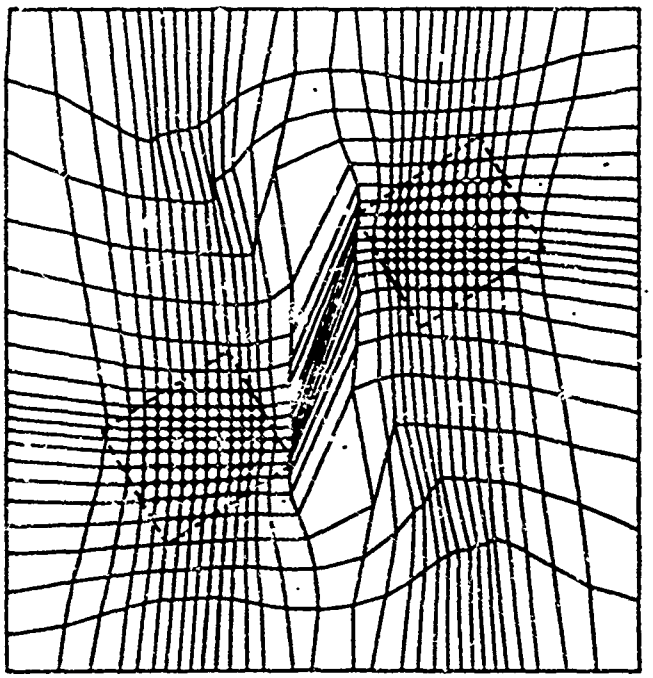


Figure 16. Distorted Mesh of Example 4.2 at t = 1.15.

Example 4.3 This problem is an uncoupled system of moving cones that pass through one another. This causes the error clusters to collide and merge, and then later separate. The problem is given in Equations (4.7), (4.8) and (4.9).

$$u_t + u_x = 0 \tag{4.7a}$$

$$v_t - v_x = 0 \tag{4.7b}$$

$$v(x,y,0) = \begin{cases} 1-16((x+1/2)^2+1.5y^2), & \text{if } (x+1/2)^2+1.5y^2 \le 1/16 \\ 0, & \text{otherwise} \end{cases} \tag{4.8a}$$

$$u(x,y,0) = \begin{cases} 1-16((x-1/2)^2+1.5v^2), & \text{if } (x-1/2)^2+1.5y^2 \le 1/16 \\ 0, & \text{otherwise} \end{cases} \tag{4.8b}$$

$$u(x,y,t) = v(x,y,t) = 0 \text{ on all boundaries of the domain} \tag{4.9}$$

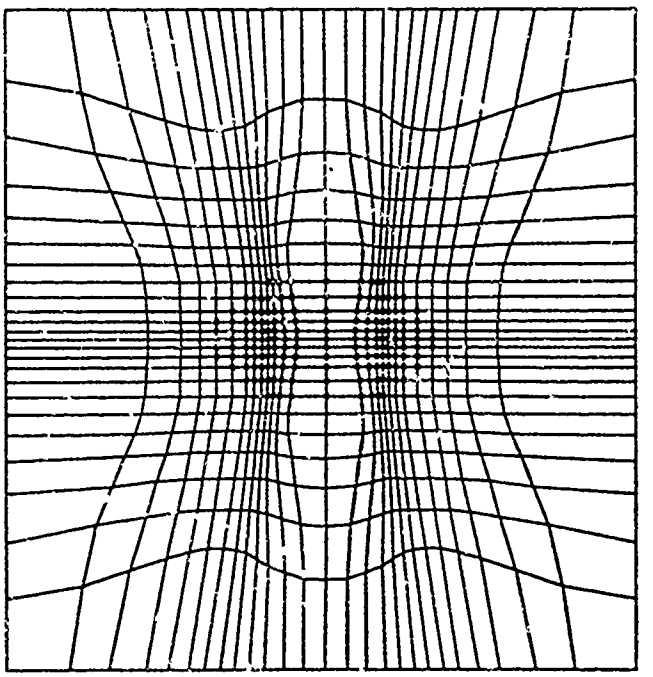


Figure 17. Mesh of Example 4.3 at t = 0.35, Clusters have merged into single cluster centered at the Origin.

Figure 3 showed the initial mesh for this problem, and Figure 17 shows the mesh at t = 0.35, just as the clusters have collided and merged. From t = 0.35 to t = 0.9, the single cluster stays centered at the origin so the mesh does not move during this time. At t = 0.9 the cones have passed completely through one another, and Figure 18 shows the separation of the error clusters and the movement of the mesh toward the boundaries. Figure 19 shows the mesh at t = 1.3. The cones and error clusters have reached the domain boundary and no further movement of the mesh will take place as the cones exit the domain.

5. DISCUSSION AND CONCLUSIONS. We have described a general two dimensional mesh moving technique based on the nodes following error propagation that is determined from the movement of clusters nodes with significantly high error. This mesh moving was tested on linear hyperbolic problems having solutions with large gradients. Even though mesh moving in two dimensions is difficult, we are encouraged by these initial results. The mesh moving algorithm was able to control the error rotation of the rotating cone in Example 4.1 and the merging and separating of error regions in Example 4.3. The distortion of the mesh in Example 4.2 showed the need for static rezoning when such distortions occur.

We are investigating ways to improve the efficiency, reliability, and robustness of the algorithm. Possible improvements include: not

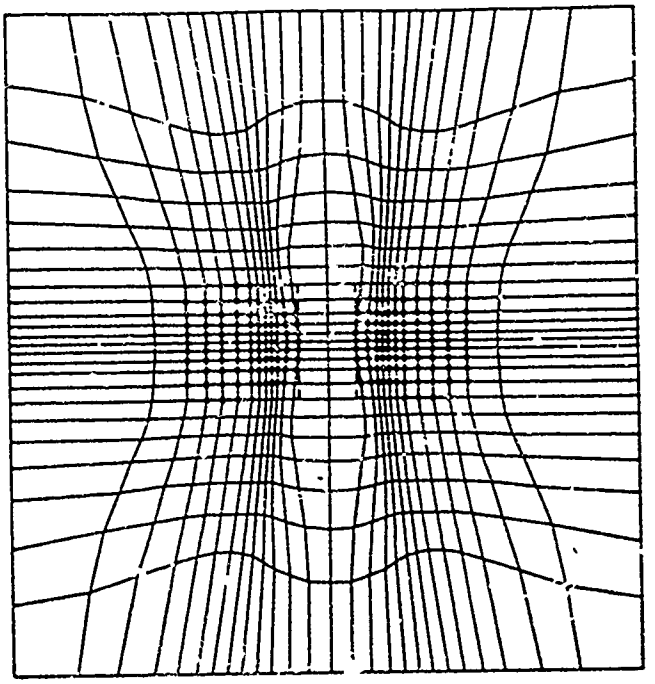


Figure 18. Mesh of Example 4.3 at t = 0.9, Clusters are separating and moving toward the Domain Boundaries.

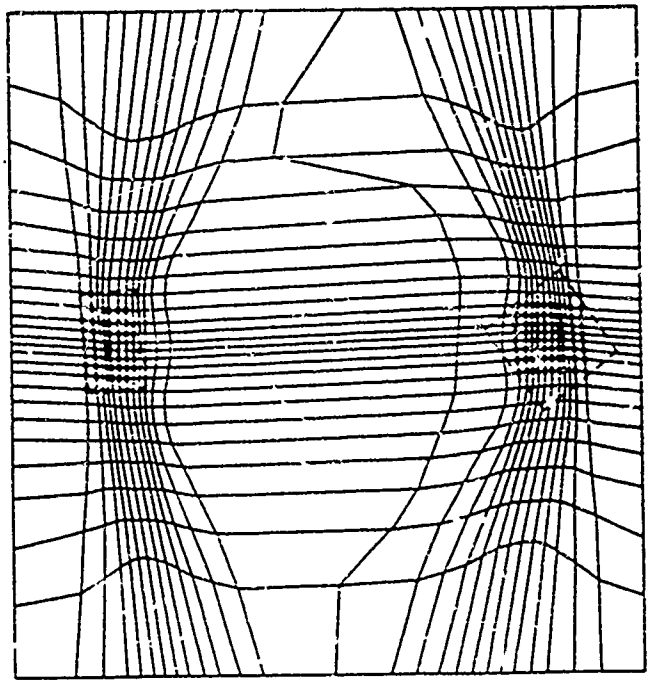


Figure 19. Mesh of Example 4.3 when Clusters reach the Domain Boundaries.

clustering at every time step and letting the mesh move at a constant velocity for several time steps, efficiently testing for mesh tangleing or distortion, using a better solver for hyperbolic equations such as the monotonic schemes of Osher [22], vanLeer [27], or Engquist [12], and using better error estimates. We intend to show the flexibility of the mesh mover by implementing it with a finite element solver for parabolic problems.

Finally, we intend to implement local mesh refinement in the algorithm. We believe that with such a combination algorithm efficiency, accuracy, and robustness can be increased.

ACKNOWLEDGEMENT. We would like to thank James Hayes, Department of Mathematics, United States Military Academy, for the use of his graphics programs that are included in the system code and were used to plot many of the figures in this article.

REFERENCES.

1. BABUSKA, I. CHANDRA, J., AND FLAHERTY, J. E. (Eds.) *Adaptive Computational Methods for Partial Differtial Equations*, SIAM, Philadelphia, 1983.

2. BELL, J. B. AND SHUBIN, G. R., 'An Adaptive Grid Finite Difference Method for Conservation Laws,' *J. Comput. Phys.*, Vol. 52, pp. 569-591, 1983.

3. BERGER, M., 'Adaptive Mesh Refinement for Hyperbolic Partial Differential Equations,' Ph.D Thesis, Computer Science Department, Stanford University, 1982.

4. BERGER, M., 'Data Structures for Adaptive Mesh Refinement,' Babuska, I., Chandra, J., Flaherty, J. E.,(eds), *Adaptive Computational Methods for Partial Differential Equations*, SIAM, Philadelphia, 1983.

5. BIETERMAN, M. AND BABUSKA, I., 'The Finite Element Method for Parabolic Equations, I. A Posteriori Error Estimation,' *Numer. Math.*, Vol 40, pp. 339-371, 1982.

6. BIETERMAN, M. AND BABUSKA, I., 'The Finite Element Method for Parabolic Equations, II. A Posteriori Error Estimation and Adaptive Approach,' *Numer. Math.*, Vol 40, pp. 373-406. 1982.

7. BRACKBILL, J. U. AND SALTZMAN, J. S., 'Adaptive Zoning for Singular Problems in Two Dimensions,' *Journal of Comp. Physics*, Vol 46, pp. 342-368, 1982.

8. COYLE, M., FLAHERTY, J. E., AND LUDWIG, R., 'On the Stability of Mesh Equidistributing Strategies for Time Dependent Partial Differential Equations,' submitted to *Journal of Comp. Physics*.

9. DAVIS, S. AND FLAHERTY, J. E., 'An Adaptive Finite Element Method for Initial-Boundary Value Problems for Partial Differential Equations,' *Siam J. Sci. Stat. Comput.*, Vol. 3, pp. 6-27, 1982.

10. DREW, D. A. AND FLAHERTY, J. E., 'Adaptive Finite Element Methods and the Numerical Solution of Shear Band Problems,' to appear in M. Gurtin (ed.) *Phase Transitions and Material Instabilities in Solids*, Academic Press, 1984.

11. DWYER, H. A., 'Grid Adaption for Problems with Separation, Cell Reynolds number, Shock-Boundary Layer interaction, and Accuracy,' AIAA Paper No. 83-0449, AIAA Twenty First Aerospace Sciences Meeting, 1983.

12. ENGQUIST, B. AND OSHER, S., 'One-sided Difference Approximations for Nonlinear Conservation Laws,' *Math. Comp.*, Vol 36, pp. 321-351, 1981.

13. FLAHERTY, J. E. AND MOORE, P. K., 'An Adaptive Local Refinement Finite Element Method for Parabolic Partial Differential Equations,' to appear in *Proc. Conf. Accuracy Estimates and Adaptive Refinements in Finite Element Computations*, Lisbon, 1984.

14. FLAHERTY, J. E., COYLE, J. M., LUDWIG, R., AND DAVIS, S. F., 'Adaptive Finite Element Methods for Parabolic Partial Differential Equations,' *Adaptive Commputational Methods for Partial Differentiel Equations*, Babuska, I., Chandra, J., and Flaherty, J. E. (Eds), SIAM, Philadelphia, 1983.

15. GELINAS, R. J., DOSS, S. K., and MILLER, K., 'The Moving Finite Element Method. Application to General Partial Differential Equations with Large Gradients,' *Journal of Computational Physics*, Vol 40, 1981.

16. GANNON, D., 'Self Adaptive Methods for Parabolic Partial Differential Equations,' Department of Computer Science, University of Illinois at Urbana-Champaign, 1980.

17. GOTTLIEB, D. AND ORSZAG, S., *Numerical Analysis of Spectral Methods Theory and Applications*, SIAM, 1977.

18. HARTEN, A. AND HYMAN, J. M., 'Self-Adjusting Grid Methods for One Dimensional Hyperbolic Conservation Laws,' *J. of Comput Phys.*, Vol 50, pp. 235-269, 1983.

19. HYMAN, J. M., 'Adaptive Moving Mesh Methods for Partial Differential Equations,' Los Alamos National Laboratory report LA-UR-82-3690.

20. MILLER, K. AND MILLER, R. N., 'Moving Finite Elements I,' *SIAM J. Num. Anal.*, Vol 18, pp. 1019-1032, 1981.

21. MILLER, K. Moving Finite Elements, II, *Siam J. Num Anal.*, Vol 18, pp. 1033-1057, 1981.

22. OSHER, S. AND CHAKRAVARTHY, S., 'Upwind Schemes and Boundary Conditions with Applications to Euler Equations in General Geometries,' *J. Comput. Phys*, Vol 50, pp. 447-481, 1983.

23. RAI, M. AND ANDERSON, D., 'The Use of Adaptive Grids in Conjunction with Shock Capturing Methods,' AIAA Paper 81-1012, June 1981.

24. RAI, M. AND ANDERSON, D., 'Application of Adaptive Grids in Fluid Flow Problems with Asymptotic Solutions,' AIAA Paper 81-0114, January 1981.

25. RAI, M. AND ANDERSON, D., 'Grid Evolution in Time Asymptotic Problems,' *Journal of Computational Physics*, Vol 43. October 1981.

26. SALTZMAN, J. S. AND BRACKBILL, J., 'Applications and Generalizations of Variational Methods for Generating Adaptive Meshes,' *Numerical Grid Generation*, 1982.

27. VAN LEER, B., 'Computational Methods for Ideal Compressible Flow,' NASA Report 172180, 1983.

28. ZIENKIEWICZ, O. C., KELLY, D. W., GAGO, J., AND BABUSKA, I., 'Hierarchical Finite Element Approaches, Error Estimates and Adaptive Refinement', *Proc. MAFELAP 1981*, April 1981.

# Numerical Simulation of Fluid Ejection from a Spinning Cylinder

Paul D. Fedele
Research Division
US Army Chemical Research and Development Center
Aberdeen Proving Ground, Maryland 21010-5423

Abstract: A computer code, based on a convective flux approximation on a finite difference Eulerian grid, was used to model the rate of fluid ejection from the opened end of an azimuthally rotating cylinder. The computer code, SOLA-VOF/CSL is described in a previous report (1). A constantly rotating cylinder, with an 80% fluid fill, is set in equilibrium solid body rotation and one end is instantaneously removed, allowing the centrifugal force to drive the fluid from the opened end. Fluid parameters have been chosen to model the behavior of water and glycerin at 25°C. The ratio of the volume ejection rate to the volume rotation rate shows similarity when expressed as a function of the cylinder rotation time. The fluid viscosity is observed to have negligible effect on the ejection rate for the spin rates of interest. This behavior is shown to be consistent with a dimensional analysis of the flows considered.

I. Introduction: The centrifugal force provides a mechanism for the distribution of a liquid from the opened end of a rapidly rotating cylinder. In 1974, Stuempfle originally studied the centrifugal distribution process, and developed an analytic model (2). To further examine such a distribution process, we have used a computer code, SOLA-VOF/CSL (1), to generate solutions to the flow problem. In the simulation, a cylinder, partially filled with fluid, is initially set in solid-body rotation. The centrifugal force holds the fluid against the wall of the cylinder and establishes a static pressure distribution which increases as the radius squared. To start the distribution process, one end of the cylinder is instantaneously removed and the pressure along the exposed surface of the fluid is set to zero. The pressure due to the centrifugal force then drives the fluid from the end of the cylinder. The resulting flow is somewhat analogus to the flow which follows the collapse of a dam.

The form of the fluid distribution and the time scale of the distribution process are the main concerns of this investigation. These are examined most conveniently by considering the rate of volume flow out of the cylinder as a function of time. Our objectives are to characterize the fluid distribution as a function of the kinematic viscosity of the fluid and the spin rate of the cylinder.

## II. Flow Model.

### A. Initial Conditions.

Figure 1 shows the conditions at the time t = 0. Due to axial symmetry

PREVIOUS PAGE IS BLANK

it is necessary to display only half of the cylinder. The cylinder, which has a length L = 33 cm and a radius R = 10 cm, is spinning with angular frequency, $\Omega$ , about the vertical axis on the left of the figure. The fluid is moving only in the azimuthal direction. The pressure is taken to be zero in the void surrounding the fluid. Since the gravitational force is much smaller than the centrifugal force in the range of variables considered, the gravitational force is neglected. In calculating the rate of volume flow out of the cylinder, fluid is considered to have left the cylinder when it passes beyond the outer radius of the cylinder wall.

B. Flow Equations.

The equations which govern the flow are the continuity equation expressing conservation of mass and the Navier-Stokes equation expressing conservation of momentum. The fluid and the flow are considered to be incompressable. The equations are

continuity
$$\frac{\partial u}{\partial r} + \frac{u}{r} + \frac{\partial v}{\partial z} = 0 \tag{1}$$

momentum
$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial r} + v\frac{\partial u}{\partial z} - \frac{w^2}{r} = -\frac{1}{\rho}\frac{\partial p}{\partial r} + \nu\left[\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} - \frac{u^2}{r} + \frac{\partial^2 u}{\partial r^2}\right] \tag{2}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial r} + v\frac{\partial v}{\partial z} = -\frac{1}{\rho}\frac{\partial p}{\partial z} + \nu\left[\frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} + \frac{\partial^2 v}{\partial z^2}\right] \tag{3}$$

$$\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial r} + \frac{uw}{r} + v\frac{\partial w}{\partial z} = \nu\left[\frac{\partial^2 w}{\partial r^2} + \frac{1}{r}\frac{\partial w}{\partial r} - \frac{w}{r^2} + \frac{\partial^2 w}{\partial z^2}\right] . \tag{4}$$

We have labeled the radial, axial and azimuthal coordinates as r, z and $\theta$ respectively, and the respective velocity components are u, v and w. The kinematic viscosity, density and pressure of the fluid are $\nu$ , $\rho$ and p, and the time is labeled t. We have assumed no azimuthal dependence. These equations are solved to give the flow of the fluid from the cylinder. The velocity components are numerically determined using an explicit finite difference method applied on a finite Eulerian mesh. The fluid pressure is determined implicitly, using an iterative scheme. Details of the solution precedure are given by Nichols, Hirt and Hotchkiss (3).

To track the free surface of the fluid, a fractional volume of fluid function is used. This function is denoted F and is specified by the equation

$$\frac{\partial F}{\partial t} + \frac{1}{r}\frac{\partial}{\partial r}(rFu) + \frac{\partial}{\partial z}(Fv) = 0 \tag{4}$$

The quantity, F, is carried with the fluid and the value of F in any particular mesh cell ranges from 1 to 0. A value of unity corresponds to a mesh cell which is entirely filled with fluid, while a value of zero corresponds to a mesh cell which is entirely empty. Mesh cells with intermediate values of F contain free surface. The Donor-Acceptor flux

approximation is used to calculate sequential values of F in the mesh cells. When the values of F are determined, free surface is extrapolated through the regions where $0 < F < 1$. This scheme is illustrated in Figure 2. In regions where $F = 0$, the pressure is taken to be zero.

## III. Results:

### A. Flow Pattern.

The flow pattern observed in the calculations is shown in Figure 3. Each computer-generated frame depicts the fluid configuration at the specified time. In the calculation shown, the fluid density was 1 gr $cm^{-3}$ and the kinematic viscosity was 5 stokes. The angular rotation rate of the cylinder was $2.89 \times 10^{2}$ $sec^{-1}$. Line segments indicate the relative magnitude and direction of the local radial and axial velocity components in each mesh cell within the fluid region. The free surface is illustrated with a solid line.

### B. Ejection Rate Analysis

1. Viscosity Effects: The effect of the fluid viscosity was examined by comparing the ejection rate for fluids of density 1 gr $cm^{-3}$ and kinematic viscosities in the range 0.009 to 1.20 stokes. The cylinder rotation rate remained constant at $5.78 \times 10^{2}$ $sec^{-1}$. The various fluid ejection rates are plotted as a function of time in Figure 4. The ejection rates show an initial peak at $2 \times 10^{-3}$ sec, followed by a somewhat fluctuating slower rise to about $1.2 \times 10^{-2}$ sec, after which the rate decays monotonically. Half of the fluid originally in the cylinder is ejected at $t = 1.75 \times 10^{-2}$ sec. The fluctuations in the ejection rate are due to the propogation of waves on the free surface of the fluid. The display of the differential quantity, the rate of fluid ejection, magnifies the observed effects of the surface disturbances.

No significant differences are observed between the ejection rates calculated with the various values of kinematic viscosity. This result can be interpreted in terms of the ratio of the centrifugal force to the viscous force, in a manner similar to the use of the Reynolds number with inertial forces. The appropriate ratio, C, is

$$C = \frac{R^2\Omega}{\nu} \qquad (6)$$

where R is the cylinder radius, $\Omega$ is the angular rotation rate and $\nu$ is the kinematic viscosity of the fluid. The range of values of the centrifugal/viscous force ratio for the calculations shown in Figure 4 is $2.9 \times 10^{3}$ to $6.4 \times 10^{6}$. In these cases, the fluid ejection process is dominated by the centrifugal force, and the viscous forces play no significant role.

To further examine the characterization of the ejection rate by the ratio C, calculations were performed with $C = 1.2 \times 10^{3}$ and with $C = 6.3 \times 10^{1}$. To obtain the low values of C, the spin rate was decreased to $6.3 \times 10^{1}$ $sec^{-1}$ and viscosity values of 5 and 100 stokes were used. Results of these calculations are shown in Figure 5. A smaller ejection rate is observed for the case where $C = 6.3 \times 10^{1}$. We conclude that the centrifugal/viscous force ratio is useful for characterizing the ejection rate. When the value of C is greater than $2.9 \times 10^{3}$, the viscosity of the fluid plays no significant part

in the ejection process. Extensive calculations at low values of C were not performed because the time increment limitations, which the code imposes to control the rapid momentum transfer at high viscosity, caused runs to execute for rather long, costly times, and viscous effects are not important in our region of interest.

2. Rotation Rate Effects: To examine the effect of rotation rate, calculations were performed for a fluid with a viscosity of 5 stokes and a density of 1 gr cm $^{-3}$. Angular spin rates of $2.89 \times 10^{2}$ to $1.16 \times 10^{3}$ $sec^{-1}$ were selected. The calculated ejection rates are shown in Figure 6. The rates show a definite dependence on the spin rate. For calculations with various spin rates, the time at which half of the fluid is ejected, $T_{1/2}$, is plotted as a function of the inverse spin rate in Figure 7. Figure 7 demonstrates that the time scale of the ejection process is inversely proportional to the spin rate.

In Figure 8, the ejection rate for the calculations shown in Figure 6, have been non-dimensionalized by dividing by the volume rotation rate, $Q_R$, given by

$$Q_R = LR^2\Omega \qquad (7)$$

and the nondimensional values have been plotted as a function of the angle of rotation, $\theta$, given by

$$\theta = \Omega t \qquad (8)$$

Before nondimensionalization, the ejection rates were averaged over a time interval of $2 \times 10^{-3}$ sec to reduce the fluctuation level. To within the limits of the smoothing process (indicated on Figure 8), the non-dimensionalization process reduces the ejection rates to the same curve. Figure 8 shows that the magnitude and time dependence of the ejection rate is characterized linearly by the spin rate. The ejection distribution and time scale can be determined from the spin rate and the functional form displayed in Figure 8.

## IV. Summary and Conclusions:

The fluid dynamics simulation code, SOLA-VOF/CSL has been used to investigate aspects of the rate of fluid ejection from the opened end of an azimuthally rotating cylinder. At angular spin rates of $5.78 \times 10^{2}$ $sec^{-1}$, fluid viscosities from 0.009 to 20 stokes showed no significant effect on the ejection rate. This behavior is consistent with the magnitude of the ratio of the centrifugal force to the viscous force. As the ratio becomes smaller, some viscous effects can be observed with the simulation code. The time dependence and magnitude of the ejection rate scales linearly with the spin rate of the cylinder.

## References


1. Hirt, C.W. and Campbell, J. R., "Modifications to SOLA-VOF for Flow Dynamics in Spinning Cylinders," ARRADCOM/CSL Report ARCSL-CR-84064. (unpublished)

2. Stuempfle, A. K., Chemical Research and Development Center, Aberdeen Proving Ground, MD. (unpublished data)

3. Nichols, B. D., Hirt, C. W. and Hotchkiss, R. S., "SOLA-VOF: A Solution Algorithm for Transient Fluid Flow with Multiple Free Boundaries," Los Alamos Scientific Laboratory Report No. LA-8355, 1980.

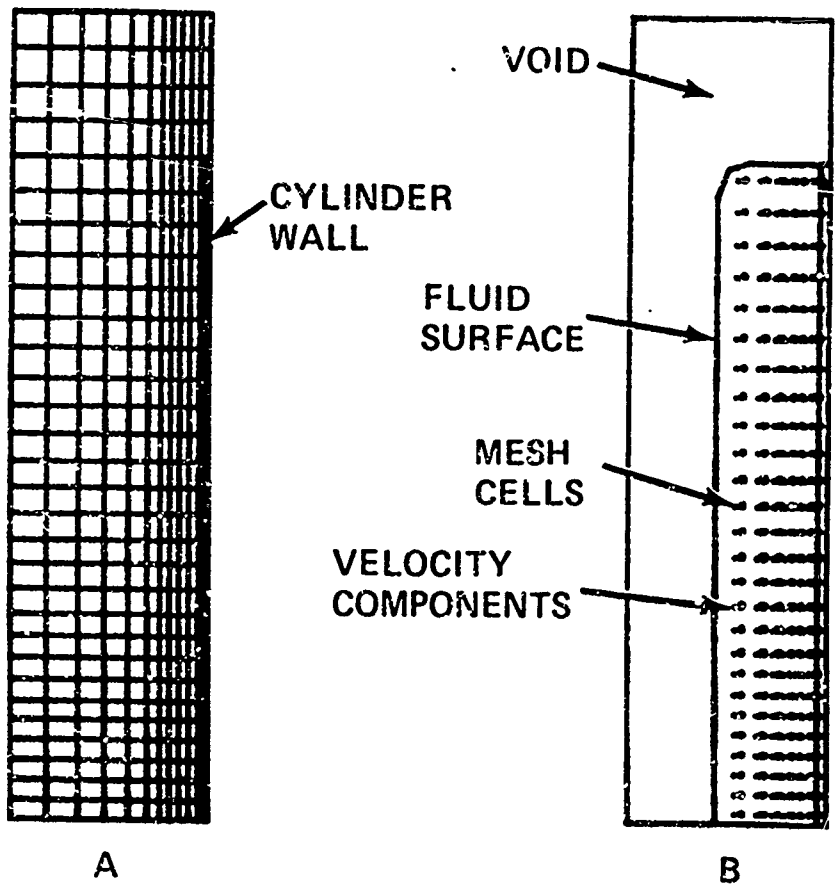



Figure 1. Initial set up of the flow problem showing (a) the Eulerian mesh and (b) the fluid configuration at time, t = 0. The radial and axial velocity components are intitially zero in all mesh cells. The cylinder is 33 cm high and 10 cm in radius. The mesh is 40 cm high and 10.5 cm in radius.

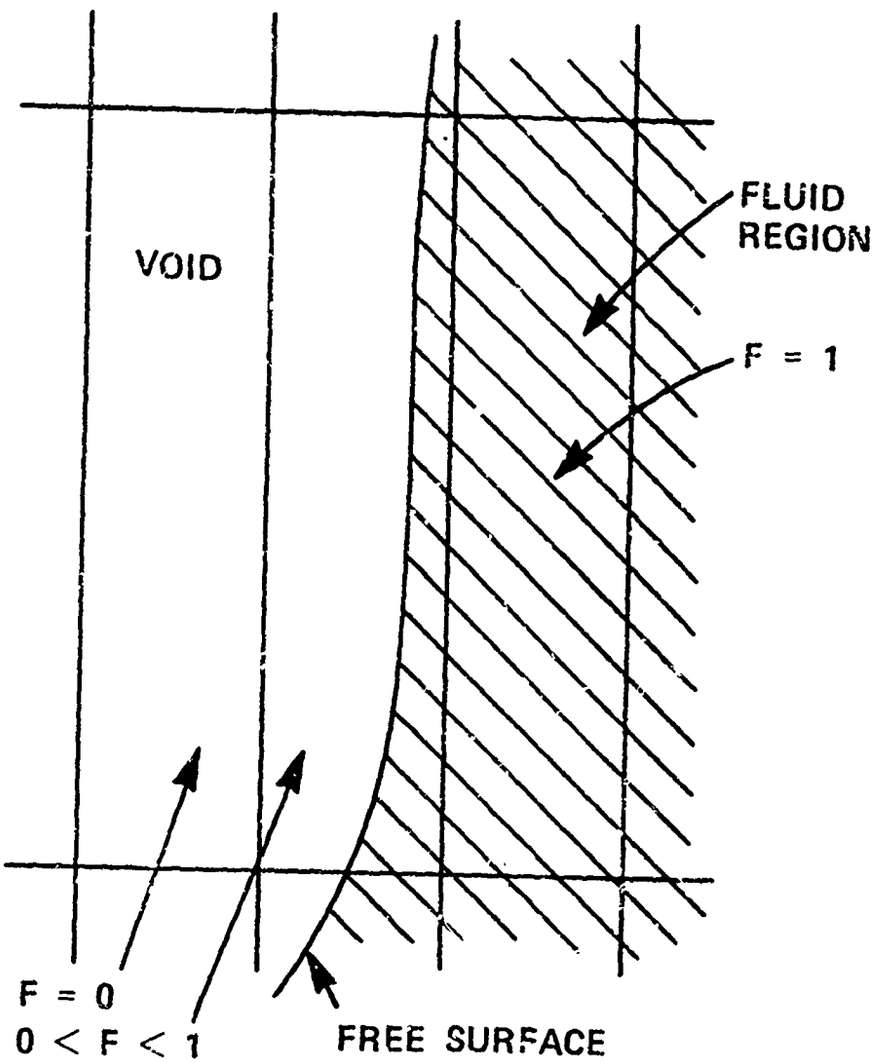



Figure 2. Extrapolation of the free surface through regions where F lies between zero and unity.

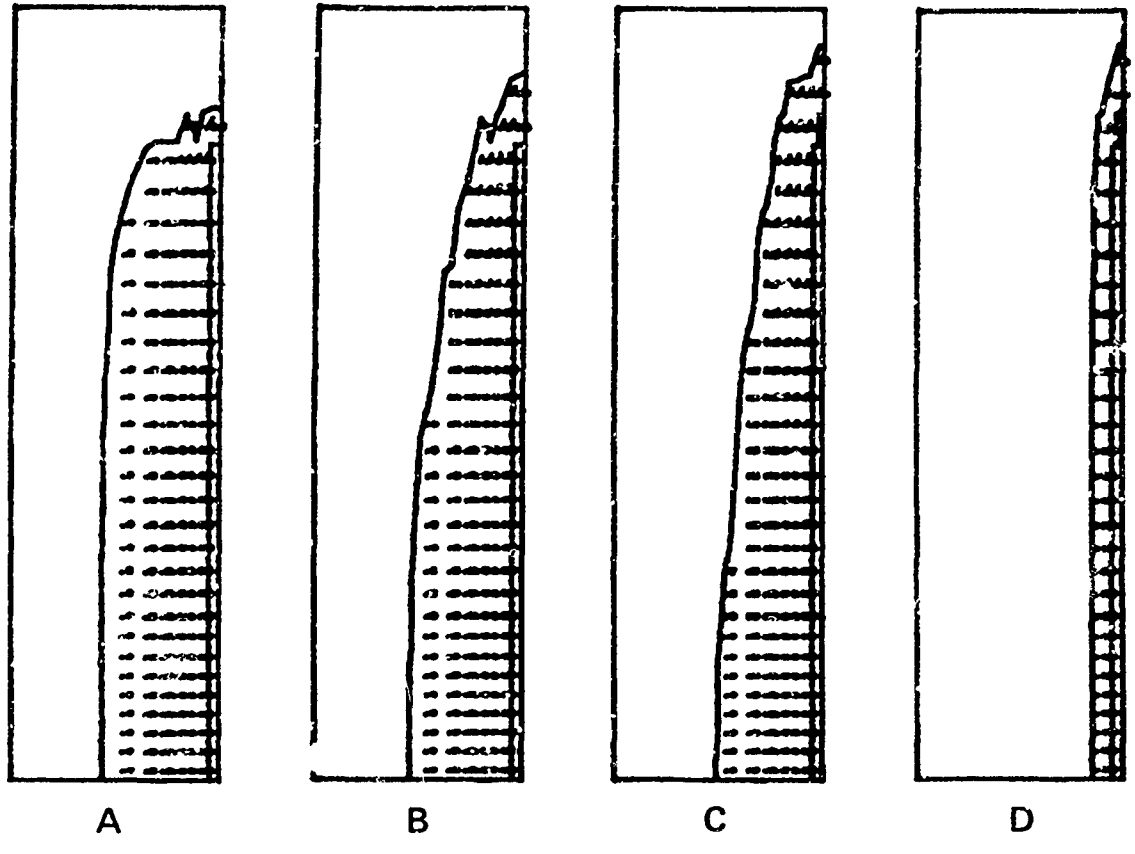



Figure 3. Flow pattern of the ejection process Times shown are (a) 3.06 x $10^{-3}$ sec, (b) 1.00 x $10^{-2}$ sec, (c) 1.51 x $10^{-2}$ sec, and (d) 6 13 x $10^{-2}$ sec. Fluid viscosity is 5 stks, the density is 1 gr $cm^{-3}$, and the angular spin rate is 2.89 x $10^{2}$ $sec^{-1}$.

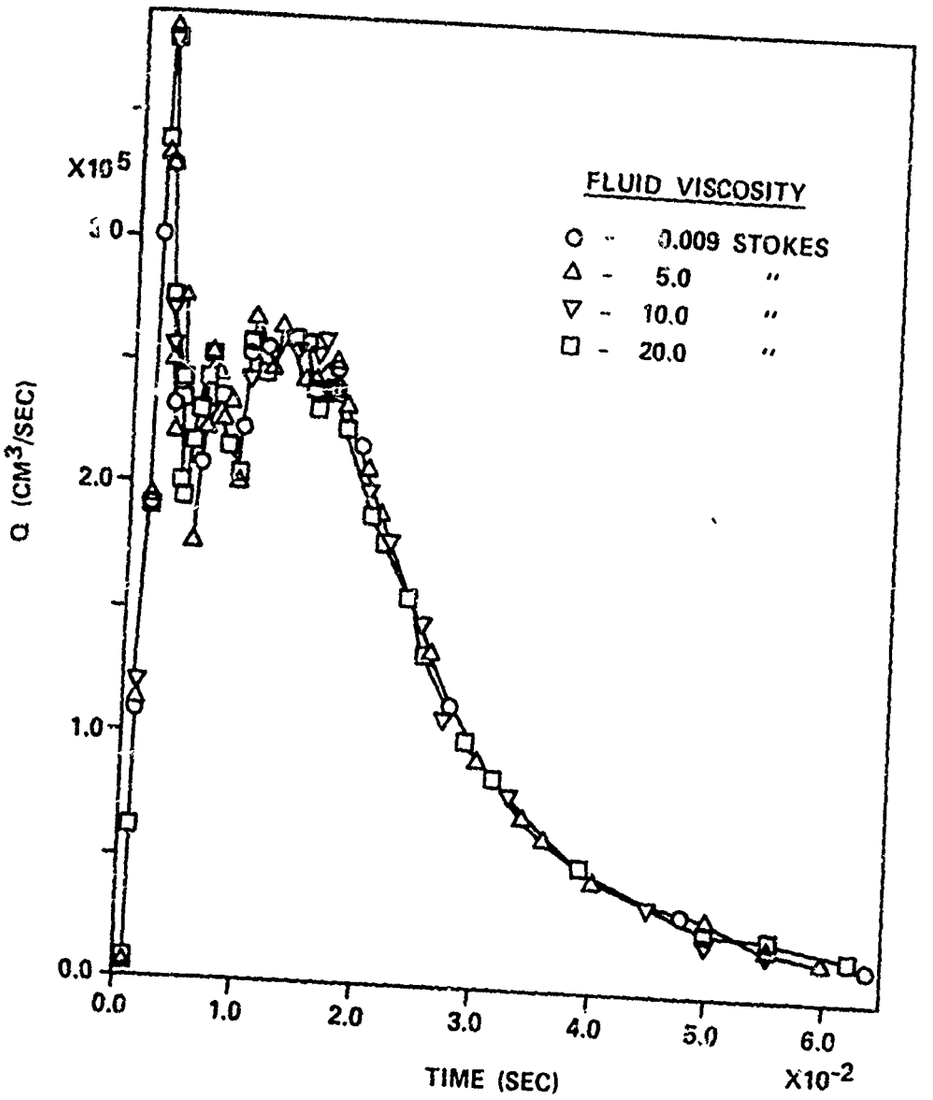



Figure 4. Fluid ejection rates, Q, for fluids with various viscosities, at a spin rate of $5.78 \times 10^2$ sec$^{-1}$.

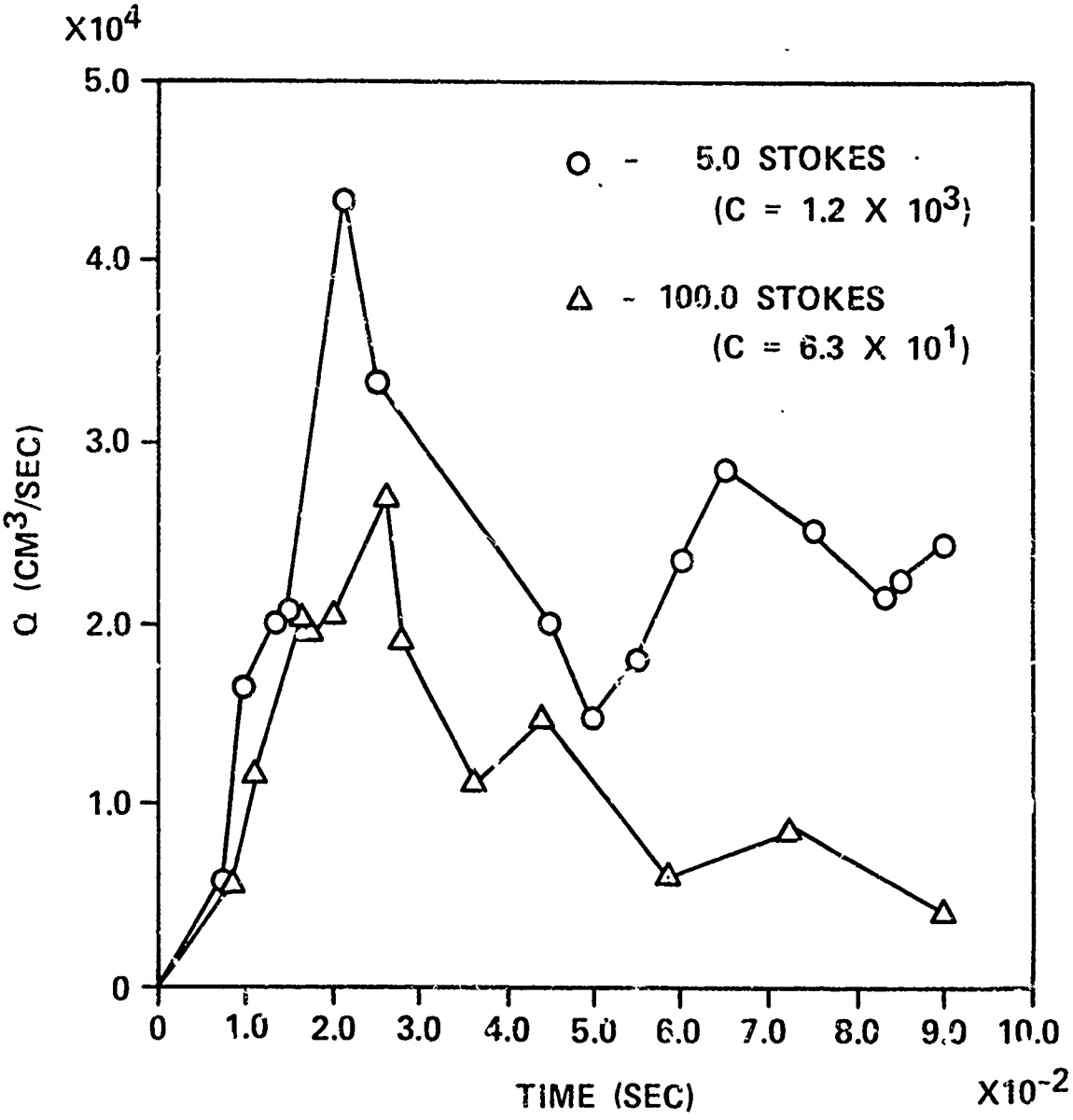



Figure 5. Fluid ejection rates for fluid viscosities of 5.0 and 100.0 stokes, at a spin rate of $6.28 \times 10^{1}$ $sec^{-1}$.

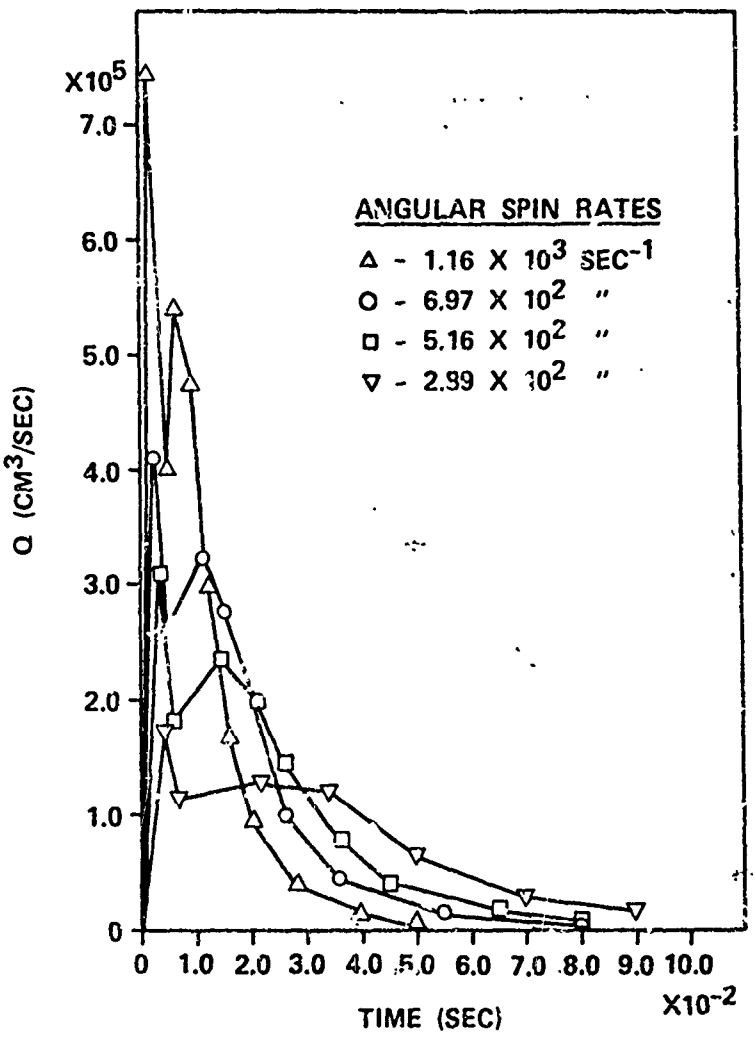



Figure 6. Fluid ejection rates, $Q_r$, for a fluid with a viscosity of 5.0 stokes, at various spin rates.

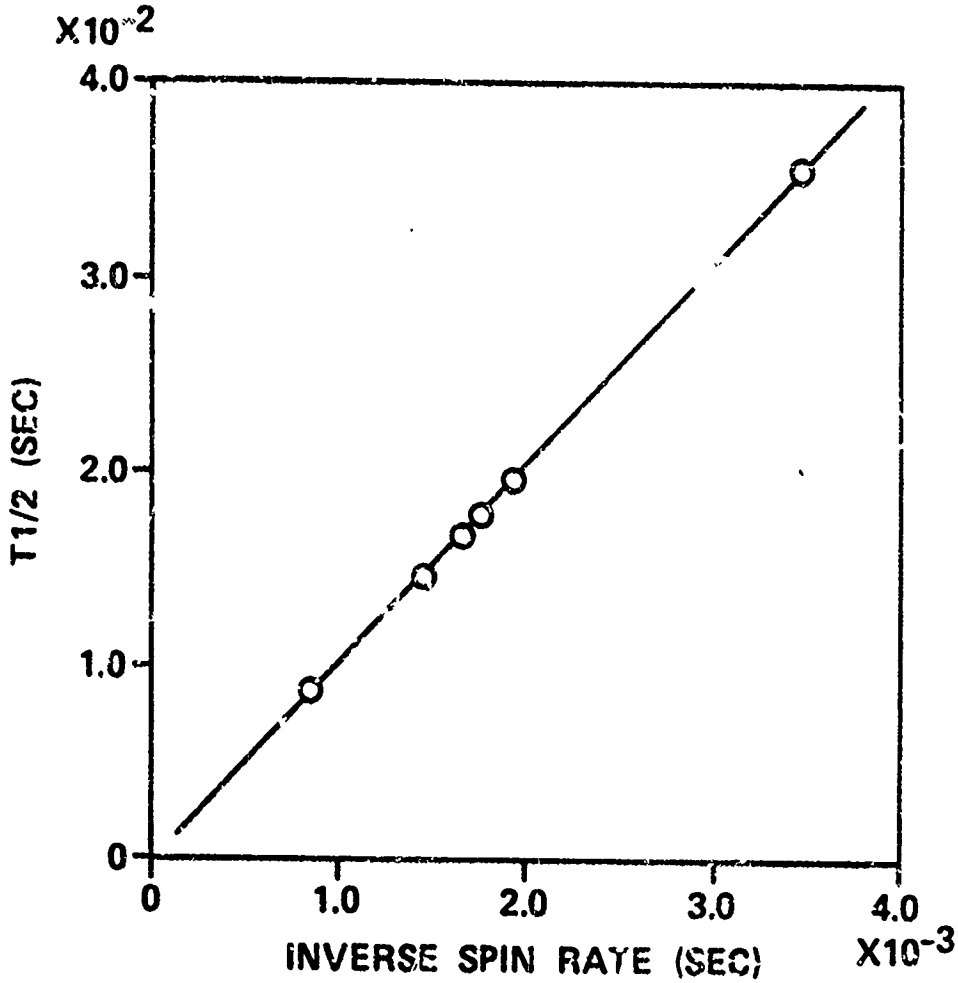



Figure 7. Half fluid ejection time, $T_{\frac{1}{2}}$, vs inverse spin rate.

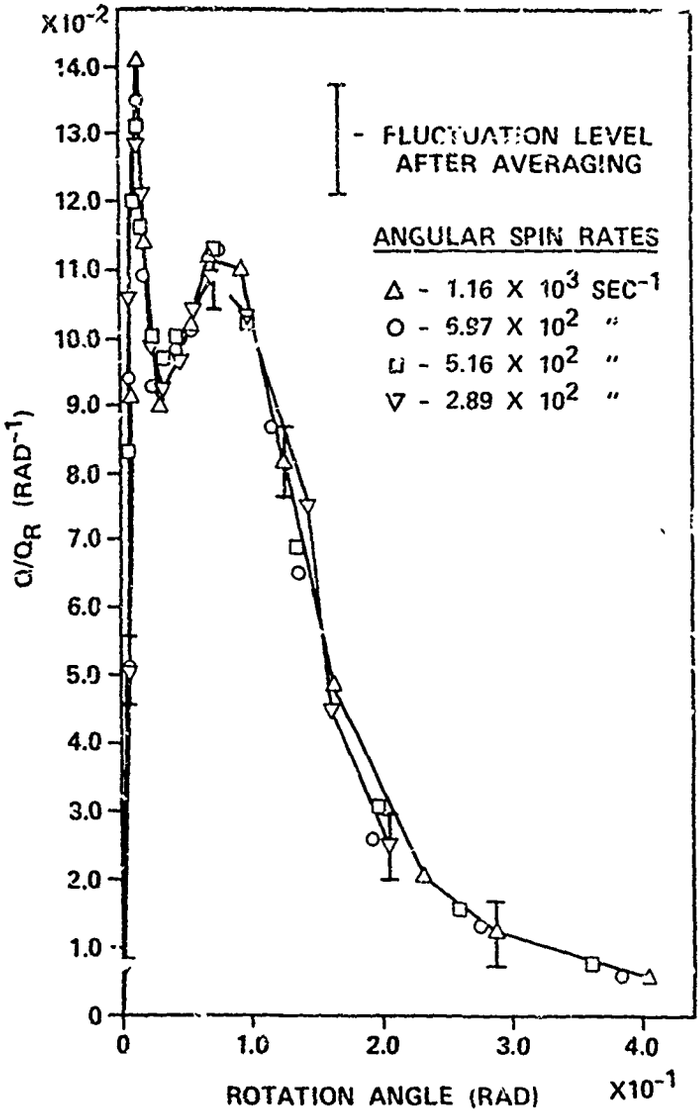



Figure 8. Nondimensional ejection rate for a 5.0 stokes fluid at various spin rates vs the rotation angle.

# A Numerical Algorithm for the Multidimensional, Multiphase, Viscous Equations of Interior Ballistics

James A. Schmitt*
*Ballistic Research Laboratory*
*Aberdeen Proving Ground, MD 21005*

ABSTRACT

A numerical method based on a linearized ADI (Alternating Direction Implicit) scheme is described in the context of the solution procedure for the nonlinear partial differential equations associated with an average two-phase (gas-solid), two-dimensional, fully viscous model of interior ballistics. This method was chosen because the linearization of the time-differenced equations within the temporal truncation error permits a non-iterative solution procedure for this implicit scheme, and the splitting of difference equations along the coordinate directions provides a block tridiagonal structure of the solution matrices. The implementation of the algorithm possesses several novel features: the algorithm is derived in the context of a moving coordinate system; the non-conservational form of the governing equations avoids both mass sources (which can be generated by grid motion) and singular solution matrices (which can arise in regions of one-phase flow in a two-phase calculation); and finally, the Jacobian type matrices (which arise from the linearization process) are determined by numerical differentiation instead of the usual analytic calculations. This numerical scheme is encoded in the DELTA computer code.

Verification of the DELTA algorithm is obtained by comparing simulations to an analytic solution of an isentropic core flow and to the two-dimensional results of an experiment.

* Currently with AT&T Bell Laboratories, Crawford Corners Road, Holmdel, NJ 07733.

## 1. INTRODUCTION

The flowing medium in a gun tube typically is a mixture of a compressible gas and burning solid propellant grains. Details of the flow are important for weapons development, but only bulk properties can be routinely measured, such as the trajectory of the projectile, the pressure history at a fixed station, the heating inside the gun tube, etc. Therefore, a need exists for a detailed mathematical model of interior ballistics two-phase flows, and an algorithm to solve the corresponding equations.

The three-dimensional mathematical model is developed carefully in Reference [1]. References [2] and [3] are shorter versions of Reference [1]. This two-phase model is based on instantaneous, finite volume, weighted averaging, and consists of nonlinear partial differential equations, constitutive laws for the averaged variables, and correlations for the interphase terms. The transient phenomena included in this model are: the convection of the phases driven by gas phase pressure, gas phase viscous stresses, turbulence, intergranular stresses, interphase drag, and interphase mass transfer due to burning of the solid grains; the change of energy in the gas due to convection, pressure, laminar and turbulence dissipation, conduction, and interphase heat transfer; and the change of the geometry and number of the burning grains. These phenomena occur within the volume defined by the gun tube and the base of an accelerating projectile. The motion of the projectile and the two-phase flow field are coupled via the gas pressure exerted on the projectile's base.

This model is specialized to the case of axial symmetry of the flow within the tube. Appendicies A and B, which are reproduced from Reference [1], list all the differential equations, constitutive laws and correlations. This axial symmetric model and numerical scheme are encoded in the DELTA computer code. The purpose of this paper is to describe in some detail the numerical algorithm (Section 2), and to present some validating computer runs of the combined model and algorithm (Section 3).

Previous multi-dimensional, multi-phase work applied to interior ballistics includes that from Paul Gough Associates, Inc. References [4-5], and Scientific Research Associates, Inc. References [6-7]. Gough's work addresses the inviscid flow during the ballistic cycle, i.e. the average gas phase viscous stresses, heat conduction, and turbulence are excluded. Thus, the phenomenology of pressure waves inside a gun tube is modelled primarily in Gough's work. The people at Scientific Research Associates considered the viscous flow phenomena, i.e. the development of boundary layers inside a gun tube. The work reported here is more similar to that of Scientific Research Associates, but deviates from it in the model, i.e. the equations, choice of dependent variable, some correlations, and in alterations in the numerical algorithm. Differences and similarities in the model are addressed in detail in Reference [1], and in the algorithm in Section 2 of this paper.

## 2. ALGORITHM

### 2.1 Governing Equations

The governing equations listed in Appendices A and B are a set of nonlinear partial differential equations which are first order in time and second order in the two spatial coordinates. The rationale for the specific form of the equations is given in Reference [1]. The general form is

$$\mathbf{y}_t = \mathbf{G}(r,z,t,\mathbf{y},\mathbf{y}_r,\mathbf{y}_z,\mathbf{y}_{rr},\mathbf{y}_{rz},\mathbf{y}_{zz}), \tag{2.1}$$

where the independent variables of time, radial coordinate and axial coordinate are denoted by $t$, $r$, $z$, respectively. The vector of dependent variables is denoted by $\mathbf{y}$, and the partial derivatives of $\mathbf{y}$ with respect to the spatial and temporal coordinates are denoted by subscripts. The components of the vector $\mathbf{y}$ can be the radial, circumferential (swirl), axial components of the gas phase velocity $(u,v,w)$, respectively; the radial, circumferential (swirl), axial components of the solid phase velocity $(u^*,v^*,w^*)$, respectively; the gas phase specific entropy $s$ ; the logarithm of gas phase pressure $q$ ; the regression distance of the solid phase $d^*$ ; the number of particles $m^*$ ; the particle surface temperature $T^*$ ; and two variables to define the turbulence in the flow field. Thus, depending on the simulation, the number of dependent variables change from a minumum of four to a maximum of thirteen. For the case of an one-phase, laminar flow simulation with swirl, the dependent vector $\mathbf{y}$ has five components, $u$, $v$, $w$, $s$ and $q$. For the case of a two-phase, turbulent flow simulation with ignition and burning of the solid phase, the dependent vector $\mathbf{y}$ has nine components, $u$, $w$, $u^*$, $w^*$, $s$, $q$, $d^*$, $m^*$, $T^*$. This assumes the absense of swirl and an algebraic turbulence model (i.e. the turbulence properties are described only by algebraic relations). The components of the vector $\mathbf{G}$ are nonlinear functions which can depend on the variables $r$, $z$, $t$, $\mathbf{y}$, $\mathbf{y}_r$, $\mathbf{y}_z$, $\mathbf{y}_{rr}$, $\mathbf{y}_{rz}$, $\mathbf{y}_{zz}$.

The spatial domain is a confined volume within a tube bounded in length by a stationary wall (the gun breech) and the base of an accelerating projectile. The radius of the tube can depend on the axial position from the breech which is denoted by $z_B$. The radial coordinate varies from the axis of symmetry to the tube wall. The radial positions of the axis and wall are denoted by $r_A$ and $r_w(z)$, respectively. The axial coordinate varies from the breech to the base of the projectile. The axial position of the breech and projectile base may have a radial dependence which is denoted by $z_B(r)$ and $z_p(r)$, respectively.

The projectile is assumed to move as a rigid body. The unsteady projectile motion is governed by the following equations:

$$w_p = \frac{dz_p(r_*,t)}{dt} , \tag{2.2}$$

$$\frac{dw_p(t)}{dt} = \frac{1}{m_{AUG}}\left\{ 2\pi\int_{r_A}^{r_w(z_p)} p(r,z_p,t)\, r\, dr - F_A - F_D - F_B \right\} , \tag{2.3}$$

$$m_{AUG} = m_p + I_m\left[\frac{\theta_R}{r_w(z_p)}\right]^2 , \tag{2.4}$$

$$v_p(r,t) = w_p(t)\,\theta_R\frac{r - r_A}{r_W - r_A} , \tag{2.5}$$

where $w_p, v_p, m_p$ denote the axial velocity, circumferential velocity and mass of the projectile, respectively. The forces that retard the motion of the projectile are those due to air resistance, friction between the projectile and tube wall, and gas leakage around the projectile, and are denoted by $F_A$, $F_D$, and $F_B$, respectively. These retarding forces are assumed to be known functions. If the tube is rifled, an additional phenomena is present which causes the projectile to rotate, and its mass to be effectively increased via equation (2.4). In this case the angle of rifling $\theta_R$ is nonzero, and the moment of inertia of the projectile $I_m$ must be given. Because the pressure $p$ is determined from the solution of governing equations of the flow field, which depends on the value of $w_p$, equations (2.1) - (2.5) represent a coupled system with a moving boundary.

### 2.2 Numerical Algorithm

We want to compute by finite difference approximations the transient values of the variables which describe the fluid dynamics of the flow in the region confined by the inner tube wall, breech, and moving projectile. One way to calculate in this expanding computational region is by an "accordion" type grid in the axial direction, i.e. the first and last axial grid points are attached to the breech and projectile, respectively, and the mesh expands as the projectile accelerates down the tube. Thus, the physical grid moves in accordance with the projectile motion. With regard to the spatial finite difference approximation, the goal is to obtain an accurate approximation to the actual physical happening. It can be shown that the finite difference approximations to the physical variables in the physical mesh are the same whether one directly differences on the physical grid, or one differences on a transformed grid and then transforms back to the physical grid. Higher accuracy in a transformed space is not meaningful if it is lost in the transformation back to the physical space. Furthermore, our physical grids will be orthogonal or nearly orthogonal. Thus, we choose to compute finite difference approximations to spatial derivatives on the physical grid. An additional advantage of our method is that the governing equations need not be transformed, and thus are simpler to understand and change in the compute code.

The finite differencing of the time derivatives can be of two generic types: implicit or explicit (See Reference [8]). The principal advantage of an implicit scheme is its superior stability properties compared to an explicit scheme. For convection-diffusion type problems like those given by equation (2.1), an explicit finite difference method has two

stability conditions, the Courant-Friedrichs-Lewy condition and the viscous stability limits In one dimension, these condition are:

$$\Delta t = \text{CFL}\,\frac{\Delta x}{c+w}\ , \tag{2.6}$$

$$\Delta t = \text{VSL}\,\frac{\rho}{2\mu}(\Delta x)^2\ ,$$

where $\Delta x$ is the spatial mesh increment, $c$ is the sound speed, $w$ is the gas velocity, $\mu$ is the viscosity and $\rho$ is the density. The constants CFL and VSL are less than or equal to one. For simulations which involve boundary layers, small grid sizes are necessary. Thus, for this type of simulation, the time step (the size of $\Delta t$) must be proportional to the square of the smallest grid increment for an explicit scheme. On the other hand, most implicit schemes have no corresponding stability conditions, and significantly larger time steps based on accuracy considerations rather than stability can be used. The basic disadvantage of implicit algorithms is that they tend to be more complicated than explicit schemes, and thus more difficult to understand and implement. In particular, applying a standard implicit scheme to system of equation (2.1), we obtain a nonlinear system of algebraic equations in the variables at the new time level. This system can be quite large and complex because it possesses an equation for each dependent variable and for each grid point in the two-dimensional computational mesh. These equations are coupled via the spatial derivatives. Iterative methods are the most common solution procedure, but they can be quite complex and time-consuming for such a general system. To mitigate these undesirable characteristics, we apply a method developed in References [9-11]. A salient feature of this method is the temporal linearization of the nonlinear terms to within the local truncation error of the finite difference approximation of the time derivative. The resulting system can then be represented in a matrix equation.

$$\mathbf{A}\mathbf{y}^n = \mathbf{b} \tag{2.7}$$

where $\mathbf{y}^n$ is the vector of unknown dependent variables at the new time level, and $\mathbf{A}$ and $\mathbf{b}$ are the matrix and vector of values at the known time level, respectively. Furthermore, the matrix $\mathbf{A}$ can be structured if we decompose the time-differenced equation of (2.1) into two systems of equations, each of which involves the spatial derivatives of the unknown variable in only one coordinate direction. This decomposition or splitting is done so that the error incurred is of the order of the local temporal truncation error, and so that the decomposed or split equations still form a consistent approximation to equation (2.1). When centered differences are used to approximate the spatial derivatives, tridiagonal matrices are obtained. Because we are dealing with a system of equations, the matrices are block tridiagonal where the size of the blocks is equal to the number of dependent variables. This method of splitting the implicitly differenced equations is called an Alternating Direction Implicit (ADI) scheme (see Reference [8]). Because we have also linearized the equations, we shall refer to this method as a linearized ADI scheme.

This general linearized ADI scheme is applied to the instantaneous, finite-volume, weighted, averaged equations of interior ballistics with several unique features: The algorithm is derived for a moving coordinate system, and has no mass source due to the

motion of the grid. The elements of the matrices derived by the linearization process are obtained directly by numerical differentiation which bypasses the tedious and error prone task of analytically deriving each element, and the subsequent coding of these complex expressions. The spatial differencing is performed directly on nonuniform distributed grids.

### *2.2.1 Algorithm for Non-Boundary, Non-Center-Line Points*

We derive the numerical scheme for the system of equations (2.1) on a moving coordinate system; i.e., the coordinates of the spatial grid system varies in time. We let the superscripts $n$ denote the new time level and $c$ the current time level. The change in the $j^{th}$ coordinate position of a spatial coordinate $x$ from level $c$ to $n$ is denoted by

$$\Delta x_j = x_j^n - x_j^c = \left(\frac{\partial x}{\partial t}\right)_{t=t^*}(t^n - t^c) = O(\Delta t), \quad t^c < t^* < t^n, \tag{2.8}$$

where $\Delta t = t^n - t^c$. A Taylor expansion of $\mathbf{y}^n = \mathbf{y}(r^n, z^n, t^n)$ about the current values at $(r^c, z^c, t^c)$ is

$$\begin{aligned}\mathbf{y}^n = \mathbf{y}^c &+ \frac{\partial \mathbf{y}^c}{\partial t}\Delta t + \frac{\partial \mathbf{y}^c}{\partial z}\Delta z + \frac{\partial \mathbf{y}^c}{\partial r}\Delta r \\ &+ \frac{1}{2}\left[\frac{\partial^2 \mathbf{y}^c}{\partial t^2}\Delta t^2 + 2\Delta t\Delta z\frac{\partial^2 \mathbf{y}^c}{\partial t \partial z} + 2\Delta t\Delta r\frac{\partial^2 \mathbf{y}^c}{\partial t \partial r}\right. \\ &\left. + 2\Delta r\Delta z\frac{\partial^2 \mathbf{y}^c}{\partial r \partial z} + \frac{\partial^2 \mathbf{y}^c}{\partial r^2}\Delta r^2 + \frac{\partial^2 \mathbf{y}^c}{\partial z^2}\Delta z^2\right] + O(\Delta t^3)\end{aligned} \tag{2.9}$$

By adding $(1 - \beta)$ times (2.9) and $\beta$ times a similar expansion to (2.9) of $\mathbf{y}^c$ expanded about $\mathbf{y}^n$, and noting that $\mathbf{y}^n - \mathbf{y}^c = O(\Delta t)$, $z^n - z^c = O(\Delta t)$, and $r^n - r^c = O(\Delta t)$, we obtain

$$\begin{aligned}&\mathbf{y}^n - \beta\Delta t\,\mathbf{G}^n - \beta\,\Delta r\left[\frac{\partial \mathbf{y}}{\partial r}\right]^n - \beta\,\Delta z\left[\frac{\partial \mathbf{y}}{\partial z}\right]^n \\ &= \mathbf{y}^c + (1-\beta)\Delta t\,\mathbf{G}^c + (1-\beta)\Delta r\left[\frac{\partial \mathbf{y}}{\partial r}\right]^c + (1-\beta)\Delta z\left[\frac{\partial \mathbf{y}}{\partial z}\right]^c \\ &+ (\beta - \tfrac{1}{2})O(\Delta t^2) + E_T(\Delta t^3), \qquad \tfrac{1}{2} \le \beta \le 1.0,\end{aligned} \tag{2.10}$$

where $E_T(\Delta t^3)$ denotes the neglected truncation error of $O(\Delta t^3)$. For a stationary grid $\Delta r = \Delta z = 0$, we obtain the standard Crank-Nicolson scheme for integration parameter $\beta = \frac{1}{2}$ and the standard fully implicit scheme for $\beta = 1.0$. Equation (2.10)

is a nonlinear system of equations in $\mathbf{y}^n$ because $\mathbf{G}^n$ is a nonlinear function. To make (2.10) a system of linear equations, we linearize $\mathbf{G}^n$ via a Taylor expansion about the current level, i.e.

$$\mathbf{G}^n = \mathbf{G}^c + \left(\frac{d\mathbf{G}}{dt}\right)^c \Delta t + E_L(\Delta t^2) \tag{2.11}$$

$$= \mathbf{G}^c + \left(\frac{\partial \mathbf{G}}{\partial t}\right)^c \Delta t + \left(\frac{\partial \mathbf{G}}{\partial r}\right)^c \Delta r + \left(\frac{\partial \mathbf{G}}{\partial z}\right)^c \Delta z$$

$$+ D^c \Delta \mathbf{y} + DR^c \Delta \mathbf{y}_r + DZ^c \Delta \mathbf{y}_z$$

$$+ DRR^c \Delta \mathbf{y}_{rr} + DRZ^c \Delta \mathbf{y}_{rz} + DZZ^c \Delta \mathbf{y}_{zz}$$

$$+ E_L(\Delta t^2),$$

where $E_L(\Delta t^2)$ denotes the neglected linearization error of $O(\Delta t^2)$. The Jacobian type matrices are denoted by $D$, $DR$, $DZ$, $DRR$, $DRZ$, $DZZ$, are evaluated at the current time level, and are defined as

$$D^c_{ij} \equiv \left(\frac{\partial \mathbf{G}_i}{\partial \mathbf{y}_j}\right)^c, \quad DR^c_{ij} \equiv \left(\frac{\partial \mathbf{G}_i}{\partial \mathbf{y}_{r_j}}\right)^c, \ldots, \quad DZZ^c_{ij} \equiv \left(\frac{\partial \mathbf{G}_i}{\partial \mathbf{y}_{zz_j}}\right)^c. \tag{2.12}$$

Upon substituting (2.11) into (2.10), we obtain the following linear system of equations in $\mathbf{y}^n$ after algebraic manipulation:

$$\mathbf{y}^n - \beta\left\{\Delta r\left[\frac{\partial \mathbf{y}}{\partial r}\right]^n + \Delta z\left[\frac{\partial \mathbf{y}}{\partial z}\right]^n + \Delta t\left[D^c\mathbf{y}^n + DR^c\left[\frac{\partial \mathbf{y}}{\partial r}\right]^n + DZ^c\left[\frac{\partial \mathbf{y}}{\partial z}\right]^n\right.\right.$$

$$\left.\left. + DRR^c\left[\frac{\partial^2 \mathbf{y}}{\partial r^2}\right]^n + DRZ^c\left[\frac{\partial^2 \mathbf{y}}{\partial r \partial z}\right]^n + DZZ^c\left[\frac{\partial^2 \mathbf{y}}{\partial z^2}\right]^n\right]\right\} \qquad (2.13)$$

$$= \mathbf{y}^c - \beta\left\{\Delta r\left[\frac{\partial \mathbf{y}}{\partial r}\right]^c + \Delta z\left[\frac{\partial \mathbf{y}}{\partial z}\right]^c + \Delta t\left[D^c\mathbf{y}^c + DR^c\left[\frac{\partial \mathbf{y}}{\partial r}\right]^c + DZ^c\left[\frac{\partial \mathbf{y}}{\partial z}\right]^c\right.\right.$$

$$\left.\left. + DRR^c\left[\frac{\partial^2 \mathbf{y}}{\partial r^2}\right]^c + DRZ^c\left[\frac{\partial^2 \mathbf{y}}{\partial r \partial z}\right]^c + DZZ^c\left[\frac{\partial^2 \mathbf{y}}{\partial z^2}\right]^c\right]\right\}$$

$$+ \Delta r\left[\frac{\partial \mathbf{y}}{\partial r}\right]^c + \Delta z\left[\frac{\partial \mathbf{y}}{\partial z}\right]^c + \Delta t\,\mathbf{G}^c + \beta\,\Delta t\left[\Delta t\left[\frac{\partial \mathbf{G}}{\partial t}\right]^c + \Delta r\left[\frac{\partial \mathbf{G}}{\partial r}\right]^c\right.$$

$$\left. + \Delta z\left[\frac{\partial \mathbf{G}}{\partial z}\right]^c\right] + (\beta - \tfrac{1}{2})\,O(\Delta t^2) + E_T(\Delta t^3) + E_L(\Delta t^3).$$

The neglected linearization error $E_{Li}$ is now $O(\Delta t^3)$ because $\mathbf{G}^n$ was multiplied by $\Delta t$ in (2.10). Thus, the linearization process does not alter the order of the temporal error.

The values $\mathbf{y}^c$, $r^c$, $z^c$, $t^c$, $r^n$, $z^n$, $t^n$ are all known before the start of the integration routine to determine the values of $\mathbf{y}^n$. Thus, the right hand side of (2.13) is a known vector. The left hand side of (2.13) can be written as a matrix with known values times a vector of unknowns, $\mathbf{y}^n$. Thus, (2.13) has the form of (2.7). The matrix $\mathbf{A}$ interconnects values of $\mathbf{y}^n$ at a grid point to all the values of its spatial neighbors via the first and second spatial partial derivatives at level $n$. If the term $\left[\frac{\partial^2 \mathbf{y}}{\partial r \partial z}\right]^n$ were absent from the left hand side of (2.13), we could decompose (2.13) into two matrix equations which are highly structured and easily solvable while still retaining only two time levels of the dependent variable vector, i.e. $\mathbf{y}^n$ and $\mathbf{y}^c$. To this end, we linearize the mixed derivative at the $n$ level about the current level and retain only the leading term:

$$\left[\frac{\partial^2 \mathbf{y}}{\partial r \partial z}\right]^n = \left[\frac{\partial^2 \mathbf{y}}{\partial r \partial z}\right]^c + E_{LA}(\Delta t). \tag{2.14}$$

Substituting (2.14) into (2.13), we obtain after some algebraic manipulations

$$\left[\mathbf{I} - \beta(D_r^n + D_z^n)\right]\mathbf{y}^n = \left[\mathbf{I} - \beta(D_r^c + D_z^c)\right]\mathbf{y}^c + L\mathbf{y}^c \tag{2.15a}$$

$$= (\beta - \tfrac{1}{2})O(\Delta t^2) + E_T(\Delta t^3) + E_L(\Delta t^3) + E_{LA}(\Delta t^2),$$

where the operators are defined as follows:

$$D_r^k \equiv \Delta t\left[\frac{\partial}{\partial r}\right]^k + \left[D^c + DR^c\left[\frac{\partial}{\partial r}\right]^k + DRR^c\left[\frac{\partial^2}{\partial r^2}\right]^k\right], \tag{2.15b}$$

$$D_z^k \equiv \Delta z\left[\frac{\partial}{\partial z}\right]^k + \left[DZ^c\left[\frac{\partial}{\partial z}\right]^k + DZZ^c\left[\frac{\partial^2}{\partial z^2}\right]^k\right], \tag{2.15c}$$

$$L\mathbf{y}^c \equiv \Delta r\left[\frac{\partial \mathbf{y}}{\partial r}\right]^c + \Delta z\left[\frac{\partial \mathbf{y}}{\partial z}\right]^c + \Delta t\,\mathbf{G}^c \tag{2.15d}$$

$$+ \beta\,\Delta t\left[\Delta t\left[\frac{\partial \mathbf{G}}{\partial t}\right]^c + \Delta r\left[\frac{\partial \mathbf{G}}{\partial r}\right]^c + \Delta z\left[\frac{\partial \mathbf{G}}{\partial z}\right]^c\right].$$

The symbol $\mathbf{I}$ represents the identity matrix, and the superscript $k$ can be either $n$ or $c$. The "lagging" of the mixed derivative (2.14) increased the error of the approximation to $O(\Delta t^2)$ for any value of the integration parameter $\beta$, but we gain a structured matrix.

We can decompose (2.15a) along coordinate directions in the following manner:

$$\left[\mathbf{I} - \beta D_r^n\right]\mathbf{y}^I = \left[\mathbf{I} - \beta D_r^c\right]\mathbf{y}^c + L\mathbf{y}^c\,, \tag{2.16}$$

$$\left[\mathbf{I} - \beta D_z^n\right]\mathbf{y}^F = \left[\mathbf{I} - \beta D_z^c\right]\mathbf{y}^c + (\mathbf{y}^I - \mathbf{y}^c). \tag{2.17}$$

Equation (2.16) constitutes the radial sweep of this Alternating Direction Implicit method because it involves only spatial derivatives in the radial direction at the new time level. One solves this equation for each fixed axial index and for radial indices varying from the axis of symmetry to the gun tube wall. When three point centered spatial finite differences are used to approximate the spatial derivatives in the radial direction, the matrix $\mathbf{I} - \beta D_r^n$ is a block tridiagonal matrix. The size of each block is equal to the number of dependent variables. The number of block rows is equal to the number of grid points from the axis of symmetry to the gun tube wall, denoted by $JRMX$. The block rows from the second grid points to one away from the wall, namely $JRMX$-1 is determined by (2.16). The entries of the first and last block rows are

determined from the conditions imposed at the axis of symmetry and wall, respectively. The right hand side of (2.16) is a known vector because it is evaluated at the current time-step. The solution of this equations is the intermediate values of the dependent variables $\mathbf{y}^I$.

Equation (2.17) constitutes the axial sweep of this two sweep scheme because it involves only the spatial derivatives in the axial direction at the new time level. One solves this equation for each fixed radial index and for axial indices varying from the breech to the base of the projectile. The matrix $\mathbf{I} - \beta D_z^n$ is a block tridiagonal matrix when three point centered spatial finite differences are used to approximate the partial derivatives in the axial direction. The size of the blocks are the same as in the radial sweep, and the number of block rows is equal to the number grids points placed from the breech to the projectile base, denoted by *JZMX*. Equation (2.17) is used to determine the entries in the block rows from the second point (one after the breech) to *JZMX*–1 (one before the projectile base). The entries in the first and last block rows are determined from the boundary conditions imposed at the breech and projectile base, respectively. Because the intermediate value $\mathbf{y}^I$ is known, the right hand side of (2.17) is known. The solution of (2.17) is the value of the dependent variables at the new time level, denoted by $\mathbf{y}^F$. The values of the solution vector $\mathbf{y}^F$ of (2.17) and those of the solution vector $\mathbf{y}^n$ of (2.15) differ by the time error introduced by the splitting (2.16)-(2.17). To determine the order of this error, we substitute (2.17) into (2.16) and obtain

$$(\mathbf{I} - \beta D_r)\Big[(\mathbf{I} - \beta D_z)\mathbf{y}^F - (\mathbf{I} - \beta D_z)\mathbf{y}^c + \mathbf{y}^c\Big] = [\mathbf{I} - \beta D_r]\mathbf{y}^c + L\mathbf{y}^c \qquad (2.18)$$

which simplies to

$$(\mathbf{I} - \beta D_r - \beta D_z)\mathbf{y}^F = (\mathbf{I} - \beta D_r - \beta D_z)\mathbf{y}^c + L\mathbf{y}^c - \beta^2 D_r D_z(\mathbf{y}^F - \mathbf{y}^c). \qquad (2.19)$$

Subtracting (2.15a) from (2.19), have

$$(\mathbf{I} - \beta D_r - \beta D_z)(\mathbf{y}^F - \mathbf{y}^n) = -\beta^2 D_r D_z(\mathbf{y}^F - \mathbf{y}^n + \mathbf{y}^n - \mathbf{y}^c) \qquad (2.20)$$

or

$$(\mathbf{I} - \beta D_r - \beta D_z + \beta^2 D_r D_z)(\mathbf{y}^F - \mathbf{y}^n) = -\beta^2 D_r D_z(\mathbf{y}^n - \mathbf{y}^c). \qquad (2.21)$$

We note that the coefficient of $\mathbf{y}^F - \mathbf{y}^n$ is $O(1)$, $D_r$ and $D_z$ are each $O(\Delta t)$ and $(\mathbf{y}^n - \mathbf{y}^c)$ is at least $O(\Delta t)$. Thus,

$$\mathbf{y}^F = \mathbf{y}^n + E_S(\Delta t^3), \qquad (2.22)$$

that is, $\mathbf{y}^F$ is equal to $\mathbf{y}^n$ to within the local truncation error of the scheme.

Equations (2.16), (2.17) with the definitions (2.15b)-(2.15d) represent the time differenced, linearized, ADI scheme. We now turned to the finite difference approximation of the spatial derivatives. The standard centered finite difference approximations to the spatial partial derivatives in the coordinate direction $x$ at the $i^{th}$ grid point are

$$\left\{\frac{\partial \mathbf{y}}{\partial x}\right\}_i \approx \frac{h_i(\mathbf{y}_i - \mathbf{y}_{i-1})}{h_{i-1}(h_i + h_{i+1})} + \frac{h_{i-1}(\mathbf{y}_{i+1} - \mathbf{y}_i)}{h_i(h_i + h_{i+1})}, \tag{2.23}$$

$$\left\{\frac{\partial^2 \mathbf{y}}{\partial x^2}\right\}_i \approx \frac{2(\mathbf{y}_{i-1} - \mathbf{y}_i)}{h_{i-1}(h_i + h_{i-1})} + \frac{2(\mathbf{y}_{i+1} - \mathbf{y}_i)}{h_i(h_i + h_{i-1})}, \tag{2.24}$$

where $h_i \equiv x_{i+1} - x_i$. However, instead of (2.23) we use

$$\frac{\partial \mathbf{y}_i}{\partial x} = \frac{(\mathbf{y}_{i+1} - \mathbf{y}_{i-1})}{(h_i + h_{i-1})}. \tag{2.25}$$

(See Reference 12). For equally spaced meshes (2.23) and (2.25) are identical. For a nonuniform spaced grid, the difference between them can be expressed as

$$\frac{1}{2}(h_{i-1} - h_i)\frac{\partial^2 \mathbf{y}}{\partial x^2}. \tag{2.26}$$

The nonuniform grids that are used in our application have the property that $h_{i-1} \geq h_i$ so that (2.26) acts as a stabilizing viscous term to the spatial differencing.

To complete the description of the interior point algorithm we discuss the determination of the Jacobian type matrices $D$, $DR$, $DZ$, $DRR$, $DZZ$ defined by (2.12). The obvious way to evaluate the elements of these matrices is to manually take the partial derivatives and code each element. This double procedure is very error prone because the right hand sides of the equations are extremely complex. If there are $NEQ$ variables, then each element of the vector $\mathbf{G}$ is a function of $6 * NEQ + 3$ arguments, in general. Furthermore, for two-phase simulations, the correlations are not fixed, but can vary substantially from simulation to simulation. In these cases new elements of the matrices would have to be determined and encoded. To avoid this, one can lag the contribution of these correlations by one time step. This is not totally desirable because one increases the local truncation error (see the discussion near (2.14)) which can be large when the correlations add significantly to the flow dynamics in a given time step, e.g., burning of the grains. An alternative to this whole procedure is to determine these matrices numerically. Consider the determination of $D_{ij} \equiv \dfrac{\partial \mathbf{G}_i}{\partial \mathbf{y}_j}$ which can be approximated by

$$D_{ij} \approx \Big[\mathbf{G}_i(r,z,t,y_1,\cdots,y_j+\delta,\cdots,y_{NEQ},\mathbf{y}_r,\cdots,\mathbf{y}_{zz}) - \mathbf{G}_i(r,z,t,y_1,\cdots,y_j-\delta,\cdots,y_{NEQ},\mathbf{y}_r,\cdots,\mathbf{y}_{zz})\Big] /(2\delta), \tag{2.27}$$

where $\delta$ is the pre-determined increment. Once these increments are obtained, the elements of the matrices can be computed trivially by repeated calls to a subroutine which computes the right hand sides of the equations. A characteristic of the $\mathbf{G}'s$ is that the terms $\mathbf{y}_r$ and $\mathbf{y}_z$ appear at most quadratically, and the terms $\mathbf{y}_{rr}$ and $\mathbf{y}_{zz}$ appear at most linearly. Thus, by using centered differences the matrices $DR$ and $DZ$ can be

obtained exactly for any value of the increment. Likewise, using one-side differences the matrices $DRR$ and $DZZ$ can be determined exactly. In these cases, a large value of the increment can be used to avoid any round off errors. On the other hand the vector $\mathbf{G}$ is a non-algebraic, nonlinear function of the vector $\mathbf{y}$. In this case one cannot obtain an exact value of the elements of $D$ for any value of the increment. We have developed a strategy to compute an increment value based on the current error estimates of $\mathbf{G}_j$ and $\mathbf{y}_j$, that is, to determine $D_{ij}$ a $\delta_{ij}$ is used. The drawback of this approach is the computing time necessary to evaluate the right hand sides as often as required.

Finally, we address the problem of artificial mass sources induced solely by the motion of a grid system. The problem was illustrated and resolved in Reference 13. There exist a standard procedure to determine if mass sources occur in a numerical scheme when the grids are moved. First one assumes a constant flow field at the current level, secondly one applies the method to compute the new time level of values on a displaced grid, and finally one determines if these new values differ from the constant values. Following this procedure, we assume that the flow field variables are constants which satisfy the partial differential equations, and assume that $\Delta r \neq 0$ and $\Delta z \neq 0$. Consequently, all the spatial derivatives at the current time level are zero and (2.13) reduces to

$$\mathbf{y}^n - \beta\Big\{\Delta r\,\mathbf{y}_r^n + \Delta z\,\mathbf{y}_z^n \tag{2.28}$$

$$+ \Delta t\Big[D\,\mathbf{y}^n + DR\,\mathbf{y}_r^n + DZ\,\mathbf{y}_z^n + DRR\,\mathbf{y}_{rr}^n + DRZ\,\mathbf{y}_{rz}^n + DZZ\,\mathbf{y}_{zz}^n\Big]\Big\}$$

$$= \mathbf{y}^c - \beta\,\Delta t\,D\,\mathbf{y}^c.$$

Using the fact that the spatial derivatives at the current level are zero, we add zero to (2.28) in a convenient form to obtain

$$\mathbf{A}(\mathbf{y}^n - \mathbf{y}^c) = 0, \tag{2.29}$$

where $\mathbf{A}$ is a matrix. If $\mathbf{A}$ is nonsingular, then the solution of (2.29) is $\mathbf{y}^n = \mathbf{y}^c$. Thus, no mass sources exist. The lack of mass sources is due to the form of the equations, i.e. $\mathbf{y}_t = \mathbf{G}$, and the algorithm. For a simpler set of equations than the one we are solving, and for a set in a conservation form which are transformed to a stationary uniform computational grid, a "Geometric Conservation Law" is needed to prevent mass sources. (See Reference 13). Our method automatically avoids this other partial differential equation, and the need to obtain its solution at every time step.

### *2.8 Algorithm for Points at the Center-Line and at the Solid Surfaces*

The method to obtain the new time level of the flow variables along the center-line is different. To maintain the axial symmetry of the flow, the physical conditions on the flow are the radial and circumferential velocity components for both the gas and particle phases must be zero, and the first partial derivatives with respect to the radial direction at the axis of symmetry of the remaining variables must be zero.

The contribution from the points on the axis of symmetry can be done in at least two ways. The first and simplest is to directly apply the symmetry conditions. For a radial sweep, the elements of the first block row of the matrix **A** and known vector **b** are the finite difference approximations of these conditions. The axial sweep along the center-line is performed after the axial sweeps along interior axial indices. The final values at the center-line are obtained using the symmetry conditions, and the final axial sweep values of the non-center-line points. The second way is more complex. Because the center-line is part of the flow field (physically a non-boundary), the governing partial differential equations are valid on the axis of symmetry. One may rewrite these equations with the symmetry conditions imposed in the equations themselves, and with the correct limit conditions as the radial coordinate goes to zero. Then, solve these new equations by exactly the some method as described for non-boundary points. Although both are coded, the simpler first option is utilized.

The boundary conditions at the solid surfaces such as the breech, tube wall and projectile base can vary substantially with the particular simulation. This makes a general discussion of boundary conditions difficult. However, we will discuss some implementations of common types of boundary conditions. The simple functional form boundary condition $y = constant$, and the simple derivative boundary condition $\frac{\partial y}{\partial x} = constant$ are the most trivial kinds. Their finite difference approximations are straightforward, if $y$ is a variable computed directly by a governing partial differential equation, i.e., one of the variables in (2.1). However, if one has a nonlinear boundary condition of the form $f(\mathbf{y})^n = 0$ or $f\left(\frac{\partial \mathbf{y}}{\partial x}\right)^n = 0$, then one must linearize function $f$ in time in the same manner as is any component of the nonlinear vector function **G** in (2.1) and (2.11).

Sometimes a nonlinear boundary condition can be reformulated as a linear one. Consider the adiabatic condition $\left(\frac{\partial T}{\partial \mathbf{n}}\right)^n = 0$ were $T$ is the temperature and **n** is the outward normal. In our method the entropy $s$ and pressure function $q$ are computed directly from the governing partial differential equations. Thus $T$ is a nonlinear function, $T = T(s,q)$. We ordinarily would expand $\left(\frac{\partial T}{\partial \mathbf{n}}\right)^n$ in a Taylor series in time to obtain a linear function in $s^n$ and $q^n$. However, for a Noble-Abel equation of state, we have a simplification. We use the chain rule to obtain

$$\left( \frac{\partial T}{\partial s}\frac{\partial s}{\partial \mathbf{n}} + \frac{\partial T}{\partial q}\frac{\partial q}{\partial \mathbf{n}} \right)^n = 0. \tag{2.30}$$

Since $\frac{\partial T}{\partial s} \neq 0$, then (2.30) can be written as

$$\left( \frac{\partial s}{\partial \mathbf{n}} + \frac{\dfrac{\partial T}{\partial q}}{\dfrac{\partial T}{\partial s}} \frac{\partial q}{\partial \mathbf{n}} \right)^n = 0. \tag{2.31}$$

By noting that the ratio of partial derivatives of temperature is a constant, (2.31) is a linear function in $s^n$ and $q^n$ Thus, (2.31) can be finite differenced and incorporated directly into the matrices of the linearized ADI method.

### *2.2.3 The Order of the Sweeps*

The order of the sweeps is mainly a bookkeeping problem, and should not have a large effect on the solution. We have used the following procedure. First, radial sweeps from the center-line to the tube wall are performed along constant axial indices to obtain intermediate values of the variables $\mathbf{y}$ using (2.16). The axial index varies from the one after the breech to the one before the projectile base. Second, axial sweeps from the breech to the projectile base are performed along constant radial indices to obtain values of the variables $\mathbf{y}$ at the new time level using (2.17). The radial index varies from the one after the axis of symmetry to the one before the gun tube wall. Third, the symmetry conditions are applied to determine the new time level values at all points on the axis of symmetry by using the results of the axial sweep (step two). Fourth, the value of the variables at the new time level at all grid points on the wall are determined by imposing the wall boundary condition using, if necessary, the results of the axial sweep (step two).

We note that the radial sweeps along the breech and projectile are avoided. The justification is that in many applications the boundary conditions at the breech and projectile do not involve radial derivatives. Consequently, the linearized version of these conditions can be expressed without the need for the determination of the intermediate values of the variables. Recall that the matrix $\mathbf{D}$ (equation (2.12)) was included in the radial sweep. (See equations (2.15b) & (2.16)). However, it could have been incorporated into the axial sweep formula. (See equations (2.15c) & (2.17)). If no radial derivatives exist for the boundary conditions, we avoid the radial sweep and incorporate the $\mathbf{D}$ matrix in the axial sweep formulation. An example of this type of boundary condition is the gas continuity equation used to determine the transformed pressure $q$ at the breech for a one-phase, viscous simulation.. Imposing the no-slip conditions $u = v = w = 0$ in the continuity equation evaluated at the breech, all radial derivatives vanish, and the above method is applied directly during the axial sweep where one-sided finite differences are used to approximate the spatial derivatives at the breech. If one uses the normal velocity equation for the determination of $q$ at the projectile base, for example, some radial derivatives remain in the viscous terms. Only by lagging them could one use the above approach.

## 3. RESULTS

The results of two simulations using the DELTA code are presented in this section. These particular calculations are selected because they can be compared to independently determined answers, and thus, give some verification of the code's accuracy and capabilities. Both cases involve a one-phase gas expansion in a constant cross-section tube closed at one end by a stationary surface called the breech, and at the other end by a movable piston called the projectile. The breech and projectile base are assumed to be flat surfaces. The initial states of the gas are uniform and quiescent, but the gas pressures are great enough to accelerate the projectile through the tube. The controlling mechanisms of the expansion flows are the same, namely the propagation of the rarefaction wave, generated by the projectile displacement, and its reflection from the breech, then the projectile, and so forth. However, the gas pressure levels differ greatly between the simulations, and the subsequent flows are in different regimes.

**TABLE 1.** Geometry and Gas Properties for the 150-mm Gun and Bicen-Whitelaw Experiments

| 150-mm Gun | Description | Bicen-Whitelaw |
|---|---|---|
| 150.0 | Bore Diameter (mm) | 76.7 |
| 1.698 | Initial Projectile Displacement (m) | 0.1773 |
| 6.0 | Maximum Travel of Projectile (m) | 0.3 |
| 50.0 | Projectile Mass (kg) | 2.54 |
| 0.001 | Covolume ($m^3/kg$) | 0.0 |
| 1.22 | Ratio of specific heats, $\gamma$ | 1.4 |
| 621.09 | Initial Pressure (MPa) | 0.28 |
| 2666.8 | Initial Temperature (K) | 293.0 |

The first simulation corresponds to the gas expansion within a 150-mm tube away from the effects of the tube wall (core-flow) under ballistic conditions. The specifications for this case are given by the first column of Table 1. The analytic solution of the one-dimensional, inviscid gas equations governing the flow within this 150-mm tube at certain positions from the breech and for specified times was obtained by Love and Pidduck. (See Reference [14].) Their solution is valid under the assumptions of isentropic expansion of each element of gas, of constant covolume, of an even integer value of the ratio $(\gamma + 1)/(\gamma - 1)$, where $\gamma$ is the ratio of specific heats, and the frictionless motion of the projectile. The solution is in terms of truncated power series, and becomes more complicated as the number of rarefaction wave reflections increase. For the one-dimensional DELTA calculation, the computational mesh (covering the enclosed cavity behind the projectile) consisted of four equidistant mesh lines parallel to the axis of symmetry and 89 uniformly spaced grid lines orthogonal to the axis of symmetry. To maintain a one-dimensional simulation for comparison to the Love & Pidduck results, the following conditions

$$u = \frac{\partial w}{\partial r} = \frac{\partial s}{\partial r} = \frac{\partial q}{\partial r} = 0$$

are imposed along both the tube wall and axis of symmetry. The boundary conditions at the breech and projectile are no-slip velocity and adiabatic walls. Love and Pidduck developed their solution with the assumption of an inviscid isentropic flow which allowed them to use a special form of the Noble-Abel equation of state, namely

$$P\left(\frac{1}{\rho} - \eta\right)^{\gamma} = P_0\left(\frac{1}{\rho_0} - \eta\right)^{\gamma} = \text{constant},$$

where subscript zero indicate their initial values. However, in the DELTA simulation viscous effects are included and the general form of the Noble-Abel equation of state. However, the special form of the equation of state used by Love and Pidduck is maintained to within less than two percent in the DELTA simulations. Thus, the non-isentropic and viscid effects included in the general framework of DELTA are minor for this one-dimensional flow, and the comparison of the analytic solution and numerical results is reasonable.

The comparison of the pressure histories at the projectile base is given in Figure 1. The change in slope in the pressure curve is due to the first reflection of the rarefaction wave at the projectile base. The magnitude of the slope discontinuity of the pressure curve decreases with time due to the equilibration of the pressures during the gas expansion. Another slope change exists theoretically at 7.137ms, although it cannot be detected even in the graph of the analytical results. Figure 2 is similar to Figure 1 except that the pressures at the breech are compared The effect of the first arrival of the wave at the breech is more obvious in both solutions.

Figures 3 and 4 show comparisons of the histories of the projectile velocity and projectile displacement from the breech, respectively. The large values of the tangent to the curve in Figure 3 indicate the extreme acceleration the projectile experiences. The agreement of the results show that the numerical solution of the partial differential equations governing the gas motion using the DELTA algorithm is correctly coupled to the proper solution of the projectile motion. Comparisons of the pressure profiles from the breech to the projectile at specified times are given in Figures 5 and 6. The ranges of pressure values on the ordinate are the smallest possible to provide accurate comparisons. Figure 5 shows a comparison of the pressure values at 2.898ms, that is, after the rarefaction wave has been reflected from the breech, then the projectile base, and is approximately halfway between them. The DELTA calculation differs by 0.6% at most from the analytic solution values. The slope discontinuity is smeared out in the numerical calculation. Figure 6 shows the pressure profiles at 10.23ms which is near tube-exit time of the projectile. The maximum discrepancy between the two results is approximately 1.2%. Because the analytic solution is a truncated power series solution, it is difficult to determine which values are more accurate.

Next we compare a DELTA simulation with laser-Doppler anemometry, time-resolved measurements of the axial velocity field inside a tube behind a slowly accelerating projectile. The experiment was performed at Imperial College, London under funding by European Research Office of the Army and the Ballistic Research Laboratory, and is reported in Reference [15]. This experiment is important to the development of DELTA

because it provides the first transient, two-dimensional measurements of a quantity to which the results of a DELTA simulation can be compared. The schematic of the apparatus is given in Figure 7. Nitrogen is the gas, and the other characteristic of the experiment are given in the second column of Table 1.

Three important conclusions of this study are: The maximum intensity level of the turbulence was approximately four percent which implies a very low level of turbulence, the tube wall boundary layer remained laminar, and the heat transfer to the tube wall was minimal. Consequently, a laminar flow simulation with adiabatic walls should approximate this experiment. The two-dimensional computational grid had 33 uniformly distributed points in the axial direction, and 19 nonuniformly distributed points in the radial direction. The radial grid was such that, while 19 points spanned the distance from the axis of symmetry ($r = 0$) to the wall ($r = 38.35$mm), 12 points were distributed from $r = 32$mm to the wall and 5 points were distributed from $r = 37.85$mm to the wall. The maximum and minimum distance between the grid points were 6.5mm and 70$\mu$m, respectively. Constant values were used for the coefficients of viscosity and thermal conductivity of nitrogen, namely, 17.07$\mu$(Pa·s) and 0.02524W/(m·K), respectively. Because the experimental apparatus was mounted vertically, the equation for the projectile motion (2.3) was changed to include its acceleration due to gravity. Because the total retarding force ($F_D + F_B + F_A$ in (2.3) ) experienced by the projectile was not determined by the experiment, no values of these forces could be assigned. Thus, a total retarding force profile versus axial displacement was obtained so that the axial velocity of the projectile determined by the DELTA code matched the experimental values as shown in Figure 8. Because the projectile velocity values agree, so must the projectile displacement values. Figure 9 compares the axial velocity profiles along the axis of symmetry at various times. Both the DELTA and experimental results show a linear profile from the zero value at the breech to the value of the projectile velocity for each time. The axial velocity histories at 76.7mm from the breech and at 0.5, 1.0, 2.0, and 3.0mm from the tube wall are compared in Figure 10. The values from the calculation are within the scatter of the experimental data at radial positions of 1, 2 and 3mm from the wall. However, the discrepancy between the values at the 0.5mm position increases with time. The same quantities are graphed in Figure 11 but at 153.4mm from the breech. The comparisons in Figure 11 show similar behavior to that in Figure 10, but with the discrepancy at the 0.5mm position considerably larger. After a discussion with the experimentalists, it was agreed that these most difficult measurements at 0.5mm from the tube wall are likely to contain errors and should be redone. We are presently awaiting accurate measurements in the sub-millimeter range.

## 4. SUMMARY

The numerical algorithm encoded in the DELTA computer code, and comparisons of its calculations to an analytic solution and experimental measures are described.

A numerical algorithm to solve the two-dimensional, axisymmetric, unsteady, finite volume, weighted averaged two-phase equations which govern certain flows inside a gun tube is discussed. These equations are in their most general form. In particular, when the flow regime is governed only by convection and pressure forces, this general form automatically gives the solution of the corresponding inviscid equations. When in the

boundary layer regime, this general form gives without any assumptions, the solution in the boundary layer where viscous forces dominate. Moreover, this approach naturally provides all the coupling between different phenomena because only one set of equations, which govern all the phenomena, is solved. Thus, phenomena that is controlled by basically inviscid flow but exists because of viscous forces, like the additive particle laden gun tube wall boundary layer which governs heat transfer to the gun tube, can be studied for the first time without assumptions on the natures of the core flow or boundary layer, the validity of heat-transfer correlations, and/or the intra-flow coupling between various regimes.

Two examples of calculations with the DELTA code are presented. These computations are limited to one-phase expansion flows in an adiabatic tube so to allow comparisons with independently obtained data. More realistic calculations that involve ballistic environments, heat transfer to the gun tube wall, and other phenomena are presented in Reference [16].

## 5. ACKNOWLEDGMENTS

The development of a large computer code such as DELTA can not be accomplished in any timely fashion by one man. During its evolution several have contributed ideas and/or programming time. Thomas Mann, Aivars Celmins, Rudi Heiser, Csaba Zoltani, Christopher Roller, and Stephen Davis made contributions to the development of the DELTA code. I gratefully acknowledge their efforts. Furthermore, this code was developed under the administrative supervision of Norman Banks. Without his support, this project could not have been started and nurtured.

## REFERENCES

[1] A.K.R. Celmins and J.A. Schmitt, "Volume Averaged Two-Phase (Gas-Solid) Interior Ballistics Equations," ARBRL-TR-2593, *USA Ballistic Research Laboratory Report,* 1984.

[2] A.K.R. Celmins and J.A. Schmitt, "Three Dimensional Modeling of Gas-Combusting Solid Two-Phase Flows," *Proceedings of the Third Multi-Phase Flow and Heat Transfer Symposium-Workshop,* T. N. Veziroglu, ed., 18-20 April 1983, Miami Beach, Florida.

[3] A.K.R. Celmins and J.A. Schmitt, "Modeling of Gas-Solid Phenomena in Interior Ballistics," *Proceedings of The Seventh International Symposium on Ballistics,* 19-21 April 1983, The Hague, The Netherlands.

[4] P.S. Gough, "Two-Dimensional, Two-Phase Modelling of Multi-Increment Bagged Artillery Charges," ARBRL-CR-00503, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1982.

[5] P.S. Gough, "Modeling of Rigidized Gun Propelling Charges," ARBRL-CR-00518, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1983.

[6] H.J. Gibeling and H. McDonald, "Development of a Two-Dimensional Implicit Interior Ballistic Code," ARBRL-CR-00451, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1981.

[7] H.J. Gibeling and H. McDonald, "An Implicit Numerical Analysis for Two-Dimensional Turbulent Interior Ballistics Flows," ARBRL-CR-00523, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1984.

[8] R.D. Richtmyer and K.W. Morton, **Difference Methods for Initial-Value Problems,** 2nd ed., Wiley-Interscience (New York, 1967).

[9] W.R. Briley and H. McDonald, "Solution of the Multi-Dimensional Compressible Navier-Stokes Equations by a Generalized Implicit Method," *J. Comp. Phy.* **24** (1977) 372-397.

[10] R.M. Beam and R.F. Warming, "An Implicit Finite-Difference Algorithm for Hyperbolic Systems in Conservation Form," *J. Comp. Phy.* **22** (1976) 87-110.

[11] W.R. Briley and H. McDonald, "On the Structure and Use of Linearized Block Implicit Schemes," *J. Comp. Phy.* **34** (1980) 54-73.

[12] E. Kalnay de Rivas, "On the Use of Nonuniform Grids in Finite-Difference Equations," *J. Comp. Phy.* **10** (1972) 202-210.

[13] P.D. Thomas and C.K. Lombard, "Geometric Conservation Law and Its Application to Flow Computations on Moving Grids," *AIAA Journal* **17** (1979) 1030-1036.

[14] E.H. Love and F.B. Pidduck, "Lagrange's Ballistic Problem," *Phil. Trans. Roy. Soc.* **222** (1921) 167-226.

[15] A.F Bicen and J.H. Whitelaw, "Velocity Characteristics of the Wake of an In-Cylinder Projectile," *Imperial College of Science and Technology Report* FS/83/19, London, England, June 1983.

[16] R. Heiser and J. A. Schmitt, "Simulations of Special Interior Ballistic Phenomena With and Without Heat Transfer to the Gun Tube Wall," *Proceedings of the Second Army Conference on Applied Mathematics and Computing,* ARO Report, 1984.

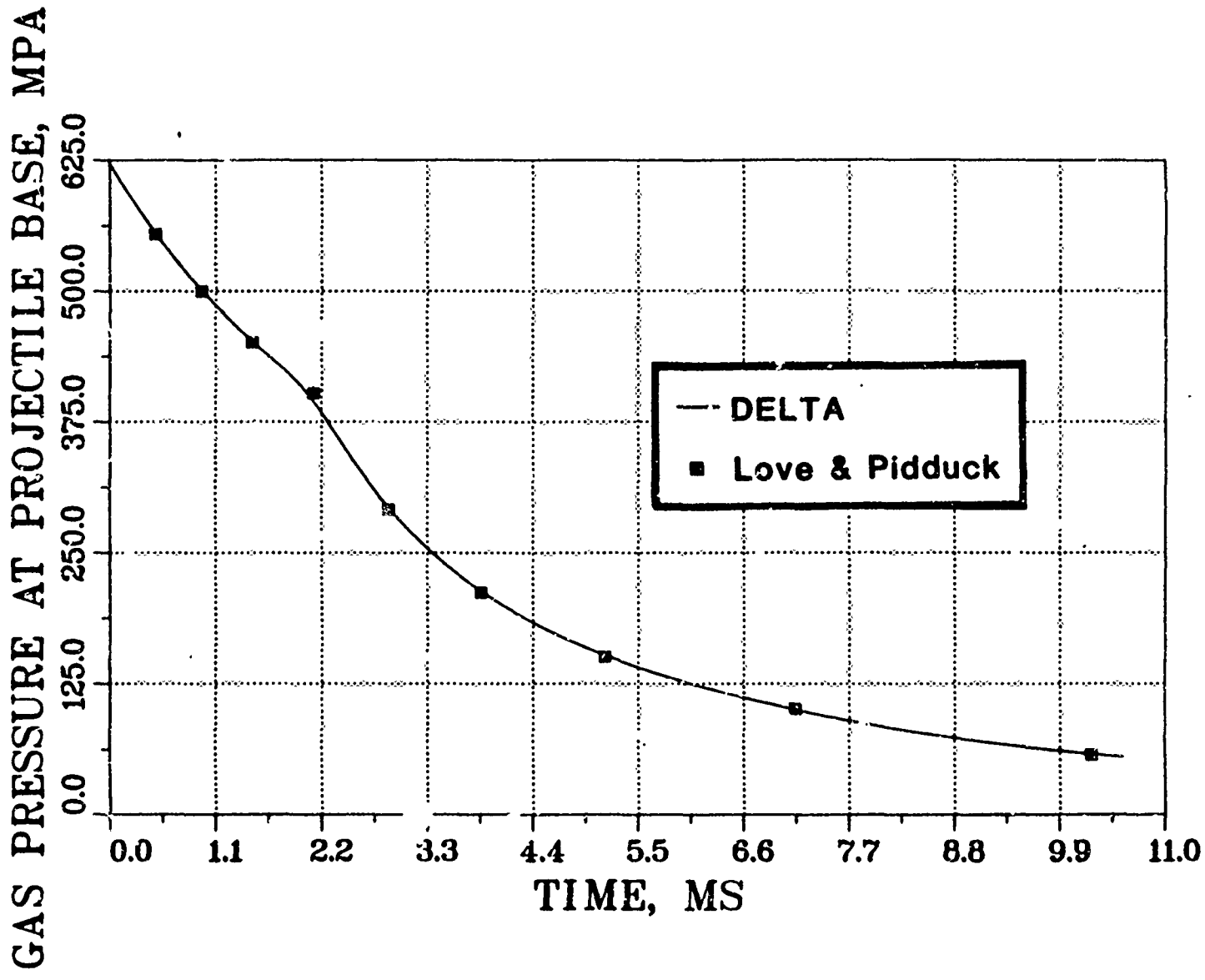



Figure 1. The comparison of the pressure histories at the projectile base.

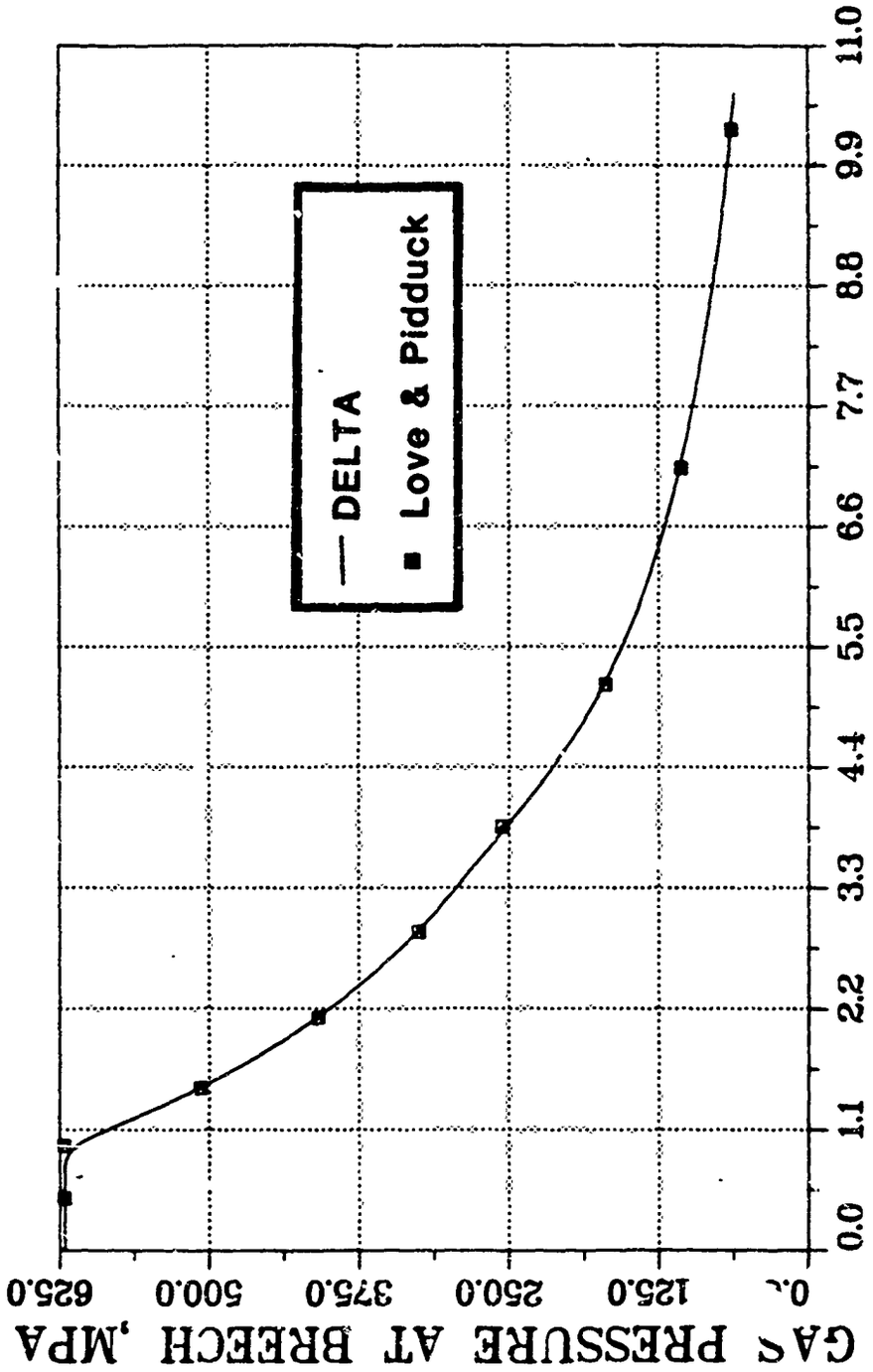



Figure 2. The comparison of the pressure histories at the breech.

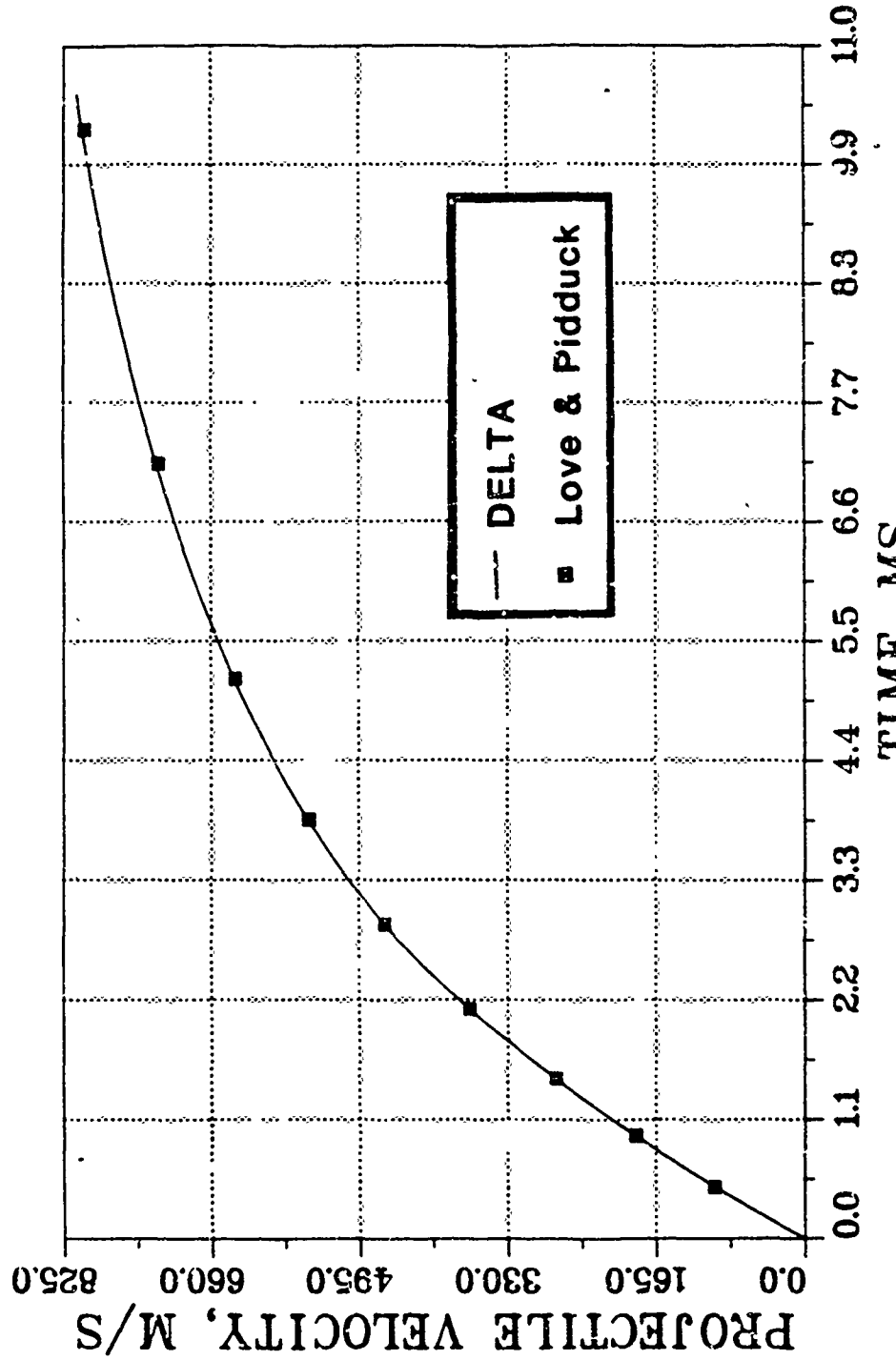



Figure 3. The comparison of the histories of the projectile velocity.

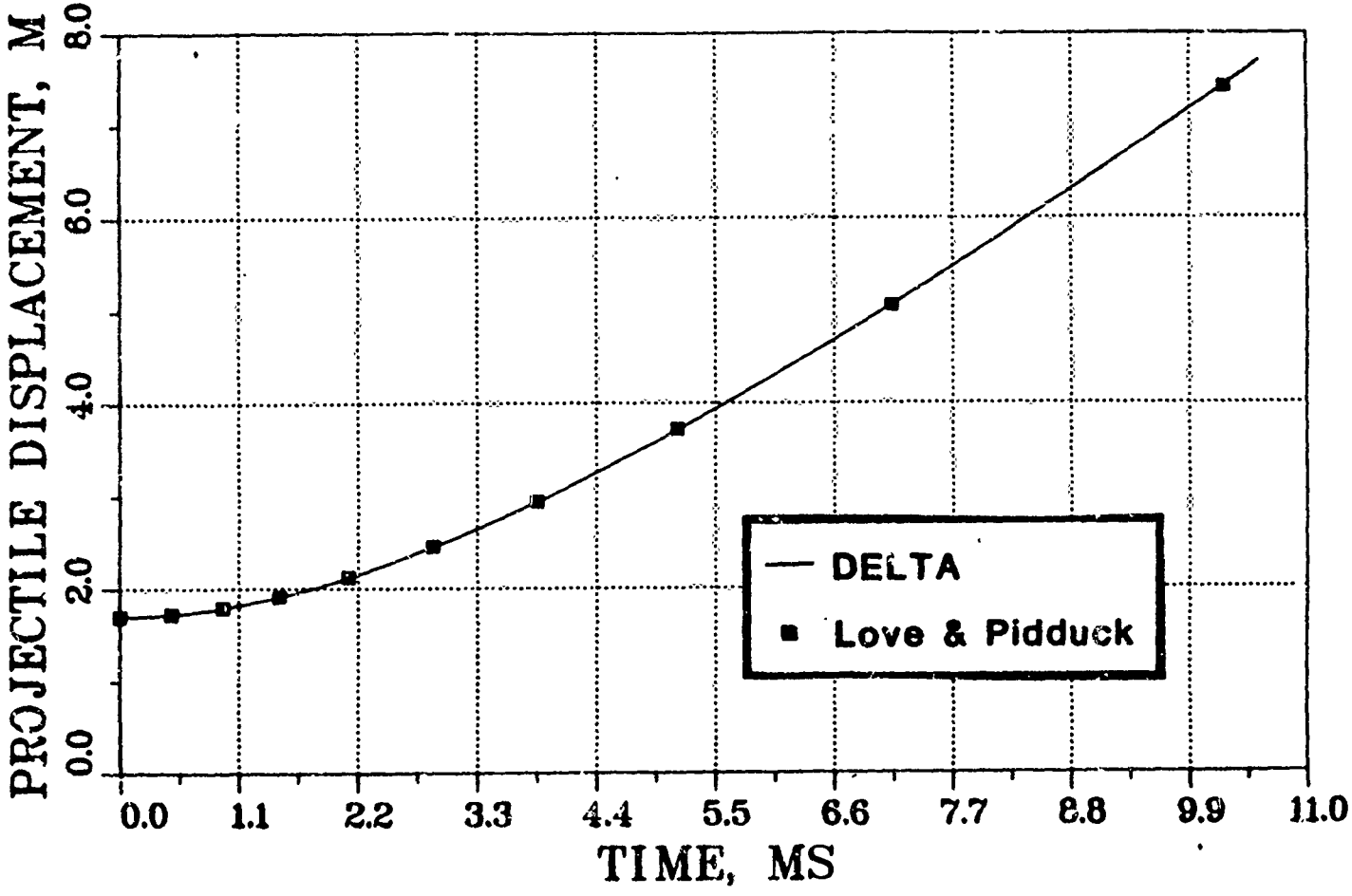



Figure 4. The comparison of the histories of the projectile displacement.

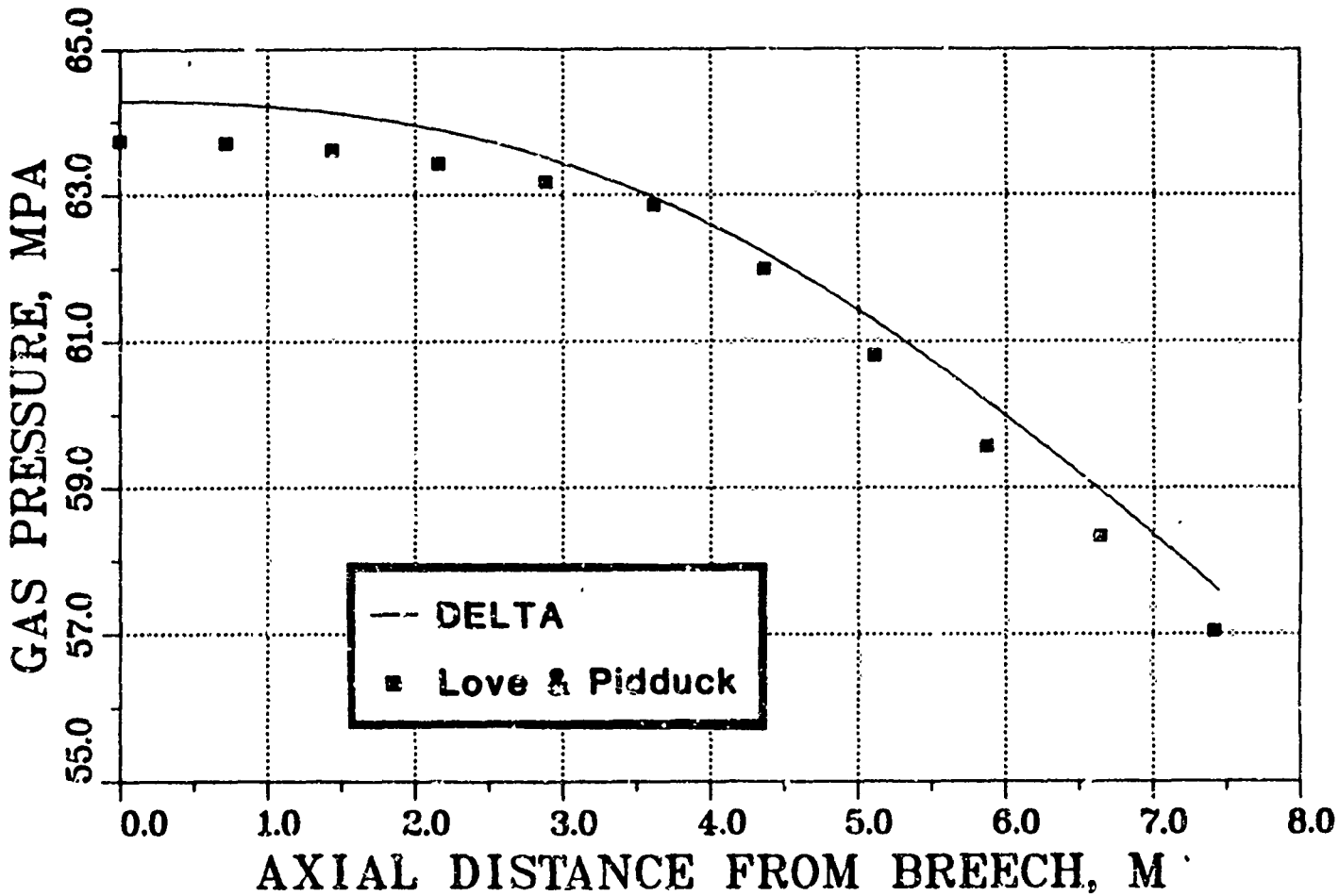



Figure 5. The comparison of the pressure profiles from the breech to the projectile at time 2.898ms.

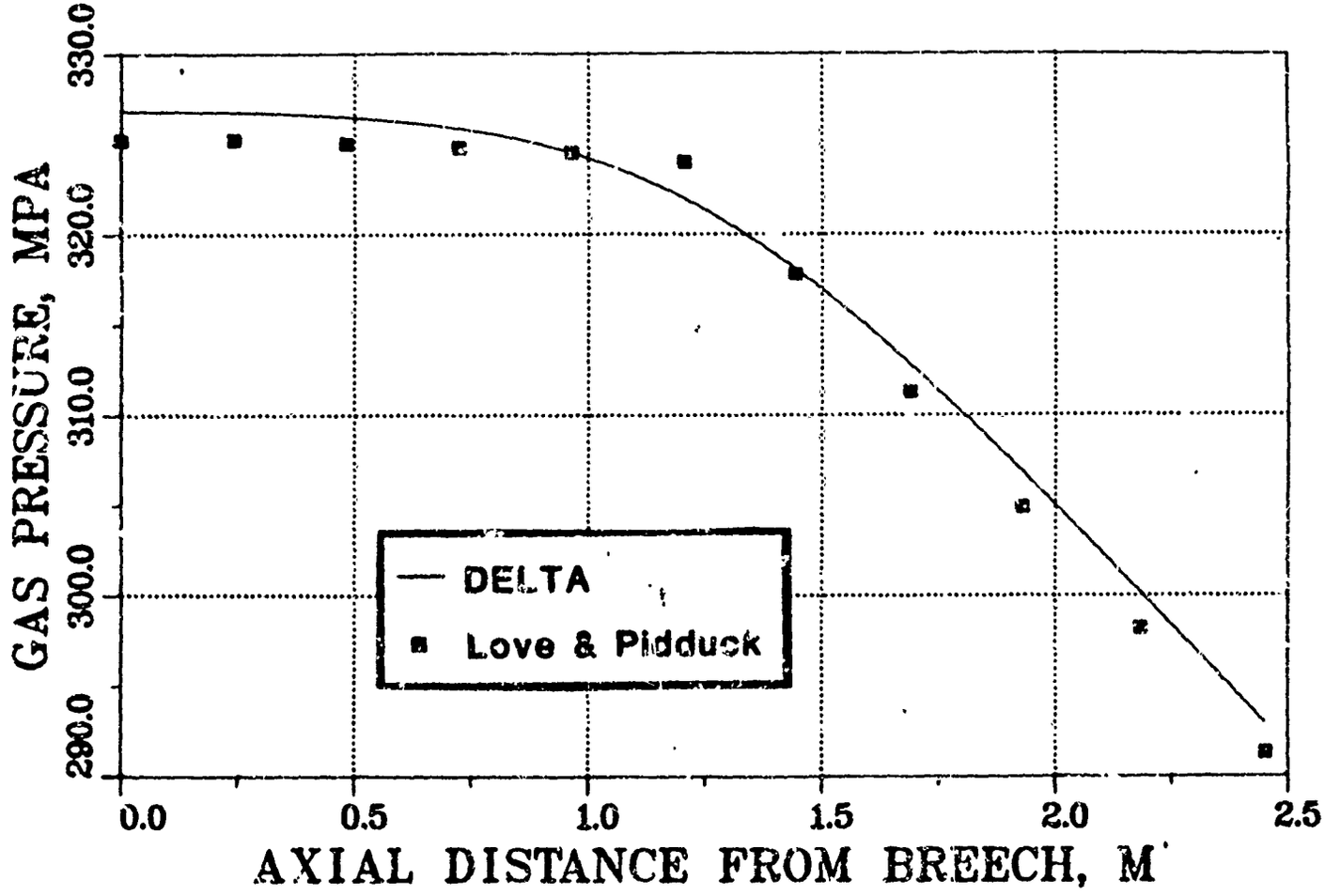



Figure 6. The comparison of the pressure profiles from the breech to the projectile at time 10.23ms.

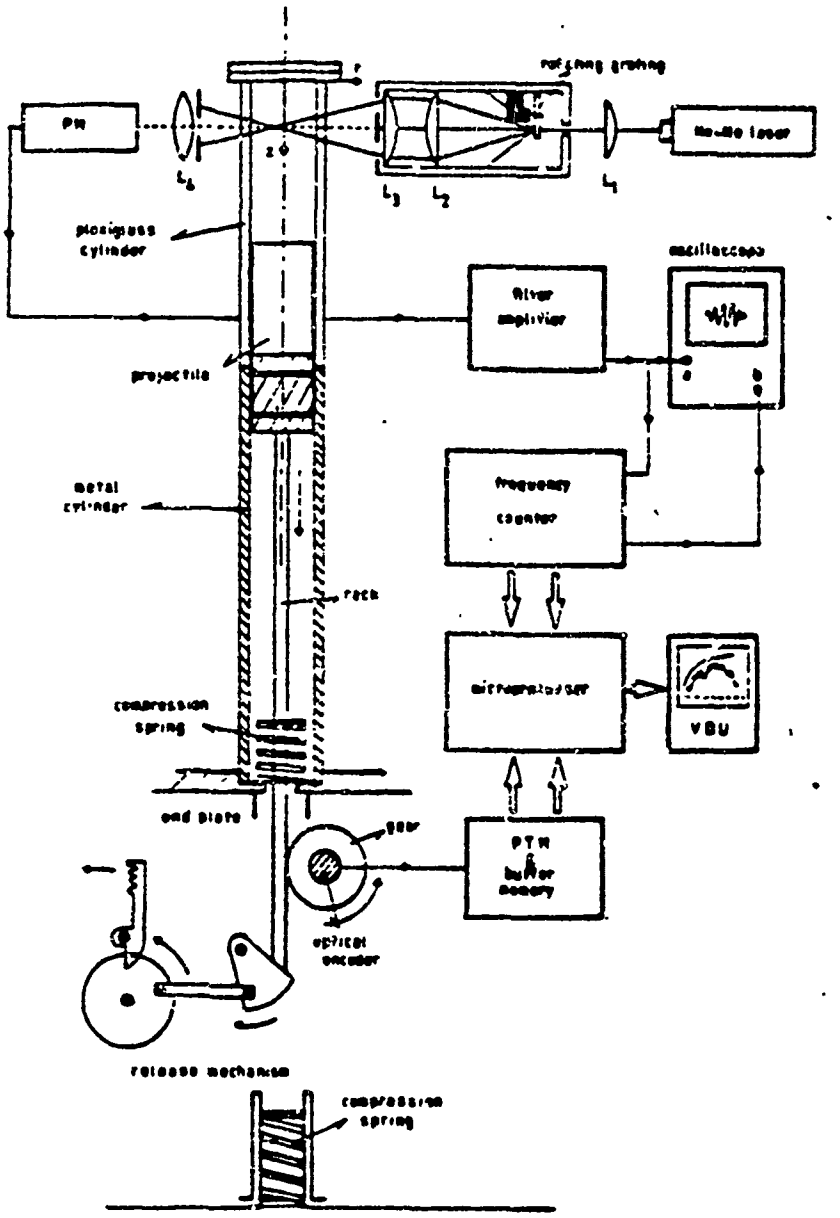



Figure 7. A schematic of the Bicen-Whitelaw experimental apparatus.

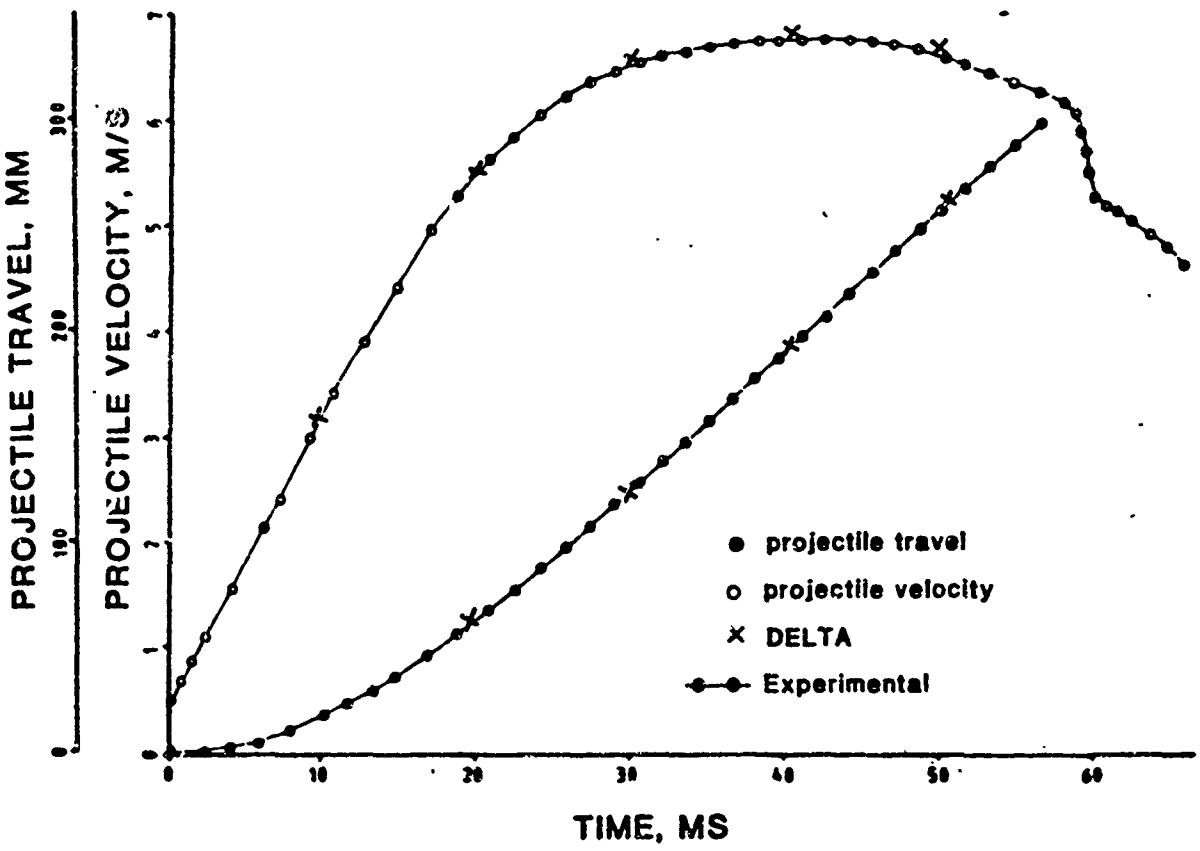



Figure 8. The comparison of the axial velocities and displacements of the projectile.

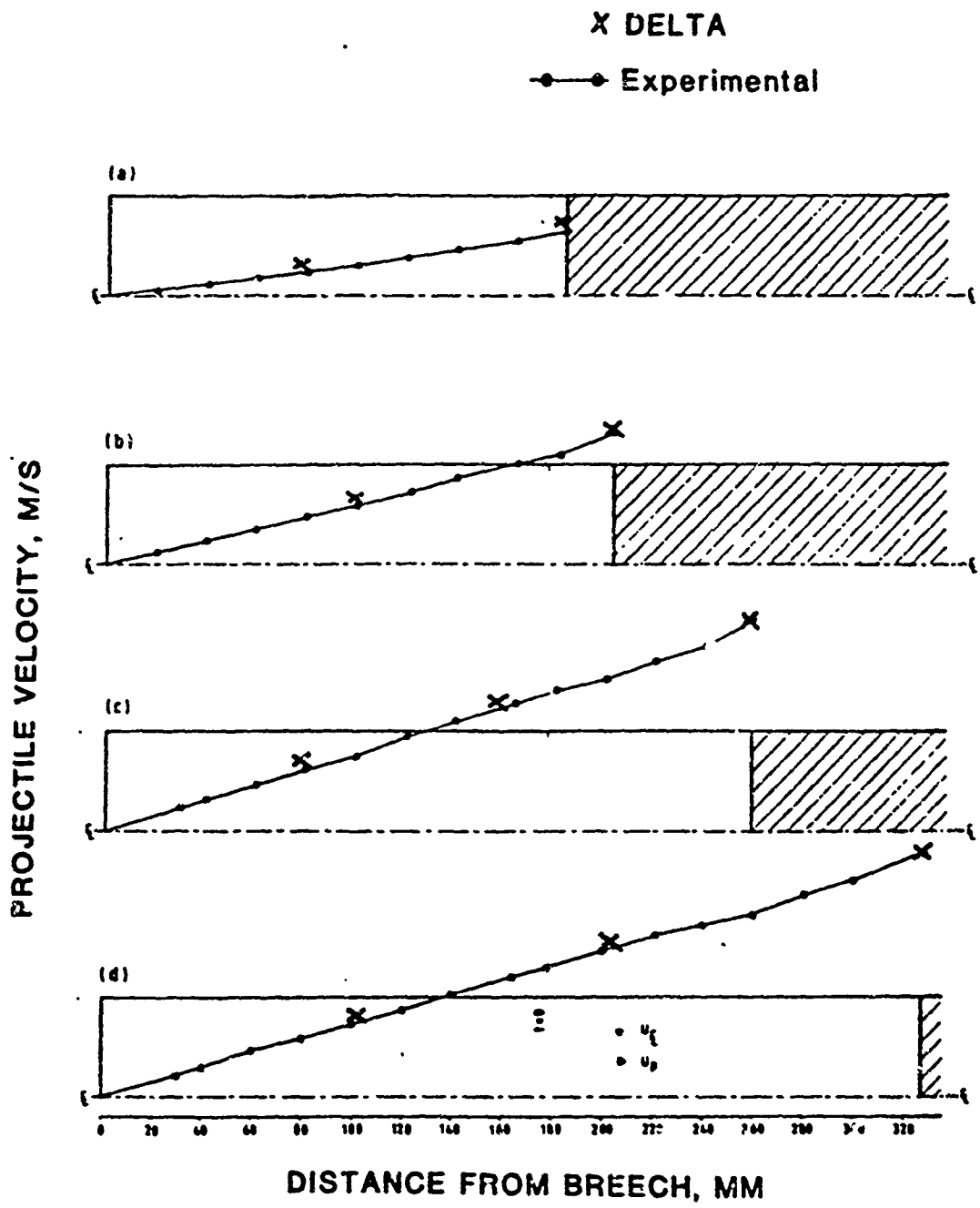



Figure 9. The comparison of the axial velocity profiles along the axis of symmetry at various times:
(a) 4.8ms (c) 22.8ms
(b) 11.6ms (d) 33.8ms

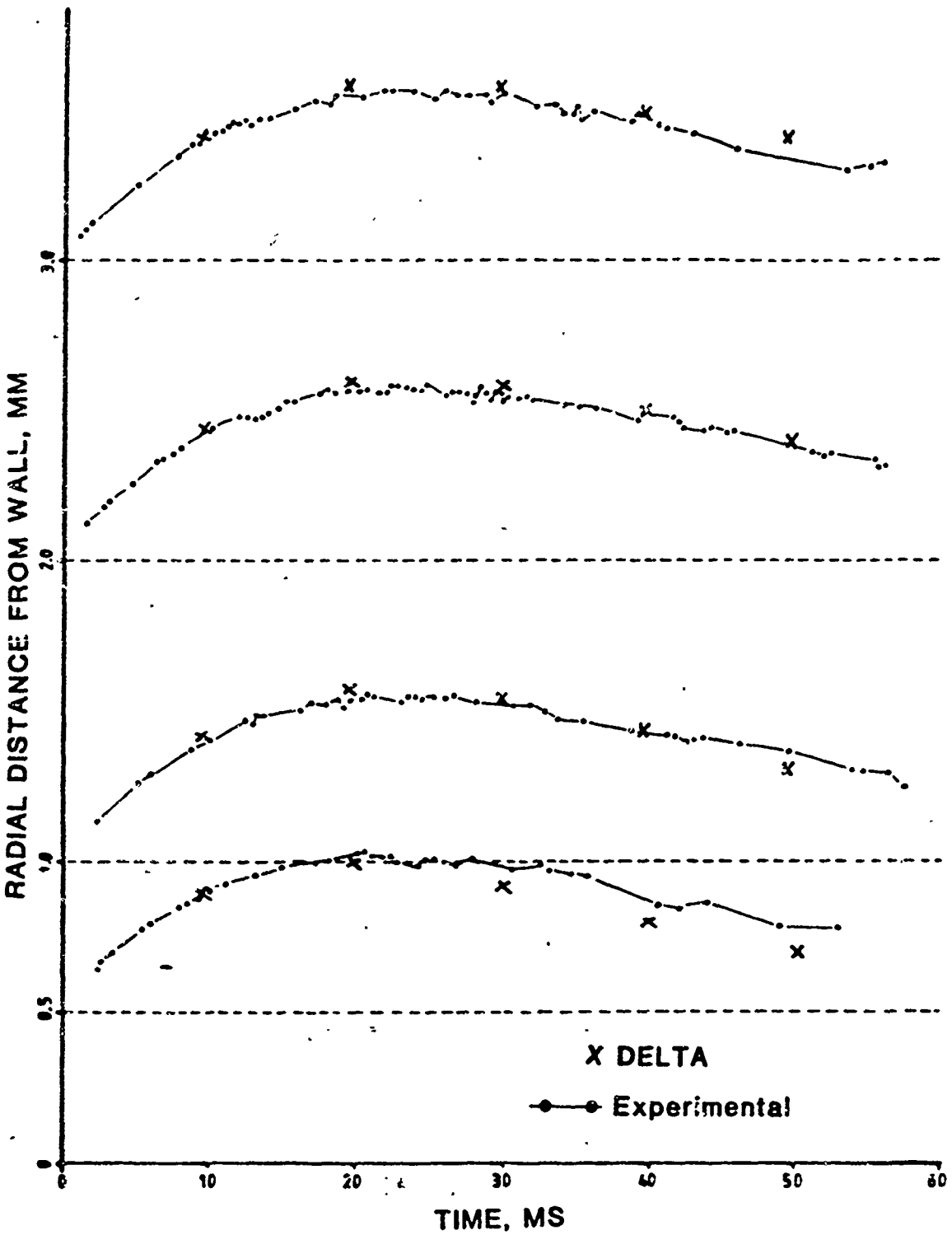



Figure 10. The comparison of the axial velocity histories at 77.6mm from the breech.

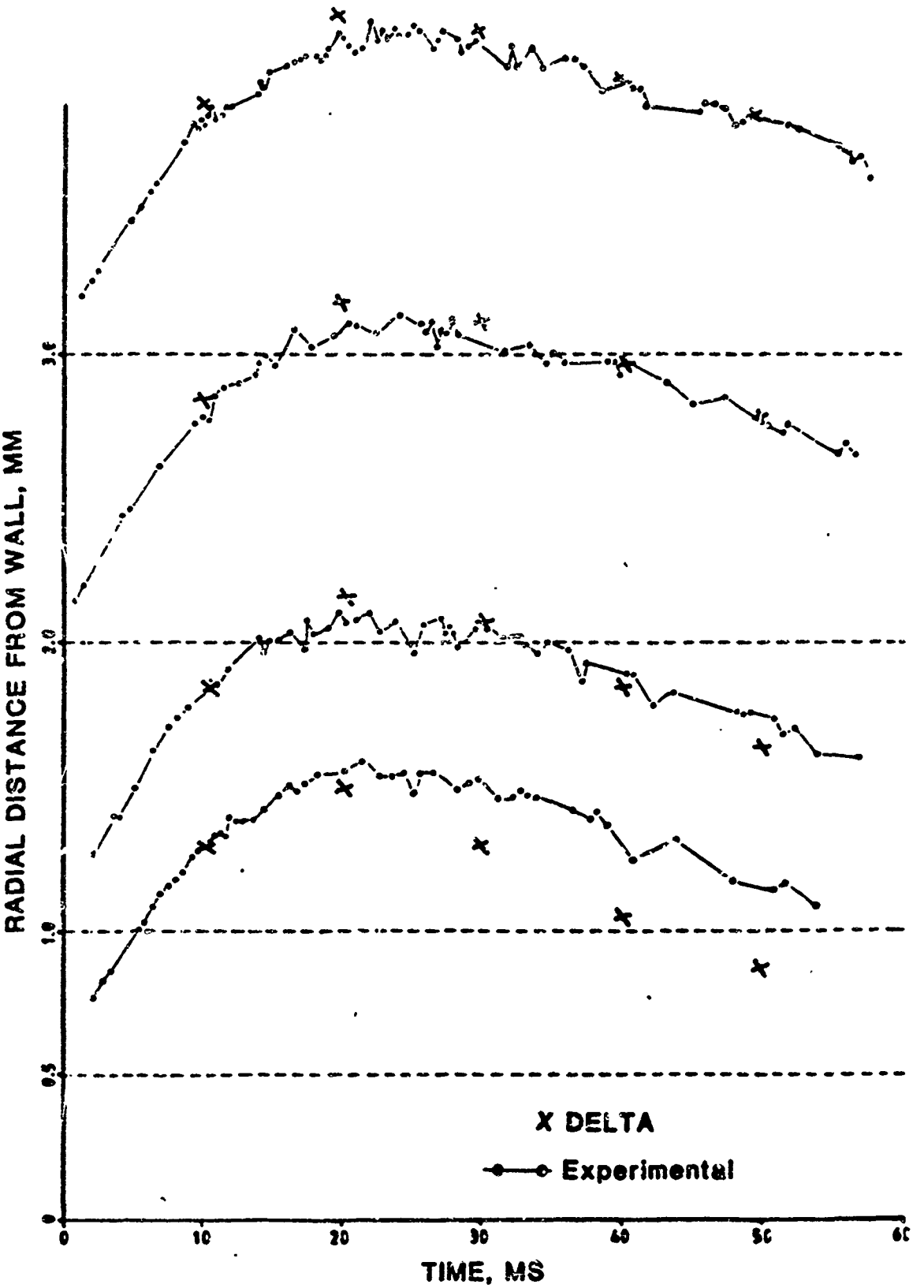



Figure 11. The comparison of the axial velocity histories at 153.4mm from the breech.

# APPENDIX A

## GOVERNING EQUATIONS FOR AXIALLY SYMMETRIC FLOWS IN CYLINDRICAL COORDINATES

This appendix contains a list of the governing equations in component form in cylindrical coordinates for the case of axially symmetric flow. The subscripted variables denote the components of a vector and not the derivative of these variables. All derivatives are written in a non-abbreviated form. The listed equations are in a form which is compatible with Eq. (4.32), Sect. 4.3. The components of the gas average velocity and the particle average velocity are

$$u = (u_r, u_\theta, u_z) \quad , \ (m/s) \quad , \tag{A.1}$$

$$\overset{*}{u} = (\overset{*}{u}_r, \overset{*}{u}_\theta, \overset{*}{u}_z) \quad , \ (m/s) \quad , \tag{A.2}$$

where the subscripts r, θ, and z refer to the radial, angular, and axial coordinate directions, respectively. The components of the gradient of a scalar f are

$$\nabla f = \{(\frac{\partial f}{\partial r})_r, (0)_\theta, (\frac{\partial f}{\partial z})_z\} \quad . \tag{A.3}$$

The divergence of a vector $F = (F_r, F_\theta, F_z)$ is

$$\nabla \cdot F = \frac{1}{r}\frac{\partial (rF_r)}{\partial r} + \frac{\partial F_z}{\partial z} \quad . \tag{A.4}$$

The independent variables are time t, radial position r, and axial position z. The dependent average variables which are computed form the governing partial differential equations are: the specific entropy s, the pressure logarithm function q, the radial gas velocity $u_r$, the circumferential gas velocity $u_\theta$, the circumferential particle velocity $\overset{*}{u}_\theta$, the axial particle velocity $\overset{*}{u}_z$, the number of particles within the averaging volume $\overset{*}{n}$, the regression distance $\overset{*}{d}$, and the surface temperature of the particles $\overset{*}{T}$.

The entropy equation is

$$\frac{\partial s}{\partial t} = - u_r \frac{\partial s}{\partial r} - u_z \frac{\partial s}{\partial z} + \frac{p}{\rho T} B + H \Gamma + \phi + \Psi \qquad (A.5)$$

where $p$, $\rho$, $T$, $H$, $\Gamma$, are given by Eqs. (B.6), (B.4), (B.2), (B.27), and (B.26), respectively. The expression for $B$ is

$$B = \frac{1}{\alpha} \left\{ (1-\alpha) \left[ \frac{1}{r} \frac{\partial (r u_r^*)}{\partial r} + \frac{\partial u_z^*}{\partial z} \right] - (u_r - u_r^*) \frac{\partial (1-\alpha)}{\partial r} - (u_z - u_z^*) \frac{\partial (1-\alpha)}{\partial z} \right\} , \qquad (A.6)$$

and the porosity $\alpha$ is given by Eq. (B.1). The dissipation function $\phi$ is

$$\phi = \frac{1}{\rho T} \bar{\phi}(E) + \frac{1}{\rho T} \phi_T + \langle \phi \rangle \quad , \qquad (A.7)$$

where

$$\begin{aligned} \bar{\phi}(E) = {} & \frac{4}{3} \mu \left[ \left( \frac{\partial u_r}{\partial r} \right)^2 + \left( \frac{u_r}{r} \right)^2 + \left( \frac{\partial u_z}{\partial z} \right)^2 - \left( \frac{u_r}{r} \frac{\partial u_r}{\partial r} + \frac{\partial u_r}{\partial r} \frac{\partial u_z}{\partial z} + \frac{u_r}{r} \frac{\partial u_z}{\partial z} \right) \right] \\ & + \mu \left[ \left( r \frac{\partial}{\partial r} \left( \frac{u_\theta}{r} \right) \right)^2 + \left( \frac{\partial u_z}{\partial r} + \frac{\partial u_z}{\partial r} \right)^2 + \left( \frac{\partial u_\theta}{\partial z} \right)^2 \right] \\ & + \lambda \left[ \frac{\partial u_r}{\partial r} + \frac{u_r}{r} + \frac{\partial u_z}{\partial z} \right]^2 \quad , \end{aligned} \qquad (A.8)$$

and $\mu$, $\lambda$, $\langle \phi \rangle$, and $\phi_T$, are given by Eqs. (B.7), (B.8), (B.13), and (B.35), respectively. The heat conduction term $\Psi$ is given by Eq. (B.15) as

$$\Psi = \Psi_{gas} + \Psi_{particle} + \Psi_{turb} \quad , \qquad (A.9)$$

where

$$\Psi_{gas} = \frac{1}{\alpha\rho T}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\alpha\kappa\frac{\partial T}{\partial r}\right) + \frac{\partial}{\partial z}\left(\alpha\kappa\frac{\partial T}{\partial z}\right)\right] \quad , \tag{A.10}$$

$$\Psi_{turb} = \frac{1}{\alpha\rho T}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\alpha\kappa_T\frac{\partial T}{\partial r}\right) + \frac{\partial}{\partial z}\left(\alpha\kappa_T\frac{\partial T}{\partial z}\right)\right.$$

$$\left. - \frac{1}{r}\frac{\partial}{\partial r}\left(r\kappa_T(T_i-T)\frac{\partial\alpha}{\partial r}\right) - \frac{\partial}{\partial z}\left(\kappa_T(T_i-T)\frac{\partial\alpha}{\partial z}\right)\right] \quad , \tag{A.11}$$

$\Psi_{particle}$, $\kappa$ are given by Eqs. (B.17), (B.14), respectively, and $\kappa_T$, $T_i$ are discussed near Eq. (B.36).

The pressure logarithm function equation is

$$\frac{\partial q}{\partial t} = -u_r\frac{\partial q}{\partial r} - u_z\frac{\partial q}{\partial z} - \frac{\rho}{\frac{\partial\rho}{\partial q}}\left(\frac{1}{r}\frac{\partial(ru_r)}{\partial r} + \frac{\partial u_z}{\partial z} + \frac{\partial e}{\partial s}\frac{1}{T}B\right)$$

$$+ \frac{1}{\frac{\partial e}{\partial q}}\left(\hat{e} - e - \frac{\partial e}{\partial s}H\right)\Gamma - \frac{\partial\rho}{\partial s}\frac{1}{\frac{\partial\rho}{\partial q}}(\Phi + \Psi) \quad , \tag{A.12}$$

where $\rho$, $e$, $T$, $B$, $\hat{e}$, $H$, $\Gamma$, $\Phi$, and $\Psi$ are given by Eqs. (B.4), (B.3), (B.2), (A.6), (B.28), (B.27), (B.26), (A.7), and (A.9), respectively.

The radial gas velocity equation is

$$\frac{\partial u_r}{\partial t} = -u_r\frac{\partial u_r}{\partial r} - u_z\frac{\partial u_r}{\partial z} + \frac{u_\theta^2}{r} - \frac{dp}{dq}\frac{1}{\rho}\frac{\partial q}{\partial r} - (u_r - u_r^*)\Gamma$$

$$- \frac{(1-\alpha)}{\alpha}(A_{drag})_r + (A_{visc})_r + (A_{turb})_r \quad , \tag{A.13}$$

where

$$(A_{visc})_r = \frac{1}{\alpha\rho}\left\{\frac{\partial}{\partial r}\left[\alpha\mu\ \frac{2}{3}\left(2\frac{\partial u_r}{\partial r} - \frac{u_r}{r} - \frac{\partial u_z}{\partial z}\right) + \alpha\lambda\left(\frac{1}{r}\frac{\partial}{\partial r}(ru_r) + \frac{\partial u_z}{\partial z}\right)\right]\right.$$

$$\left. + \frac{\partial}{\partial z}\left[\alpha\mu\left(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r}\right)\right] + 2\alpha\mu\ \frac{\partial}{\partial r}\left(\frac{u_r}{r}\right)\right\} \quad , \tag{A.14}$$

and p, ρ, Γ, α, μ, and λ are given by Eqs. (B.6), (B.4), (B.26), (B.1), (B.7), and (B.8), respectively. The radial component of the drag $(A_{drag})_r$ is given by the radial component of Eq. (B.20). The radial component of acceleration due to turbulence $(A_{turb})_r$ could be given by the radial component of Eq. (B.34) which is Eq. (A.14) with μ and λ replaced by $\mu_T$ and $\lambda_T$.

The circumferential gas velocity equation is

$$\frac{\partial u_\theta}{\partial t} = - u_r \frac{\partial u_\theta}{\partial r} - u_z \frac{\partial u_\theta}{\partial z} - \frac{u_r u_\theta}{r} - (u_\theta - u_\theta^*)\Gamma$$

$$- \frac{(1-\alpha)}{\alpha}(A_{drag})_\theta + (A_{visc})_\theta + (A_{turb})_\theta \quad , \tag{A.15}$$

where

$$(A_{visc})_\theta = \frac{1}{\alpha\rho}\left\{\frac{\partial}{\partial r}\left[\alpha\mu\ r\frac{\partial}{\partial r}\left(\frac{u_\theta}{r}\right)\right] + \frac{\partial}{\partial z}\left[\alpha\mu\ \frac{\partial u_\theta}{\partial z}\right] + 2\alpha\mu\ \frac{\partial}{\partial r}\left(\frac{u_\theta}{r}\right)\right\} \tag{A.16}$$

and α, Γ, μ, λ, and ρ are given by Eqs. (B.1), (B.26), (B.7), (B.8), and (B.4), respectively. The circumferential component of the drag $(A_{drag})_\theta$ is given by the circumferential component of Eq. (B.20). The circumferential component of the acceleration due to turbulence $(A_{turb})_\theta$ could be given by the circumferential components of Eq. (B.34) which is Eq. (A.16) with μ and λ replaced by $\mu_T$ and $\lambda_T$.

The axial gas velocity equation is

$$\frac{\partial u_z}{\partial t} = - u_r \frac{\partial u_z}{\partial r} - u_z \frac{\partial u_z}{\partial z} - \frac{dp}{dq}\frac{1}{\rho}\frac{\partial q}{\partial z} - (u_z - u_z^*)\Gamma$$

$$- \frac{(1-\alpha)}{\alpha}(A_{drag})_z + (A_{visc})_z + (A_{turb})_z \quad , \tag{A.17}$$

where

$$(A_{visc})_z = \frac{1}{\alpha\rho}\left\{\frac{\partial}{\partial r}\left[\alpha\mu\left(\frac{\partial u_r}{\partial z}+\frac{\partial u_z}{\partial r}\right)\right]+\frac{\alpha\mu}{r}\left[\frac{\partial u_r}{\partial z}+\frac{\partial u_z}{\partial r}\right]\right.$$

$$\left.+\frac{\partial}{\partial z}\left[\alpha\mu\frac{2}{3}\left(2\frac{\partial u_z}{\partial z}-\frac{\partial u_r}{\partial r}-\frac{u_r}{r}\right)+\alpha\lambda\left(\frac{\partial u_z}{\partial z}+\frac{\partial u_r}{\partial r}+\frac{u_r}{r}\right)\right]\right\} \quad , \tag{A.18}$$

and $p$, $\rho$, $\Gamma$, $\alpha$, $\mu$, and $\lambda$ are given by Eqs. (B.6), (B.4), (B.26), (B.1), (B.7), and (B.8), respectively. The axial component of the drag $(A_{drag})_z$ is given by the axial component of Eq. (B.20). The axial component of the acceleration due to turbulence $(A_{turb})_z$ could be given by the axial component of Eq. (B.34) which is Eq. (A.18) with $\mu$ and $\lambda$ replaced by $\mu_T$ and $\lambda_T$.

The components of the solid phase velocity equation are the radial solid phase velocity equation

$$\frac{\partial u_r^*}{\partial t} = -u_r^*\frac{\partial u_r}{\partial r}-u_z^*\frac{\partial u_r^*}{\partial z}+\frac{u_\theta^{*2}}{r}-\frac{dp}{dq}\frac{1}{\rho^*}\frac{\partial q}{\partial r}+\frac{\rho}{\rho^*}(A_{drag})_r+(A_{stress})_r \quad , \tag{A.19}$$

the circumferential solid phase velocity equation

$$\frac{\partial u_\theta^*}{\partial t} = -u_r^*\frac{\partial u_\theta}{\partial r}-u_z^*\frac{\partial u_\theta^*}{\partial z}-\frac{u_r^* u_\theta^*}{r}+\frac{\rho}{\rho^*}(A_{drag})_\theta \quad , \tag{A.20}$$

and the axial solid phase velocity equation

$$\frac{\partial u_z^*}{\partial t} = -u_r^*\frac{\partial u_z^*}{\partial r}-u_z^*\frac{\partial u_z^*}{\partial z}-\frac{dp}{dq}\frac{1}{\rho^*}\frac{\partial q}{\partial z}+\frac{\rho}{\rho^*}(A_{drag})_z+(A_{stress})_z \tag{A.21}$$

where $p$ and $\rho$ are given by Eqs. (B.6) and (B.4), respectively. The density of the solid phase $\rho^*$ is assumed constant. The components of the accelerations due to drag, $A_{drag}$, and intergranular stress, $A_{stress}$, are given by the components of Eqs. (B.20) and (B.23), respectively.

The particle number equation is

$$\frac{\partial n^*}{\partial t} = -\frac{1}{r}\frac{\partial}{\partial r}(rnu_r^{**})-\frac{\partial}{\partial z}(nu_z^{**}) \quad . \tag{A.22}$$

The regression distance equation is

$$\frac{\partial \overset{*}{d}}{\partial t} = - \overset{*}{u}_r \frac{\partial \overset{*}{d}}{\partial r} - \overset{*}{u}_z \frac{\partial \overset{*}{d}}{\partial z} + \langle \dot{d} \rangle \quad , \tag{A.23}$$

where the burning rate correlation $\langle \dot{d} \rangle$ is given by Eq. (B.25).

The surface temperature equation is

$$\frac{\partial \overset{*}{T}}{\partial t} = - \overset{*}{u}_r \frac{\partial \overset{*}{T}}{\partial r} - \overset{*}{u}_z \frac{\partial \overset{*}{T}}{\partial z} \langle \dot{T} \rangle \quad , \tag{A.24}$$

where the correlation $\langle \dot{T} \rangle$ for the rate of change of grain surface temperature is discussed in Section 4.7.10.

# APPENDIX B

## CORRELATION MODEL FORMULAS

This appendix contains a list of correlation model formulas. The formulas are discussed in detail in Section 4.7. The terms listed in this appendix are in a form compatible with Eq. (4.32), Section 4.3, and Appendix A.

The porosity or gas volume fraction (Section 4.2.1) is given by

$$\alpha = 1 - v_p(\overset{*}{d})\overset{*}{m}/VG \quad . \tag{B.1}$$

The equations of state (Section 4.7.1) are

$$T(p,s) = T_R\left(\frac{p}{p_R}\right)^{(\gamma-1)/\gamma}\exp\left(\frac{M}{R}\frac{\gamma-1}{\gamma}s\right) \quad , \quad K \quad , \tag{B.2}$$

$$e = \frac{1}{\gamma-1}\frac{R}{M}T \quad , \quad J/kg \quad , \tag{B.3}$$

$$\rho = \left(\frac{R}{M}\frac{T}{p} + \eta\right)^{-1} \quad , \quad kg/m^3 \quad , \tag{B.4}$$

$$a^2 = \gamma\frac{p}{\rho}\frac{1}{1-\eta\rho} \quad , \quad m^2/s^2 \quad , \tag{B.5}$$

where $R = 8.3143$ J/(mol·K) is the universal gas constant, $M$ (kg/mol) is the molar mas and $\eta$ ($m^3$/kg) is the covolume. The pressure logarithm function $q$ is defined by (Section 4.2.2)

$$q = q_1[\ln(p/p_1) + 1] \quad , \quad Pa, \quad \text{or} \quad p = p_1\exp\left(\frac{q}{q_1} - 1\right) \quad , \quad Pa \quad . \tag{B.6}$$

The shear viscosity coefficient $\mu$ and the bulk viscosity coefficient $\lambda$ are (Section 4.7.2)

$$\mu = \mu_o + \mu_1\frac{T^{1.5}}{\mu_2 + T} \quad , \quad Pa\cdot s \quad , \tag{B.7}$$

$$\lambda = \lambda_o + \lambda_1\frac{T^{1.5}}{\lambda_2 + T} \quad , \quad Pa\cdot s \quad . \tag{B.8}$$

The acceleration by viscosity is modeled by (Section 4.7.2)

$$A_{visc} = \frac{1}{\alpha\rho} \nabla \cdot \{\alpha[2\mu E + (\lambda - \frac{2}{3}\mu)\ (\text{trace } E)\ I]\}, \quad m/s^2 , \qquad (B.9)$$

where E is the strain rate tensor computed using the average velocities, i.e.,

$$E = 0.5\ (\nabla u + (\nabla u)^T) \quad . \qquad (B.10)$$

The heat dissipation function term is modeled by (Section 4.7.3)

$$\phi = \frac{1}{\rho T} \bar{\phi}(E) + \langle\phi\rangle + \frac{1}{\rho T} \phi_T \quad , \quad W/(kg \cdot K) \quad , \qquad (B.11)$$

where

$$\bar{\phi}(E) = 2\mu\ \text{trace}(E^2) + (\lambda - \frac{2}{3}\mu)\ (\text{trace } E)^2 \quad , \quad W/m^3 \quad , \qquad (B.12)$$

$$\langle\phi\rangle = \frac{1}{\rho T} |u-u^*|^2 (\frac{m^*}{4 \cdot VG})^{2/3} \pi^2 (\frac{5}{3}\mu + \frac{1}{2}\lambda) \quad , \quad W/(kg \cdot K) \quad , \qquad (B.13)$$

and $\phi_T$ is given by Eq. (B.35).

The thermal conductivity coefficient $\kappa$ is modeled by (Section 4.7.4)

$$\kappa = \kappa_0 + \kappa_1 \frac{T^{1.5}}{\kappa_2 + T} \quad , \quad W/(m \cdot K) \quad . \qquad (B.14)$$

The heat conduction term in the governing equations is modeled by (Section 4.7.4)

$$\Psi = \Psi_{gas} + \Psi_{particle} + \Psi_{turb} \quad , \quad W/(kg \cdot K) \quad , \qquad (B.15)$$

where

$$\Psi_{gas} = \frac{1}{\alpha\rho T} \nabla \cdot (\alpha\ \kappa\ \nabla T) \qquad (B.16)$$

and

$$\dot{Q}_{particle} = \begin{cases} -\frac{1}{\alpha\rho T}\frac{\overset{*}{m}}{VG} s_p\left[h_c(T-\overset{*}{T}) + h_r(T-\overset{*}{T})\right] & \text{, before ignition} \\ 0 & \text{, after ignition} \end{cases} \tag{B.17}$$

with

$$h_c = \frac{\kappa}{\overset{*}{D}_p/2} + 0.2\left(\frac{\gamma}{\gamma-1}\frac{R}{M}\frac{(\kappa 2\rho)^2|u-\overset{*}{u}|^2}{\mu\overset{*}{D}_p/2}\right)^{1/3} \quad , W/(m^2\cdot K) \quad , \tag{B.18}$$

and

$$h_r = \overset{*}{\varepsilon}\,\sigma_{SB}(T+\overset{*}{T})\,(T^2+\overset{*}{T}^2) \quad , W/(m^2\cdot K) \quad . \tag{B.19}$$

In Eq. (B.19), $\overset{*}{\varepsilon}$ is the particle emissivity, $\sigma_{SB} = 5.67032\cdot 10^{-8}\ W\cdot m^{-2}K^{-4}$ is the Stephan-Boltzman constant, and $\overset{*}{T}$ is the average grain surface temperature. The turbulent heat flux within the gas, $\dot{Q}_{turb}$ is given by Eqs. (B.36) and (B.37).

The <u>acceleration term due to the drag</u> between gas and particles is modeled by (Section 4.7.5).

$$A_{drag} = \begin{cases} A_{Ergun} & \text{for } \alpha<0.65 \quad , \\ 4[(\alpha-0.65)A_{Reynolds} + (0.9-\alpha)A_{Ergun}] & \text{for } 0.65<\alpha<0.9 \quad , \\ A_{Reynolds} & \text{for } 0.9<\alpha \quad , \end{cases} \tag{B.20}$$

where

$$A_{Ergun} = (u-\overset{*}{u})\frac{\overset{s}{\rho}_p}{v_p}\frac{2}{3}\frac{1}{\alpha^2}\left[1.75\,|u-\overset{*}{u}| + 150\,(1-\alpha)\frac{\overset{*}{\mu}}{\rho\overset{*}{D}_p}\right] \quad , \quad m/s^2 \quad . \tag{B.21}$$

and

$$A_{Reynolds} = (u-\overset{*}{u}) \frac{a_p}{v_p} \left[0.2 \left|u-\overset{*}{u}\right| + 12 \frac{\mu}{\overset{*}{\rho} D_p}\right] \quad , \quad m/s^2 \quad . \tag{B.22}$$

The acceleration term due to intergranular stress is modeled by (Section 4.7.6)

$$A_{stress} = - \overset{*}{a}^2 \frac{1}{1-\alpha} \nabla(1-\alpha) \quad , \quad m^2/s^2 \quad , \tag{B.23}$$

where $\overset{*}{a}(\alpha)$ is a sound speed function for the particulate phase. The function is modeled by

$$\overset{*}{a}(\alpha) = \begin{cases} a_{sp}\left(\frac{\alpha_1 - \alpha_o}{\alpha - \alpha_o}\right)\left(\frac{\alpha_2 - \alpha}{\alpha_2 - \alpha_1}\right) & \text{for } \alpha_o < \alpha < \alpha_2 \\ 0 & \text{for } \alpha_2 < \alpha \end{cases} \tag{A.24}$$

The burning rate is modeled by (Section 4.7.7)

$$\langle \dot{d} \rangle = \begin{cases} 0 & \text{for } \langle \overset{*}{T} \rangle < T_{ingition} \\ B_o + B_1 p^{B_2} \quad , \quad m/s, & \text{for } \langle \overset{*}{T} \rangle > T_{ignition} \end{cases} \quad . \tag{B.25}$$

The source term $\Gamma$ is (Section 4.7.8)

$$\Gamma = \frac{1}{\alpha} \frac{\overset{*}{\rho}}{\rho} \frac{\overset{*}{m}}{VG} s_p \langle \dot{d} \rangle \quad , \quad 1/s \quad . \tag{B.26}$$

The enthalpy factor H of the source term (Section 4.7.8) is defined by

$$H = \frac{1}{T} \left[ (\hat{e} + p/\overset{*}{\rho}) - (e + p/\rho) \right] \quad , \quad J/(kg \cdot K) \quad , \tag{B.27}$$

where $\hat{e}$ is

$$\hat{e} = \frac{1}{\gamma-1}\frac{R}{M}T_{flame} = \frac{1}{\gamma-1}g_a I_p \quad . \quad J/kg \quad , \qquad (B.28)$$

with $g_a = 9.80665\ m/s^2$ being the standard acceleration.

The particle geometry enters the equations as the four functions $v_p(\overset{*}{d})$, $s_p(\overset{*}{d})$, $\overset{*}{D}_p(\overset{*}{d})$ and $a_p(\overset{*}{d})$. We provide the formulas that define these functions for spherical, cylindrical, and tubular grains.

For a spherical grain with initial diameter $\overset{*}{D}_o$ one defines

$$\left.\begin{aligned} R &= \max\,(0,\ ((\overset{*}{D}_o - 2\overset{*}{d})/2) \quad , \\ v_p &= \frac{4}{3}\pi R^3 \quad , \\ s_p &= 4\pi R^2 \quad , \\ a_p &= \pi R^2 \quad , \\ \overset{*}{D}_p &= 2R \quad , \end{aligned}\right\} \qquad (B.29)$$

A solid cylindrical grain may be described by its initial diameter, $\overset{*}{D}_o$, and height, $L_o$. Let

$$\left.\begin{aligned} R &= (\overset{*}{D}_o - 2\overset{*}{d})/2 \quad , \\ L &= L_o - 2\overset{*}{d} \quad , \end{aligned}\right\} \qquad (B.30)$$

If either $R < 0$ or $L < 0$, then the grain has been burnt. If both quantities are positive, then we define

$$v_p = \pi L R^2 \quad ,$$

$$s_p = 2\pi R(R+L) \quad ,$$

$$a_p = (2RL\pi R^2)/2 \quad , \qquad \text{(B.31)}$$

$$\overset{*}{D}_p = (2R+L)/2 \quad .$$

A tubular grain may be defined by its initial height and the initial outer and inner diameters, $\overset{*}{D}_o$ and $\overset{*}{d}_o$, respectively. Let

$$R = (\overset{*}{D}_o - 2\overset{*}{d})/2 \quad ,$$

$$r = (\overset{*}{d}_o + 2\overset{*}{d} \quad . \qquad \text{(B.32)}$$

$$L = \overset{*}{L}_o - 2\overset{*}{d}$$

The grain is completely burnt if either $R-r<0$ or $L<0$. If both of these quantities are positive, then the grain geometry functions are

$$v_p = \pi(\overset{*}{D}_o + \overset{*}{d}_o)\ (R-r)L/2 \quad ,$$

$$s_p = \pi\ (\overset{*}{D}_o + \overset{*}{d}_o)(R-r+L) \quad ,$$

$$a_p = (2RL + \pi(R^2 - r^2))/2 \quad , \qquad \text{(B.33)}$$

$$\overset{*}{D}_p = (2R+L)/2 \quad .$$

We consider a detailed study of turbulence models for interior ballistics flows to be outside the scope of this report. Hence, the correlation models are quite elementary and are listed in this report only for completeness. The acceleration by the gas phase turbulent stress tensor $A_{turb}$ and the turbulent heat dissipation function $\phi_T$, could have the same form as $A_{visc}$ (Eq. (B.9)) and $\bar{\phi}(Z)$, (Eq. (B.12)), respectively, but different viscosity coefficients, that is,

$$A_{turb} = \frac{1}{\alpha\rho} \nabla\cdot\{\alpha[\, 2\, \mu_T E + (\lambda_T - \tfrac{2}{3}\mu_T)\ (\text{trace } E)\ I]\} \quad , \ m/s^2 \quad , \ (B.34)$$

$$\Phi_T = 2\, \mu_T \text{ trace } (E^2) + (\lambda_T - \tfrac{2}{3}\mu_T)\ (\text{trace } E)^2 \quad , \ W/m^3 \quad . \ (B.35)$$

where $\mu_T$ and $\lambda_T$ denote the viscosity coefficients for turbulent flows. The manner in which these coefficients are determined strongly depends on the particular turbulence model one uses and, hence, will not be given. As discussed in Section 4.7.6, the solid phase turbulent stress tensor $\overset{*}{\Pi}_T$ is set to zero. The turbulent heat flux vector $Q_T$ is modeled by Ishii[13] and Gibeling et al.[5] as

$$Q_T = -\kappa_T[\nabla T - \frac{\nabla\alpha}{\alpha}\ (T_i - T)] \quad , \quad W/m^2 \quad , \qquad (B.36)$$

where $T_i$ is an average temperature on the interface (a function of T and $\overset{*}{T}$) and $\kappa_T$ is given by an algebraic formula involving an effective viscosity and Prandtl number. The corresponding model of $\Psi_{turb}$ in Eq. (B.15) is

$$\Psi_{turb} = -\frac{1}{\alpha\rho T} \nabla\cdot(\alpha Q_T) \quad , \quad W/(kg\cdot K) \quad . \qquad (B.37)$$

# NOTE ON EVAPORATION IN POROUS MEDIA

R E Meyer*
Mathematics Research Center
University of Wisconsin
Madison, WI 53706

ABSTRACT. Factors are discussed which govern evaporation of liquid in the small capillaries of a porous medium. Attention is directed to sheet-like aggregates from which the vapor can escape with little obstruction. Marked temperature gradients are then found to be confined to close neighborhoods of the menisci and evaporation is shown to proceed in statiscally quite unstable configurations under a dynamic balance of surface tension, local evaporation rate and viscous shear. Estimates of evaporation rates and fluid velocities are given. The results discourage constitutive theories for porous media because mere size of capillaries, independently of shape and chemistry, is found to change the physical processes underlying macroscopic behavior.

I. INTRODUCTION. The following study was prompted by recognition that little is known about the physics of evaporation in fabrics beyond the guess that the rate of heat supply may equal the rate of latent-heat expenditure. Fabrics come in a great variety of very different structures and as a first step, a structure characteristic of "typical" porous media is here envisaged in which the solid matrix is threaded by an irregular network of interconnecting, small capillaries along each of which the capillary bore varies greatly over relatively small distances. The immediate challenge is then to isolate some of the many interacting, physical processes in a single capillary in order to distinguish those which really govern evaporation there; macroscopic descriptions must needs reflect the insights thereby gained.

A key restriction that helps in dividing the difficulties is to focus attention on sheet-like media which are thin in one direction, like fabrics, because the escape of the vapor is then relatively unobstructed and consideration of the processes in the vapor can be postponed (to Section VIII). At first sight, evaporation might be expected to be controlled by the manner of heat supply, but for capillaries of realistically small size, most forms of heat supply have similar effects because heat transfer across the capillary wall is then always important. Since the physical signposts diverge, unless one be quite specific, attention is restricted to liquids similar to water, to pressures and temperatures typical of the outdoors, and to throat diameters of about $10^{-4}$ to $10^{-2}$ cm. A final dividing step is to start with an unrealistic configuration of geometrical and thermal symmetry in which only a single meniscus needs to be considered (Section III, IV).

Analysis of the simple thermal balances for that case shows that significant temperature gradients can occur only very close to menisci (Section IV), and a rough estimate of evaporation emerges (Section IV).

---

* on leave at Department of Mathematics, Imperial College, London SW7 2BZ



PREVIOUS PAGE IS BLANK

It shows the symmetrical case to be normally unstable (Section V). Evaporation is, in fact, found to proceed in statically grossly unstable configurations under a dynamic balance depending drastically on viscous shear. The "Haines Jumps" in the foreground of earlier accounts [1 - 3] can occur only in much larger passages than would appear realistic for soils, oil recovery or fabrics. Instead, the normal state in evaporation is one of slow liquid motion leaving the small menisci almost stationary, while most of the mass-evaporation occurs at the large ones (Section VI). These results furnish a basis (Section VII) for statistical estimates of macroscopic evaporation rates, provided enough is known about the statistical distribution of capillary throat sizes. Such knowledge appears to be an absolute prerequisite for any useful treatment of fluid motion in porous media because viscous shear depends so violently on throat size. As a result, if two media have the same chemistry and identical shape for their respective void passages, but differ in mere geometrical scale, then microscopic dynamic balances can be quite different. This discourages constitutive theories of porous media, of which invariance to geometric scale is a basic premise.

There are other caveats, for instance, if the vapor-air mixture must pass through small throats, the significant gasdynamical processes must be anticipated to change the evaporation rate drastically. On the other hand, there are also many bits of luck, which make a realistic fluid mechancis of porous media more accessible. In particular, the extreme magnitudes of relevant combinations of physical parameters will come to explain, by and by, why errors by only a factor 2, or so, will be treated so cavalierly.

II. STATIC EQUILIBRIUM. When the escape of the air-vapor mixture is unobstructed, the pressure throughout that gas is effectively the ambient pressure, $P_a$. Admittedly, since surface tension promotes evaporation, thermal equilibrium between liquid and vapour requires a gas pressure at their interface which depends on the menisci's curvature. Helmholtz' analysis [4], however, shows this to be a threshhold effect, and for realistic capillary bores, the thermal conditions here envisaged are well below the threshhold [5]. With the apparent contact angle $\beta$ measured as in Figure 1, the pressure on the liquid side of a meniscus is therefore

$$P_\ell = P_a - 2\sigma/a, \qquad (1)$$

where $a$ denotes the local capillary radius and $\sigma$ sec $\beta$ the surface tension. It will be assumed that $0 < \beta < \pi/2$, for otherwise, surface tension would have kept the liquid out of the porous matrix. Body forces will be neglected. Since it is prohibitive to take account of all the shapes of void-passage cross-sections liable to occur in a porous matrix, $a$ will be defined by $\pi a^2$ = area of cross-section; the error in (1) is then one of those treated cavalierly.

By (1), a connected liquid column is in static equilibrium if, and only if, $a$ takes the same value at all the menisci bounding the column. The equilibrium is stable if, and only if, $a$ does not decrease with distance along the capillary measured toward the gas side at any of those menisci (Fig 2).

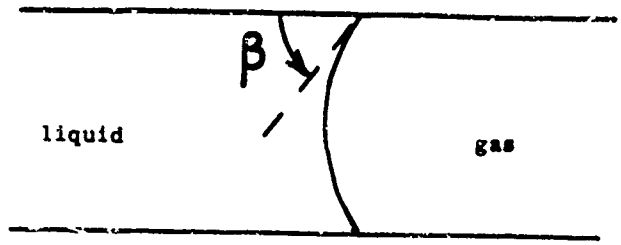



Figure 1



Figure 2a. Stable equilibrium

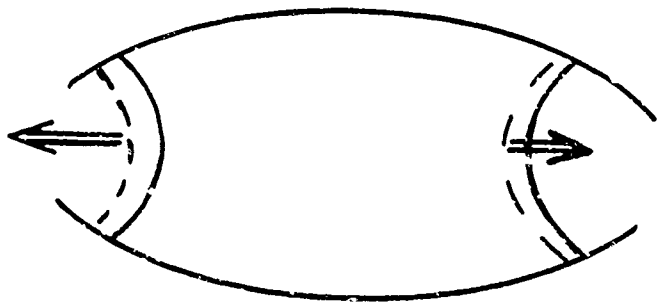


Figure 2b. Unstable equilibrium

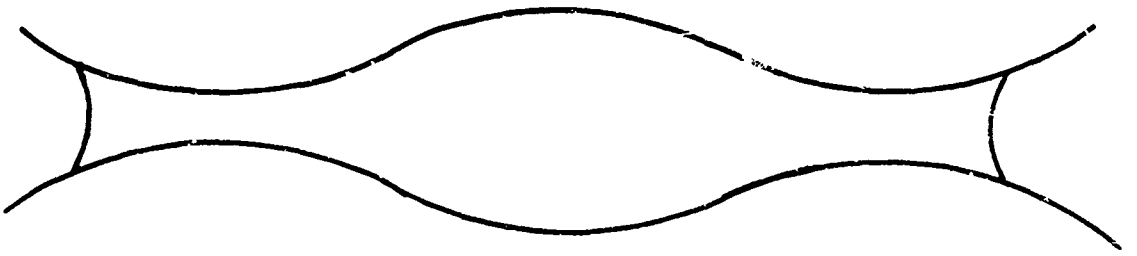


Figure 3

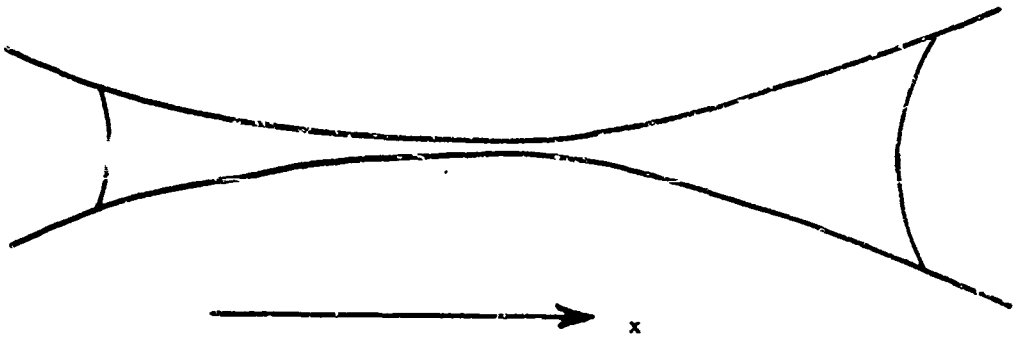



Figure 4

Those are merely local conditions, however. All kinds of void-passage shapes can be expected in a porous matrix. Figure 3 indicates one with a liquid column in a position of stable equilibrium. Suppose, however, that this liquid volume be reduced by, say, 30%, e.g., by evaporation. The remaining liquid column could not then find a stable equilibrium position anywhere in the passage segment shown : it must move to find a new, stable position elsewhere. The need for such "Haines jumps" [1] in the liquid configurations has dominated the literature on the physics of fluids in porous media [1 - 3] : on kinematic considerations, they would be expected to be sudden and frequent so that fluid motion could appear as a continuous process only on a long-term average [3]. Analysis of the dynamics (Section VI), however, will show that such notions can be relevant only to unrealistically large passages.

III. THERMAL BALANCE. Prolonged evaporation must depend on an external heat supply, and it might be anticipated that, not only the rate, but also the manner, of that supply has a major influence on the evaporation process. However, since the ratio of capillary volume to surface area is proportional to the capillary bore, heat transfer across the capillary wall is always important in small capillaries. That transfer will be found in section IV to cause an adjustment of temperatures in the fluid and solid that reduces the influence of the manner of heat supply. To fix the ideas, it will be envisaged that a reservoir supplies it to the solid matrix. in the first place, so as to heat it gradually, from the original, uniform temperature $T_0$ of the whole medium, which may be considered to be known, to a level $T_1$ at which it is then maintained. Specifically, $T_0$ may be the outdoor temperature, and $T_1$ may be near body-temperature, perhaps $T_1 \approx T_0 + 50°F$. The properties of the solid are outside the scope of this study and will be assumed uniform. This does not imply a temperature field constant in time and uniform in space, because evaporation makes the menisci act like moving heat sinks of changing strength, but the space-average $T_w$ of the capillary wall temperature will change only slowly with time. The stronger, local variation in capillary wall temperature can be accounted for approximately by rough adjustments [5] and meanwhile, the considerations can be simplified by ignoring the difference between $T_w$ and the local temperature of the capillary wall. That will lead to estimates of evaporation at menisci in terms of $T_w$, whence estimates of the time development of $T_w$ by more global considerations will follow.

Since little can be known about the shapes of realistic void-passage cross-sections, the description of evaporation will be simplified greatly by representing quantities in the fluid by their averages over a capillary cross-section and ignoring the errors resulting from use of somewhat different averages in different contexts. The thermal description is also greatly simplified, if no liquid motion couples the processes at different menisci. That is possible in the idealized case of a capillary of radius $a(x)$ even in distance $x$ along the capillary, if the temperature field is similarly even and liquid fills a capillary segment between menisci at $x = \pm \ell$, with $a'(\ell) > 0$ for static stability (Fig 2a). Such symmetry can

persist, at least in principle, and will be assumed in this Section and the next one.

The local thermal balance per unit length is then

$$\pi a^2 \rho_\ell c_\ell \frac{\partial T_\ell}{\partial t} = \frac{\partial}{\partial x}\left(\pi a^2 \lambda \frac{\partial T_\ell}{\partial x}\right) + 2\pi a h (T_w - T_\ell),$$

where the lefthand side represents the local rate of increase of liquid heat content; $\rho$ and $c$ denote density and heat capacity, respectively, and the suffix $\ell$ distinguishes liquid properties. The first term on the righthand side represents the contribution from heat conduction in the liquid, and the last term, that from heat transfer across the capillary wall on the most rudimentary, conventional model of a heat transfer rate per unit wall area and unit temperature difference represented by a constant transfer coefficient h. With the insignificant further approximation of neglect of variations in $\lambda$, $\rho_\ell$ and $c_\ell$, the balance becomes

$$\frac{\partial T_\ell}{\partial t} = \frac{\kappa}{a^2} \frac{\partial}{\partial x}\left(a^2 \frac{\partial T}{\partial x}\right) + \frac{2h}{\rho_\ell c_\ell a} (T_w - T_\ell) \qquad (2)$$

where $\kappa$ denotes the usual heat diffusivity, $\kappa = \lambda / (\rho_\ell c_\ell)$. The liquid therefore experiences a typical heat conduction process with variable effective diffusivity, on account of the capillary shape, and with heat transfer, but without convection, on account of the symmetry.

A somewhat different balance arises at a meniscus. Since the liquid and gas are there envisaged in dew-point equilibrium to begin with, and since the gas pressure remains at the ambient level $p_a$ until Section VIII, any heat reaching the meniscus will result meanwhile in evaporation, but not [2] in a change of the local temperature $T_M$ from its original level $T_0$. Conduction through the liquid column in $x < \ell$ contributes heat to the meniscus at the rate

$$-\pi a_m^2 \lambda (\partial T_\ell / \partial x)_{x=\ell} \quad ,$$

if the meniscus is at $x = \ell(t)$ and $a_m$ denotes the capillary radius $a(\ell)$ there. Conduction through the gas in $x > \ell$ does not contribute comparably because its heat conductivity is smaller. If the meniscus is markedly curved, heat reaches it also by direct transfer across a short segment of the capillary wall and radial heat conduction. The wall area from the meniscus contact line to the position of its apex (Fig 1) is approximately $2\pi a_m^2 \alpha_2$ with $0 < \alpha_2 = \sec\beta - \tan\beta < 1$. The rate of heat transfer across

this area is

$$2\pi a_m^2 \alpha_2 h_M (T_w - T_M) ,$$

where $h_M$ denotes a value of the transfer coefficient h adjusted [5] to compensate for the error made by confusing the local wall temperature at the meniscus with its level further away.

Let S denote a short capillary segment of fixed length which is stationary in a frame moving with the local, liquid velocity and which contains the meniscus at present. The gas pressure and temperature are constant in S, and if $D\ell/Dt$ denotes the velocity of the meniscus in that frame, the mass-rate of evaporation in S is

$$m = -\pi a_m^2 \rho_\ell D\ell/Dt . \qquad (3)$$

The mass of gas in S does not increase at a significant rate because the density ratio $\rho_g/\rho_\ell$ is about $10^{-3}$, in the circumstances envisaged, so that vapour leaves S at the same mass-rate, and the net rate of mass loss in S is also m, since no liquid enters S. By the First Law, the net rate of liquid enthalpy loss in S equals the rate of vapor enthalpy loss from it less the rate of heat addition by transfer and conduction into S,

$$\pi a_m^2 [2\alpha_2 h_M (T_w - T_M) - \lambda \partial T_\ell/\partial x] = mL , \qquad (4)$$

where L is the latent heat per unit mass.

IV. THERMAL LAYER. The use of these balances requires a nondimensional notation, and it is not obvious whether a single length scale X can be representative of the temperature field throughout a liquid column. Even without attention to cross-sectional shape, the capillary is described by the two functions a(x) and a'(x)/a(x) of normally quite different magnitued and each of which may vary by orders of magnitude along a capillary. To avoid confusion, let attention be confined first to a liquid column segment adjacent to the initial meniscus position $x = \ell(0)$ and short enough to occupy only a capillary segment characterized by a single triplet of scales $a_0$ of the capillary radius, G, of $a(x)/a'(x)$ and X, of the unknown temperature variation. But, there is another thermal length scale,

$$\Lambda = (\tfrac{1}{2}\lambda a_0/h)^{\frac{1}{2}} ,$$

of decisive significance because $\Lambda^2/X^2$ represents the ratio of the thermal diffusion scale to the transfer scale. How is X related to the other three length scales?

The present model can give no information on how $\Lambda$ compares with $a_0$ and G, but a more detailed calculation [5] shows that $\Lambda$ must be anticipated to be of the order of the capillary radius, so that

$$\Lambda/a_0 = [\lambda/(2a_0 h)]^{\frac{1}{2}} = \gamma_h$$

is a parameter of order unity; a rough estimate [5] is $\gamma_h \approx \frac{1}{2}$. Accordingly, $X = a_0$, unless this be found to imply that $T_\ell(x)$ can vary only on a longer

scale. The natural temperature-difference scale is $T_w - T_M = \Delta$, and if

$$T_w - T_\ell(x,t) = T(x,t)\,\Delta \quad ,$$

t is measured in units of a time scale $\tau$, a/a' in units of G, and x, a and $\ell$, in units of $X = a_0$, then the nondimensional form of (2) and (4) is

$$\frac{a_0^2}{\kappa\tau}\frac{\partial T}{\partial t} - \frac{2a'}{a}\frac{a_0}{G}\frac{\partial T}{\partial x} = \frac{\partial^2 T}{\partial x^2} - \frac{1}{\gamma_h 2a}\,T \qquad (5)$$

and

$$\frac{\partial T}{\partial x} + \frac{\alpha_3}{\gamma_h^2} = -\frac{a_0^2}{\varepsilon\kappa\tau}\frac{d\ell}{dt} \quad \text{at } x = \ell(t),$$

respectively, because

$$T(\ell,t) = 1$$

in this notation and because the liquid is at rest; here $\alpha_3 = \alpha_2 h_M/h$ and

$$\varepsilon = c_\ell \Delta/L << 1$$

in the circumstances here envisaged, eg, $\varepsilon = 1/20$ for water and $\Delta = 50°F$.

Two different time scales have emerged from the balances. The shorter, $a_o^2/\kappa$, characterizes transients arising from imbalance of heat transfer and conduction which might be anticipated, eg, in a Haines jump. It is normally a rather small fraction of a second, eg, if $\kappa = 10^{-3}cm^2/sec$ and $a_0 = 10^{-2}cm$, then $a_0^2/\kappa = 10^{-1}sec$. The motion of the meniscus, on the other hand, is on the longer time scale

$$a_0^2/(\kappa\varepsilon) \quad .$$

The first question must be whether a relatively stable evaporation process is possible, and to examine this, the time scale $\tau$ must be identified with $a_0^2/(\kappa\varepsilon)$. If also $a_0/G << 1$, as one would usually expect, then the lefthand side of (5) becomes unimportant by comparison to the righthand terms, in which $a^{-1} \approx 1$, to the same approximation. The balances then imply

$$T = \exp\frac{x-\ell(t)}{\gamma_h}, \qquad \ell(t) = \ell(0) - \frac{\gamma_h+\alpha_3}{\gamma_h^2}\,t \qquad (6)$$

[except when the capillary segment under scrutiny contains x = 0, in which

case

$$T = 2^{-\ell(t)/\gamma_h} \cosh(x/\gamma_h)].$$

This solution describes a very short meniscus layer, of thickness equal to the thermal length scale $\Lambda = \gamma_h a_0$, in which virtually the whole process of heat transfer to, and heat conduction in, the liquid takes place.

This conclusion destroys the analysis sketched so far because one of its main premises -- that thermal balances can be formulated in terms of cross-sectional averages -- cannot apply to precisely the short capillary segment containing the curved meniscus and all significant temperature gradients! Any tenable analysis of the temperature field must account for the geometry of the meniscus, but that depends mainly on matters accessible only to vague speculation, at best, namely the shape of the capillary cross-section and the apparent contact angle.

If a tenable analysis could be performed, on the other hand, it would necessarily lead to the same dimensional groups and would therefore also predict a dimensional meniscus velocity of the form

$$\gamma \frac{\kappa\epsilon}{a_0} = \gamma \frac{\lambda(T_w - T_M)}{a_0 \rho_\ell L}$$

and a dimensional evaporation time of the form

$$\gamma^{-1} \ell_0 a_0/(\kappa\epsilon)$$

for a capillary segment of length $a_0$. What has no rational support is the value $(\gamma_h + \alpha_3)/\gamma_h^2$ of the nondimensional coefficient $\gamma$ predicted by (6).

Thought about extreme cases indicates however, that the correct value of $\gamma$ cannot plausibly be far from order unity and indeed, that $\gamma = 3$, 4 or 5 cannot usually be very wrong. To fix the ideas, therefore, the value $\gamma = 4$ will be adopted speculatively for illustration. For water (with $\kappa \approx 0.0014$ cm$^2$/sec) and $\epsilon = 1/20$, various capillary radii and lengths then give roughly the meniscus velocites and evaporation times listed in the following table.

TABLE 1

| | $a_0$(cm) | | | | |
|---|---|---|---|---|---|
| | $10^{-2}$ | $10^{-3}$ | | $10^{-4}$ | |
| $4\epsilon\kappa/a_0$ | $3 \times 10^{-2}$ | $3 \times 10^{-1}$ | | 3 | cm/sec |
| $\ell_0$ | $10^{-1}$ | $10^{-1}$ | $10^{-2}$ | $10^{-2}$ | $10^{-3}$ | cm |
| $\ell_0 a_0/(4\epsilon\kappa)$ | 3 | $\frac{1}{3}$ | $\frac{1}{30}$ | $3 \times 10^{-3}$ | $3 \times 10^{-4}$ | sec |

In sum, the analysis has been wrong in everything but its result. Most of all, what has been proven, if only by contradiction, is that the temperature variation associated directly with evaporation from a meniscus must be quite local. It follows that the main results are also independent of some other premises. There can be no significant, direct thermal interaction between different menisci even if the liquid moves on the time scale $a_0^2/(\kappa\epsilon)$; the meniscus velocity should then be interpreted as that of the meniscus relative to the adjacent liquid and $a_0$ must represent the scale of the capillary radius $a_m$ at the instantaneous meniscus position. The manner of heat supply, moreover, can have less direct influence on the local process than might have been thought at first : apart from the local dip in solid temperature at the meniscus (accounted for by a correct value of $\gamma$) the solid temperature background of a capillary must reflect the macroscopic scale of the solid matrix as a whole and must therefore appear effectively uniform on the length scale $a_0$ of the local evaporation process.

In the first place, (6) applies only to a capillary segment in which the radius differs by less than an order of magnitude from that at the initial meniscus position. Once evaporation has cleared that segment, however, an analogous calculation with a different scale $a_0$ applies to the next segment. Since narrow segments are seen to clear in a much shorter time than wide ones, it does not appear worth entering here upon the refinement of replacing $a_0$ from the start by the capillary radius $a_m(t)$ at the meniscus.

V. INSTABILITY. Before evaporation, all menisci bounding a connected liquid column must be of the same size (Section II), but in a realistic porous medium $a'(x)$ cannot also be expected to have the same value at different such menisci. The meniscus velocity (Section IV) then takes the same value at all the initial menisci positions, but if they all started to move with it, static equilibrium would be lost promptly. Surface tension acts towards restoring it, but the differences in meniscus velocity relative to the liquid act in the opposite sense. Stability of evaporation therefore poses a question different from that of static stability (Section II).

To examine it, consider a liquid column bounded by two menisci at which the gas pressure, $p_g$, and temperature, $T_M = T_0$, remain the same, but at which the capillary radii differ. To illuminate the distinction from static stability, suppose $a'(x)$ is monotone over the whole liquid-filled capillary segment, which contains a throat (Fig 4). Denote the meniscus positions by $x = \ell_+$ and $x = \ell_- < \ell_+$ and the capillary radii there, by $a(\ell_+) = a_+$ and $a(\ell_-) = a_- < a_+$, respectively; ie, the smaller meniscus is at the lefthand end of the liquid column. By (1), surface tension generates a pressure difference

$$p(\ell_+) - p(\ell_-) = 2\sigma \frac{a_+ - a_-}{a_+ a_-}$$

driving the liquid towards the left.

If $a'(x)$ is not too large, the viscous shear generated by the ensuing liquid motion sets up a pressure gradient related approximately by Poiseuille's formula

$$Q = -\frac{\pi\rho}{8\mu} a^4 \frac{dp}{dx}$$

to the mass-flow rate Q (counted towards the right), which is independent of x, by mass conservation. Since liquid density variation is insignificant,

$$p(\ell_+) - p(\ell_-) = -\frac{8\mu Q}{\pi\rho} \int_{\ell_-}^{\ell_+} [a(x)]^{-4}\, dx,$$

and since the main contribution to this integral arises from the throat region, it promotes clarity to write the integral as $\ell_t/a_t^4$ in terms of the throat radius $a_t$ and a "throat length" $\ell_t$. The mass-flow rate generated by surface tension is then

$$Q = -\frac{\pi\rho\sigma}{4\mu} \frac{a_t^4}{a_+ a_-} \frac{a_+ - a_-}{\ell_t}$$

and the corresponding cross-sectional averages of liquid velocity are $Q/(\pi a_+^2)$ at the right meniscus and $Q/(\pi a_-^2)$, at the left one.

Evaporation, on the other hand, retracts the menisci into the liquid with velocities $\gamma\kappa\varepsilon/a_+$ and $\gamma\kappa\varepsilon/a_-$, respectively (Section IV). The center of the liquid column therefore shifts at the rate

$$\frac{d}{dt} \frac{\ell_+ + \ell_-}{2} = \frac{\kappa\varepsilon}{2} \frac{a_+ - a_-}{a_+ a_-} \left[ \gamma - \left( \frac{a_t^2}{a_+ a_-} \right)^2 \frac{a_+^2 + a_-^2}{4Ce\, a_0 \ell_t} \right],$$

where

$$Ce = \varepsilon\kappa\mu/(\sigma a_0)$$

is a capillary number based on evaporation velocity and is normally very small, if the apparent contact angle is not close to 90° (Fig 1); for water, eg, $4a_0 Ce \sim 10^{-8}$ cm. The factor of $Ce^{-1}$ in the last bracket, however, tends to be even smaller, as long as $a_+$ and $a_-$ remain large

compared to the throat radius $a_t$. During such a phase of evaporation, the liquid column therefore shifts towards the right, ie, in the direction opposite to that suggested by surface tension alone. Any statically stable liquid configuration with $a_m >> a_t$ is therefore unstable under evaporation.

If the menisci were found close to the throat, on the other hand, in the last stage of evaporation, then the second term in the last bracket would be large, the liquid column would move leftward, and the smaller meniscus would move away from the throat.

In sum, most of the evaporation must be anticipated to occur in liquid configurations that are not static equilibria, and there might be preferred positions for the smallest menisci.

VI. DYNAMIC BALANCE. For a realistic impression of evaporation in porous media, one must therefore consider liquid configurations far from static equilibrium, for instance, such as indicated in Figure 5, which envisages a situation that might be seen in a snapshot of a "Haines Jump" after evaporation has made one meniscus clear a throat. The disparity of the meniscus sizes then generates a marked pressure difference driving the liquid towards the smaller meniscus, and an unsteady liquid motion must be anticipated. There are two very small time scales, namely the liquid column length divided by the sound speed in it and the time scale $a_t^2/\nu$ of viscous diffusion of shear from the capillary wall in the throat region (Fig 5), which is about $10^{-4}$sec in water, if $a_t \sim 10^{-3}$cm. The evaporation time scales (Section IV) are much longer, and the full viscous shear must therefore be expected to have been established, particularly in the throat region. The drastic degree to which this viscous shear in small capillary throats will be seen presently to control evaporation illustrates the reason for the prominence of Darcy's law in porous fluid mechanics.

The pressure imbalance drives a mass-flow rate $Q(t) < 0$, since it is directed towards the smaller meniscus (Fig 5). At the same time, the menisci retract into the liquid with their respective evaporation velocities $\gamma\kappa\varepsilon/a_-$ and $\gamma\kappa\varepsilon/a_-$ (Section IV). It will promote clarity, and help to distinguish the more generic case from that discussed in the preceding Section, to exploit the disparity in meniscus sizes (Fig 5) to the degree of neglecting $a_-$ against $a_+$. The larger meniscus is then considered to move just with the velocity $dx_+/dt = Q/(\pi\rho a_+^2)$, but the smaller, to move with the velocity

$$dx_-/dt = \gamma\kappa\varepsilon/a_- + Q/(\pi\rho a_-^2) \qquad (7)$$

relative to a fixed frame. The pressure difference is now approximated as $p_+ - p_- = 2\sigma/a_-$ and the total shear stress is $-8\mu\, Q\ell_t/(\pi\rho a_t^4)$, in

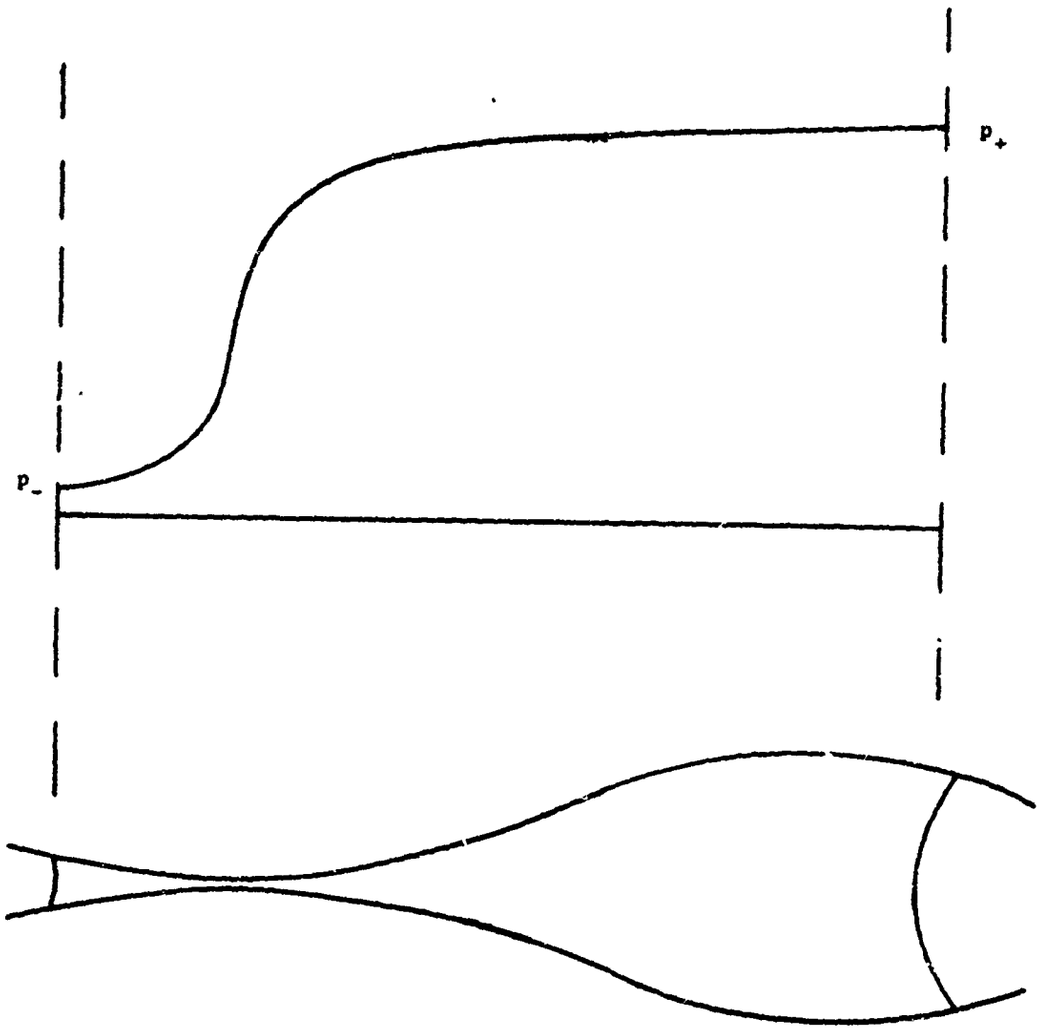



Figure 5

terms of the "throat length" $\ell_t$ of Section V. In small capillaries, the pressure drop and viscous shear cannot come into significant imbalance, and therefore,

$$2\sigma/a_- = -8\mu\, Q\ell_t/(\pi a_t^4).$$

This is the microscopic version of Darcy's law for liquid motion driven through a small capillary by surface tension $\sigma\sec\beta$ at menisci of disparate size. Since it relates Q to $a_-$, (7) may be written

$$a_- \frac{dx_-}{dt} = \gamma\varepsilon\left[1 - \frac{\sigma}{4\gamma\varepsilon\kappa\mu\ell_t}\left(\frac{a_t^2}{a_-}\right)^2\right],$$

and since $da/dx < 0$ at the position $x = x_-(t)$ of the smaller meniscus (Fig 5), this shows the approximate dynamics to be represented by an equation of the structure

$$c_1 \frac{d}{dt}(a_-^2) = \frac{c_2}{a_-^2} - 1, \qquad 1/c_1 = -2\gamma\varepsilon a'(x_-) > 0, \tag{8}$$

and with increasing time, $a_-^2$ must approach

$$c_2 = \sigma a_t^4/(4\gamma\varepsilon\kappa\mu\ell_t).$$

This does not represent a strict equilibrium because $T_w$, and therefore also $\varepsilon$, may change slowly with time, and in any case, the liquid keeps moving, but only a minor drift of the smaller meniscus results therefrom. In terms of a hybrid capillary number

$$Ct = \varepsilon\kappa\mu/(\sigma a_t)$$

based partly on evaporation and partly, on throat radius, the smaller meniscus remains close to the position where

$$a_-/a_t \sim (4\gamma Ct\ell_t/a_t)^{-\frac{1}{2}}. \tag{9}$$

The capillary radius at the smaller meniscus is thus seen to depend most of all on the throat radius, indeed, to be proportional to $a_t^2$, on account of the dominance of viscous shear in small capillary throats. For a rough impression, for water, $\varepsilon = 1/20$, $a_t = 10^{-3}$cm and $\sec\beta = 2$, the hybrid capillary number is $Ct \sim 10^{-5}$; a value 15 of $4\gamma$ is unlikely to be wrong by a large factor, and if $\ell_t \sim 20a_t$, then the last formula

predicts $a_-/a_t \sim 20$. Figure 5, accordingly, gives a reasonable impression of a typical configuration. Table I gives an impression of the liquid velocity : for $a_t = 10^{-3}$cm, eg, it is only about $10^{-2}$cm/sec at the smaller meniscus, and the velocity $dx_+/dt$ of the larger meniscus relative to a fixed frame is even less.

Most of the mass-evaporation, on the other hand, occurs at the larger meniscus (Fig 5) because its area is larger; by (3), it is

$$m = \pi\gamma\rho\kappa\varepsilon a_+ = \pi\gamma a_+ \lambda(T_w - T_M)/L \quad . \tag{10}$$

For water and $a_+ = 1$mm, eg, it would be about $10^{-7}$gr/sec. In turn, an impression of the dependence of the wall-temperature level $T_w$ upon the external heat supply begins to emerge, because the rate $Q_h$ of that supply per meniscus is mL. In a liquid column bounded by just two menisci of disparate size, the smaller expends relatively little of this, so that (10) shows the external heat supply to such a column to depend only on $T_w$ and on the capillary bore at the larger meniscus.

The critical importance of capillary size merits re-emphasis here. The pressure difference due to surface tension is proportional to $a_m^{-1}$, and so is the meniscus velocity relative to the liquid, but the pressure drop due to viscous shear is proportional to $a_t^{-4}$. Accordingly, if two porous samples be compared which are identical in regard to chemistry and to shape of the void-passages, but differ by a factor 10 in the size of those passages, then the dynamic balances for them differ by, essentially, a factor $10^3$ and therefore, the fluid physics in the two samples may be quite different. That contrasts strongly with constitutive theories of porous-media mechanics, of which invariance under mere change of geometric scale is a main premise. It appears doubtful, therefore, that a substantive description of fluid mechanics in porous media is obtainable without some insight into the fluid dynamics on the microscopic level, whence macroscopic behavior must spring.

VII. MACROSCOPIC IMPLICATIONS. For an impression of global evaporation in a porous medium, a thermodynamically steady phase may need to be distinguished from an initial, transient phase. The later phase is characterized by essential equality, at any time, of the rates of external heat supply and of latent-heat expenditure. The global mass-rate of evaporation is then immediately known and a lower bound $\tau_0$ of total evaporation time can be deduced as that which evaporation of the initial liquid mass would need under such conditions.

Whether such a late phase occurs at all, or at the other extreme, whether the transient phase is of no importance, must be judged from comparison of $\tau_0$ with the time scale of the transient phase. The latter may be one of three scales, of which the first, $\tau_1$, characterizes the rate at which the external reservoir can communicate heat to the porous aggregate and a second, $\tau_2$, characterizes the rate at which heat can be distributed through the solid matrix. Neither of these is within the scope of this account, but the time scale $\tau_3$ of liquid response to the heat supplied to it can be predicted on the basis of the present results, if adequate knowledge of the statistical distribution of void-passage sizes is at hand.

Indeed, if even a relatively small number of large passages thread the porous aggregate, then all the "action" will occur in them, and Haines Jumps may be there observable, while the rest of the medium remains essentially inert until the by-passes have cleared to an extent making them effectively a part of the outer boundary of the medium.

On the other hand, if enough passage throats of sufficiently small size are distributed sufficiently well through the aggregate, then they will anchor the liquid and permit only creeping motion. The time scale $\tau_3$ would then be expected to be essentially that of transition from static stability to dynamic balance, which (8) shows to be

$$\tau_3 = c_1 a_0^2/(\kappa\varepsilon) = (a_0/\varepsilon)^2/[2\gamma\kappa|a'(x_-)|].$$

Since it is seen to depend most of all on $a_0$ and $\varepsilon$, a useful, macroscopic estimate of this scale requires both a judicious choice of $\varepsilon$ between 0 and its level in the steady stage, and also a statistically valid measure $\pi a_t^2$ of the cross-sectional areas of small throats, whence the corresponding measure $a_0$ of capillary radii at the small menisci can be deduced by (9). For a very rough impression (which may well be misleading in regard to specific cases), if $\varepsilon = 1/50$, $a_0 = 10^{-2}$ cm and $|a'(x_-)| = 10$ were appropriate, then for water, $\tau_3 \sim 4$ min.

If $\tau_3$ should turn out to be much longer than $\tau_0$, $\tau_1$ and $\tau_2$, the microscopic dynamics of the main phase of evaporation would be that described in Section V. The stability analysis there given is not linearized, and its results therefore apply to the whole transition from static stability to dynamic balance. Of course, once the initial configuration of static stability has been left well behind, the much simpler approximation of Section VI becomes adequate. The translation here of the microscopic description into a predictive algorithm on the macroscopic scale

may be premature, however, because its usefulness is likely to depend critically on a more precise knowledge of the statistical distribution of void-passage sizes than appears available to-date for any real sample.

VIII. VAPOR TRANSPORT. For evaporation in dynamic balance, the time-dependence of the processes in the air-vapor mixture may also be expected to amount to no more than a slow drift leaving the processes quasi-steady. Transfer from the capillary wall will heat this gas, but if the pressure differences in it are insignificant because no small throats obstruct its passage, then the attendant density change will not be worth accounting for, at the temperature levels here envisaged. Accordingly, the volume flow rate $\pi a^2 u$ of gas will also be considered constant along the passage.

The mass evaporated consists of vapor, but the gas flow moves the air-vapor mixture and must therefore be accompanied by diffusion of vapor and air into each other. The gas at the meniscus must be at its dew point, so that the partial vapor pressure there is the saturation pressure at the meniscus temperature $T_M$. For water vapor, e.g., that partial pressure is about 1/40 (or 3/40) at $T_M \approx 293$ (or 313)° K, and the gas even at the meniscus then consists almost entirely of air. Since the partial vapor density is even smaller [4], the diffusion of the vapor is adequately approximated by the standard, linear model of Fick's law, $j_v = -\vartheta \nabla \rho_v$, for the vapor flux. With unsteadiness already neglected, the same mass-flow rate of vapor must cross every capillary cross-section, so that $\pi a^2 u \rho_v - \pi a^2 \vartheta d\rho_v/dx$ is independent of x. It follows that

$$\frac{\rho_v - \rho_{vm}}{\rho_{ve} - \rho_{vm}} = e^{(\xi - \xi_e)/q},$$

where subscripts m and e distinguish respective values at meniscus and capillary exit,

$$\xi = a_m^2 \int_{x_m}^{x} [a(s)]^{-2}\, ds,$$

and $q = \vartheta/u_m$ is a diffusion-length scale based on the gas velocity $u_m$ at the meniscus. Most of the diffusion therefore occurs within a $\xi$-distance q of the exit, by contrast to the heat transfer, most of which occurs fairly close to the meniscus.

The process can be radically different, however, if the gas must pass through a small throat. Let u denote again the cross-sectional average of the gas velocity and let subscripts $\ell$, g, v, a, m and t distinguish reference to the liquid, gas, vapor, air, meniscus and throat, respectively. Then from the estimate of meniscus velocity relative

to the liquid in Section IV, $a_m u_m = \gamma\varepsilon\mu_\ell\rho_\ell/\rho_v$, approximately, and since this is independent of meniscus size, so is the Reynolds number $Re_m = a_m u_m/\nu_a$ of the gas-flow at the meniscus. Since $\nu_a \approx 0.15$ cm$^2$/sec and $\rho_\ell/\rho_v \approx 10^3$ for water, Table 1 (Section IV) indicates $Re_m = 2$ to be a rather typical value. The mass-flow rate $m = \pi a^2 \rho_g u$ is independent of x in near-steady evaporation, and apart from the influence of density changes, the local Reynolds number $Re = au/\nu_a$ of the gas-flow varies in proportion to $a_m/a(x)$. For most plausible values of $a_m/a_t$, the gas-flow therefore remains laminar even in a throat, and the pressure drop can again be estimated from Poiseuille's formula,

$$a\, dp/dx = -8\mu_a m/(\pi a^3 \rho_g) = -8\mu_a a_m^2 u_m/a^3,$$

where $8\mu_a a_m u_m$ is independent of meniscus size and typically, $\approx 5 \times 10^{-7}$ gr when $\mu_a = 2 \times 10^{-7}$ gr sec/cm$^2$.

If now $a_m = 10^{-2}$ cm, to fix the ideas, then $a_m/a_t = 10$ yields a value of $5 \times 10^{-3}$ atm for $|a\, dp/dx|$ at the throat, and the pressure drop is insignificant. If $a_m/a_t = 100$, however, then the estimate suggests a value of 5 atm for $|a\, dp/dx|$ at the throat, and not only the estimate, but clearly also, most of the premises and assumptions of this Note, collapse. If the evaporation estimates remained valid when the gas must pass through very small throats, they would imply major gasdynamical effects in such throats, the work expended on them would play a major role in the thermodynamic balances, and the meniscus temperature $T_M$ could not be expected to be close to the initial, ambient temperature $T_0$; the physics of evaporation would be quite different from that here described. Accordingly, the microscopic physics of evaporation may depend rather drastically on whether a porous medium is formed into a sheet-like or ball-like aggregate...

The great sensitivity of quantitative estimates to void-passage size suggests that a profitable discussion of fluid mechanics in porous media may need to relate to quite specific circumstances. In particular, if thought returned to fabrics, it appears unlikely that the same microscopic physics could describe evaporation from wool, gore-tex or pile.

Acknowledgements. The author is indebted to Mr. A.K. Stuempfle and Drs. E.B. Dussan V. and B. Miller for helpful discussions. This work was sponsored by the United States Army under Contract DAAG29-80-C-0041 and partially supported by the National Science Foundation under Grant MCS-8215064.

## REFERENCES


[1] W.B. Haines, Studies in the physical properties of soils, Part V, J. Agri. Sci., Camb. 20, 1930, 97.

[2] J.P. Heller, The drying through the top surface of a vertical porous column, Soil Sci. Soc. Am. Proc. 32, 1968, 778.

[3] N.R. Morrow. Physics and thermodynamics of capillary, Ind. Engng. Ch[illegible]nd. (int.) Edn. 62, 1970, 32.

[4] J.H. Keenan, Thermodynamics, John Wiley, New York, 1941.

[5] R.E. Meyer, Note on evaporation in capillaries, IMA J. Appl. Math. 32, 1984, 235.

# JET-CONTAMINANT INTERACTION IN CONFINED GEOMETRIES

Lang-Mann Chang
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005

ABSTRACT. A numerical simulation is presented for investigation of the early phase of the flow interaction between a water jet and a chemical contaminant residing in cavities of a wall and in corners of two perpendicular walls. Such a interaction often occurs in surface decontamination processes. The flow model for this analysis is a two-dimensional, two-fluid flow governed by the unsteady Navier-Stokes equations. The equations were solved via finite difference schemes using the SOLA-VOF code. Computer plots of the flow development are presented. The results show that an inclined jet is more effective than a normal jet for decontaminating these confined geometries. In all flow cases studied, the impact pressure on the impingement wall far exceeds the corresponding steady-state dynamic pressure of the jet.

I. INTRODUCTION. Utilization of liquid jet spray is one of the most practical and most effective means for decontaminating Army vehicles in chemical warefare or for surface cleaning in the industry. The procedure is to use the force produced by the turning of jet stream at the impingement to displace the contaminant.

For a plane wall decontamination using a water jet spray, Chang [1] has characterized the interaction of the jet with the contaminant. In many areas there exist cavities under a surface or corners formed by two perpendicular walls, as depicted in Fig. 1. The chemical contaminant under consideration may reside in these confined geometries in the form of a drop or a layer of fluid covering the entire bottom surface of the geometries.

The flow interaction involves two fluids, the jet fluid (water) and the contaminant, and there is an interface in between, presenting a complex two-fluid problem. To simplify the analysis, we treated the interaction as a two-dimensional flow. The flow field is governed by the unsteady Navier-Stokes equations which were solved numerically via finite difference schemes using the SOLA-VOF computer code [2].

The emphases of this study are on the early evolution of the contaminant drop and the magnitude of the impact pressure on the bottom surfaces of the confined geometries. Computer plots of the flow process are presented. The effect of the angle of jet incidence on the flow is discussed. The results obtained provide useful data for the design of efficient jet sprays used for chemical decontamination.

II. FLOW MODELS AND GOVERNING EQUATIONS. Figs. 2 and 3 are the models of the pre-impingement configurations corresponding to



PREVIOUS PAGE IS BLANK

the schematics presented in Fig. 1. The shaded areas are the regions initially covered by the contaminant and the rest of the space in the geometries is filled with water. The dimensions of the contaminant drops are 3 mm by 0.6 mm, representing the average size of the drops deployed on vehicles. The dimensions of the cavity, however, are representive. A single water jet with a steady and uniform velocity is directed to impinge on the upper surface of the water at an angle of incidence, $\theta$. Two angles, $\theta=45^{\circ}$ and $\theta=90^{\circ}$, have been considered. The jet width $D_j$ is 1.83 mm with which a jet can perform decontamination effectively and efficiently [1]. In Figs. 2 and 3 there is a thin water layer covering the contaminant. This layer may exist practically and was found helpful in reducing the numerical instability problem encountered at the water-contaminant interface. Without the layer the stationary and highly viscous contaminant will be in direct contact with the high-speed jet fluid and, as a consequence, there is a great shear stress at the interface.

Fig. 4 shows the flow region and its necessary boundary conditions for the flow analysis in a cavity. An outflow condition is specified at the upper boundary so that the fluids can flow out the region after the start of the jet flow. The setup for the flow analysis in a corner is essentially the same except that an additional outflow condition is prescribed at the left wall.

The governing equations of the above flow models are the continuity equation

$$\frac{1}{\rho c^2}\frac{\partial p}{\partial t} + \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{1}$$

and the momentum equations

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + \nu\left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right] \tag{2}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial y} + \nu\left[\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}\right] \tag{3}$$

where t is time, u and v are the x-component and the y-component of the flow velocity, respectively. The density $\rho$, the sound speed c, and the kinematic viscosity $\nu$ are assumed to be constant. Based on the jet width and the jet velocities used in this study, the Reynolds numbers are between 20 and 2000. Within this range, Eqs. (2) and (3) are felt to be appropriate without considering turbulence effects.

For the tracking of the water-contaminant interface, we define a function F, called the fractional volume of fluid function, satisfying the relation

$$\frac{\partial F}{\partial t} + u\frac{\partial F}{\partial x} + v\frac{\partial F}{\partial y} = 0 \qquad (4)$$

The value of F in a computational mesh cell is equal to the fractional volume of the cell occupied by the contaminant. Then the value of F is one in cells full of the contaminant and zero in cells containing only water. Cells with F values between zero and one contain an interface, as illustrated in Fig. 5.

In order to adapt the SOLA-VOF code to the present problem involving two fluid with distinct viscosities, we use the following viscosity relation

$$\nu = \nu_c F + (1-F)\nu_w \qquad (5)$$

where $\nu$ is the average kinematic viscosity of the fluid mixture in a cell. $\nu_c$ and $\nu_w$ are the kinematic viscosities of the contaminant and water, respectively. Since the densities of the two fluids are different, the density of the fluid mixture in a cell is approximated as

$$\rho = \rho_c F + (1-F)\rho_w \qquad (6)$$

where $\rho_c$ and $\rho_w$ are the densities of the contaminant and water, respectively.

III. COMPUTATIONAL RESULTS. The jet velocities $V_j$ chosen for computations are 5 and 10 m/s (producing steady-state dynamic pressures of 12.4 kPa and 55 kPa, respectively) which are practical for chemical decontamination. The viscosity of the contaminant under consideration is $\nu_c=10\,\nu_w$ and the density is $\rho_c=1.07\,\rho_w$.

The computational results to be presented are in two parts: flow patterns and the impact pressure on the bottom walls of the confined geometries.

## Flow Patterns

Fig. 6 shows the flow generated by a water jet impinging on the upper surface of a cavity filled with water and with a contaminant drop initially located at the right corner, as seen in Fig. 2a. It is noted that the computer plots have been magnified three times in the vertical direction in order to provide a clear flow visualization. The plots in the left and the right columns correspond to the 45°- and 90°-impingement, respectively. The flow direction of the main stream in the

cavity is strongly influenced by the angle of jet incidence. In the 45°-impingement, the main stream moves toward the left wall and then turns and exits the cavity, while in the 90°-impingement the main stream exits the cavity adjacent to the entrance of the jet stream. As a result, the contaminant subjected to the 45°-impingement has experienced a larger displacement along the bottom wall than the contaminant subjected to the 90°-impingement. In addition, there is still a small amount of contaminant stagnating at the right corner at 0.3 ms in the 90° case. It is, therefore, obvious that a jet impinging at 45° has more cleaning power for decontaminating the cavity. Fig. 7, which was obtained by using the technique of embedding marker particles in the region initially covered by the contaminant, shows another view of the evolution of the contaminant drop.

In the case that a contaminant drop is initially located at the left corner of a cavity as depicted in Fig. 2b, Figs. 8 and 9 also indicate that the jet impinging at 45° is more effective for cleaning the cavity. There is an interesting flow phenomenon to be noticed in the 45° case. The upward flow velocity near the left wall is greater than the velocity slightly far away the wall. It is attributed to the flow impingement on that wall.

Fig. 10 shows the results corresponding to Fig. 2c in which an impingement takes place in the central part of the cavity. In the 90°-impingement, the main water stream induced by the jet flow does not reach the end walls. Therefore, there is little movement in the contaminant along the right end wall.

Fig. 11 displays the flow development corresponding to the configuration in Fig. 2d in which the cavity is full of contaminant with a thin water layer on the top. At times up to 0.27 ms, the flow patterns for the two angles of jet incidence are similar. However, in the 90° case the movement of the contaminant along the right end wall slows down as time progresses due to a faster development of viscous layer along the wall.

Next, we examine the results for the other kind of confined geometry investigated: the corners shown in Fig. 3. Figs. 12 and 13 present the flow patterns developed from Fig. 3a and Fig. 3b, respectively. The main stream is moving freely to the left because the left end is open to the flow. Though the flow patterns for the two angles of jet incidence are similar in general, the 45° jet exhibits slightly better performance since it cleans up the contaminant in the right corner faster.

## Pressure Distribution on the Bottom Wall

The impact pressure on the bottom surfaces of the subject confined geometries is another important datum to be determined. In some critial areas of a vehicle, such as optical windows, the pressure applied without causing damage is limited to a certain level. Figs. 14 through 17 present the results of ratios of the instantaneous impact pressure to the corresponding steady-state

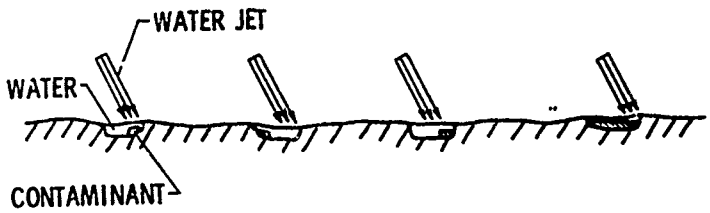
dynamic pressure which is $1/2(\rho_w V_j^2)$ for various flow configurations. Fig. 14 is the result for the configuration in Fig. 2a, showing that the pressure ratio can reach 13. In the configuration in Fig. 2a, the pressure ratio shown in Fig. 15 is even higher, approximately 19. In the corner flows shown in Figs. 3a and 3b, the corresponding pressure ratios displayed in Figs. 16 and 17, respectively, are relatively lower. It is a result of the change of boundary condition from a noflow to an outflow condition. We also notice that the $45^\circ$-impingement produces a slightly higher impact pressure in Figs. 14 and 15 than the $90^\circ$-impingement does.

IV. CONCLUSIONS. Computer plots have been generated to show the detailed flow development in the early phase of the jet-contaminant interactions in cavities and corners. Based on the flow patterns, a jet impinging at an appropriate inclined angle, say $45^\circ$, is more effective than a normal jet for decontaminating such confined geometries. The instantaneous pressure rise on the bottom surfaces of the confined geometries can far exceed the corresponding steady-state dynamic pressure of the jet. In general, the pressure rise is higher in cavities than in corners.

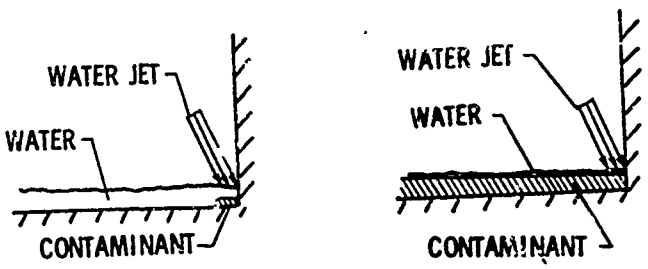
## REFERENCES


1. L. Chang,"Characterization of Jet-Contaminant Interaction Flow in Chemical Decontamination," U.S. Army Ballistic Research Laboratory Technical Report (in press).

2. B. Nicholas, C. Hurt, and R. Hotchkiss,"SOLA-VOF: A Solution Algorithm for Transient Fluid Flow with Multiple Free Boundaries," Los Alamos Scientific Laboratory Report No. LA-8355, 1980.




Contaminant in Cavities




Contaminant in Corners

Fig. 1 Contaminant in Confined Geometries

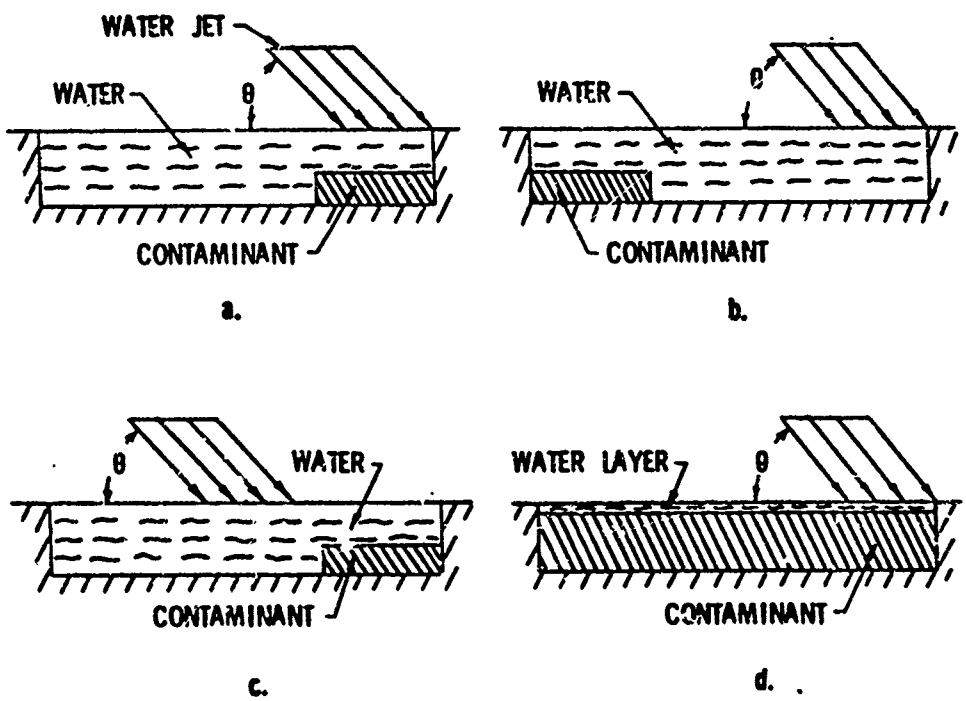
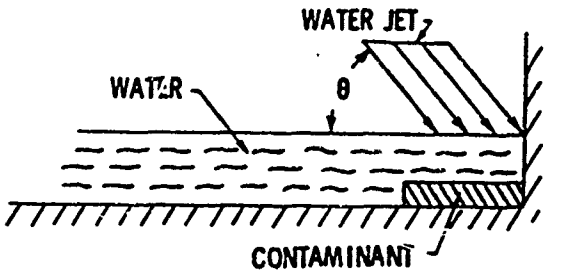
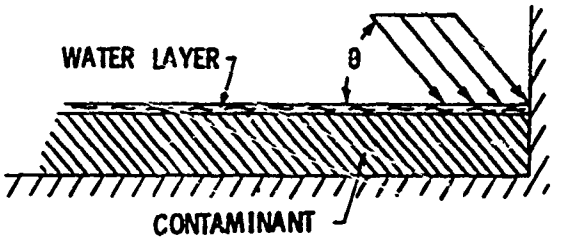



Fig. 2 Pre-impingement Flow Configurations in Cavities




a.




b.

Fig. 3 Pre-impingement Flow Configurations in Corners

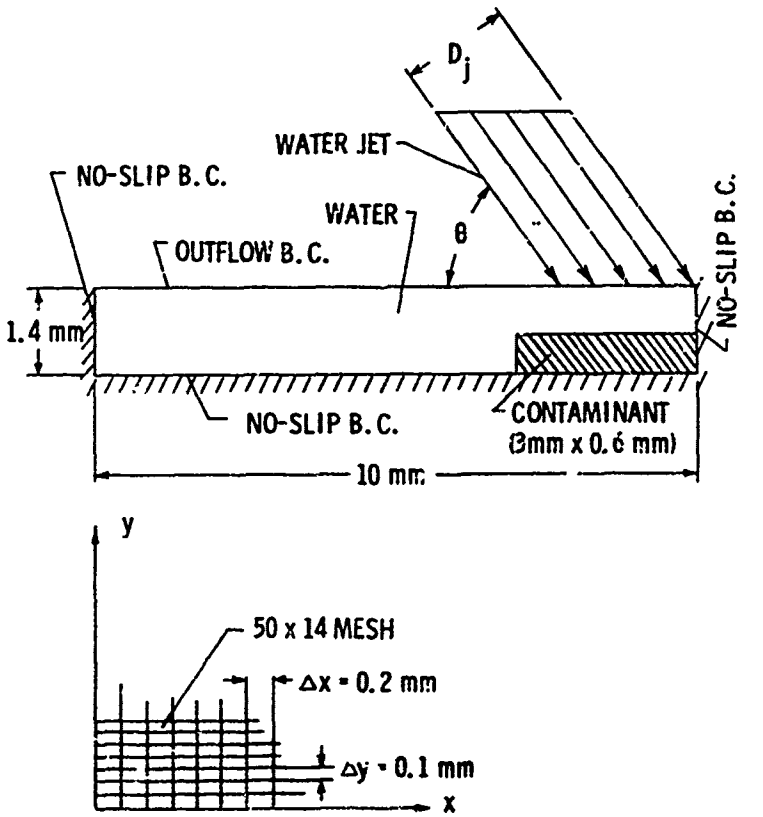



Fig. 4 Flow Region and Boundary Conditions for a Cavity

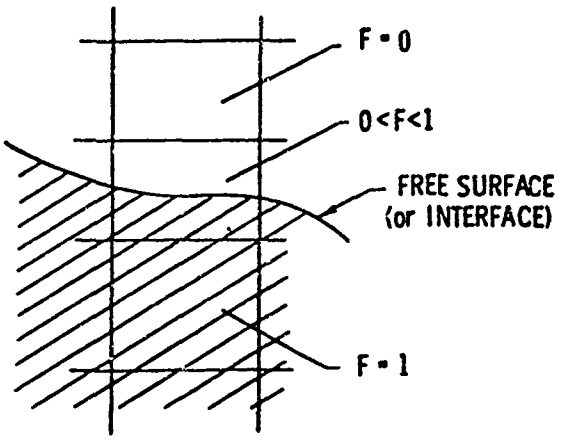



Fig. 5 Free Surface (or Interface) Across a Mesh Cell

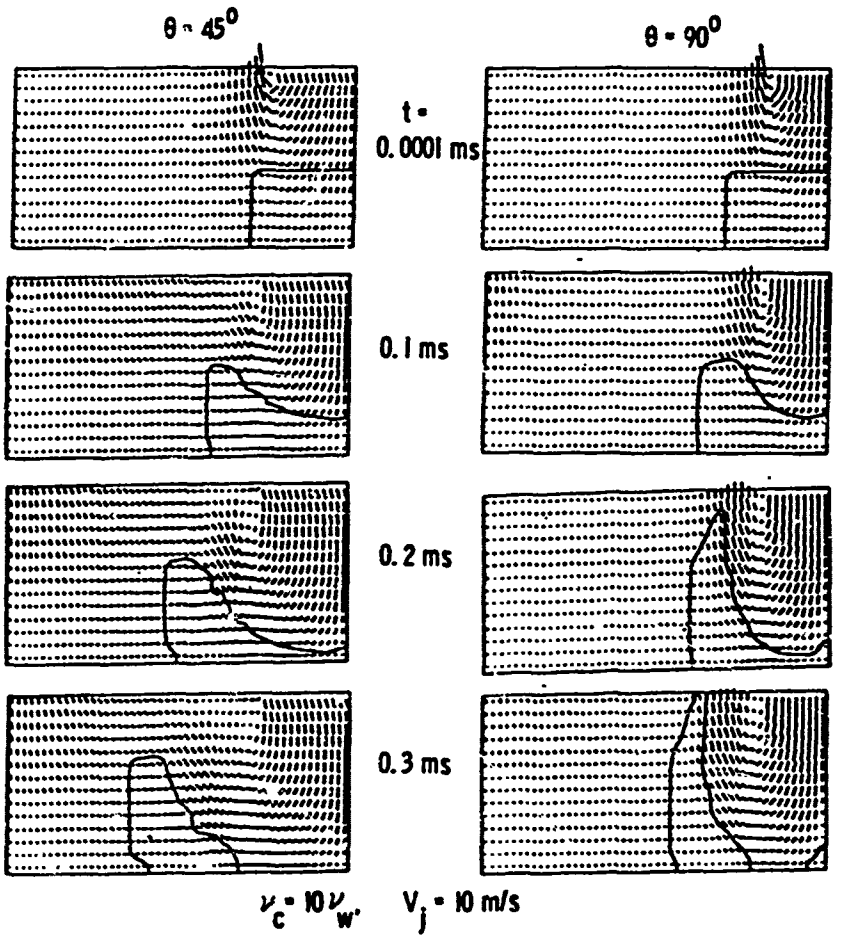



Fig. 6 Flow Development (corresponding to Fig. 2a)

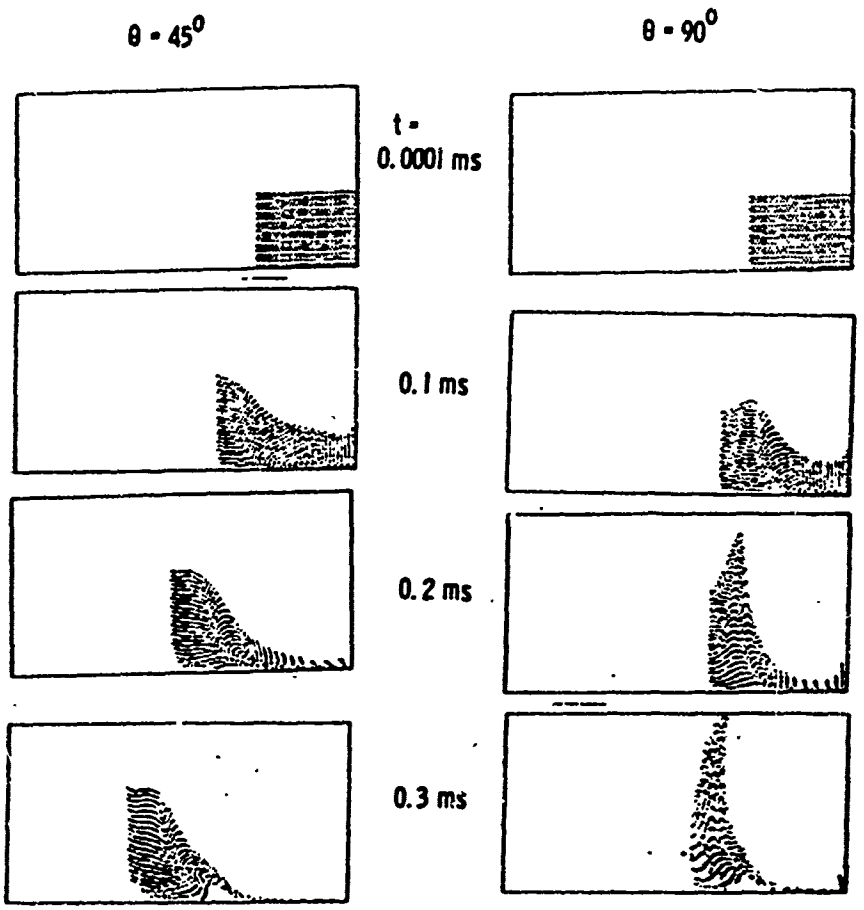



Fig. 7 Evolution of Drops (corresponding to Fig. 6)

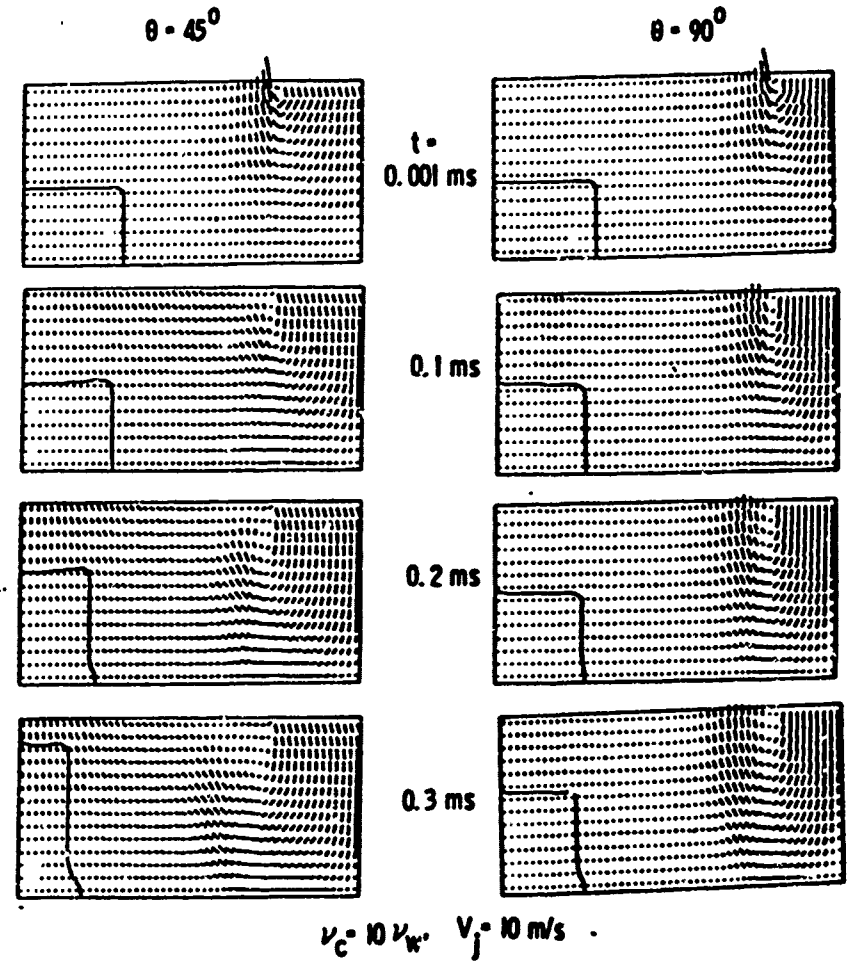



Fig. 8 Flow Development (corresponding to Fig. 2b)

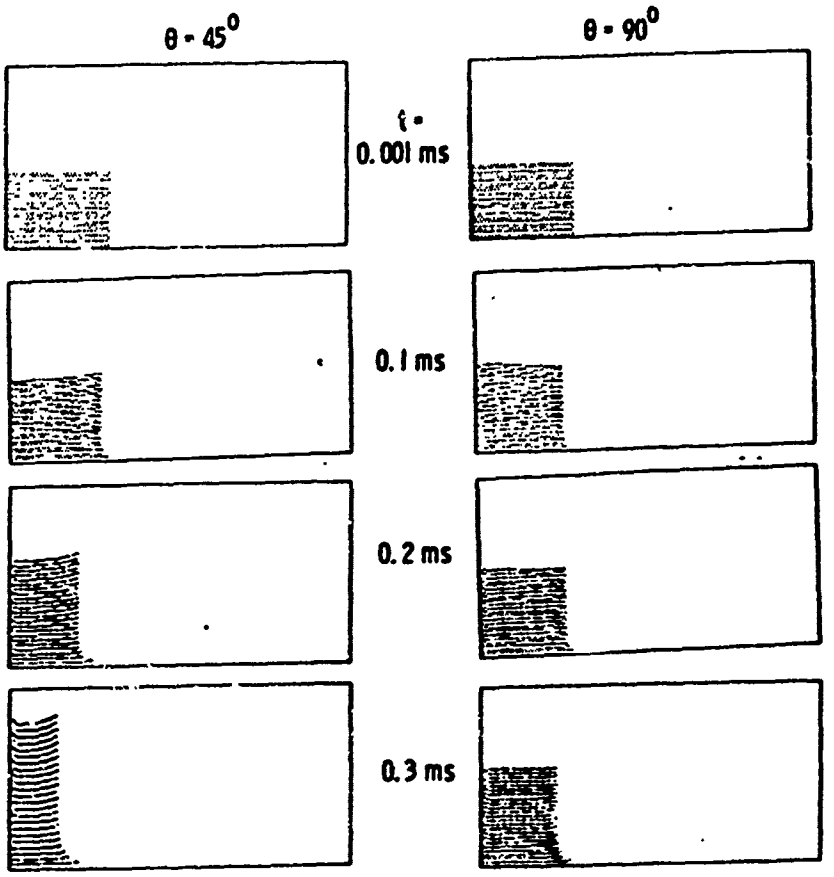



Fig. 9 Evolution of Drops (corresponding to Fig. 8)

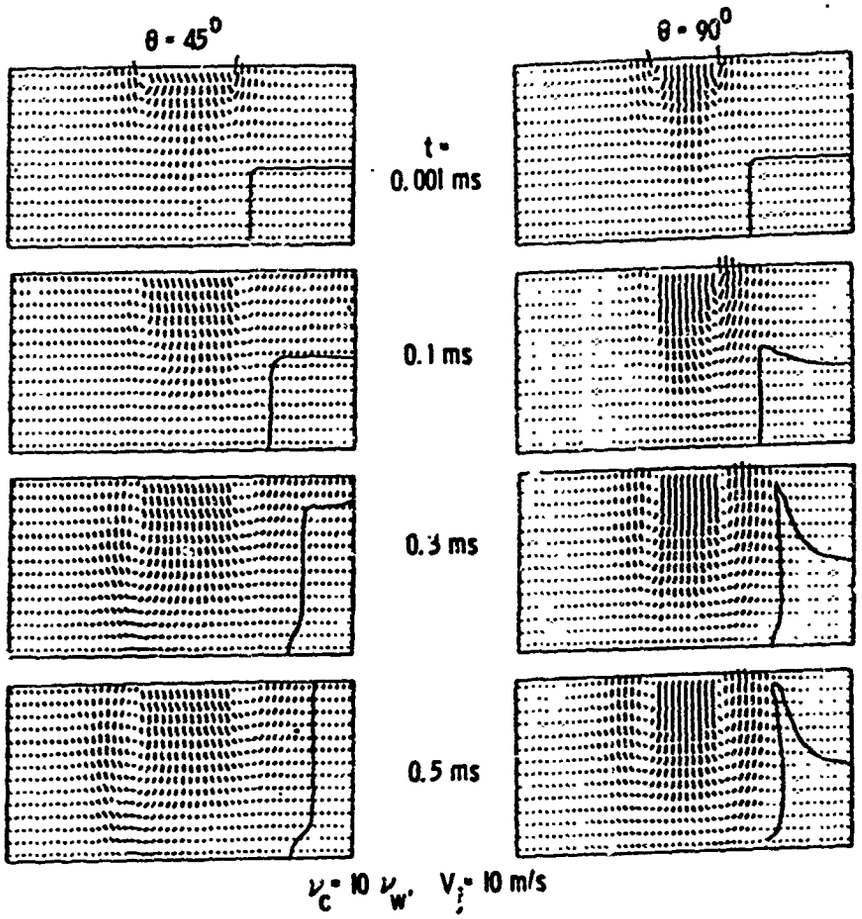



Fig. 10 Flow Development (corresponding to Fig. 2c)

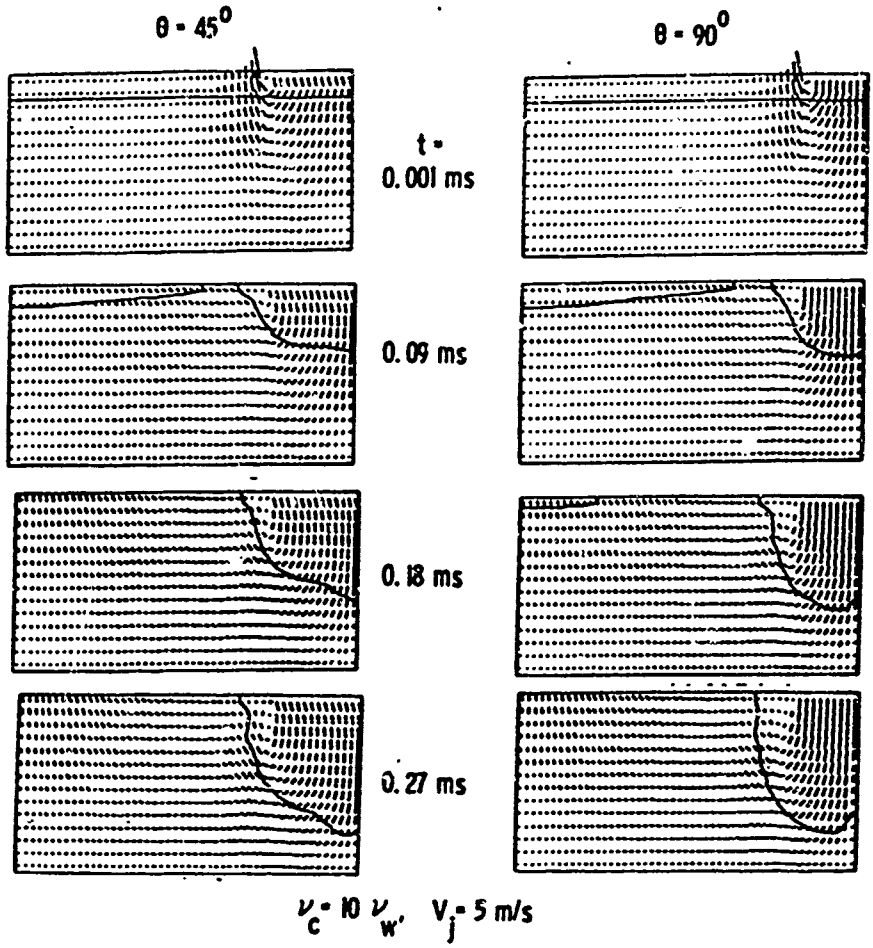



Fig. 11 Flow Development (corresponding to Fig. 2d)

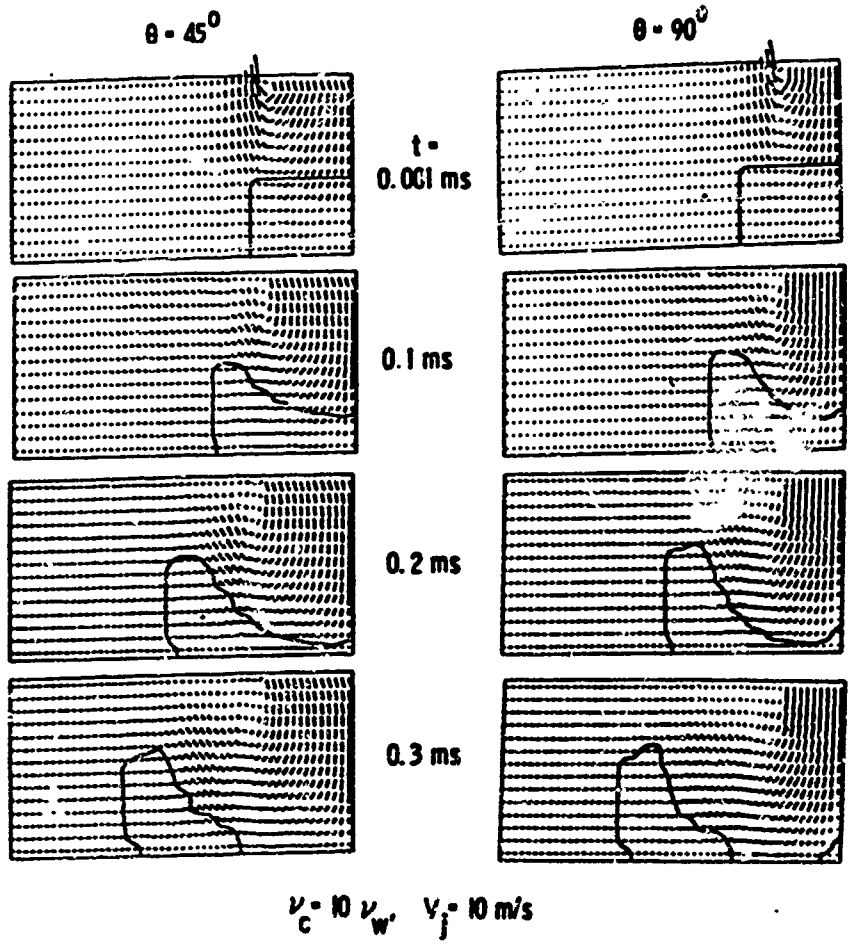



Fig. 12 Flow Development (corresponding to Fig. 3a)

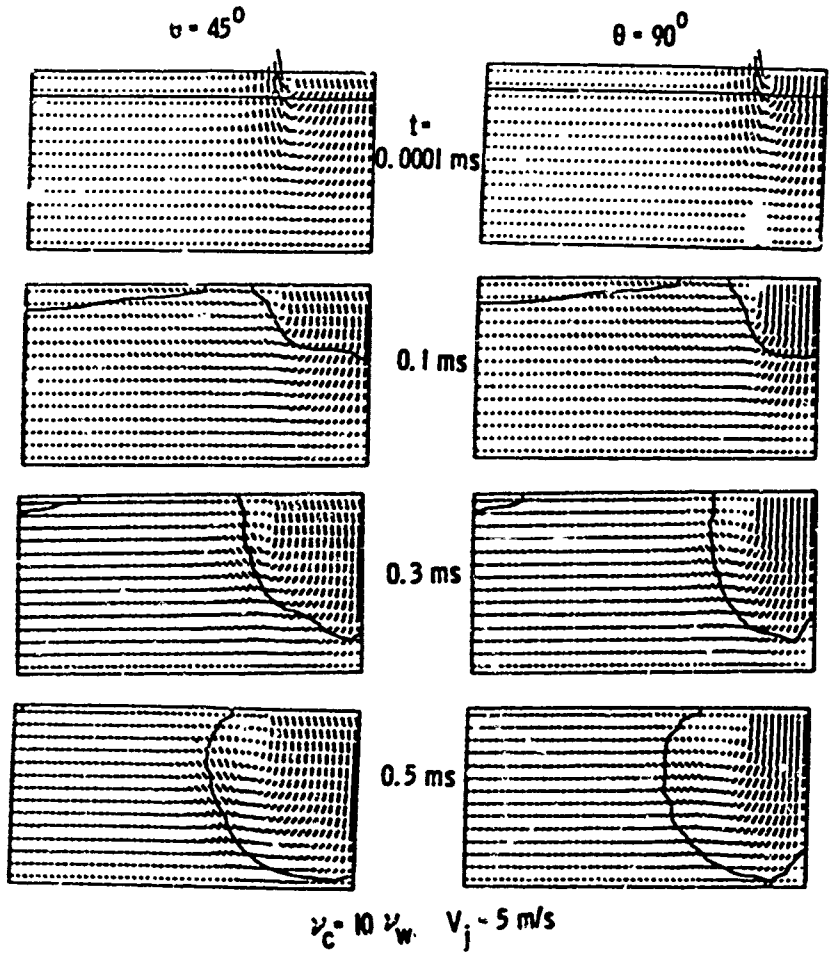



Fig. 13 Flow Development (corresponding to Fig. 3b)

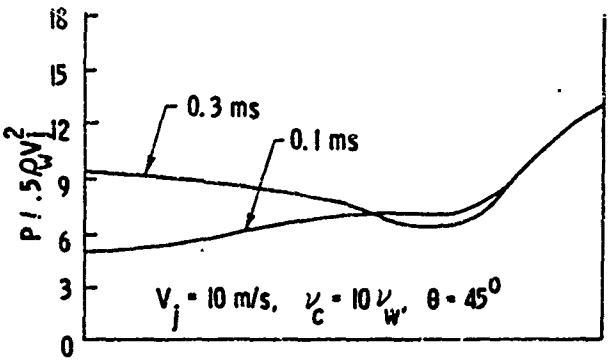



Fig. 14 Pressure Distribution on Bottom Wall Resulting from the Impingement of Fig. 2a

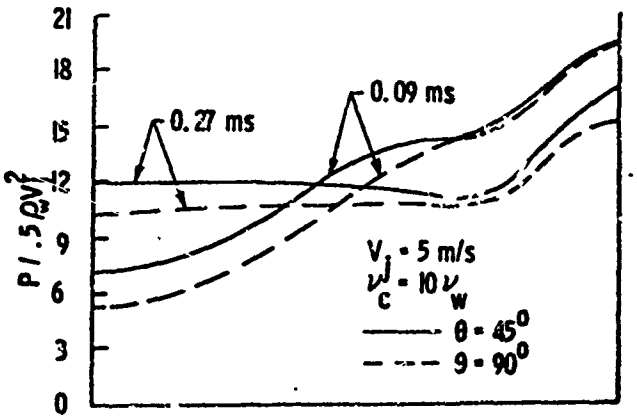



Fig. 15 Pressure Distribution on Bottom Wall Resulting from Impingement of Fig. 2d

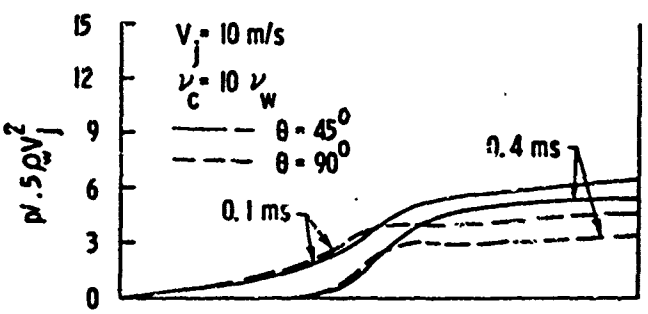



Fig. 16 Pressure Distribution on Bottom Wall Resulting from the Impingement of Fig. 3a

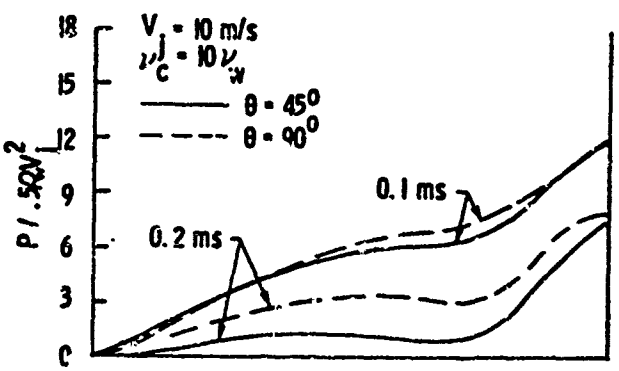



Fig. 17 Pressure Distribution on Bottom Wall Resulting from Impingement of Fig. 3b

# A METHODOLOGY FOR THE DEVELOPMENT OF FIRE CONTROL EQUATIONS FOR GUNS AND ROCKETS FIRED FROM AIRCRAFT

Harold J. Breaux

Computer Techniques & Analysis Branch
Systems Engineering & Concepts Analysis Division
United States Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005

## ABSTRACT

The accurate firing of unguided projectiles (bullets or rockets) from aircraft leads to a requirement for rapid computation of the launch vector needed to assure the projectiles striking a given target. The computation of this laying vector and fuze time is the function of the on-board fire control system. The fire control system includes sensors which measure target range and velocity, aircraft attitude, position and velocity, and atmospheric conditions. These measurements are fed to an on-board fire control computer which in real time, typically at 50 Hz, must compute anew the laying vector appropriate for the rapidly varying variables which influence the ballistic trajectory. Six-degree-of-freedom models, which are normally used in laboratory ballistic modeling and simulation, are computationally too slow and otherwise cumbersome to be implemented for real time fire control. A methodology for developing an alternative--simplified, yet very accurate, model is described in detail.

## 1. INTRODUCTION

The fundamental objective of fire control ballistics is the provision of a means for aiming and fuze setting of a rocket or projectile so as to best assure placement of accurate and effective fire on a selected target. This objective was accomplished historically by the publication, and use in the field, of the classical "firing table", a book of numbers tabulated so as to provide an easy means for determining azimuth and elevation settings for guns and rockets. As computer technology has evolved, the firing table has been relegated to manual backup and the field computer computes gun and rocket settings in real time. Modern weapon systems such as computer controlled air defense guns, and helicopters and tanks which fire on the move encounter a fire control problem characterized by very rapidly changing variables which drive the ballistics. As a result, a wholly new level of difficulty and sophistication is placed on the problem of fire control ballistic prediction.

The fundamental "generic" models used extensively for prediction of launch and exterior ballistic performance have come to be known as Six-Degree-of-Freedom (SDF) models. Such models embody the equations of motion for both translational and rotational displacement of a projectile or rocket. The development, refinement, maintenance and modifications for new technology, of this type model, has long been performed in Defense Department Laboratories, and in particular, Army laboratories such as BRL. See, e.g., *Lieske* and *McCoy* [1] and *Barnett* [2]. The SDF model is a natural extension of the three-degree-of-freedom (TDF) model, long a mainstay of ballistic fire control prediction. This latter model represents only the translational aspects of projectile motion and is of an order of magnitude less difficult in computational labor. However, TDF models may, in some circumstances, be insufficiently accurate because they do not model the interaction of translational motion with the aerodynamic effects associated with yaw and pitch along with the yaw and pitch interaction with spin.

[1] R. F. Lieske and R. L. Mc Coy, " Equations of Motion of a Rigid Projectile," BRL Report No. 1244, March 1964.

[2] D. Barnett, "Trajectory Equations for a Six-Degree-of-Freedom-Missile Using a Fixed-Plane Coordinate System," Technical Report 3391, Picatinny Arsenal, Dover, New Jersey, June 1966.

Artillery projectiles, like the classical spinning top in mechanics, have very high frequency motion associated with spin, precession and nutation. A fundamental computational requirement of numerical integration, (the technique used for solving these models) is that oscillatory variables need to be sampled at several points within each oscillation to maintain accuracy and stability. The result of this is that small steps in time are required in the forward marching process of integration. Accordingly, complete SDF calculations for spinning artillery or automatic cannon projectiles is time consuming, relatively expensive, and is done sparingly on laboratory computers. An alternative to the SDF and TDF models was developed by *Lieske* and *Reiter*[3] which has been found extremely useful for firing table computation and implementation in some ground based fire control computers. Their successful approach was to develop a modified TDF model, MTDF, which incorporated an explicit estimate of the "yaw of repose" and its effects into the translation equations without having to integrate the high frequency motion of the full SDF models. This model has resulted in greatly reducing the computational burden referred to above while incurring only a slight loss in accuracy. It is now a standard model in the repertoire of ballistic mathematical tools widely used in the United States and the NATO defense community.

Helicopter fire control requires the use of ballistic models as described above. The SDF model is needed for rockets and the MTDF (and SDF) model are needed for automatic cannon. Unfortunately, for reasons detailed herein, these models, can only serve an intermediate role, albeit an important one, in the process of developing an on board--real time ballistic prediction model. A fundamental problem in attack aircraft fire control is that of maintaining the "timeliness" of the ballistic solution. In a turning, climbing maneuver, the aircraft velocity components, the geometrical relations between target and cannon/rocket and other variables-- all change rapidly. Note also that these models really solve the "inverse" of the fire control problem. In fire control one specifies terminal conditions such as target location with respect to the launch platform. The solution desired consists of the departure attitude needed to strike the target and the time of flight. Accordingly, one would need to solve the problem iteratively, i.e., guess at a trial solution, and continually readjust the departure angles until the desired terminal conditions are satisfied. Meanwhile, if the aircraft is in a maneuver, the problem has changed because the variables driving the ballistics have changed since the iterations were initiated. The consequence of this is that a ballistic solution that is not computed instantaneously, (or nearly so), is old and obsolete before the munition can be fired. Despite the recent and continuing revolution in computer technology, the embedded computers in aircraft fire control systems are small in memory capacity, and are not fast enough to iteratively solve the models described above at the required frequency.

Modern attack helicopters such as the Cobra and Apache are armed with an automatic cannon and a family of 2.75 inch or Hydra 70 rockets. By the flip of a switch, the pilot can select the munition he wishes to fire. Accordingly, the model(s) embedded in the on board fire control computer must be able to key on this switching process and compute the solution for the munition selected. In fact, it is possible, and sometimes desirable, to be able to fire both the gun and rockets simultaneously. The requirement for representing ten or more rocket types, each having differing weights, measures, aerodynamics, staging and fuzing characteristics-- all add to the need for developing a common general model. Furthermore, Army aircraft such as the Cobra and Apache have differing modes of articulating rocket pods It is highly desirable that the rocket ballistics be developed independently of the method of pod articulation. See Appendix F. The logical strategy evolves for developing a procedure which makes the general model applicable to a specific munition by selective retrieval of a pool of constants (based on switch position). Some constants of the model, however, may differ from one aircraft type to another. See Appendix E. This necessity for providing a capability for many munitions tends to reduce the storage capacity available for the collection of instructions related to the model itself. The requirement for an accurate ballistic solution, for such a family of munitions, varying types of aircraft, air defense guns, tanks, or other moving gun platforms, that can be computed cyclically, in real time, during an engagement, leads to the need for the methodology described herein.

[3] R. F. Lieske and M. L. Reiter, "Equations of Motion for a Modified Point Mass," Ballistic Research Laboratories Report No. 1314, March 1965.

## 2. GENERAL DESCRIPTION OF THE MATHEMATICAL METHODOLOGY

A perspective for defining the mathematical problem can be obtained by examination of Figure (2-1). Therein an attack helicopter in flight is depicted engaging a ground target. While the scenario

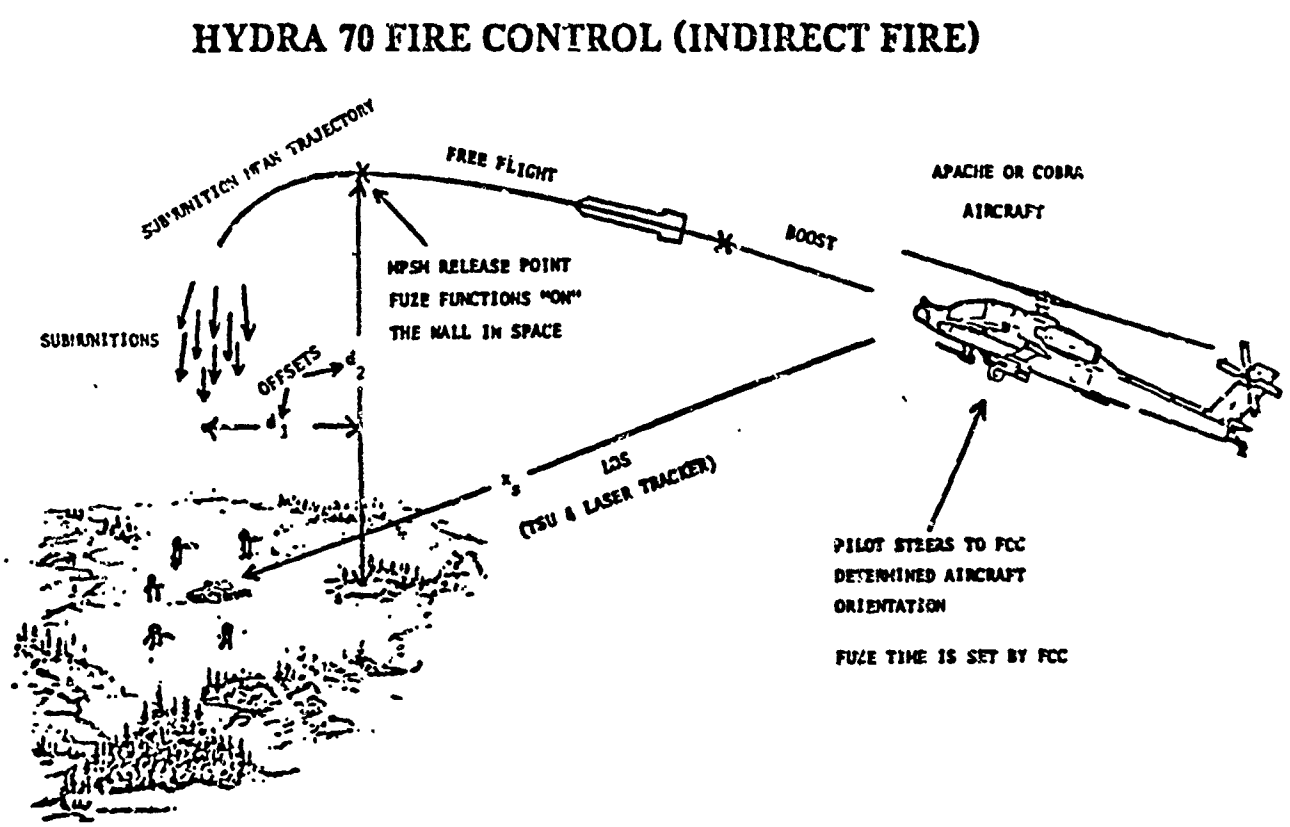



Figure (2-1). Typical attack helicopter scenario for rocket firing.

depicts the use of a Hydra-70 rocket, the employment of an automatic cannon would be similar. The telescopic sight unit, (TSU), is continually maintaining the line of sight, (LOS), which can be viewed as a vector connecting a reference point on the aircraft to a selected point on the target. On board sensors continually measure the vector velocity and attitude orientation of the aircraft along with the LOS range, $x_s$ , between target and aircraft. For moving target, sensors coupled with mathematical "filters" continually estimate the vector velocity of the target. Target motion, along with other considerations, determine components $x_m$ and $x_l$ which along with $x_s$ define where the projectile should be in "one time of flight". Environmental factors such as wind, air density and temperature are also provided by sensors on the aircraft. The temperature of the munition is also known, either by use of a magazine thermometer or by assuming that the munition has the same temperature as the environment. Required information on downwash due to the rotation of the helicopter blades is obtained by means described in Appendix D. Components of the gravity vector are made available by reference to orientation of coordinate axes that are aligned with the local gravity vector.

The above described information is fed, at typically 50 Hz, to the Fire Control Computer, (FCC). The primary function of the FCC is to compute the angular settings (the attitude of departure) and for

some munitions, fuze time, which best assures accurate placement of fire on the target. The objective is thus seen to be the development of a collection of formulae, to implement in the FCC, which accomplishes this task. Toward this end it is useful to review two procedures, used previously, that can serve as building blocks toward development of a more generalized procedure.

## Global Fitting Approach

A perspective on the global fitting *approach*[3] can be obtained from the work of Chandler, Baker and Dinjar at the US Army Redstone Arsenal and C. Masaitis and H. Breaux at the Ballistic Research Laboratory. That work is reviewed (and references listed) by *Breaux*[4]. The objective of that work was to find an alternative to the SDF models for use in fire control with ground based missiles such as the Redstone, Jupiter, Pershing and Lance. For these stationary-at-launch, ground systems, certain complexities unique to the moving platform are not present and the computational speed factor is not as critical. A computational cycle time of seconds, or even tens of seconds, is tolerable. However, at the longer ranges of these systems, other complexities enter such as those associated with earth rotation and curvature. Nevertheless, a basic procedure employed in that work remains as the cornerstone of current methodology. That concept is depicted in Figure (2-2) below and consists of approximating one model

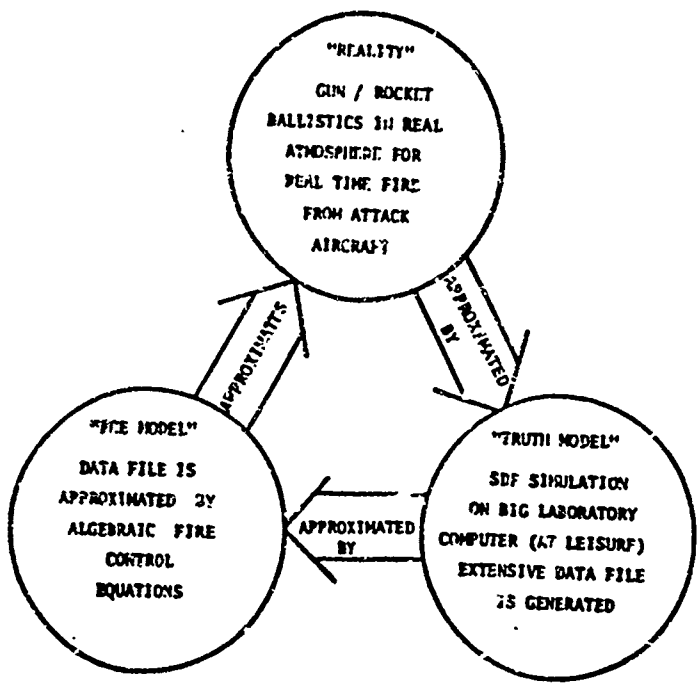



Figure (2-2). Triangle of Approximations for Development of Fire Control Ballistics.

[3] Reference to unpublished work in BRL related to attack helicopters is cited in the Acknowledgment.

[4] H. Breaux, "The Computation of Firing Tables for Guided Missiles," BRL Report No. 1348, November 1966.

with another model. First a "truth" model must be developed--both as a means for assessing actual ballistic performance and then to make possible the development and accuracy assessment related to the fire control model. This model must represent all the munitions of interest-- fired with launch conditions representing the aircraft environment. Speed of computation is not a significant factor here and this model is normally implemented in a laboratory computer.

The global fitting approach consists of basically five phases.

(A) Development of the SDF Model.
(B) Specification of a candidate fire control equation (FCE) model to be fitted.
(C) Development of a data base of trajectory calculations which span the expected range of all variables (and mixtures) in the planned deployment scenarios.
(D) Fitting of the model to obtain values for all coefficients.
(E) Computer validation of the model.

Phase (A) generally consists of obtaining all weights and measures, physical characteristics, and aerodynamic functions that are appropriate to the munition(s) of interest and processing the data into a form acceptable to the SDF program. Special features of the munition ballistics or launch dynamics may require modification of the SDF program. This is a very critical phase in that everything done hereafter hinges on the adequacy and accuracy of this model. Phase (B) is the most difficult aspect of the problem and is the core of the methodology to which a large part of this paper is devoted. Phase (C) consists of designing a matrix of conditions and use of the SDF program to compute the corresponding trajectories. The fitting process, Phase (D), is itself a part of the methodology. However, the procedure designed by *Breaux* [4], will be used herein to form a hybridized methodology which will be described. Phase (E) is designed to test the execution of the model in a form similar to it's field employment and simultaneously compare its predicted results with the SDF "truth model".

Phase (E) should not be confused with field validation. In field validation the results of the effort, namely the ballistic FCE's, are programmed into the Fire Control Computer and actual live tests conducted. Poor performance in this phase can frequently occur due to improper programming of the FCE's, poor sensor performance, or a poorly designed or specified SDF model. If the latter is found to be true, the SDF model must be fixed and the process described above must be *repeated*[5].

Closed Form Solution Approach

A solution approach has been employed by *Norwood* [6], for airborne gunfire, that makes use of the "Method of Siacci". In this approach, the computational problem is reduced to a collection of integrals. Norwood's approach was made more practical by *Benokratis* [6], who obtained closed form expressions for the Siacci integrals by approximating the projectile drag by three connected line segments. This approach has the benefit of "permanence" in that the approximate solution is there for all time-for all projectiles that satisfy the flat fire assumption and which can be adequately approximated by a TDF model. The drag curve must also permit adequate approximation by the three line segments. The disadvantage is that this technique does not incorporate ballistic effects that arise from factors that only the SDF model

[5] Now that the methodology and related software has been developed, reprocessing for a revised or new SDF data base can be done routinely and very quickly.

[6] Norwood, John M., "A Review of the Method of Siacci Trajectory Computation for Airborne Gunfire," ARL-TM-71-27, Applied Research Laboratories, The University of Texas at Austin, Austin, Texas, August 1971.

[6] Benokraitis, Vitalius, "A Closed Form Siacci Function Method for Trajectory Calculations," Journal of Ballistics, V. 6, No.1, 1982, pp. 1326-1347.

could predict. Ballistic effects due to aerodynamic "drift" and "jump" and secondary drag effects related to initial yaw and yaw rate must be added to the model by extraneous methods which require essentially the same effort as the global fitting approach. This method can represent the free flight phase of a rocket but not the boost phase. Extension of such a model to include more complex ballistics generally leads to the need for some form of fitting and in turn necessitates all the stages of the global fitting approach.

The methodology developed herein can be viewed as a hybrid method which combines features of both the global fitting technique and the closed form approximation. It can also be viewed as a generalization and extension of the earlier work done by the *author*[4]. The key aspect of the work described therein was the use of step-wise regression (least squares) to facilitate the process of model building. This process can be be exhibited by simple example. Assume that a process, whose outcome is represented by w, is dependent on the variables x,y,z. Assume that a very accurate model exists that permits the specification of x,y and z and then computes w. However, the computation is so expensive and time consuming that it can only be done sparingly and at facilities remote from where the need exists. This suggests the need and possibility of developing an approximating model which is easier to evaluate. In attempting such an effort consider the following two approaches to global fitting.

The Empirical Approach to Global Fitting

Here no specific information on the underlying mathematical model is assumed known other than the dependence of w on x, y and z. A candidate linear model is specified by the formula

$$w = a_0 + a_1 x + a_2 y + a_3 z + a_4 x y + a_5 x z + a_6 y z$$

$$+ a_7 x y^2 + a_8 x z^2 + a_9 z^2 y + \cdots \qquad (2\text{-}1)$$

Numerous types of linear models could be employed, however, a polynomial model is assumed for illustration. By providing a data base of w versus x,y and z for an appropriate mix and range of the variables and the above model to a stepwise regression procedure as described by *Breaux*[7,8], or *Hocking*[9] one might arrive at a suitable model. The result of such a regression procedure is the sequential listing of submodels, each differing from the previous submodel, by one term taken from the above collection of terms. The first submodel is the one term alone which best approximates w. The second model contains two terms, including the one term model plus a second term which when added to the first provides the best approximation. The process continues in this fashion but simultaneously seeks to weed out terms that are no longer useful due to cross correlations between clusters of variables that have been introduced into the model. Variations in strategy and computational sequencing are discussed at length by *Hocking*[9]. By examining the program output which includes the progression of the variance of residuals in w, the correlation coefficient, the "t" values on the regression coefficients, etc., and performing some experimentation on model definition, one can generally build a suitable model by this process.

The Physical Approach to Global Fitting

In the physical approach one first recognizes that the problem at hand embodies a field of knowledge to include a collection of literature. Such a literature indicates that the process leading to the outcome w has a "closed form" result, dependent on x,y, and z, if certain idealizing assumptions are made. One such outcome might be

[7] Breaux, H. J., "On Stepwise Multiple Linear Regression," BRL Report No. 1369, August 1967.
[8] Breaux, H. J., "A Modification of Efroymson's Technique for Stepwise Regression Analysis," Comm. of the ACM, 1968, Vol 11, No. 8.
[9] Hocking, R. R., "The Analysis and Selection of Variables in Linear Regression," Biometrics, V. 32, March 1976, pp 1-49.

$$w = A_0 z + B_0 \exp(\alpha z y/z) . \qquad (2\text{-}2)$$

This might suggest that a more intelligently defined model would proceed from Eq. (2-2) as a base. For this simple model one might use non-linear least squares and try to fit $A_0$, $B_0$ and $\alpha$ as free parameters. When the model is complex, containing many variables and terms, it is generally better to linearize the model and proceed as follows: In Eq. (2-2), for example, expand non linear terms in a Taylor series

$$w = a_0 + a_1 z + a_2 z y/z + a_3 (z y/z)^2 + a_4 (z y/z)^3 + \cdots \qquad (2\text{-}3)$$

As before, this candidate model and the data base is processed by the stepwise regression process. The resulting model, developed by the latter procedure, will, in general, be superior to one developed by the empirical process. Any insight, born of experience and knowledge of the process, helps toward better defining the candidate model, used as a point of departure, when thereafter employing a fitting process.

The fire control problem addressed herein is analogous to the above simple example with one very important difference. The example concerned itself with one process and one approximating model was needed. Hence, either of the two approaches, if found to lead to an adequate model, could be viewed as successful. The attack helicopter fire control problem, by contrast, can be viewed as representing ten or more related processes. If the problem is addressed by the empirical approach described above, intuition and experience indicates that the result will be as many separate models as their are rocket/munition *types* *.

The approach taken will be one of examining all components of the launch and flight process, the boost and free flight coupling, and practical idealizations which provide a physical basis for simplified closed form solutions leading to an intelligent definition of the candidate model. These components, will be processed by the stepwise regression procedure by use of a data base of calculations from an SDF program to determine their adequacy. The approach is thus seen to be a hybridized one which includes obtaining closed form approximations and then employs the features of the global fitting approach as described above.

Note: Due to its length, only Sections 1. and 2. of this paper have been included in the Proceedings. Information on the availability of the complete document can be obtained by writing to:

Director
USA Ballistic Research Laboratory
ATTN: DRXBR-TSD
Aberdeen Proving Ground, Md. 21005

* This aspect was found to be critically important in the currently on-going Fire Control Systems Integration Program for Cobra. The magnitude of effort, and resulting time required, to provide for as many as nine different rocket types, forced an early "freeze" in the basic structure of the model long before work was completed. The basic model was then programmed, validated for specific rocket types, and then coefficients for the remaining members of the family were added as they were obtained. This permitted timely progress on the overall program.

# SINGULAR VALUE DECOMPOSITION FOR SOLUTION OF* DIFFERENTIAL-ALGEBRAIC EQUATIONS OF MECHANICAL SYSTEM DYNAMICS

Neel K. Mani
Edward J. Haug
Center for Computer Aided Design
College of Engineeri.
University of Iowa
Iowa City, Iowa 52242

ABSTRACT. A computer-based method for solution of non-linear, constrained differential-algebraic equations of motion of mechanical systems is developed. The differential equations of motion and non-linear holonomic constraint equations are written in terms of a maximal set of Cartesian generalized coordinates, to facilitate the formulation of constraints and forcing functions. Singular Value Decomposition of the constraint Jacobian matrix is used to generate a coordinate transformation that defines a new set of generalized coordinates that are naturally partitioned into independent and dependent sets, with several desirable properties. This information is used to construct a reduced system of independent differential equations of motion that can be integrated using standard numerical integration algorithms. It is also shown that the method speeds the iterative solution of dependent generalized coordinates from constraint equations. A physically reasonable method is presented to determine when the choice of independent generalized coordinates needs to be changed. A tracked vehicle example is presented to illustrate the method and its advantages over other methods of solution.

1. INTRODUCTION. The availability of dynamic analysis codes for high speed digital computers has greatly increased the dimension of systems that can now be analyzed. The most convenient method of defining the kinematics of large scale systems is in terms of a maximal set of Cartesian generalized coordinates, which must satisfy kinematic constraints. Such a coordinate system may, however, not be best suited for solution of the equations of motion. This paper presents a singular value decomposition algorithm that transforms the Cartesian generalized coordinates into a system of coordinates that is well suited for solution of the equations of motion. Such an approach permits easy user representation of complex mechanical systems, without loss of accuracy during solution of the system equations of motion.

To develop and illustrate the theory, planar dynamic systems are considered. The method is equally applicable to spatial dynamics. A typical rigid body (denoted body i) in the plane is shown in Fig. 1, with a centroidal body-fixed coordinate system represented by $\xi_i$ and $\eta_i$ axes. The X and Y axes are an inertial Cartesian coordinate system. The position and orientation of the body in the X-Y plane are

* Research supported by US Army Research Office
Grant No. DAAE07-82-G-4004.

PREVIOUS PAGE IS BLANK

specified by giving the coordinates $x_i$ and $y_i$ of the body centroid and the angle $\phi_i$ that the $\xi_i$ axis makes with the global X axis. Thus, for body i, a generalized coordinate vector $q^i$ is defined as $q^i = [x_i \; y_i \; \phi_i]^T$. The corresponding generalized force on body i is $Q^i = [Q_x^i \; Q_y^i \; Q_\phi^i]^T$, where $Q_x^i$ and $Q_y^i$ are the X- and Y- components of force acting at the origin of the $\xi_i - \eta_i$ coordinate system and $Q_\phi^i$ is the torque acting on the body. The complete vector of generalized coordinates for the system is given by $q = [q^{1^T} \; q^{2^T} \; \ldots \; q^{NB^T}]^T$, $q \in R^n$, and $Q = [Q^{1^T} \; Q^{2^T} \; \ldots \; Q^{NB^T}]^T$, where NB is the number of bodies in the system and $n = 3NB$.

Kinematic constraints between bodies in the system can be written [1] as

$$\Phi \equiv [\phi^{1^T}(t,q) \; \phi^{2^T}(t,q) \; \ldots \; \phi^{NF^T}(t,q)]^T = 0 \tag{1}$$

where there are m constraint equations. To ensure that the constraints are independent, it is required that the mxn Jacobian matrix $\Phi_q = [\partial\Phi_i/\partial q_j]$ be of full row rank.

The kinetic energy of the system can be written as

$$T = \frac{1}{2}\dot{q}^T M \dot{q} \tag{2}$$

where M is the mass matrix of the system. The mass matrix of body i can be written as $M^i = \text{Diag}\,[m_i \; m_i \; J_i]$, where $m_i$ and $J_i$ are the mass and moment of inertia of body i. The system mass matrix can then be written as

$$M = \text{Diag}[M^1 \; M^2 \; \ldots \; M^{NB}] \tag{3}$$

Lagrange's equations of motion for a constrained system with workless constraints are [2,3]

$$M\ddot{q} + \Phi_q^T \lambda = Q \tag{4}$$

where $\lambda$ is a vector of Lagrange multipliers. The initial conditions at time $t = t_0$ are specified as

$$\left.\begin{aligned} q_{(t_0)} &= q^0 \\ \dot{q}_{(t_0)} &= \dot{q}^0 \end{aligned}\right\} \tag{5}$$

where the initial position $q^0$ and velocity $\dot{q}^0$ must be consistent with system constraints.

Differentiating the constraint equations of Eq. 1 with respect to time yields the velocity equation

$$\Phi_q \dot{q} + \Phi_t = 0 \tag{6}$$

Differentiating Eq. 6 with respect to time yields the acceleration equation

$$\Phi_q \ddot{q} = -2(\Phi_{tq})\dot{q} - \Phi_{tt} - (\Phi_q \dot{q})_q \dot{q} \tag{7}$$

Equation 4 is combined with Eq. 7 to give a system of matrix equations for accelerations and Lagrange multipliers that can be written as

$$\begin{bmatrix} M & \Phi_q^T \\ \Phi_q & 0 \end{bmatrix} \begin{bmatrix} \ddot{q} \\ \lambda \end{bmatrix} = \begin{bmatrix} Q \\ -2(\Phi_{tq})\dot{q} - \Phi_{tt} - (\Phi_q \dot{q})_q \dot{q} \end{bmatrix} \tag{8}$$

This combined system of equations of motion, the kinematic constraints of Eq. 1, and the initial conditions of Eq. 5 completely determine dynamic response of the system. Equations 8 and 1 represent a system of mixed Differential-Algebraic equations (DAE).

## 2. EXISTING METHODS FOR SOLVING SYSTEMS OF MIXED DIFFERENTIAL-ALGEBRAIC EQUATIONS.

Systems of DAE can not normally be solved by couvention numerical methods that are used to solve ordinary differential equations [4]. Several methods have been developed to solve sets of DAE. Gear [5] has proved that his stiff integration algorithm can be used to solve certain kinds of DAE. Certain other DAE systems can be converted into equivalent forms that are solvable by Gear's method. There exist, however, systems of DAE that can not be solved by this method. In Refs. 6 and 8, this idea has been successfully employed in analysis and optimal design of mechanical and electrical networks.

A second approach, due to Baumgarte [9], uses ideas from feedback control theory to construct a modified differential equation that implicitly accounts for constraint equations. A difficulty in this approach, however, is the selection of certain factors that influence the accuracy of solutions. Proper values of these factors are not known and experts are in disagreement over suitable values. Furthermore, there is no way to impose positive error control on constraint violations for this method.

The major drawback of the two methods described so far is that all n generalized coordinates must be integrated, whereas only n-m of them are independent. Furthermore, m is often large and n-m is small. Thus, it would be attractive to integrate for only n-m independent generalized coordinates. Solution of the remaining m dependent generalized coordinates from constraint equations can then ensure positive error control on constraint violations. The n-m generalized coordinates that are computed by integrating the equations of motion are called "independent generalized coordinates" and the remaining coordinates are called "dependent generalized coordinates". Wehage and Haug [1] developed an algorithm to pick an acceptable choice of independent generalized coordinates. LU decomposition of the constraint Jacobian matrix is used to identify independent and dependent coordinates. This algorithm performs satisfactorily, but may lead to poorly conditioned matrices, requiring considerable care to obtain accurate solutions. It is then required that the LU factorization be repeated and a new set of independent generalized coordinates selected. The result is an increase in computer time and greater propagation of integration error than desired.

A broader class of vectors zI of independent generalized coordinates can be generated from the vector q of physical generalized coordinates by the transformation

$$zI = VI\ q \tag{9}$$

where VI is an (n-m)×n transformation matrix. In the algorithm of Ref. 1, matrix VI is a Boolean matrix. Therefore, only individual generalized coordinates are chosen as independent. Use of a Boolean transformation matrix to pick independent generalized coordinates from q is, however, restrictive. It is desirable to enlarge the set of possible independent generalized coordinates by allowing matrix VI to be more general. Such a choice of VI can allow linear combinations of physical generalized coordinates to be selected as independent generalized coordinates. If the rows of VI are mutually orthogonal, the resulting linear combinations of individual generalized coordinates will be mutually independent. Specific forms of such a transformation can be chosen, based on properties of the DAE under consideration and desired properties of the transformed coordinates.

3. USES OF SINGULAR VALUE DECOMPOSITION. Singular value decomposition (SVD) has been extensively used to solve multivariate regression problems that may have multicollinearity between independent variables [10,11]. This method, known as Principal Components Analysis, defines a new set regressors as linear combinations of the original independent variables to most closely follow the pattern made by data points. The new regressors also account for dependence between the original independent variables. SVD has also been used for solution of Least Squares Problems in which linear combinations of the unknowns are used as a new set of unknowns to solve the problems [12].

Very recently, Singh [13] has introduced an algorithm using SVD to solve equations of motion of dynamic mechanical systems.

4. BASIC PROPERTIES OF SINGULAR VALUE DECOMPOSITION. The mxn Jacobian matrix $\Phi_q$, with $m < n$, may be decomposed [12,14] in the form

$$\Phi_q = U^T D V \tag{10}$$

where U and V are orthogonal matrices of dimension m and n, respectively, and the mxn matrix D has the form

$$D = \left[\begin{array}{cccc|c} \varepsilon_1 & & & & \\ & \varepsilon_2 & & 0 & \\ & & \ddots & & 0 \\ & & & & \\ & 0 & & \varepsilon_m & \end{array}\right] \tag{11}$$

The last n-m columns of D are zeros and the $\varepsilon_i$ are called the singular values of matrix $\Phi_q$, ordered so $\varepsilon_1 > \varepsilon_2 > \ldots \varepsilon_m > 0$.

From Eq. 10,

$$\Phi_q \Phi_q^T = U^T D V V^T D^T U = U^T D D^T U \tag{12}$$

Therefore,

$$\Phi_q \Phi_q^T U^T = U^T D D^T U U^T = U^T D D^T \equiv U^T \Lambda$$

where $\Lambda$ is the diagonal matrix $DD^T = \text{Diag}\,[\varepsilon_1^2, \varepsilon_2^2, \ldots, \varepsilon_m^2]$. This implies that columns of $U^T$ (rows of U) are orthonormal eigenvectors of the symmetric matrix $\Phi_q \Phi_q^T$ and the $\varepsilon_i^2$ are the corresponding eigenvalues. Similarly,

$$\Phi_q^T \Phi_q = V^T D^T U U^T D V = V^T D^T D V \tag{13}$$

Therefore,

$$\Phi_q^T \Phi_q V^T = V^T D^T D V V^T = V^T \Omega$$

where $\Omega$ is the diagonal matrix $D^T D$. Hence, columns of $V^T$ (rows of V) are the orthonormal eigenvectors of the symmetric matrix $\Phi_q^T \Phi_q$ [14].

The singular values of a matrix are very stable with respect to perturbations of its elements [12]. They are also the most reliable indicators of the conditioning of the matrix [14]. Gaussian elimination, on the other hand, can not detect ill-conditioning or rank degeneracies due to small errors in entries of matrices. Consider, for example, two matrices $A_n$ and $B_n$ of dimension n such that

$$b_{ii} = 1;\ a_{ii} = 1,\ i = 1, \ldots, n$$

$$b_{ij} = 1;\ a_{ij} = -1,\ j = i + 1, \ldots, n;\ i = 1, \ldots, n$$

$$b_{ij} = 0;\ a_{ij} = 0,\ j = 1, \ldots, i - 1;\ i = 2, \ldots, n$$

These upper triangular matrices are of rank n. However, matrix $A_n + E_n$ is of rank n-1, where $E_n$ is an n dimensional matrix with the nth element in the first column equal to $-2^{2-n}$ and all other elements equal to zero. On the other hand, matrix $B_n + E_n$ is of rank n. Therefore, diagonal entries of an upper or lower triangular matrix do not indicate existence of ill-conditioning. Since small errors enter into matrix entries due to truncation of numbers in computer representation, rank degeneracy of a matrix is a serious problem in numerical computation. The most reliable method to determine the rank of a matrix is to look at its smallest singular values. For matrix $A_n$ when n is equal to 25, the two smallest singular values are 1.5003342770 and $8.94 \times 10^{-8}$, but for matrix $B_n$ the two smallest singular values are 0.503818[illegible]889 and 0.5009501377. The number of zero singular values indicates the order of rank deficiency. A very small singular value indicates ill-conditioning of the matrix [12,14].

Even though the singular values are the square roots of eigenvalues of the matrix $\Phi_q^T \Phi_q$, they must not be determined this way, since small eigenvalues will be reduced to rounding error proportions and be lost in the computing process. A singular value decomposition algorithm due to Golub and Kahan [14] is based on a special adaptation of the QR algorithm. The singular value decomposition is computed in two stages. In the first stage, matrix $\Phi_q$ is transformed into a bidiagonal matrix B by a sequence of Householder transformations. The second stage of computation is application of a specially adapted QR algorithm to compute the singular value decomposition of B. The first part of this algorithm is direct and the second part is iterative. This algorithm is available in the form of computer software in Refs. 15 to 17.

5. USE OF SINGULAR VALUE DECOMPOSITION FOR GENERALIZED COORDINATE PARTITIONING. One may define a new variable z, such that

$$z = Vq \tag{14}$$

This is an orthogonal transformation of coordinates that gives a new vector z of generalized coordinates for the system. Since matrix V is orthogonal, hence non singular, there exists a one-to-one correspondence between z and q, locally. From now on, generalized coordinates q will be referred to as physical coordinates and generalized coordinates z will be called composite coordinates.

Taking the first time derivative of Eq. 14, with the transformation matrix V held constant,

$$\dot{z} = V\dot{q} \tag{15}$$

The time derivative of Eq. 15, gives

$$\ddot{z} = V\ddot{q} \tag{16}$$

Consider a perturbation $\delta z$ (virtual displacement) of z that is to be consistent with the constraint equations $\phi(z) = 0$,

$$\Phi_q \delta q = \Phi_q V^T \delta z = 0 \tag{17}$$

where $\delta q = V^T \delta z$ from Eq. 14. From Eq. 10 and the fact that V is orthogonal, Eq. 17 becomes

$$U^T D \delta z = 0 \tag{18}$$

Since U is orthogonal, one may premultiply Eq. 18 by U and use the form of D in Eq. 11 to obtain

$$[\varepsilon_1 \delta z_1 \;\; \varepsilon_2 \delta z_2 \;\; \cdots \;\; \varepsilon_m \delta z_m]^T = 0 \tag{19}$$

This shows that $\delta z_{m+1}, \ldots, \delta z_n$ can not be computed from Eq. 18. Hence, $\delta z_{m+1}, \ldots, \delta z_n$ can be evaluated only from the differential equations of motion. Therefore, $z_{m+1}, \ldots, z_n$ are selected to be the independent generalized coordinates for solution of the equations of motion and $z_1, \ldots, z_m$ are, be default, the dependent generalized coordinates that must be computed from the constraint equations. Since the kinematic constraint equations are nonlinear, an iterative technique such as the Newton-Raphson method is required to solve for the dependent generalized coordinates.

An important point in using generalized coordinate partitioning for solution of DAE is to obtain a criterion that determines if a set of independent generalized coordinates is acceptable from a computational point of view. Each redefinition of independent generalized coordinates requires restarting the predictor-corrector method [18] used for their integration, which requires substantial computational effort. On the other hand, not using the appropriate choice of independent generalized coordinates introduces avoidable computational errors.

6. PROPERTIES OF GENERALIZED COORDINATE PARTITIONING USING SINGULAR VALUE DECOMPOSITION. Considering only constraints for which $\Phi_t = 0$, Eq. 6 can be written as

$$\dot{\Phi} \equiv \Phi_q \dot{q} = 0 \tag{20}$$

Premultiplying Eq. 20 by U and using the definition of $\dot{z}$ in Eq. 15,

$$D\dot{z} = 0$$

Due to the special form of matrix D (Eq. 11), this is

$$[\varepsilon_1 \dot{z}_1 \quad \varepsilon_2 \dot{z}_2 \quad \cdots \quad \varepsilon_m \dot{z}_m]^T = 0 \tag{21}$$

Since the $\varepsilon$'s are not zero for a Jacobian matrix with full row rank, Eq. 21 implies

$$[\dot{z}_1 \quad \dot{z}_2 \quad \cdots \quad \dot{z}_m]^T = 0 \tag{22}$$

Since orthogonal transformations preserve norm,

$$||\dot{z}||_2 = ||\dot{q}||_2 \tag{23}$$

and since $\dot{z}_1, \ldots\ldots \dot{z}_m$ are equal to zero, Eq. 23 is

$$\sum_{i=m+1}^{n} \dot{z}_i^2 = \sum_{i=1}^{n} \dot{q}_i^2 \tag{24}$$

If one defines Unit Mass Kinetic Energy of a system as the kinetic energy when the system has unit masses and unit moments of inertia, Eq. 24 indicates that the unit mass kinetic energy due to independent composite coordinates captures the entire unit mass kinetic energy of the system. Thus, one can locally generate all motion in the system

with a smaller number of coordinates (equal in number to the number of degrees of freedom). This property can be used to develop a criterion that determines when to redefine independent composite coordinates. As time progresses, after a new set of independent composite coordinates has been selected, the dependent composite coordinates acquire small non-zero values, resulting in a reduction of unit mass kinetic energy captured by the independent composite coordinates. Therefore, if the unit mass kinetic energy captured by the independent composite coordinates is less than a predetermined fraction of the total unit mass kinetic energy of the system, a new set of independent composite coordinates is defined.

From Eqs. 14 and 15, the vector of virtual displacements $\delta q$ and the vector of velocities $\dot{q}$ can be expressed in terms of the rows of V:

$$\delta q = \sum_{j=m+1}^{n} \delta z_j V_j^T \tag{25}$$

and

$$\dot{q} = \sum_{j=m+1}^{n} \dot{z}_j V_j^T \tag{26}$$

Thus, displacement is confined to a subspace of $R^n$ that is spanned by $V_{m+1}^T$, ..., $V_n^T$ and velocities along $V_1^T$, ..., $V_m^T$ axes are zero. Therefore, to first order, the independent composite coordinates are Lagrangian coordinates that locally generate all system information.

Matrix V of Eq. 10 may be partitioned into submatrices VI and VD, representing the independent and dependent portions of matrix V,

$$V = \begin{bmatrix} VD \\ \hline VI \end{bmatrix} \begin{matrix} m \\ n - m \end{matrix} \tag{27}$$

It can be shown that rows of matrix VI are orthogonal to rows of the Jacobian matrix [19]. Therefore, integrating for independent composite coordinates moves the system along a tangent hyperplane of the constraint surface.

Consider, the slider crank mechanism shown in Fig. 2(a), initially in position ABC. Let $\phi_1$, the angular coordinate of body 1, be selected as the independent generalized coordinate for integration of the equations of motion. Prediction of $\phi_1$ moves body 1 to position A'B' (Fig. 2(a)), breaking revolute joints at A' and B'. Iterative solution of the dependent generalized coordinates from constraint equations moves the system to position AB"C", which involves considerable movement of all bodies. Unless an efficient and accurate prediction of initial estimates for dependent generalized coordinates is made prior to iterative solution, a large number of iterations is

required and, under some conditions, the iterative method may fail to converge.

During integration for independent composite coordinates, all physical coordinates are advanced very close to their true values (Fig. 2(b)), along a tangent hyperplane to the constraint surface. Therefore, the estimate of dependent composite coordinates is accurate and either none or just one Newton Raphson iteration is required for convergence.

To compute physical coordinates and velocities after independent composite positions and velocities have been computed, matrix equations of the form

$$\begin{bmatrix} \Phi_q \\ \hline VI \end{bmatrix} X = b \tag{28}$$

need to be solved [1,19]. It can be shown that the coefficient matrix of
Eq. 28 can not be ill-conditioned if the Jacobian matrix is not ill-conditioned [19].

It has been observed, during simulation of several mechanical systems, that composite generalized coordinates remain an acceptable choice of independent coordinates for a longer period of simulation time than do physical generalized coordinates based on the LU partitioning.

7. ALGORITHM FOR SOLUTION OF EQUATIONS OF MOTION. In this section, an algorithm to solve the equations of motion of dynamic mechanical systems using the method developed in Sec. 5 is presented. Generalized coordinates are partitioned into independent and dependent parts, using the SVD method of Sec. 5. Subroutine DE, which is based on an Adams-Bashforth predictor-corrector [18] method, is employed to integrate for the independent composite coordinates. Newton Raphson iteration is used to compute physical coordinates from the independent composite coordinates at every time step. This information is used to compute physical velocities from the independent composite velocities and finally to solve for the vector of physical accelerations. The algorithm is as follows:

Let $i$, initially 0, be an indicator of the current time step; i.e., $i = 0$ implies $t = t_0$.

(1) Read initial position, velocity, and other system data. Construct the Jacobian $\Phi_q$. Consider the n-m user supplied positions v to be accurate and correct the position vector $q^0$ using Newton Raphson iteration from Eqs. 29 and 27. Velocity vector $\dot{q}^0$ is computed from Eq. 31, considering the n-m user supplied velocities v to be accurate. Boolean matrix B fixes the user specified positions and velocities to prescribed values,

$$\left[\begin{array}{c}\Phi_q \\ \hline B\end{array}\right] \Delta q^i_{(k)} = \left[\begin{array}{c}-\Phi(q^i_{(k)}) \\ \hline v - Bq^i_{(k)}\end{array}\right] \tag{29}$$

$$q^i_{(k+1)} = q^i_{(k)} + \Delta q^i_{(k)} \qquad k = 1,\ldots \tag{30}$$

$$\left[\begin{array}{c}\Phi_q \\ \hline B\end{array}\right] \dot{q}^i = \left[\begin{array}{c}0 \\ \hline \vdots\end{array}\right] \tag{31}$$

Here, k is an iteration counter. Equation 29 computes successive corrections to $q^i_{(k)}$ until all kinematic constraints are satisfied, to required accuracy.

(2) Factor $\Phi_q$, using singular value decomposition of Sec. 4, as

$$\Phi_q = U^T D V \tag{32}$$

Partition V into independent and dependent portions (See Sec. 5),

$$V = \left[\begin{array}{c}VD \\ \hline VI\end{array}\right] \tag{33}$$

(3) Calculate $\ddot{q}$ and $\lambda$ from the acceleration equation (Eq. 8, presented here for reference),

$$\left[\begin{array}{cc} M & \Phi_q^T \\ \Phi_q & 0 \end{array}\right] \left[\begin{array}{c}\ddot{q} \\ \lambda\end{array}\right] = \left[\begin{array}{c} Q \\ -(\Phi_q \dot{q})_q \dot{q} - \Phi_{tt} - 2(\Phi_{qt})\dot{q} \end{array}\right] \tag{34}$$

(4) Calculate independent composite position $zI^i$, velocity $\dot{z}I^i$, and acceleration $\ddot{z}I^i$ (using the definition of zI from Eq. 14) as

$$[zI^i \; \dot{z}I^i \; \ddot{z}I^i] = [VI][q^i \; \dot{q}^i \; \ddot{q}^i]$$

(5) Integrate $[\dot{z}I, \ddot{z}I]$, with $[zI^i, \dot{z}I^i]$ as initial conditions, to get $[zI^{i+1}, \dot{z}I^{i+1}]$ at $t_{i+1} = t_i + \Delta t$, using Adams-Bashforth predictor-corrector integration.

(6) Predict $q^{i+1}$ and correct it iteratively by Newton Raphson iteration, using Eqs. 35 and 36, until constraints are satisfied to required accuracy,

$$\begin{bmatrix} \Phi_q \\ VI \end{bmatrix} \Delta q_{(k)}^{i+1} = \begin{bmatrix} -\Phi(q_{(k)}^{i+1}) \\ zI - VIq_{(k)}^{i+1} \end{bmatrix} \tag{35}$$

$$q_{(k+1)}^{i+1} = q_{(k)}^{i+1} + \Delta q_{(\kappa)}^{i+1}, \qquad k = 1,... \tag{36}$$

(7) Compute $\dot{q}^{i+1}$ from the velocity equation (Eq. 6 presented here for reference),

$$\begin{bmatrix} \Phi_q \\ VI \end{bmatrix} \dot{q}^{i+1} = \begin{bmatrix} 0 \\ \dot{z}I^{i+1} \end{bmatrix} \tag{37}$$

(8) Compute acceleration $\ddot{q}$ and Lagrange multiplier $\lambda$ from Eq. 34.

(9) Set $i = i + 1$ and $t_{i+1} = t_i + \Delta t$. If $||\dot{z}I||_2$ is less than a predetermined fraction of $||\dot{q}||_2$ (See Eq. 24), matrix VI needs to be updated. If so, repeat steps 2-9. If VI is still acceptable, repeat steps 3-9.

8. NUMERICAL EXAMPLE. To illustrate the method developed, dynamics of an M113 tracked personnel carrier going over a bump is simulated. The vehicle is modeled by 11 bodies, as shown in Fig. 3. The global coordinate system is located at the front of the vehicle, as shown in Fig. 4. The local coordinates of the wheels and the chassis are located at their centers of mass and are initially oriented parallel to the global X-axis. The road arms, however, are assumed to have their centers of mass at the points where they are attached to the chassis by revolute joints. The $\xi$-axes of their local coordinates are parallel to the axes of the road arms (Fig. 4).

Body 1 is the chassis, with weight 22,449 lb and pitch moment of inertia 133,000 lb-in.-$s^2$. Bodies 2-6 are the wheels, each with weight 180 lb and moment of inertia 90 lb-$ft^2$. The wheels are attached to the chassis by road arms (bodies 7-11). Bodies 7, 8, and 11 each have weight 44 lb and moment of inertia 15.86 lb-$ft^2$. Bodies 9 and 10 each have weight 22 lb and moment of inertia 7.93 lb-$ft^2$.

Torsional springs, each with a spring rate of 70,000 in.-lb/rad., are attached between the chassis and road arms. In addition, a bump stop prevents contact between the leading road arm (body 7) and the chassis. The bump stop is activated if body 7 moves by an angle greater than 0.2090 radians, counterclockwise. The stiffness of the spring that models the bump stop is $5.76 \times 10^6$ lb/in. Fluid dampers are connected between the chassis and bodies 7, 8, and 11, with damping characteristics shown in Fig. 5. To prevent adjacent wheels from penetrating one another, logical springs, each with stiffness 10,000 lb/in. in compression and zero in tension, are modeled between them. The track is modeled as a series of spring-dampers, with spring constant $10^5$ lb/in. and damping coefficient 41.7 lb-sec/in. A pretension of 2,250 lb is applied to the track.

The vehicle is initially at rest in the configuration given by Table 1.

Table 1. Initial Configuration of M113

| Body No. | Position | | |
|---|---|---|---|
| | x | y | $\phi$ |
| 1 | 6.424830 | 3.594392 | 0.0098366 |
| 2 | 2.207953 | 1.142995 | -0.0002644 |
| 3 | 4.380040 | 1.137461 | -0.0004955 |
| 4 | 6.555254 | 1.138714 | -0.0007750 |
| 5 | 8.729923 | 1.140114 | -0.0010999 |
| 6 | 10.90677 | 1.145697 | -0.0015061 |
| 7 | 1.293782 | 1.642426 | -0.5000145 |
| 8 | 3.481177 | 1.663943 | -0.5298527 |
| 9 | 5.668571 | 1.685460 | -0.5525503 |
| 10 | 7.855965 | 1.706978 | -0.5754019 |
| 11 | 10.04336 | 1.728495 | -0.5937436 |

To simulate the vehicle moving over a semi-circular obstacle of 8 in. radius, the x coordinate of the chassis is held fixed and the terrain is moved under it at a speed of 20 ft/sec. The VI matrix at the initial time is given in Table 2.

The simulation was carried out for a period of 2 seconds. The position, velocity, and acceleration of the vertical coordinate of the chassis are plotted versus time in Figs. 6-8.

Table 2. Matrix VI for the M113

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 1

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | -0.086930 | -0.073470 | -0.017800 | 0.022600 | 0.000000 | -0.026900 | -0.088370 | 0.000000 | 0.015130 |
| 0.022100 | 0.000000 | -0.038190 | -0.037160 | 0.000000 | 0.013300 | 0.047310 | 0.000000 | -0.014300 | 0.029000 |
| -0.070160 | -0.014200 | 0.012900 | -0.024200 | -0.014000 | -0.031360 | 0.028400 | -0.013900 | -0.019200 | 0.017700 |
| -0.013700 | -0.035300 | 0.046300 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 2

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.010900 | 0.004510 | -0.010200 | -0.011700 | 0.000000 | 0.069840 | 0.020000 | 0.000000 | -0.050470 |
| 0.009590 | 0.000000 | -0.028300 | -0.033400 | 0.000000 | 0.020500 | 0.041600 | 0.000000 | 0.008803 | 0.085670 |
| -0.022200 | 0.008706 | 0.095540 | 0.011600 | 0.008609 | 0.010500 | -0.010830 | 0.008511 | 0.011500 | -0.051500 |
| 0.008414 | 0.012500 | 0.033700 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 3

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | -0.011000 | -0.038234 | 0.015300 | 0.024100 | 0.000000 | 0.029300 | 0.044200 | 0.000000 | 0.085900 |
| 0.060840 | 0.000000 | 0.046770 | -0.025890 | 0.000000 | 0.097470 | 0.027190 | 0.000000 | -0.01[illegible]70 | -0.067930 |
| 0.033800 | -0.015890 | -0.085940 | 0.058700 | -0.015720 | -0.010400 | 0.018600 | -0.015540 | -0.012200 | 0.011000 |
| -0.015360 | -0.014000 | 0.019400 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 4

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.022900 | -0.000513 | 0.016200 | 0.053000 | 0.000000 | -0.081020 | 0.093780 | 0.000000 | 0.046320 |
| 0.030600 | 0.000000 | -0.011300 | 0.055590 | 0.000000 | -0.013700 | 0.025760 | 0.000000 | -0.001001 | 0.023200 |
| 0.032700 | -0.000990 | 0.023000 | -0.015200 | -0.000979 | 0.022900 | 0.086500 | -0.000968 | 0.022800 | -0.019700 |
| -0.000957 | 0.022700 | -0.023300 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 5

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.017200 | 0.013280 | -0.062260 | -0.058530 | 0.000000 | 0.049650 | 0.017300 | 0.000000 | -0.046330 |
| 0.045210 | 0.000000 | 0.025800 | 0.055000 | 0.000000 | 0.013300 | 0.038000 | 0.000000 | 0.026120 | 0.010300 |
| -0.017700 | 0.025830 | 0.013300 | 0.045240 | 0.025540 | 0.016200 | -0.013100 | 0.025250 | 0.019100 | 0.041000 |
| 0.024960 | 0.022000 | 0.018500 | | | | | | | |

Table 2.(cont.)

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 6

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.027520 | 0.021430 | -0.094710 | -0.033200 | 0.000000 | 0.080330 | 0.030930 | 0.000000 | 0.035500 |
| 0.052100 | 0.000000 | 0.016740 | 0.021640 | 0.000000 | -0.054190 | -0.034450 | 0.000000 | 0.041840 | -0.082460 |
| -0.027300 | 0.041380 | -0.035580 | 0.073990 | 0.040920 | 0.011310 | 0.057500 | 0.040460 | 0.058190 | -0.041830 |
| 0.039990 | 0.010500 | -0.016200 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 7

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.050580 | -0.052500 | -0.028700 | -0.017680 | 0.000000 | 0.021240 | 0.041400 | 0.000000 | -0.017800 |
| -0.032260 | 0.000000 | -0.081360 | 0.002777 | 0.000000 | -0.025100 | -0.036600 | 0.000000 | -0.010200 | 0.032000 |
| -0.036900 | -0.010100 | 0.020500 | 0.023300 | -0.010000 | 0.090280 | -0.014200 | -0.099080 | -0.024550 | 0.031270 |
| -0.097950 | -0.013900 | -0.026300 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 8

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.100000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 9

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.100000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 10

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.100000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | | | | | | | |

Table 2.(cont.)

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 11

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.100000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | | | | | | | |

LINEAR COMBINATION COEFFICIENTS FOR INDEPENDENT COORDINATE 12

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.100000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 0.000000 | 0.000000 | 0.000000 | | | | | | | |

For this problem, solution using the SVD partitioning method did not require redefinition of the independent coordinate for the entire simulation period. Furthermore, no Newton-Raphson iterations were required to compute physical coordinates from the independent composite coordinates. When LU partitioning was used to solve this problem, independent generalized coordinates were redefined several times. During integration for independent generalized coordinates, using subroutine DE [18], the tightest error tolerance that can be imposed for that set of equations is determined. If the user specified tolerance is less than this value, subroutine DE increases the error tolerance. For integration of the independent composite coordinates, subroutine DE increased the error tolerance to 0.0032. The maximum allowed error tolerance computed by subroutine DE for integrating the independent physical coordinates was 0.0128, four times the tolerance for solution by the SVD method. This suggests that, for this problem, the independent composite coordinates are better than independent physical coordinates, from an integration point of view.

## References


1. Wehage, R. A., and Haug, E. J., "Generalized Coordinate Partitioning for Dimension Reduction in Analysis of Constrained Dynamic Systems," ASME J. of Mechanical Design, Vol. 104, January 1982, pp. 247-255.

2. Goldstein, H., "Classical Mechanics," Addison-Wesley Publishing Company, Inc.

3. Greenwood, D. T., "Principles of Dynamics," Prentic Hall, Inc., Englewood Cliffs, New Jersey.

4. Petzold, L., "Differential-Algebraic Equations are Not ODE's," SIAM J. of Scientific and Statistical Computing, Vol. 3, No. 3, 1982.

5. Gear, C. W., "Differential-Algebraic Equations," Computer Aided Analysis And Optimization of Mechanical System Dynamics (ed. E. J. Haug), Springer Verlag, Heidelberg, 1984.

6. Orlandea, N., Chase, M. A., and Calahan, D. A., "A Sparsity Oriented Approach to Dynamic Analysis of Mechanical Systems," Parts I and II, ASME J. of Engineering for Industry, Ser. B., Vol. 99, 1977, pp. 773-784.

7. Haug, E. J., Wehage, R. A., and Barman, N. C., "Design Sensitivity Analysis of Planar Mechanisms and Machine Dynamics," ASME J. of Mechanical Design, Vol. 103, July 1981.

8. Hachtel, G. D., Brayton, R. K., and Gustavson, F. G., "The Sparse Tableau Approach to Network Analysis and Design," IEEE Transactions on Circuit Theory, Vol. CT-18, No. 1, January 1971.

9. Baumgarte, J., "Stabilization of Constraints and Integrals of Motion in Dynamic Systems," Computer Methods in Applied Mechanics and Engineering, North Holland Publication Company, 1972, pp. 1-16.

10. Wonnacott, R. J., and Wonnacott, T. J., "Regression: A Second Course in Statistics," Wiley, 1981.

11. Mandel, J., "Use of Singular Value Decomposition in Regression Analysis," The American Statistician, Vol. 36, No. 1, February 1982.

12. Lawson, C. L., and Hanson, R. J., "Solving Least Squares Problems," Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1974.

13. Singh, R. P., and Likins, P. W., "Singular Value Decomposition For Constrained Dynamical Systems," ASME J. of Applied Mechanics, to appear.

14. Golub, G., and Kahan, W., "Calculating the Singular Values and Pseudo-Inverse of a Matrix," SIAM J. of Numerical Analysis, Ser. B., Vol. 2, No. 2, 1965.

15. "IMSL Users Guide," International Mathematical and Statistical Libraries Inc., Houston, Texas.

16. "LINPACK Users Guide," Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

17. "Harwell Subroutine Library," AERE Harwell, Oxon, U. K.

18. Shampine, L. F., and Gordon, M. K., "Computer Solution of Ordinary Differential Equations: The Initial Value Problem," W. J. Freeman, San Francisco, California, 1975.

19. Mani, N. K., "Use of Singular Value Decomposition for Analysis and Optimization of Mechanical System Dynamics," Ph.D. Thesis, The University of Iowa, Iowa City, July 1984.

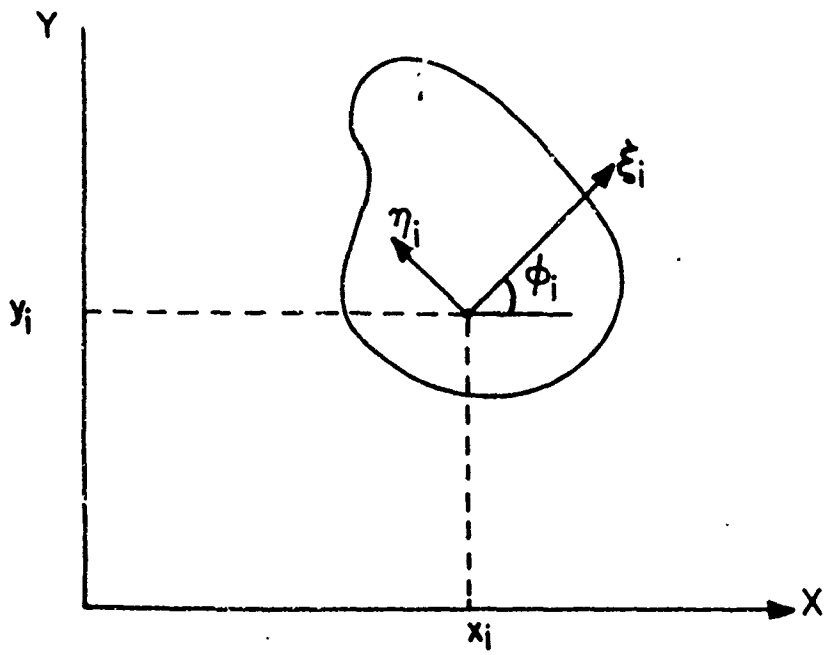
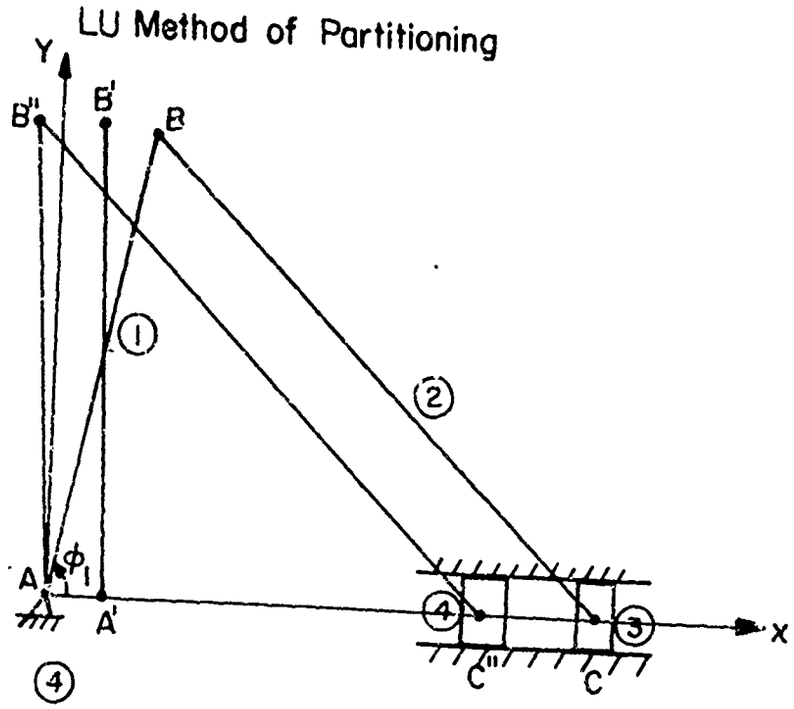



Figure 1. Coordinate System in Two Dimensions



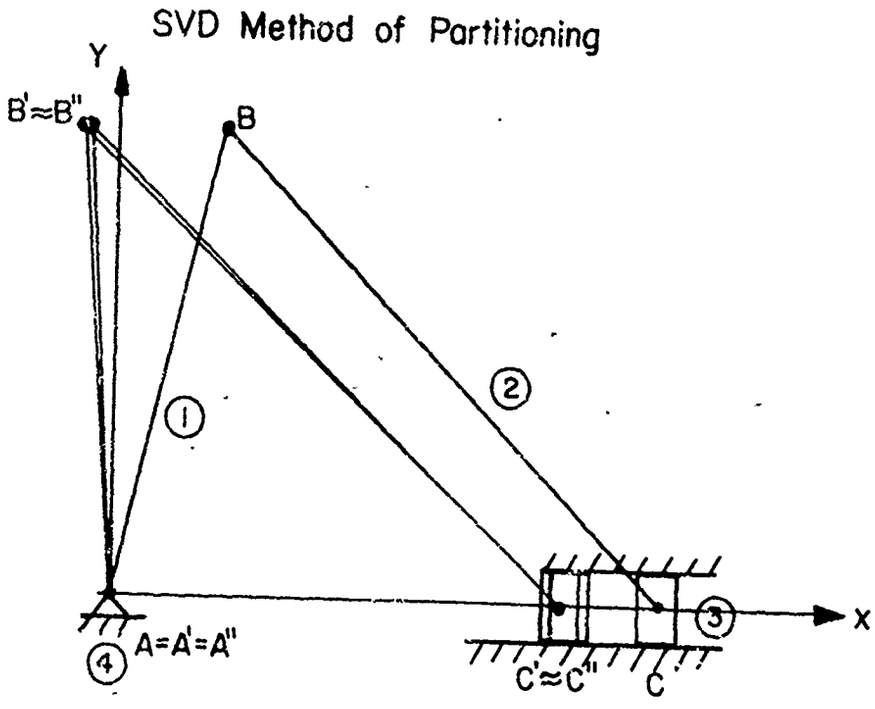




Figure 2. Graphical Representation of Partitioning Methods

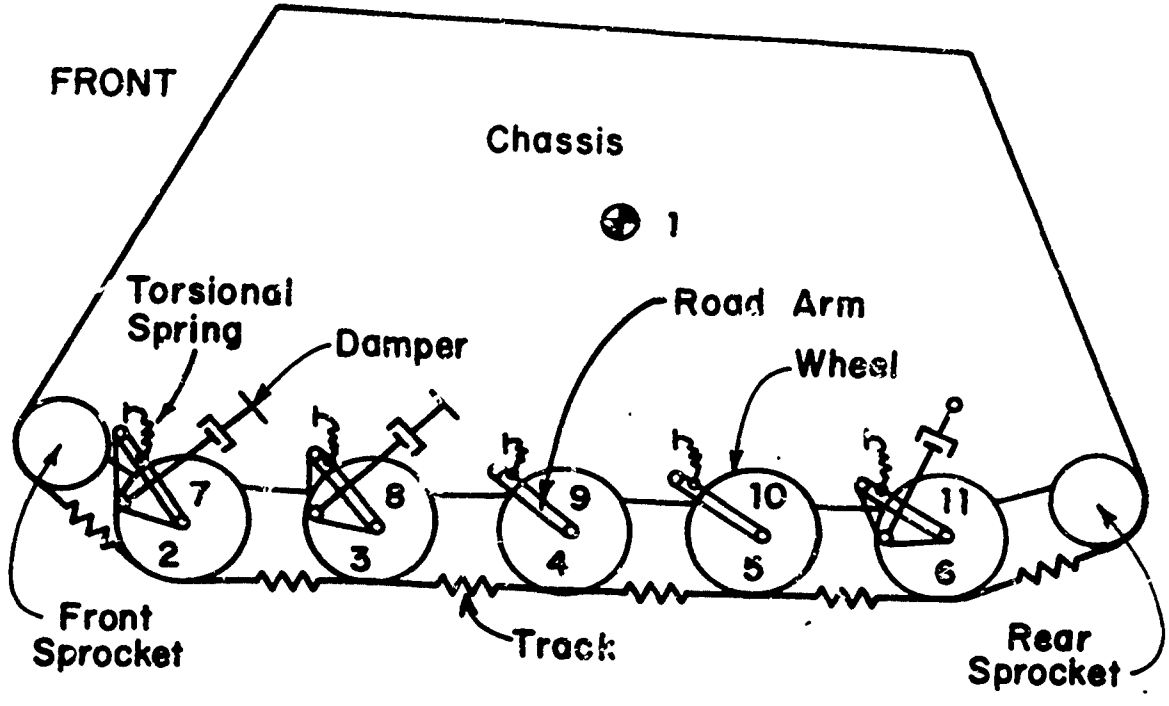



Figure 3. M113 Armor Vehicle Model

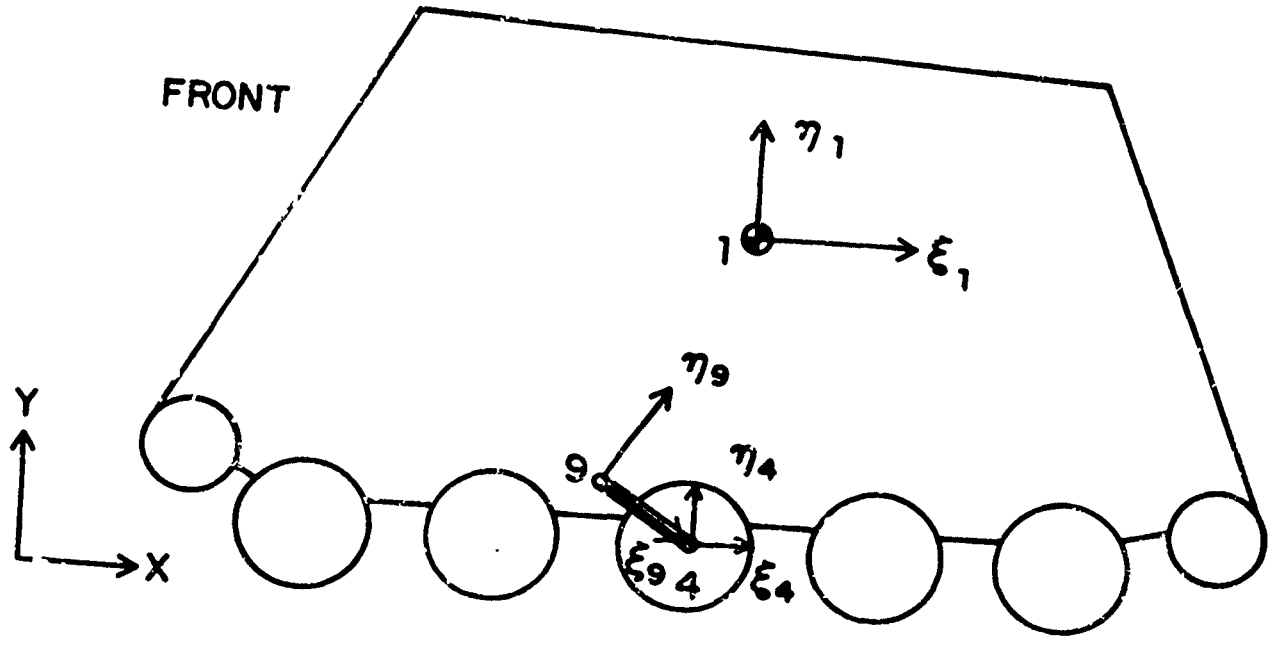



Figure 4. Coordinate System for M113 Problem

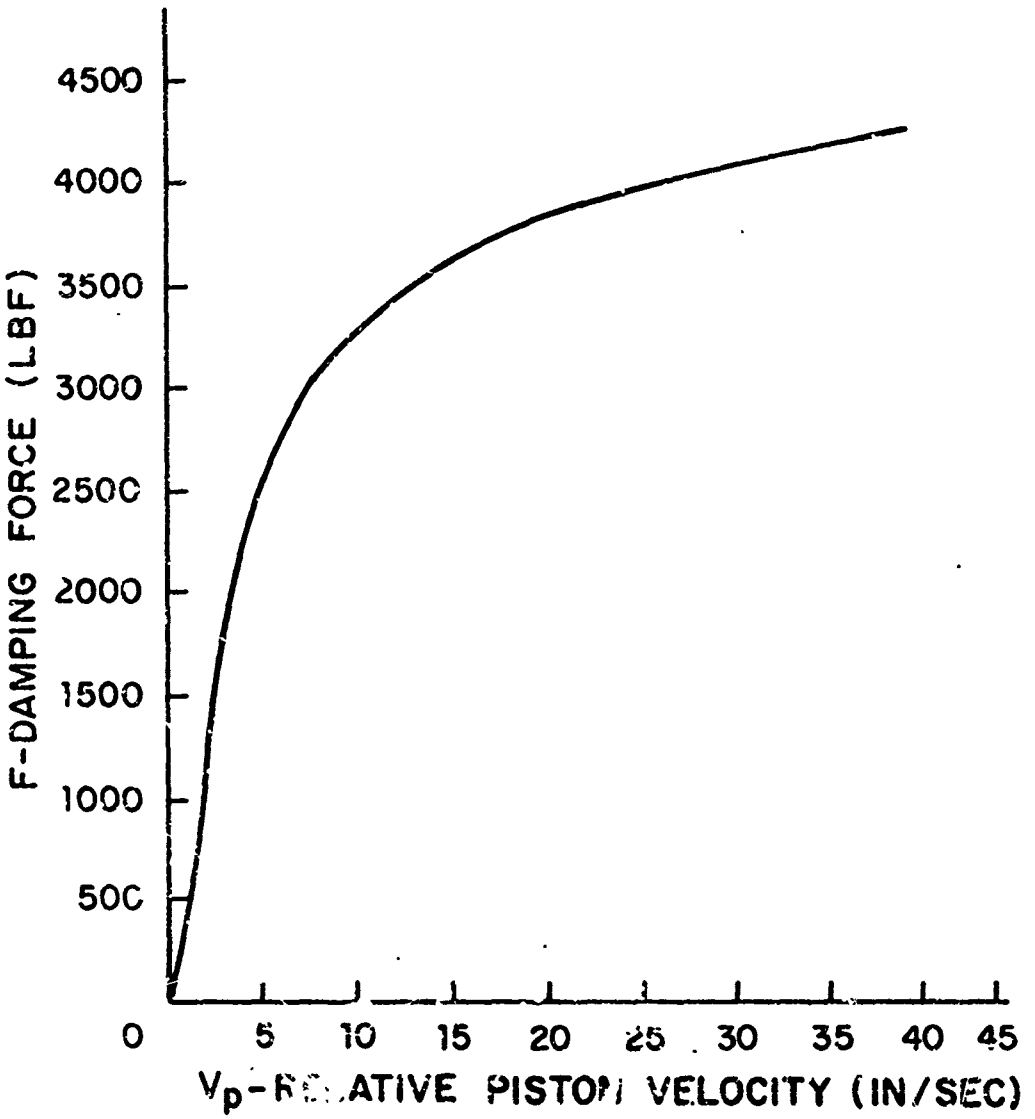



Figure 5. Damper Characteristics for M113 Vehicle

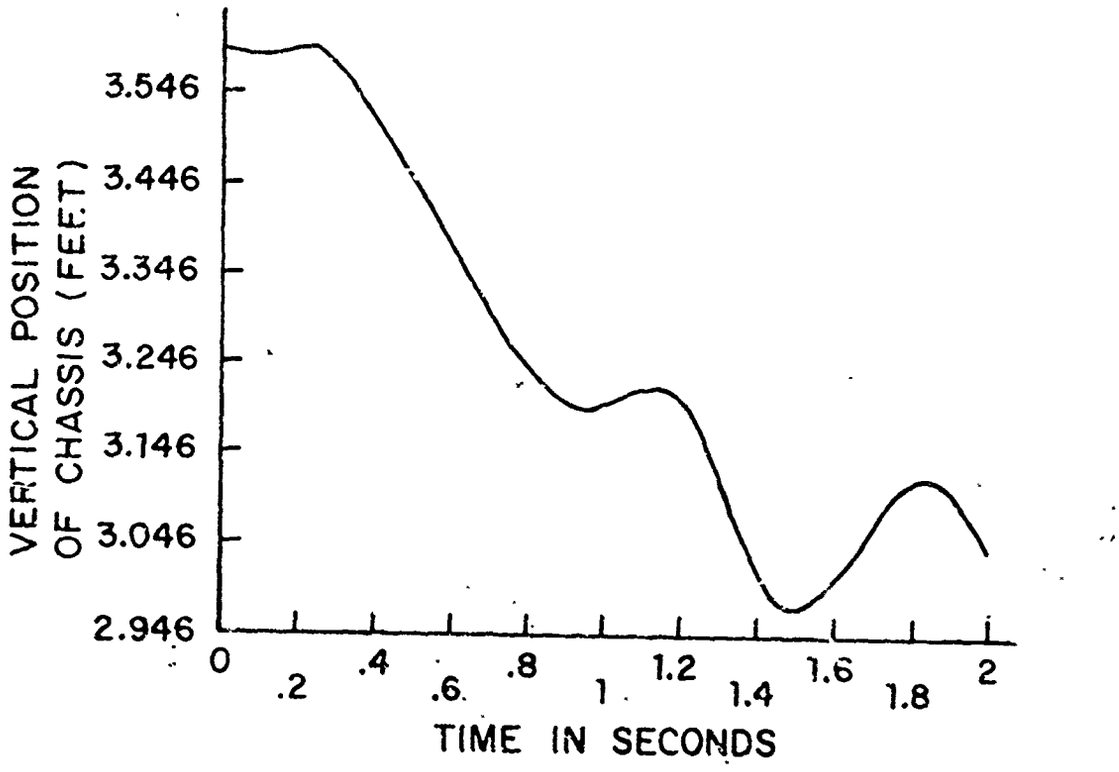



Figure 6. Vertical Position of the Chassis of M113

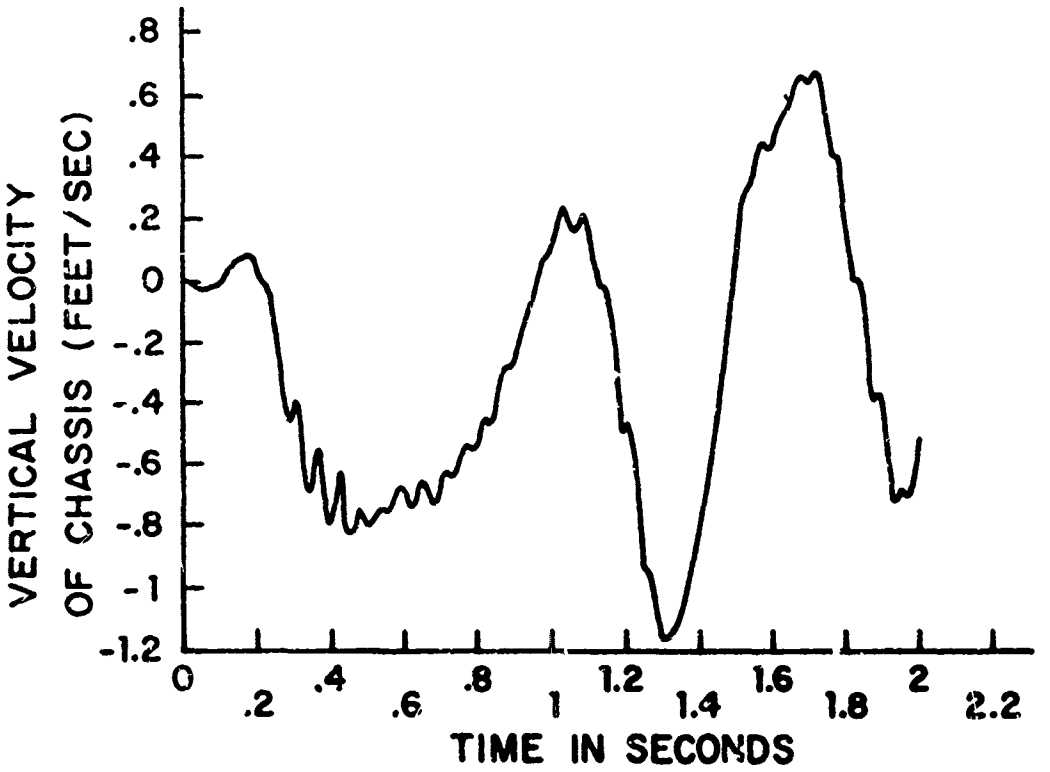



Figure 7. Vertical Velocity of the Chassis of M113

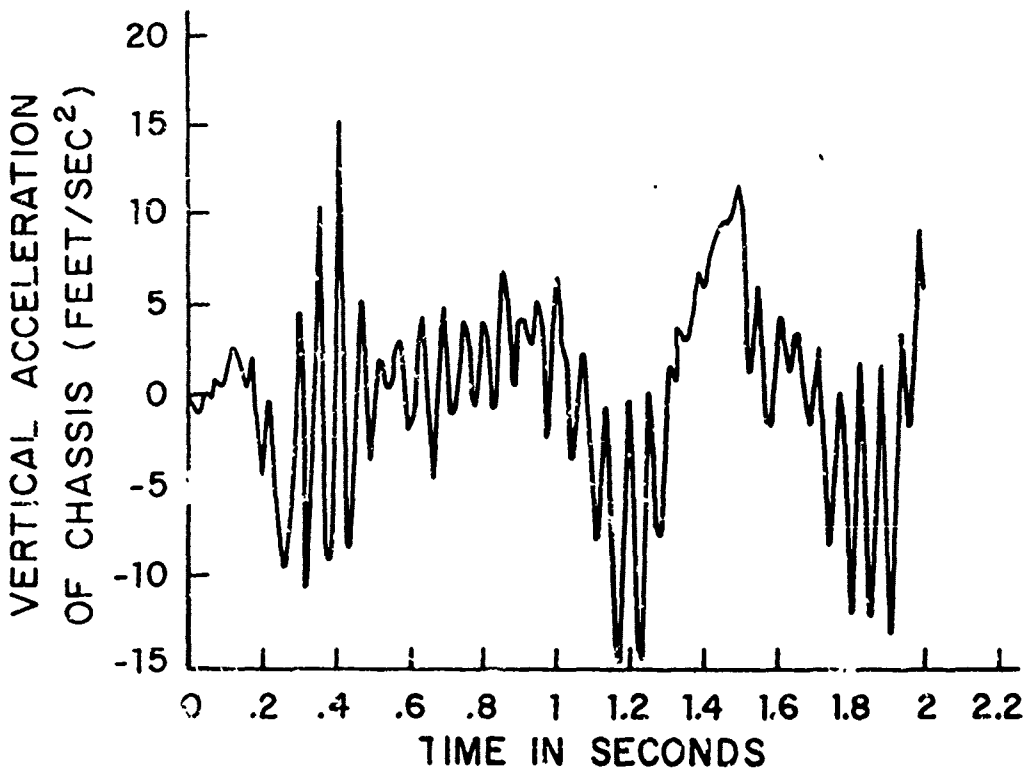



Figure 8. Vertical Acceleration of the Chassis of M113

# APPLICATION OF SCREW CALCULUS TO THE EVALUATION OF MANIPULATOR WORKSPACE

L.M. Hsia
Department of Mechanical Engineering
California State University
Los Angeles, California 90032

Ting W. Lee[1]
Department of Mechanical and Aerospace Engineering
Rutgers, The State University of New Jersey
New Brunswick, New Jersey 08903

ABSTRACT. In this paper, an analytical techique is presented that is based on screw calculus and dual-number matrices to derive the kinematic equations and the workspace formulations of robotic manipulators. A computational procedure for the quantitative evaluation of workspace volume is also developed. Several examples are chosen to demonstrate the usefulness and the effectiveness of the approach.

1. INTRODUCTION. One of the basic problems encountered in manipulator design is the determination of the shape of workspace and its characteristics. The workspace, which is the zone of operation of a manipulator, is the space associated with possible positions and orientations of the last link of a manipulator. A knowledge on the workspace of a manipulator can provide a measure of the efficiency of the design. Therefore, the investigation on workspace is of fundamental interest.

There have been few investigations of the subject on the record. A summary is given in references [1,2]. Several approaches have been used dealing with computational development of workspaces, such as iterative (Kumar and Waldron [3], Sugimoto and Duffy [4]; Tsai and Soni [5]); grid-scanning (Lee and Yang [1]); and lately, a non-iterative generation scheme (Hansen, Gupta and Kazerounian [2]) and the technique based on Gauss' divergence theorem (Jou and Waldron [8]). Most investigators involve the use of conventional [4x4] matrices method of Hartenberg and Denavit [5] for the analytical representation and generation of workspace. Two problems are of major concern: one is computational efficiency; the other is the limitation of most conventional methods which are applicable only to manipulators with revolute joints. Manipulators with prismatic joints, which are common, and other special kinematic pairs such as cylindrical and screw pairs are very difficult to evaluate as far as the workspace is concerned. In most of these cases, methods are either not available, or they are computationally too inefficient to be useful.

An analytical technique is presented in this investigation. It is based on screw calculus and dual-number matrices, to derive the kinematic equations and the workspace formulations of robotic manipulators. The application of the method to the study of the kinematics and dynamics of manipulation is relatively

---

[1] Present address: Department of Mechanical Engineering, State University of New York at Stony Brook, Stony Brook, New York 11794

PREVIOUS PAGE IS BLANK

new [6,7]. However, its application to the representation of manipulator workspace has not been explored.

The method of screw calculus offers many advantages, especially in dealing with three-dimensional kinematics. In these cases the effort to derive closed form analytical expressions is often impaired by laborious or even insurmountable algebraic manipulations. The use of compact screw notations and dual formulation facilitate substantially the effort on algebraic manipulations. Consequently, concise expressions can be formulated, which would provide valuable geometric insight into the rigid body under constrained spatial motion. In addition, there is significant computational advantage. Dual number operations not only facilitate analytic work, but also make computer programming simpler and more efficient via dual subroutines, namely, incorporating algebraic operations of dual numbers and trigonometric functions of dual angles into computer subroutines. By declaring all the motion variables to be complex, we are able to perform algebraic manipulations on the rotational part (i.e., real part) and translational part (i.e., imaginary part) of each motion variable at the same time in a dual equation.

In this paper, a new workspace generation technique, using the optimum path search, is proposed. The outlining of the projection of workspace on a specified plane involves the determination of a minimum distance between the end-effector of a manipulator and a specified target point on the plane and the search is converted into an optimization problem. An optimization technique is used, which is the FMFP code of Fletcher and Powell [9]. The algorithm thus developed is a partial-scanning method to generate the workspace. It is shown that the algorithm ensures a significant reduction of scanning points and provides improved computational efficiency in especially complicated cases as compared to the conventional scanning technique [1]. The algorithm developed here is applicable to manipulators having not only the revolute joints but also the prismatic and cylindrical joints. With slight extension, it can include also the screw joints. An algorithm for non-revolute type of joints for the quantitative evaluation of workspace is seldom seen, at least to the author's knowledge.

In the following, we first begin with a description of the underlining principles of the new technique and how the algorithm works. Then a comparative study is performed with a previously established technique [1] on how is computation time affected with accuracy. Finally, the algorithm is applied to evaluate several workspace boundary profiles and the results, whenever possible, are compared with previously published data [1,10].

2. MANIPULATOR DESCRIPTION. A manipulator with n joints in series can be represented schematically as shown in Fig. 1. The motion of joint n consists of a rotation of $\theta_n$ about axis $Z_n$, a translation along $Z_n$, a translation along the current $X_n$ axis of $a_n$, and finally a rotation of $\alpha_n$ about the current $X_n$ axis.

The purposes of $\theta_n$ and $b_n$ are a function of the joint type. If the joint is a revolute, then $b_n$ is a constant and $\theta_n$ is variable. If the joint is prismatic, then $\theta_n$ is constant and $b_n$ is variable (Fig. 2).

The transformation that performs the task of moving one joint to the next is represented by the product of homogeneous transformation matrices, such as in Paul [11]

$$A_n = \text{Rot}(Z,\theta_n)\ \text{Trans}(0,0,b_n)\ \text{Trans}(a_n,0,0)\ \text{Rot}(X,\alpha_n) \qquad (1)$$

where

$\text{Rot}(Z,\theta_n)$ - Rotate about Z-axis $\theta_n$

$\text{Trans}(0,0,b_n)$ - Translate $b_n$ in the Z direction

$\text{Trans}(a_n,0,0)$ - Translate $a_n$ in the X direction

$\text{Rot}(X,\alpha_n)$ - Rotate about X-axis $\alpha_n$

Conventionally, the [4x4] matrix form of the Denavit and Hartenberg [12] is used, $A_n$ can be expressed as

$$A_n = \begin{bmatrix} \cos\theta & -\sin\theta\cos\alpha & \sin\theta\sin\alpha & a\cos\theta \\ \sin\theta & \cos\theta\cos\alpha & -\cos\theta\sin\alpha & a\sin\theta \\ 0 & \sin\alpha & \cos\alpha & b \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (2)$$

In this investigation, screw calculus and dual-number matrices [13,14] are used to derive the kinematic equations and the workspace formulations of manipulators. There are several advantages for this approach. The compact screw notations and dual formulation facilitate substantially the effort on algebraic manipulations and provide not only concise expressions but also significant computational advantage. Using dual-number operations, we have

$$\tilde{A}_n = \text{ROT}(Z,\tilde{\theta}_n)\ \text{ROT}(X,\tilde{\alpha}_n) \qquad (3)$$

where

$\tilde{A}_n$ - Dual Transformation Matrix

$\tilde{\theta}_n$ - Dual Variable Angle

$\tilde{\alpha}_n$ - Dual Constant Angle

Dual angles describe both motions of rotation and translation. $\tilde{\theta}_n$ and $\tilde{\alpha}_n$ relate to the previously defined variables as follows:

$$\tilde{\theta}_n = (\theta_n, b_n) \text{ or } \theta_n + \varepsilon b_n$$

$$\tilde{\alpha}_n = (\alpha_n, a_n) \text{ or } \alpha_n + \varepsilon a_n \tag{4}$$

The transformation can now be rewritten in [3x3] dual matrix notation as follows:

$$\tilde{A}_n = \begin{bmatrix} \cos\tilde{\theta} & -\sin\tilde{\theta}\cos\tilde{\alpha} & \sin\tilde{\theta}\sin\tilde{\alpha} \\ \sin\tilde{\theta} & \cos\tilde{\theta}\cos\tilde{\alpha} & -\cos\tilde{\theta}\sin\tilde{\alpha} \\ 0 & \sin\tilde{\alpha} & \cos\tilde{\alpha} \end{bmatrix} \tag{5}$$

The dual multiplication and dual trigonometry functions are defined by,

| | | |
|---|---|---|
| Multiplication: | $\hat{a}\,\hat{b} = a\,b + \varepsilon(a_0 b + b_0 a)$ | (6) |
| Dual sine: | $\sin(\hat{a}) = \sin(a) + \varepsilon\, a_0 \cos(a)$ | (7) |
| Dual cosine: | $\cos(\hat{a}) = \cos(a) - \varepsilon\, a_0 \sin(a)$ | (8) |

where, letters with angle symbols on top signify dual numbers

$\hat{a} = (a, a_0)$ or $a + \varepsilon\, a_0$

$\hat{b} = (b, b_0)$ or $b + \varepsilon\, b_0$

$\varepsilon$ = The dual unit having the property $\varepsilon^2 = 0$

$a$ = Real part of a

$a_0$ = Dual part of a

$b$ = Real part of b

$b_0$ = Dual part of b

Once the transformation for one joint has been defined, the transformation which represents the relation from the last joint to the ground of a 6-joint manipulator is given as:

$$T = A_1\, A_2\, A_3\, A_4\, A_5\, A_6 \tag{9}$$

An additional transformation, E, is used to represent the relation from the last joint to the hand. This transformation is a constant.

$$\text{(Ground Coordinates)} = TE(\text{Hand Coordinates}) \tag{10}$$

If the [4x4] method is chosen, the location of the hand, given the value of all joint variables, can be determined by entering the vector $(0,0,0,1)^T$ of the hand coordinate in Eq. (10).

If the dual number method is used, a technique entitled "The Transference Principle [15]" must be applied to the transformation TE to obtain this information.

Computer programming involving dual numbers is efficient via dual subroutines, namely, incorporating algebraic operations of dual numbers and trigonometric functions of dual angles into computer subroutines. By declaring all the motion variables to be complex, we are able to perform algebraic manipulations on the rotational part (i.e., the real part) and the translational part (i.e., the imaginary part) of each motion variable at the same time in a dual equation.

3. BASIC APPROACH. The basic approach involves the search for an optimum path between the end-effector of a manipulator and a target position on a specified plane in which the workspace projection is desired. Subsequently, this minimum distance, which is obtained through an optimization technique, is then compared with the actual position of the target to determine the accessibility of the manipulator. The objective is to provide an intelligent partial scanning, with significant reduction of scanning points, to gain computational efficiency.

The method of screw calculus and [3x3] dual-number matrix [13,14] is the basis for the analytical formulation of this investigation. This allows one to take full advantage of the compactness and clarity of dual formulation in describing the kinematic characteristics of the motion of manipulators.

In the following, the basic concept on the generation of workspace boundary and volume is discussed, then the optimum path search technique and the optimization problem are formulated.

### 3.1 On the Generation of Boundary Profile and Volume

The Boundary Profile is defined as the image of the workspace produced by fixing the first joint and passing this workspace through a plane that is perpendicular to the motion produced by the first joint.

For a manipulator with first joint revolute, the boundary profile would be produced by taking all the points in the workspace and mapping them onto the Y-Z plane as follows:

| 3-D points in workspace | | 2-D Y-Z plane | |
|---|---|---|---|
| (X,Y,Z) | ==> | $I = \sqrt{X^2 + Y^2}$ | (11) |
| | | $J = Z$ | |
| | | (I,J) | |

This is a circular projection on the Y-Z plane about $Z_1$ [1].

For a manipulator with first joint prismatic the mapping procedure is,

$$(X,Y,Z) \quad ==> \quad \begin{matrix} I = X \\ J = Y \\ (I,J) \end{matrix} \qquad (12)$$

This is a parallel projection on the X-Y plane.

Once the area of this projection is known, an approximation of the workspace volume can be established.

If the first joint is a revolute, the volume is approximated by integrating the area as a volume of revolution about $Z_1$ over the limits of the first joint motion. If the first joint is prismatic, then the volume is simply the area times the total joint excursion.

This approximation may result in error due to the undetermined end shape. Fortunately, most manipulators usually have the first two or three links coplanar and these links determine the majority of the workspace. Error is then due to the last smaller links resulting in small error.

### 3.2 The Optimum Path Search Technique

As it was mentioned earlier, the present method for workspace generation follows the grid-scanning approach of Lee and Yang [1]. However, there is a basic difference in performing the scanning. The new method uses an optimization technique which facilitates the scanning process and reduces the number of scanning points. Consequently, it is a partial-scanning technique. In the following the new method is compared with the full-scanning technique of Lee and Yang [1] and the basic formulations of the Optimum Path Search Technique are presented.

#### (1) The Scanning Technique of Lee and Yang [1]

This technique which is explained in the following consists of scanning through all the joint variables by means of nested loops, using appropriately small step sizes to ensure that the points generated are dense enough to fill all points inside the workspace projection.

- Define an array that represents a plane passing through the workspace.
- Scan through all combinations of joint variables and map the location of the hand onto the array.
- Find the border of the workspace mapped on the array and trace it. This is for graphic display.
- Perform volume calculation.

One way to estimate the step sizes for each joint is to configure the manipulator in its "longest" position from the joint of interest. The step size is determined by the largest joint displacement that will result in the motion of the hand to move from one array entry to the next. This calculation is repeated for each joint. Often is the case that less than all the joint variables will play a role in defining the primary workspace. An example of this would be a manipulator, such as a Unimation PUMA, without an end effector

specified and also having a 3-degrees-of-freedom wrist. This manipulator requires only two joint variables to be scanned through. Remember that the first joint is taken into account by the mapping procedure mentioned in Section 3.1.

It is worth noting that when this technique was published it was exclusively for manipulators with revolute joints. The technique works quite well with prismatic joints by making the appropriate changes in the transformation matrices.

(2) The Optimum Path Search Technique

The new method differs from the preceding method, primarily because the workspace generation technique is based on an optimum path search as explained in the following:

- Also define an array that presents a plane passing through the workspace
- Rather than scanning through the joint angles and projecting the position on the array, specify a location on the array and perform inverse calculations to obtain joint angles by solving the following optimization problem:

With the two techniques outlined in some details a comparison can now be made and is summarized in Table 1.

The Optimum Path Search Technique uses a recursive method to determine the minimum of the objective function. This can require a great deal of computer time, especially since points outside of the workspace are rechecked a number of times (3-5) before being identified as such. However, with this rather substantial penalty of large computer time, comes the advantage of having to only locate points around the border of the workspace. This results in computational time to scale with the accuracy in terms of a length scale. An example of this would be if two runs were made, one with a 40x40 array and a second with a 80x80 array, the 80x80 will take approximately twice as long as the 40x40. This is because the perimeter of the workspace has doubled in terms of array entries.

If analysis is to be performed with the joint variable scanning process, then doubling the grid size results in a multiplication of the computational time. This factor is 2 raised to the number of joints active in determining the workspace. This is due to reduction of step size. If two joint variables are scanned then the computation time would increase by 4. For a very general manipulator having five joints active in defining the workspace, the calculation time would be increased by a factor of 32.

For analysis of the workspace, the technique of choice will depend on accuracy and the number of active variables present. With increased accuracy (i.e., larger array size), there will be a point at which the optimization technique will become advantageous.

At present the best way to establish what technique is to be used is to run a particular manipulator with both techniques using a small array such as 10x10. Note the time for each run and use the calculation scaling laws mentioned to

determine which technique will take the least time for a desired accuracy. By using the trail run, effects of the number of active variables on the optimization process are also incorporated.

The Optimum Path Search Technique does have advantages that the Grid-Scanning process does not. For instance, a most significant one is that the former technique lends itself to secondary workspace analysis. If the workspace with a specific end-effector orientation is required over a plane or surface, simply introduce a constraint pertaining to orientation. This property cannot be achieved with the Grid-Scanning process.

4. THE WORKSPACE GENERATION TECHNIQUE BASED ON OPTIMUM PATH SEARCH. The basic approach, as it was outlined in the previous section, involves the solution of the following optimization problems.

The Optimization Problem

Minimize: Distance of the hand to the desired location

Subject to: All physical constraints, such as joints within limits of rotations and translations, etc.

If the distance falls within the array entry, the point is inside the workspace. Otherwise, it is outside of the workspace.

To perform the minimization of the distance function subject to joint constraints, an objective function was developed using the exterior method to be applied to an unconstrained optimization algorithm, the conjugate direction minimization technique of Fletcher and Powell or the FMFP code [5]. This method introduces penalty functions which increase the value of the objective function if a constraint is violated.

Formulation using the exterior method is as follows [16],

$$\begin{aligned} &\text{Minimize:} \quad f(\tilde{U}) = (T_x-H_x)^2 + (T_y-H_y)^2 + (T_z-H_z)^2 \\ &\text{Subject to:} \\ &\text{Limits of Joints:} \quad \phi_i(U) \le 0 \quad \text{with} \quad i = 1 \text{ to } j \end{aligned} \tag{13}$$

where $f(U)$ denotes the objective function for the constrained optimization problem, $U$ denotes the argument vector or design vector, $T(T_x,T_y,T_z)$ and $H(H_x,H_y,H_z)$ denotes the target vector and the position vector of the hand or end-effector, respectively.

The unconstrained optimization problem is defined as

$$\text{Minimize:} \quad F(\tilde{U}) = f(\tilde{U}) + C_1 E(U) \tag{14}$$

where $C_1$ is a weighting constant, $E$ is the constraint function

$$E(\tilde{U}) = \sum_{i=1}^{j} [\phi_i(\tilde{U}) + |\phi_i(\tilde{U})|]^2$$

An algorithm written in FORTRAN language for generating the workspace of a general manipulator having prismatic, cylindrical or screw joints as well as revolute joints has been developed. The principal and details of the algorithm are given in Ref. [16].

5. CASE STUDIES. Four problems which are believed to be representative are chosen to demonstrate the effectiveness as well as the capability of the workspace generation algorithm. The first two problems provide some indications of how well the prediction of the algorithm compares with both the Grid-Scanning Technique of Lee and Yang [1] as well as the theoretical result on a special case, the 3R Manipulators taken from Gupta and Roth [10]. The other two problems deal with industrial robots having the prismatic as well as the revolute joint. They are used to demonstrate the capability of the algorithm to handle manipulators with joints other than the revolutes. The quantitative workspace information of these two industrial robots presented here is believed to appear first time on public record.

Example 1: A 3R Manipulator [10]. This example is used to verify the accuracy of the algorithm. The kinematic parameters of this manipulator are given in Table 2. A comparison is given in the following:

| | Workspace Volume [in.$^3$] | Percent Error |
|---|---|---|
| This investigation | 1255.9 | 4.4% |
| Analytical result [10] | 1202.7 | -- |
| The Scanning Technique [1] | 1215.7 | 1.1% |
| The bounds on array*: | $Y_{min} = 2.0$, $Y_{max} = 18.0$, $Z_{min} = -2.0$ and $Z_{max} = 2.0$ | |

*The bounds on array denote the physical scaling of the array which gives the mathematical representation of the plane in which the workspace projection is made.

Figure 3 gives the computer graphics output of the circular projection of this manipulator.

Example 2: A 3R Manipulator. This example is used to verify that the void handling features of the algorithm worked correctly.

The kinematic parameters table is similar to that of Example 1, except $\alpha_1$ is changed to $\alpha_1 = 20°$ as shown in Table 2. Consequently, a change is resulted and a void appears in the workspace, as shown in Fig; 4.

The Bounds on array: $Y_{min} = -2.0$, $Y_{max} = 18.0$, $Z_{min} = -2.0$ and $Z_{max} = 2.0$.

Workspace volume: 2400.96 $in^3$.

Example 3: The Bendix AA 160-CNC Manipulator. Table 3 gives the kinematic parameters of this manipulator which involves one prismatic joint.

Bounds on array: $Y_{min} = -70$, $Y_{max} = 70$, $Z_{min} = -70$, and $Z_{max} = 70$.

Workspace volume: 350036.0 $in^3$.

Figure 5 gives the graph of workspace cross-section.

Example 4: A GCA Gantry Manipulator. This manipulator, given kinematically in Table 4, contains three prismatic joints. There are four variables which are active in defining the workspace: $\theta_2$, $\theta_3$, $\theta_4$ and $\theta_5$. Because of the large number of active variables involved, the problem lends itself well to the new technique. Since the workspace is intuitively easy to see, it provides as another verification of the technique.

Bounds on array: $Y_{min} = -200.0$, $Y_{max} = 200.0$, $X_{min} = -200.0$ and $X_{max} = 200.0$

Workspace volume: 1038600.0 $in^3$.

The workspace cross-section is shown in Figure 6.

6. CONCLUSIONS. An algorithm for the generation and evaluation of a manipulator workspace based on optimum path search has been presented. Numerical examples were given and results displayed on a number of industrial robots having prismatic as well as revolute joints. Perhaps this is the first workspace algorithm based on the [3x3] dual-number matrix formulation, rather than the conventional [4x4] matrix method and applicable to manipulators with joints other than the revolute type. Besides the advantage of computational efficiency dealing with especially complicated manipulator geometries, the technique also lends itself more effectively over the previous full-scanning technique [1] for some potential applications, such as the determination of secondary workspace and the control over work area of interest during specific applications--for instance, cutting workspace in a specific plane such as the plane that a conveyor passes through [8]; the reachable area over a surface not a plane such as during assembly of an object; and accessable areas in the last two kinds of problems with specific end-effector orientation, for example, when manipulating

objects on a conveyor or installing parts. The investigation of these subjects which is of interest and of practical concern represents a continuing work of the authors.

7. ACKNOWLEDGEMENTS. The second author is grateful to the U.S. Army Research Office for the support of this research through Contract DAAG29-81-K-006 to Rutgers University.

8. REFERENCES.

[1] Lee, T.W. and Yang, D.C.H., "On the Evaluation of Manipulator Workspace," Journal of Mechanisms, Transmissions and Automation in Design, Trans. ASME. Vol. 105, No. 1, March 1983, pp. 70-77.

[2] Hansen, J.A., Gupta, K.C. and Kazerounian, S.M.K., "Generation and Evaluation of the Workspace of a Manipulator," Int'l Journal of Robotics Research, Vol. 2, No. 3, Fall 1983, pp. 22-31.

[3] Kumar, A. and Waldron, K.J., "The Workspace of a Mechanical Manipulator," Journal of Mechanical Design, Trans. ASME, Vol. 103, No. 3, July 1981.

[4] Sugimoto, K. and Duffy, J., "Determination of Extrême Distances of a Robot Hand - Part I," Journal of Mechanical Design, Trans. ASME, Vol. 103, No. 3, July 1981, pp. 631-636.

[5] Tsai, Y.C. and Soni, A.H., "Accessible Region and Synthesis of Robot Arms," Journal of Mechanical Design, Trans. ASME, Vol. 103, No. 4, Oct. 1981, pp. 803-811.

[6] Pennock, G.R. and Yang, A.T., "Dynamic Analysis of a Multi-Rigid-Body Open-Chain System," Journal of Mechanisms, Transmissions and Automation in Design, Trans. ASME, Vol. 105, No. 1, March 1983, pp. 28-34.

[7] Featherstone, R., "The Calculation of Robot Dynamics Using Articulated-Body Inertias," Int'l. J. of Robotics Research, Vol. 2, No. 1, 1983, pp. 13-30.

[8] Jou, T.M. and Waldron, K.J., "Geometric Design of Manipulators Using Interactive Computer Graphics," Proceedings of 6th World Congress of the International Federation of Theory of Machines and Mechanisms, New Delhi, India, Dec. 15-20, 1983.

[9] Fletcher, R. and Powell, M.J.D., "A Rapidly Convergent Descent Method for Minimization," British Computer Journal, Vol. 6, 1963, pp. 163-168.

[10] Gupta, K.C. and Roth, B., "Design Considerations for Manipulator Workspace," ASME Journal of Mechanical Design, Vol. 104, No. 4, Oct. 1982, pp. 704-711.

[11] Paul, B., Robot Manipulators: Mathematics, Programming and Control, MIT Press, Cambridge, MA, 1981, p. 279.

[12] Denavit, J. and Hartenberg, R.S., "A Kinematic Notion for Lower-Pair Mechanisms Based on Matrices," ASME Journal of Applied Mechanisms, June 1955, pp. 215-221.

[13] Dimentberg, F.M., The Screw Calculus and Its Applications in Mechanics, (Izdat. "Nauka." Moscow, USSR, 1965) English translation: AS680993, Clearinghouse for Federal and Scientific Technical Information, April 1968.

[14] Yang, A.T., "Calculus of Screws," Basic Question of Design Theory, North Holland Publishing, Amsterdam, 1974, pp. 266-281.

[15] Hsia, L.M. and Yang, A.T., "On the Principle of Transference in Three-Dimensional Kinematics," ASME Journal of Mechanical Design, Vol 103, No. 3, July 1981, pp. 652-656.

[16] Cwiakala, M., On the Kinematics and Dynamics and Computer-Graphics Modeling of Mechanisms, Master degree thesis, College of Engineering, Rutgers, The State University of New Jersey, New Brunswick, NJ, Oct. 1984.

## LIST OF FIGURES

Figure 1 Geometric description of a manipulator

Figure 2 Geometrical relationships between joints n and n-1 ($a_{n-1}$ = common normal, $b_{n-1}$ = axial distance and $\alpha_{n-1}$ = twist angle)

Figure 3 Circular projection of the workspace for Example 1

Figure 4 Circular projection of the workspace for Example 2

Figure 5 Circular projection of the workspace for the Bendix AA-160 CNC Manipulator

Figure 6 Projection of the workspace on the X-Y plane for the GCA Gantry Manipulator

## LIST OF TABLES

Table 1 A comparison of the Optimum Path Search Technique of this investigation with the Scanning Technique of Lee and Yang [1].

Table 2 Kinematic parameters for the 3R Manipulator

Table 3 Kinematic parameters for the Bendix AA 160-CNC Manipulator

Table 4 Kinematic parameters for the GCA Gantry XR Manipulator

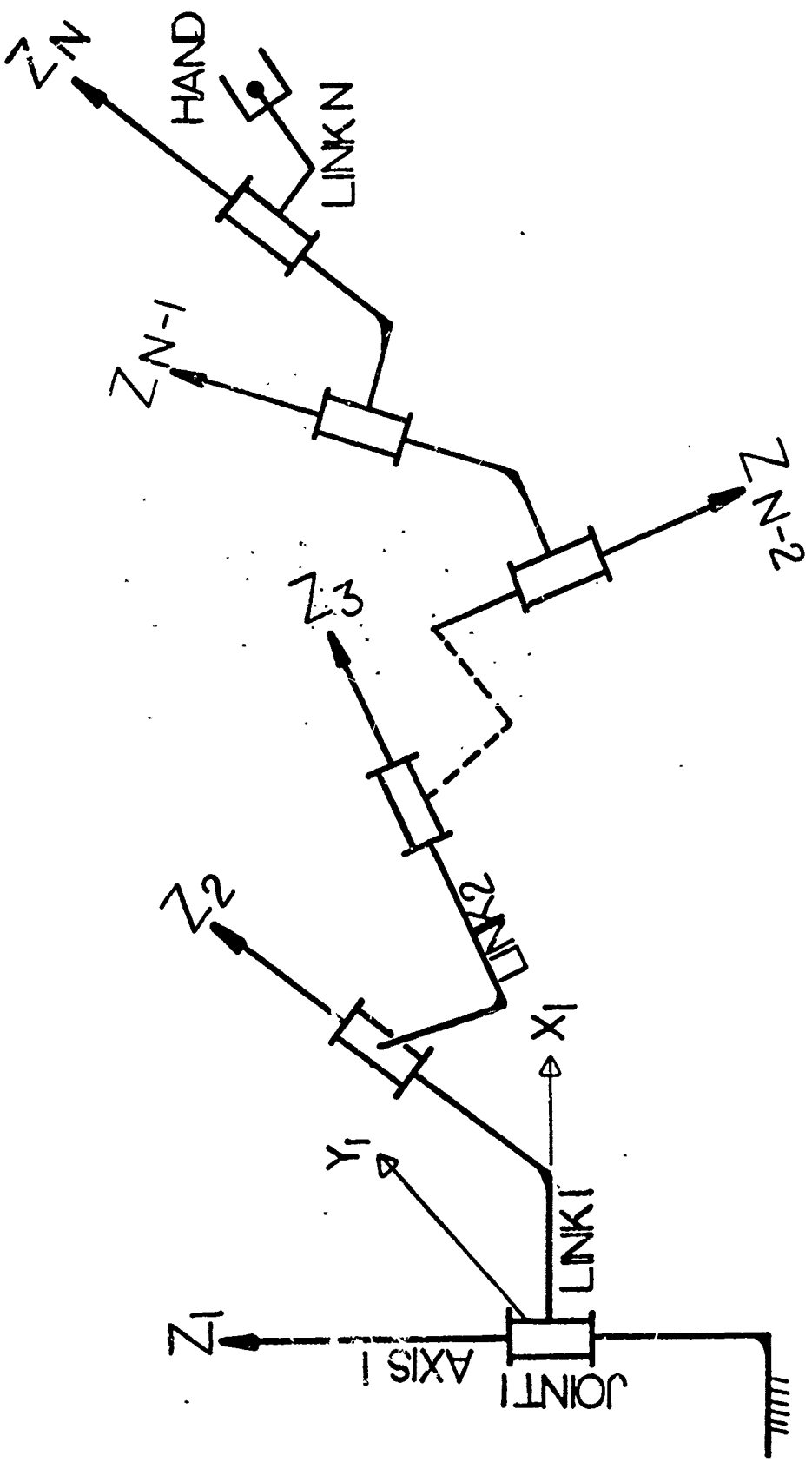



Figure 1 Geometric description of a manipulator

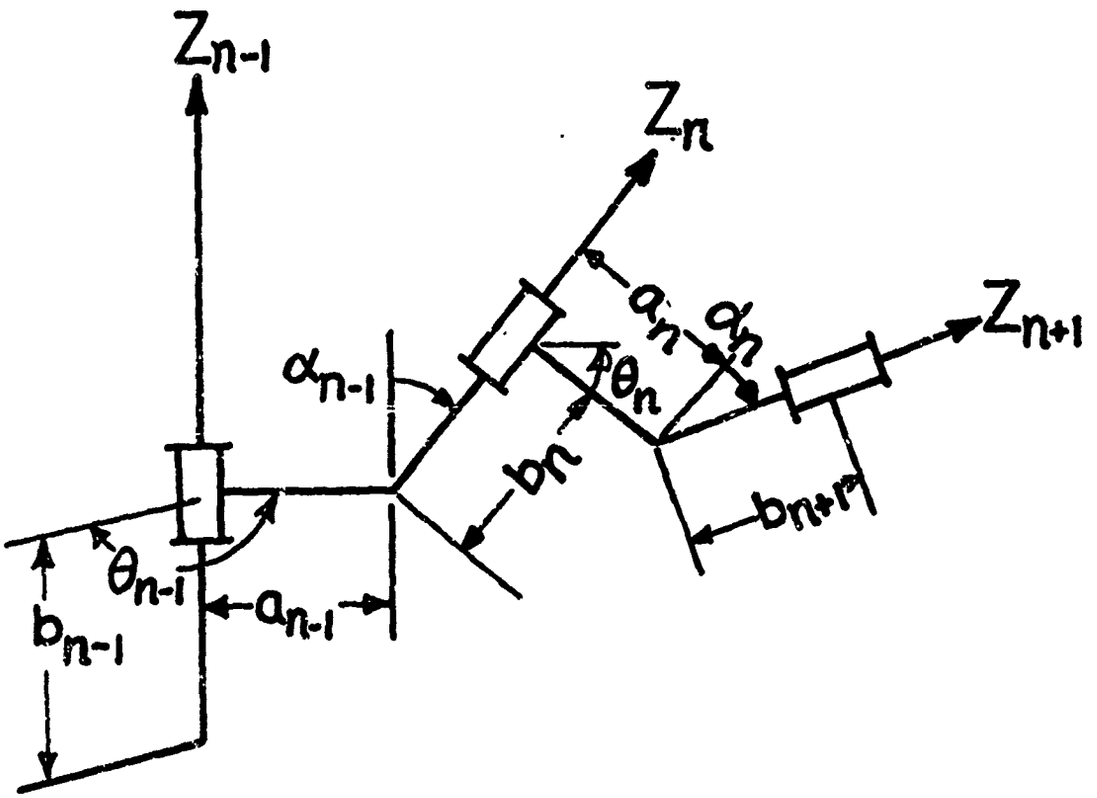



Figure 2 Geometrical relationships between joints n and n-1 ($a_{n-1}$ = common normal, $b_{n-1}$ = axial distance and $\alpha_{n-1}$ = twist angle).

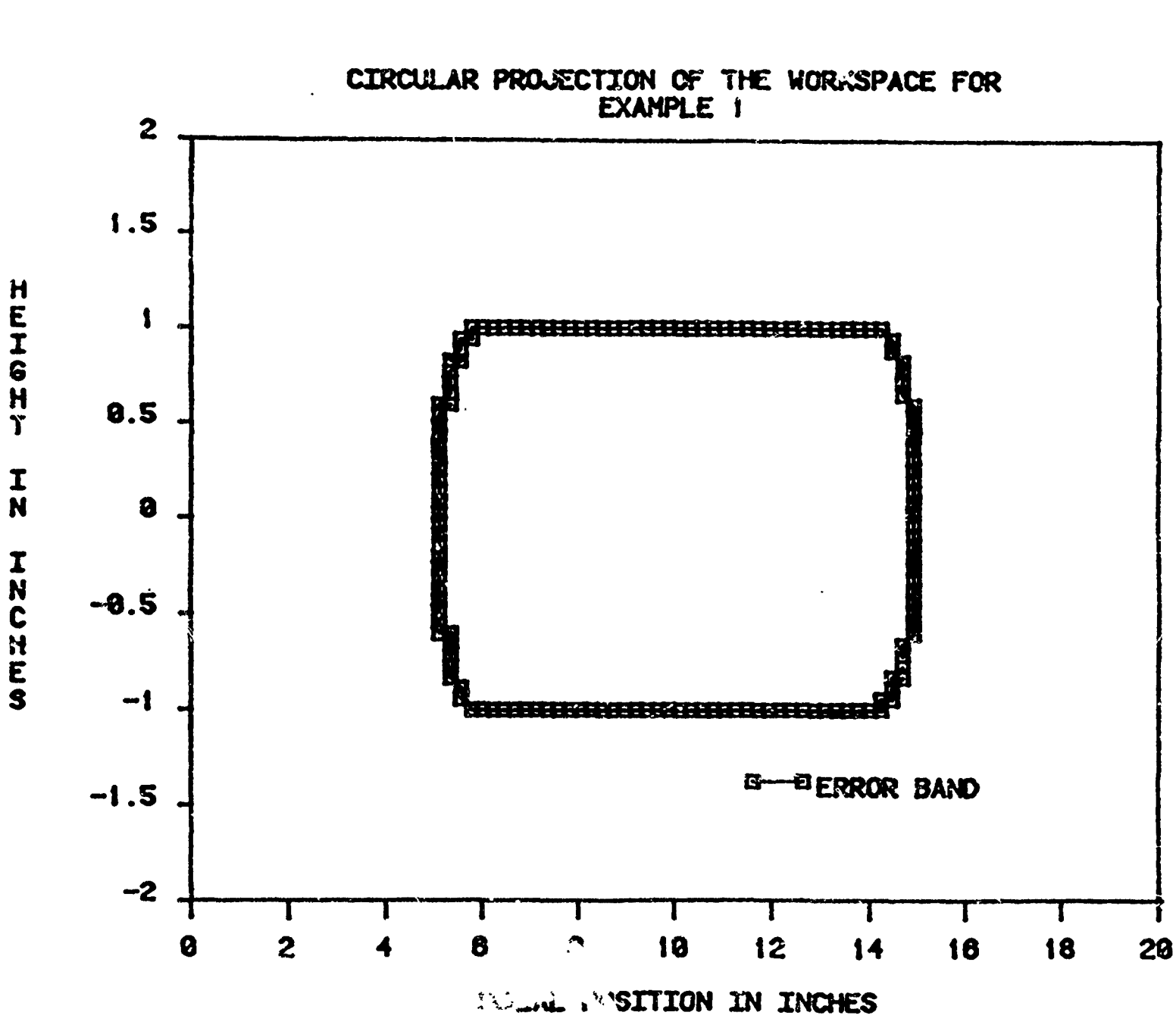



Figure 3 Circular projection of the workspace for Example 1

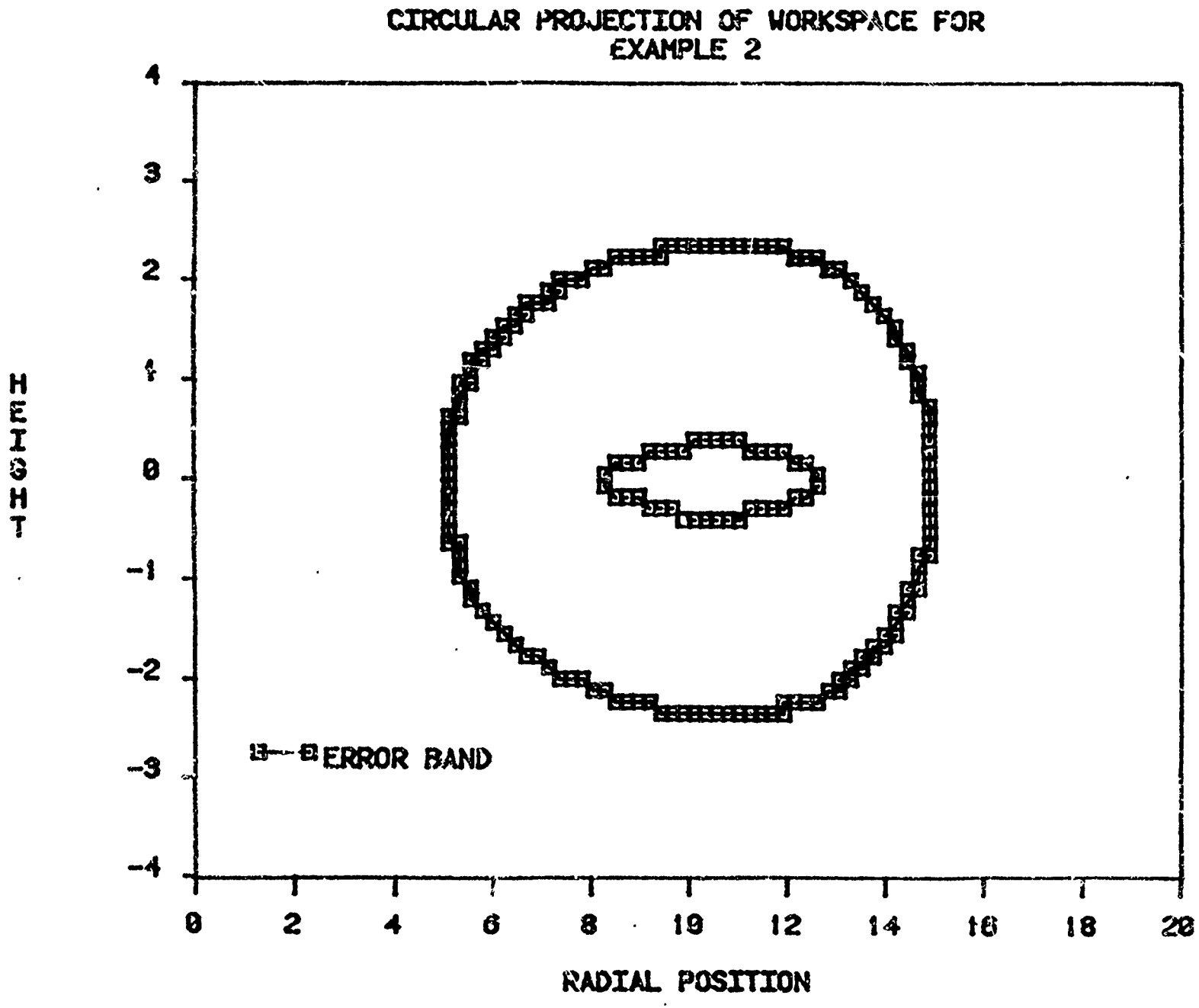



Figure 4 Circular projection of the workspace for Example 2

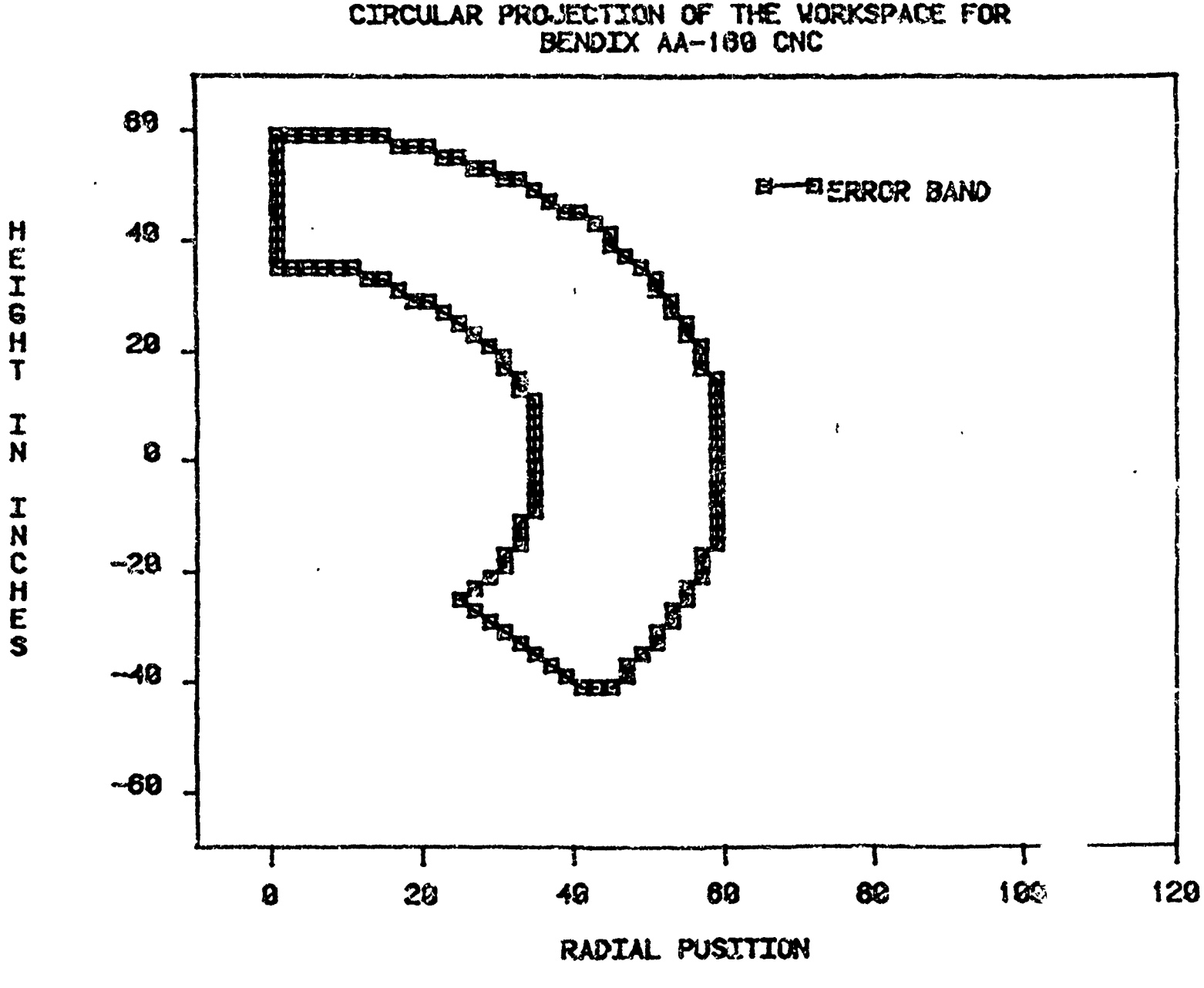



Figure 5 Circular projection of the workspace for the Bendix AA-160 CNC Manipulator

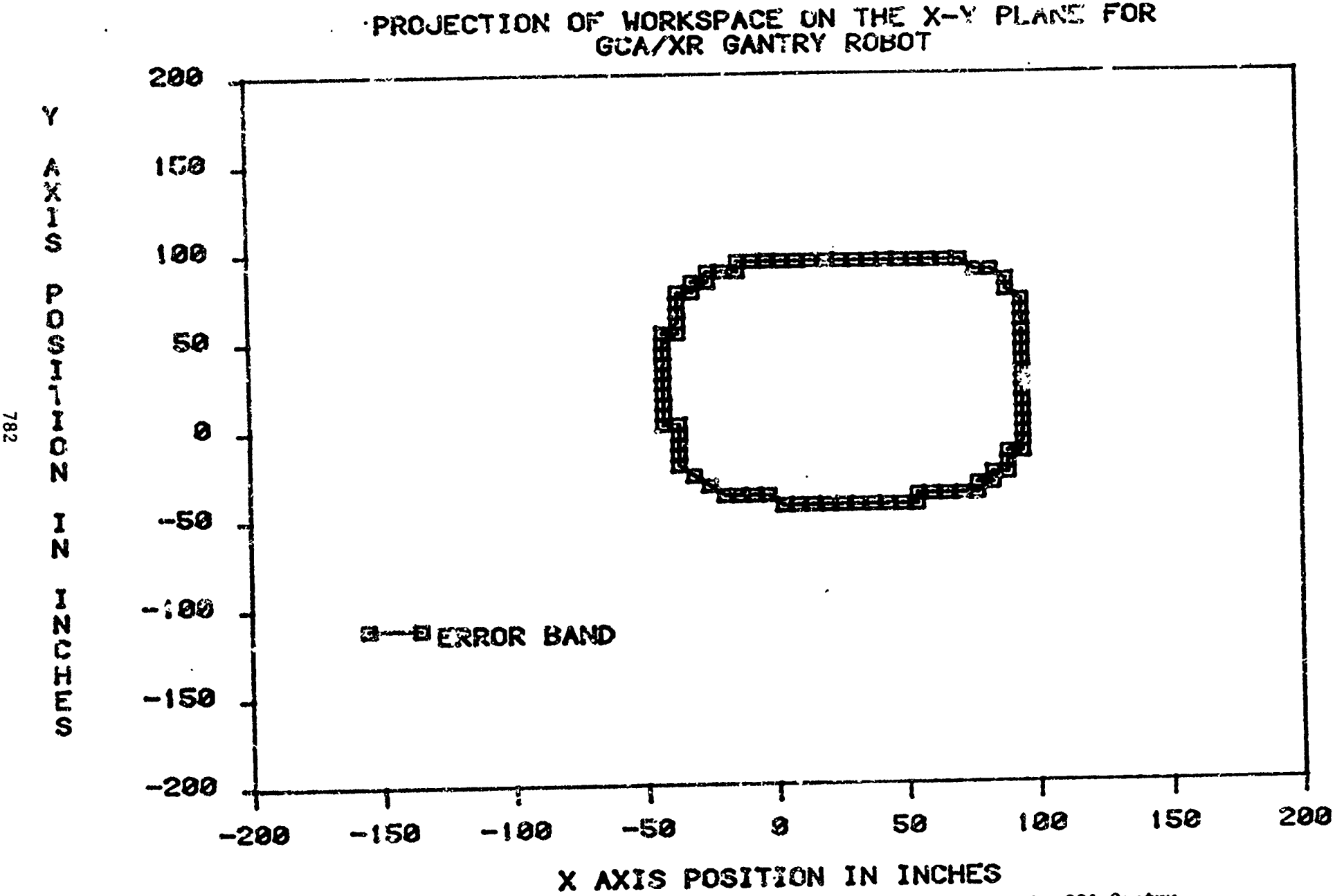



Figure 6 Projection of the workspace on the X-Y plane for the GCA Gantry Manipulator

Table 1 A comparison of the Optimum Path Search Technique of this investigation with the Scanning Technique of Lee and Yang [1].

| Property | The Scanning Technique -- Lee and Yang [1] | The Optimum Path Search -- This investigation |
|---|---|---|
| Simplicity in Programming | Yes | No |
| Computational time if grid is doubled* | $2^n$ | 2 |
| Information about reconfiguration of manipulator** | No | Yes |
| Application of analysis over a non-planar surface | Not easy | Easier |
| Application to secondary workspace | No | Yes |

* n = number of active joints to describe workspace

** It is possible to specify a path through the workspace, determining if the path can be achieved, and if at any point reconfiguration was necessary to continue.

Table 2 Kinematic parameters for the 3R Manipulator

| Joint No. n | $(\alpha_n, a_n)$ | Joint Type | Constant Portion of θ | Lower Limit | Upper Limit |
|---|---|---|---|---|---|
| 1 | (20°, 10.0) | R | 0 | -200° | 200° |
| 2 | (90°, 4.0) | R | 0 | -200° | 200° |
| 3 | (0°, 1.0) | R | 0 | -200° | 200° |
| 4 | (-90°, 0) | R | 0 | -200° | 200° |
| 5 | (90°, 0) | R | 0 | -200° | 200° |
| 6 | (0°, 0) | R | 0 | -200° | 200° |
| Hand Parameters: (α,a) = (0°,0); (θ,b) = (0°,0). | | | | | |

Table 3 Kinematic parameters for the Bendix AA 160-CNC Manipulator

| Joint No. n | $(\alpha_n, a_n)$ | Joint Type | Constant Portion of θ | Lower Limit | Upper Limit |
|---|---|---|---|---|---|
| 1 | (90°, 0) | R | 0 | -95° | 95° |
| 2 | (-90°, 0) | R | 0 | -135° | 135° |
| 3 | (0°, 0) | D | 0° | 36.0" | 60.0" |
| 4 | (-90°, 0) | R | 0 | -95° | 95° |
| 5 | (90°, 0) | R | 0 | -110° | 110° |
| 6 | (0°, 0) | R | 0 | -180° | 180° |
| Hand Parameters: (α,a) = (0°,0); (θ,b) = (0°,0). | | | | | |

Table 4. Kinematic parameters for the GCA Gantry XR Manipulator

| Joint No. n | $(\alpha_n, a_n)$ | Joint Type | Constant Portion of $\theta$ | Lower Limit | Upper Limit |
|---|---|---|---|---|---|
| 1 | (-90°, 0) | D | 0 | 0. | 56". |
| 2 | (-90°, 0) | D | -90° | 0. | 56". |
| 3 | (0°, 0) | D | 0 | 0. | 56". |
| 4 | (-90°, 0) | R | 0 | -200° | 200° |
| 5 | (90°, 0) | R | 0 | -200° | 200° |
| 6 | (0°, 0) | R | 0 | -200° | 200° |
| Hand Parameters: $(\alpha,a) = (0°,0)$; $(\theta,b) = (0°,0)$ | | | | | |

# RECURSIVE GRADIENT ESTIMATION USING SPLINES FOR NAVIGATION OF AUTONOMOUS VEHICLES

C. N. Shen*
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Command
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. Terrain gradient estimation is needed for navigation of an autonomous vehicle in climbing the hills. The in-path and cross-path terrain slopes are estimated from the set of corresponding range slopes. A two-dimensional recursive smoothing algorithm using polynomial splines in the third dimension is developed for this purpose. Approximations are introduced in the sub-optimal system so that the computation time increases only linearly with the size of the two-dimensional data.

I. INTRODUCTION. The successful development of an autonomous vision system for mobile vehicles would be of considerable value and importance to defense and related fields. Numerous reports and studies currently recommend artificial intelligence/robotics applications which require autonomous vehicles. Essential to these robotic vehicles is an adequate and efficient computer vision system. A potentially more successful approach, other than TV pictures and photographs, would be to develop a three-dimensional system employing a laser rangefinder.

A range matrix describing a certain scanned area of the terrain in front of the mobile robot can be used to estimate the slopes of the terrain. The in-path and cross-path slopes of the terrain are evaluated by a slope estimation scheme. These slope informations along the passible corridors are utilized to determine a safer and more accurate path for the mobile robot vehicle to travel.

The mobile robot vehicle is equipped with data acquisition and decision making devices for its autonomous navigation over rough terrain. A laser rangefinder can be operated by emitting laser pulses and measuring the time of flight of a pulse between the instant it was transmitted and the instant the reflected pulse is received. This time of flight is related to the distance between the transmitter and the point on the terrain from which the pulse is reflected. The terrain is scanned by changing the azimuth and elevation angles of the laser beam in a discrete fashion. The measurements are then available in the form of a NXM 'range-matrix'.

*The author is also employed by Rensselaer Polytechnic Institute, where he holds the title Professor in the Electrical Computer and Systems Engerineering Department.

PREVIOUS PAGE IS BLANK

The slope estimation problem dealt with in this section is that of obtaining smoothed estimates of function values and particularly their derivatives from a finite set of inaccurate measurements in two-dimensions. In one approach we can identify the dynamic equations of the underlying system, or estimate the distributions for the quantities of interest and then apply optimal estimation algorithms. In some engineering problems the stochastic system may not be identified easily and in these situations, spline smoothing has proved to be a useful alternative.

In this paper, we obtain the smoothed estimates of the slopes by utilizing a two-dimensional smoothing algorithm. For the problem of smoothing a finite set of noise corrupted data of an unknown function, it is proposed to obtain the smoothed estimate by fitting a two-dimensional approximating function to the data set, for a set of measurements corrupted by a white noise process.

II. HISTORICAL REVIEW. By noting the fact that original signals such as visual scenes are in analog form, techniques were developed which reconstruct analog signals from discrete data by utilizing interpolation or approximating functions. Frequency domain interpretation of the interpolation process was reported in Reference [1]. Also, B-spline interpolates [2-4] were used [5] in restoring a continuous signal from a set of digitized data. For one-dimensional noise corrupted data generated by unknown systems, Reinsch [6] utilized natural cubic splines [2-4] along with least squares constraints to solve the problem of curve plotting. Hou and Andrews [7] constructed continuous-discrete image and utilized spline basis functions along with the least squares constraints for image restoration. Because of their non-recursiveness, the algorithms in References [6,7] are involved with complex computations and cannot be implemented on line. Recently, by using a reproducing kernel Hilbert space approach, Weinert [8,9] et al, developed a structural correspondence between spline interpolation and linear least squares smoothing of a particular random process.

In recent years, two-dimensional recursive filters have drawn much attention because of the need for processing images or other two-dimensional information. Previous efforts [10-12] to achieve a truly recursive two-dimensional filter were of only limited success because of the difficulty in establishing a suitable two-dimensional recursive model as well as the high dimension of the resulting matrix and state vector. Recently, by using a two-dimensional recursive model obtained from a two-dimensional spectral factorization technique [13], Woods and Radewan [14] developed a two-dimensional Kalman vector processor and a two-dimensional Kalman scalar processor. The above mentioned time-domain design techniques assumed or identified a two-dimensional stationary discrete system model at the beginning of their problem formulation. On the other hand, Reinsch [6] interpreted a one-dimensional data smoothing problem as an optimal curve-fitting problem arising in approximation theory and proposed a nonrecursive smoothing algorithm using smoothing splines. For a two-dimensional image restoration problem, Hou and Andrews [7] followed the approach taken by Reinsch [6], and extended it to a two-dimensional problem, in a nonrecursive manner. On this paper, we develop a two-dimensional recursive smoothing algorithm. Compared to its nonrecursive counterpart, this recursive algorithm will require less

computational complexity and memory space. Especially, the amount of computation needed at each iteration is independent of the size of the two-dimensional data.

III. PROBLEM FORMULATION FOR ONE-DIMENSIONAL APPROXIMATION. From the viewpoint of approximation theory, when a set of discrete observation data is noise free, spline interpolation provides a means of optimally reconstructing an unknown original signal. When the observation data are corrupted by noises, and if the form of the original continuous signal is known, then we can use least squares estimation techniques to approximate the original signal. In this paper, we are dealing with a problem in which an unknown signal is approximated by smoothing splines from a set of noise corrupted observation data. Specifically, an unknown signal $f(\xi)$ is approximated by a polynomial spline $s(\xi)$ which minimizes the objective function:

$$J^* = \sum_{n=1}^{N} [s(\xi_n) - m_n]^T R_n^{-1} [s(\xi_n) - m_n] + \{ \sum_{n=2}^{N} \rho_n \int_{\xi_{n-1}}^{\xi_n} [s^k(\xi)]^2 d\xi \} \qquad (1)$$

where

$m_n$ is an observation data;
$m_n = f(\xi_n) + v_n$, for $n = 1, 2, \ldots, N$;
$v_n$ is a white observation noise process with error covariance $R_n$;
$R_n \overset{\Delta}{=} E\{v_n \cdot v_n^T\}$
$\rho_n > 0$ is a smoothing parameter; and
$s^k$ is the $k^{th}$ derivative of $s(\xi)$.

At this point, it is worthwhile to note the physical role of the smoothing parameter $\rho_n$ as follows: (a) when $\rho_n$ becomes very small $\rho_n \to 0^+$, the resultant approximating function will pass through each data point and become an interpolation function; (b) when $\rho_n$ assumes a very large value, $\rho_n \to \infty$, minimization of the objective function in Eq. (1) corresponds to fitting a straight line to a data set using least squares criterion. Thus, it can be said that the smoothing parameter controls resolution in a tradeoff of the smoothness of the restored function.

A. Choice of Approximating Function. As has been noted, it is desired to develop a recursive algorithm whose results are sufficiently close to those obtained by directly minimizing the global problem as given by the criterion in Eq. (1). Fundamental problems encountered in developing a recursive algorithm which generates approximating functions are:

(1) feasibility of recursive structures,
(2) feasibility of numerical calculations.

Regarding the first problem, it has been noted from References [1-5] that some of the approximating functions such as polynomial splines and piecewise Hermite polynomials have finite support. That is, a resultant approximating function for one section is mostly affected by its neighboring data points. Thus, a recursive structure with one or more sample delays would result in

sufficiently close results to nonrecursive ones.

For the second point, out of a certain set of functionals, an optimal solution to Eq. (1) is an L-spline [2-4]. L-spline is a piecewise polynomial of degree 2k-1, and has 2k-2 continuous derivatives in the region $[\xi_1, \xi_N]$. Here, we propose to restrict our approximating functions to piecewise Hermite polynomials [3] of degree 2k-1, which have k-1 continuous derivatives in the region $[\xi_1,\xi_N]$. Advantages in using piecewise Hermite polynomials are as follows. Define

$$x_i \overset{\Delta}{=} [s(\xi), s'(\xi),\ldots,s^{k-1}(\xi)]^T\Big|_{\xi=\xi_i}, \quad i = 1,\ldots,N$$

then a piecewise Hermite polynomial $s(\xi)$ is completely determined by $x_i$, $i = 1,2,\ldots,N$. For the purpose of clarity in discussion, only the case of k=2 is treated in the following. A piecewise cubic Hermite polynomial is represented as:

$$s(\xi) = \begin{cases} s_{1,2}(\xi) & \text{for } \xi_1 \leq \xi \leq \xi_2 \\ \vdots & \vdots \\ s_{N-1,N}(\xi) & \text{for } \xi_{N-1} \leq \xi \leq \xi_N \end{cases} \tag{2}$$

where

$$s_{k-1,k}(\xi) = [\phi_{k,1}(\xi)\ \psi_{k,1}(\xi)\ \phi_{k,0}(\xi)\ \psi_{k,0}(\xi)][x_k^T,\ x_{k-1}^T]^T \tag{2a}$$

$$x_k = [s(\xi_k), s'(\xi_k)]^T,\ k = 1,\ldots,N \tag{2b}$$

$$\phi_{k,1}(\xi) = (\xi-\xi_{k-1})^2[(\xi_k-\xi_{k-1}) + 2(\xi_k-\xi)]/(\xi_k-\xi_{k-1})^3 \tag{2c}$$

$$\psi_{k,1}(\xi) = (\xi-\xi_{k-1})^2(\xi-\xi_k)/(\xi_k-\xi_{k-1})^2 \tag{2d}$$

$$\phi_{k,0}(\xi) = (\xi_k-\xi)^2[(\xi_k-\xi_{k-1}) + 2(\xi-\xi_{k-1})]/(\xi_k-\xi_{k-1})^3 \tag{2e}$$

and

$$\psi_{k,0}(\xi) = (\xi-\xi_{k-1})(\xi_k-\xi)^2/(\xi_k-\xi_{k-1})^2 \tag{2f}$$

B. Smooth Integral As Quadratics at Node Points. Thus, it becomes natural that the smoothing integral in Eq. (1) is expressed in terms of $x_i$'s, $i = 1,2,\ldots,N$. With some manipulations in algebra, the smoothing integral for k = 2 in Eq. (1) is represented in a quadratic form as derived in Appendix A.

$$\int_{\xi_{n-1}}^{\xi_n} ||s''(\xi)||^2\, d\xi = (x_n - A^*x_{n-1})^T B^{-1}(x_n - A^*x_{n-1})$$

$$= \begin{bmatrix} \bar{x}_{n-1} \\ x_n \end{bmatrix}^T \cdot \begin{bmatrix} \bar{C}_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \cdot \begin{bmatrix} \bar{x}_{n-1} \\ x_n \end{bmatrix} \tag{3}$$

where

$$x_n \triangleq [s(\xi),\ s'(\xi)]^T\Big|_{\xi=\xi_n} \tag{4}$$

$$A^* = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} \frac{\Delta^3}{3} & , & \frac{\Delta^2}{2} \\ \frac{\Delta^2}{2} & , & \Delta \end{bmatrix} \tag{5a}$$

$$\Delta = \xi_n - \xi_{n-1}, \quad \text{for } n = 2,\ldots,N \tag{5b}$$

$$C_{11} \triangleq A^{*T}B^{-1}A^* \ , \quad C_{12} \triangleq C_{21}^T \triangleq -A^{*T}B^{-1}, \quad C_{22} \triangleq B^{-1} \tag{5c}$$

A nonrecursive solution for the minimization of the objective function in Eq. (1) can be obtained by taking the gradient of $J^*$ with respect to $[x_1, x_2,\ldots,x_N]^T$ and setting it to zero. However, this approach will require solutions of a set of 2N simultaneous equations. To avoid this computational problem, we developed a recursive algorithm which requires inversion of 2 x 2 matrices only.

IV. RECURSIVE ALGORITHM. Given a set of initial values for the mean $\hat{x}_1$ and its error covariance $P_1$, where $P_1 = E\{(x_1-\hat{x}_1)(x_1-\hat{x}_1)^T\}$, by using Eqs. (3) and (4) the objective function in Eq. (1) becomes

$$J_N = \sum_{n=2}^{N} [(Hx_n-m_n)^TR_n^{-1}(Hx_n-m_n)] + (x_1-\hat{x}_1)^TP_1^{-1}(x_1-\hat{x}_1) + \sum_{n=2}^{N} \rho_n(x_n-A^*x_{n-1})^TB^{-1}(x_n-A^*x_{n-1}) \tag{6}$$

where $H = (1,0)$.

Let the solutions to the above optimization problem be $[\hat{x}_1^*,\hat{x}_2^*,\ldots,\hat{x}_N^*]$. If $\hat{x}_{p|q}$ is defined as the estimate of $x_p$ obtained by minimizing Eq. (6) with N = q, then $\hat{x}_i^*$ can be written as $\hat{x}_{i|N}$. Here, it is proposed to approximate a nonrecursive solution $\hat{x}_i^* = \hat{x}_{i|N}$ by $\hat{x}_{i|i+\ell}$, where $\ell = 0$ and $\ell = N$. As has been mentioned before, due to a local base property of the polynomial splines, the smoothed estimate $\hat{x}_{i|i+\ell}$ would be sufficiently close to the nonrecursive solution $\hat{x}_i^*$ for $\ell = 1$, 2, or 3.

A. Filtering. From the definition, a filtered estimate $\hat{x}_{i|i}$ would be obtained by minimizing the objective function in Eq. (6) with $N = i$. By taking the gradient of $J_i$ with respect to $[x_1, x_2, \ldots, x_i]$, we have a set of simultaneous equations as follows:

$$
\begin{bmatrix}
G_j & \rho_j C_{12} & & & \underline{0} \\
\rho_j C_{21} & \rho_j C_{22} + \rho_{j+1} C_{11} + H^T R_{j+1}^{-1} & \rho_{j+1} C_{12} & & \\
 & \rho_{j+1} C_{21} & \ddots & \rho_{l-2} C_{12} & \\
 & & \rho_{k-2} C_{21} & \rho_{k-2} C_{22} + \rho_{k-1} C_{11} + H^T R_{k-1}^{-1} H & \rho_{k-1} C_{12} \\
\underline{0} & & & \rho_{k-1} C_{21} & \rho_{k-1} C_{22} + H^T R_k^{-1} H
\end{bmatrix}
\begin{bmatrix} x_j \\ x_{j+1} \\ \vdots \\ x_{k-1} \\ x_k \end{bmatrix}
=
\begin{bmatrix} d_j \\ H^T R_{j+1}^{-1} m_{j+1} \\ \vdots \\ H^T R_{k-1}^{-1} m_{k-1} \\ H^T R_k^{-1} m_k \end{bmatrix}
\tag{7}
$$

where $j = 1$ and $k = i$, and $G_j$ and $d_j$ are defined by

$$G_j \overset{\Delta}{=} \rho_j C_{11} + E_j^{-1} \tag{7a}$$

$$d_j \overset{\Delta}{=} E_j^{-1} \hat{x}_{j|j} \tag{7b}$$

$$\hat{x}_{j|j} = E_j[H^T R_j^{-1} m_j + \rho_{j-1} C_{21} G_{j-1}^{-1} d_{j-1}], \quad \hat{x}_{1|1} \overset{\Delta}{=} \hat{x}_1 \tag{7c}$$

$$E_j^{-1} = \rho_{j-1} C_{22} + H^T R_j^{-1} H - (\rho_{j-1} C_{21}) G_{j-1}^{-1} (\rho_{j-1} C_{12}), \quad E_1 \overset{\Delta}{=} P_1 \tag{7d}$$

It should be noted that the matrix on the left side of Eq. (7) is diagonally dominant and positive definite. Here, we are interested in solving Eq. (7) for $x_1$. Thus, by eliminating the first four equations, i.e., the variable $x_1$, Eq. (7) is reduced to the same form as itself with $j = 2$, $k = i$, and $\hat{x}_{2|2}$ and $E_2$ are calculated by Eqs. (7c) and (7d). Note that the quantity $E_j$ defined by Eq. (7d) is called pseudo error covariance, because $P_{k|k} = E\{(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T\}$ cannot be computed in a recursive manner directly.

By applying this reduction method repeatedly, the original equations in Eq. (7) are reduced to the same form as itself with $j = i-1$, $k = 1$ and every $\hat{x}_{j|j}$ and $E_j$ are computed by Eqs. (7c) and (7d) recursively. Solving Eq. (7) with $j = [=]$, $k = i$ in terms of $x_i$, we obtain a recursive estimate algorithm as

$$\hat{x}_{i|i} = E_i[H^T R_i^{-1} m_i - \rho_{i-1} C_{21} G_{i-1}^{-1} d_{i-1}] \tag{8}$$

Equation (8) is rearranged as

$$\hat{x}_{i|i} = E_i H^T R_i^{-1} m_i + F_i \hat{x}_{i-1|i-1} \tag{9}$$

where

$$F_i = -\rho_{i-1} E_i C_{21} (\rho_{i-1} C_{11} + E_{i-1}^{-1})^{-1} E_{i-1}^{-1} \tag{9a}$$

and $E_i$'s are computed by Eq. (7d) recursively. Equation (9) above is the desired filtering equation which computes $\hat{x}_{i|i}$ from the previous estimate $\hat{x}_{i-1|i-1}$ and the present measurement $m_i$. From the viewpoint of smoothing spline, the recursive filtering algorithm can be interpreted as follows. The estimate of $x_i$ obtained by fitting cubic splines to the measurement data $m_n$, $n = 2, 3, \ldots$, with the initial values $\hat{x}_1$ and $E_1 = P_1$ is the same as the estimate of $x_i$ obtained by fitting a cubic polynomial to the measurement $m_i$ with the initial values at stage $i-1$, $\hat{x}_{i-1|i-1}$, and $E_{i-1}$. In fact, the above interpretation comes from the mathematical derivations in Eq. (7) through Eq. (9).

With reference to Eqs. (7) and (8), each iteration of the recursive filtering algorithm can be interpreted as fitting a cubic polynomial to the previous estimate $\hat{x}_{i-1|i-1}$ and the present measurement $m_i$ in the region $[\xi_{i-1}, \xi_i]$.

B. Smoothing. A smoothed estimate $\hat{x}_{i|i+\ell}$ is defined as the estimate of $x_i$ obtained by solving the minimization problem in Eq. (6) with $N = i+\ell$. In fact, this can be interpreted as fitting a polynomial spline to the first $i+\ell$ data and obtaining the function value and its derivative from the approximating function at the node i. In our formulation, this corresponds to solving the simultaneous equations of the same form as Eq. (7) with $j = 1$, $k = i+\ell$ and the quantities $G_1$ and $d_1$ are defined in Eqs. (7a) and (7b). By using the same reduction method as before, the result would be the same form as itself (Eq. (7) with $j = 1$ and $k = i+\ell$).

For the case of a one-sample delay, $\hat{x}_{i|i+1}$ is obtained by solving the simultaneous equations in Eq. (7) with $j = 1$, $k = i+1$, and which yields

$$\hat{x}_{i|i+1} = V_i E_i^{-1} \hat{x}_{i|i} + K_i m_{i+1} \tag{10}$$

where

$$V_i = [-\rho_i C_{12}(\rho_i C_{22} + H^T R_{i+1}^{-1} H)\rho_i C_{21} + G_i]^{-1} \tag{10a}$$

$$K_i = -V_i \rho_i C_{12}(\rho_i C_{22} + H^T R_{i+1}^{-1} H)^{-1} H^T R_{i+1}^{-1} m_{i+1} \tag{10b}$$

and $E_i$ is defined as before.

Equation (10) is the desired smoothing algorithm, in which the smoothed estimate $\hat{x}_{i|i+1}$ is obtained by updating the filtered estimate $\hat{x}_{i|i}$ with the measurement $m_{i+1}$. The smoothing procedure described above implies the following. The smoothed estimate of $x_i$ obtained by fitting cubic splines to $m_n$, $n = 2, 3, \ldots, i+\ell$ with the initial values $\hat{x}_1$ and $E_1 = \rho_1$ is the same as the smoothed estimate of $x_i$ obtained by fitting splines to $m_n$, $n = i+1, \ldots, i+\ell$ with the filtered estimate $\hat{x}_{i|i}$ and $E_i$.

A recursive procedure to obtain smoothed estimate $\hat{x}_{i|i+1}$ is summarized as follows:

Part 1. Obtain the filtered estimates $\hat{x}_{i|i}$, $i = 2, \ldots, N$ by using Eqs. (7d), (9), and (9a).

Part 2. Smoothed estimates $\hat{x}_{i|i+1}$, for $i = 1, \ldots, N-1$ are obtained by using Eqs. (10), (10a), and (10b).

As was mentioned earlier in this section, the smoothed estimate $\hat{x}_{i|i+\ell}$ is an approximation to the nonrecursive solution $x_i^* = \hat{x}_{i|N}$. Thus, as the number of delays, $\ell$, increase, we will get a better approximation to $x_i^*$. For $\ell = 2, 3, \ldots$, only the smoothing part is modified by solving the simultaneous equations in Eq. (7) with $k = i+\ell$. In fact, the smoothing algorithm developed by far is a fixed-lag smoothing algorithm, which is suitable for an on-line implementation. If the situation does not require an on-line implementation,

we can also derive a fixed-interval smoothing algorithm with observation set M = $\{m_1,\ldots,m_N\}$. In this case, the resultant smoothed estimates become exactly the same as the nonrecursive estimates.

V. SIMULATION RESULTS. For a continuous signal

$$f(\xi) = \sin(\xi) \tag{11}$$

measurements are obtained at discrete points:

$$m_i = f(\xi_i) + v_i \quad , \quad i = 1,\ldots,100 \tag{12}$$

where $v_i$ is white Gaussian measurement noise,

$$R_i = E\{v_i v_i^T\} = 0.000025, \text{ and } \Delta\xi = \xi_n - \xi_{n-1} = 2\pi/100 = 0.062832 \tag{13}$$

Function values and the first derivatives at discrete nodes are estimated from the measurements $m_i$, $i = 1,\ldots,100$ by the three schemes below:

1. Difference quotients method.
2. Recursive smoothing algorithm with $\ell = 1$: Eq. (10).
3. Nonrecursive smoothing by cubic splines as described in Reference [6].

Table 1 shows the mean-square errors from the three schemes above.

TABLE 1. MEAN-SQUARE ERRORS

| | Difference Quotients | Recursive Smoothing by using Eq. (10) | Nonrecursive Smoothing by Cubic Splines, Ref. [6] |
|---|---|---|---|
| $\varepsilon_0$ | $2.5\times10^{-5}$ | $0.8\times10^{-5}$ | $0.57\times10^{-5}$ |
| $\varepsilon_1$ | $4.30\times10^{-1}$ | $0.12\times10^{-1}$ | $0.1007\times10^{-1}$ |

where

$$\varepsilon_0 = \frac{1}{100}\sum_{i=1}^{100}(f(\xi_i) - \hat{x}_{i|i+1}(1))^2,$$

$$\varepsilon_1 = \frac{1}{100}\sum_{i=1}^{100}(f'(\xi_i) - \hat{x}_{i|i+1}(2))^2 \tag{14}$$

From Table 1, it is noted that both smoothing algorithms are successful in reducing the error in the estimated states. The error in the first derivative is decreased by more than 10 db. Moreover, Table 1 shows that the performance of the two smoothing schemes are comparable. However, it should be emphasized that the recursive algorithm developed in this paper is much

simpler than the nonrecursive spline smoothing.

VII. PROBLEM FORMULATION FOR TWO-DIMENSIONAL APPROXIMATION. When the observation data are noise corrupted and the underlying system is unknown, it is proposed to approximate the original signal by spline functions which minimize a certain objective function. Thus, from a set of discrete measurements $m_{i,j}$ corrupted by white noise process $v_{i,j}$

$$m_{i,j} = f(\xi_i,\eta_j) + v_{i,j} \quad , \quad i = 1, 2,\ldots,N, \; j = 1, 2,\ldots,N \tag{15}$$

the original two-dimensional signal $f(\xi,\eta)$ defined in the region of $(\xi,\eta)$ is approximated by a spline function $s(\xi,\eta)$ which minimizes the following objective function:

$$J = \sum_{j=1}^{N} \sum_{i=1}^{N} [s(\xi_i,\eta_j) - m_{i,j}]^T R_{i,j}^{-1} [s(\xi_i,\eta_j) - m_{i,j}]$$

$$+ \rho\left[\int_{\eta_1}^{\eta_M} \int_{\xi_1}^{\xi_N} z(\xi,\eta)\,d\xi\,d\eta\right] \tag{16}$$

where $\rho > 0$ is the smoothing parameter; $R_{i,j}$ is the observation error covariance; and $z(\xi,\eta)$ is a certain smoothness measure of $s(\xi,\eta)$ at $(\xi,\eta)$.

A. Choice of an Approximating Function. In this paper, we are interested in obtaining smoothed estimates of function values and the first derivatives in both $\xi$ and $\eta$ directions. Here, we propose to restrict our approximating functions to piecewise bicubic Hermite polynomials which have continuous first derivatives in both $\xi$ and $\eta$ directions.

Define

$$x_{i,j} \overset{\Delta}{=} \left[s(\xi,\eta) \quad , \quad \frac{\partial s}{\partial \xi}(\xi,\eta) \quad , \quad \frac{\partial s}{\partial \eta}(\xi,\eta), \right.$$

$$\left. \frac{\partial^2 s}{\partial \xi \partial \eta}(\xi,\eta)\right]^T \Big|_{(\xi,\eta) = (\xi_i,\eta_j)} \tag{17}$$

for $i = 1, 2, \ldots,N$ and $j = 1, 2,\ldots,M$. Then, a piecewise bicubic Hermite polynomial is completely defined by $x_{i,j}$, $i = 1, 2,\ldots,N$ and $j = 1, 2,\ldots,M$, as follows:

$$s(\xi,\eta) = s_{i,j}(\xi,\eta) \quad , \quad \text{for } \xi_i \leq \xi \leq \xi_{i+1} \tag{18}$$

and

$$\eta_j \leq \eta \leq \eta_{j+1} \tag{19}$$

where

$$s_{i,j}(\xi,\eta) = \sum_{\ell=0}^{1}\sum_{m=0}^{1} \begin{bmatrix} \phi_\ell(\xi)\cdot\phi_m(\eta) \\ \psi_\ell(\xi)\cdot\phi_m(\eta) \\ \phi_\ell(\xi)\cdot\psi_m(\eta) \\ \psi_\ell(\xi)\cdot\psi_m(\eta) \end{bmatrix}^T \cdot x_{i+\ell,j+m} \tag{20}$$

B. Choice of the Smoothness Measure $z(\xi,\eta)$. Here, we present three examples of the smoothness measures and compare their physical implications.

1. Gaussian curvature: In Reference [15], the mean curvature of a surface at $(\xi,\eta)$ is defined as:

$$(0.5)\nabla^2 s(\xi,\eta) \tag{21}$$

Noting the Euler's theorem [16] that the sum of two curvatures in perpendicular directions at a point is constant, the square of $\nabla^2 s(\xi,\eta)$ in Eq. (21) would be a reasonable measure for the smoothness of a surface:

$$z(\xi,\eta) = [\frac{\partial^2}{\partial\xi^2} s(\xi,\eta) + \frac{\partial^2}{\partial\eta^2} s(\xi,\eta)]^2 \tag{22}$$

2. A variation from the Gaussian curvature: With reference to Eq. (22), an interesting case occurs when the two principal curvatures are equal and of opposite sign. The mean curvature in this case is zero. This is the so-called "saddle point" and every surface element of such a membrane is "pure twist." An appropriate smoothness measure would be changed to:

$$z(\xi,\eta) = [\frac{\partial^2}{\partial\xi^2} s(\xi,\eta)]^2 + [\frac{\partial^2}{\partial\eta^2} s(\xi,\eta)]^2 \tag{23}$$

3. In Reference [7], Hou and Andrews suggested to use $||\nabla^4 s(\xi,\eta)||^2$ as a smoothness measure for a surface. The physical interpretation of the quantity $\nabla^4 s(\xi,\eta)$ is found in a plate bending theory [16]; an unloaded plate can bend only in a biharmonic function $\omega$ where

$$\nabla^4\omega = 0 \tag{24}$$

The one which minimizes the objective function for bicubic Hermite polynomials is given in Appendix B.

C. Smoothing Integral. Now, it is needed to determine the function $s(\xi,\eta)$ which minimizes the objective function J in Eq. (16). It is noted that the smoothing integral in its present form gives difficulties in finding an explicit solution. By evaluating the integrals of the derivatives of basis functions and applying some algebraic manipulations, these smoothing integrals are converted to quadratic forms as follows:

$$\rho \int_{\eta_1}^{\eta_M}\int_{\varepsilon_1}^{\varepsilon_N} [z(\xi,\eta)]d\xi d\eta = \sum_{j=1}^{M-1}\sum_{i=1}^{N-1}\int_{\eta_j}^{\eta_{j+1}}\int_{\xi_i}^{\xi_{i+1}} [z(\xi,\eta)]d\xi d\eta$$

$$= \sum_{j=1}^{M-1}\sum_{i=1}^{N-1} (x_{i,j}^T,\ x_{i+1,j}^T,\ x_{i,j+1}^T,\ x_{2+1,j+1}^T)\cdot$$

$$C\cdot(x_{i,j}^T,\ x_{i+1,j}^T,\ x_{i,j+1}^T,\ x_{i,j+1}^T)^T \qquad (25)$$

where C is a 16 by 16 matrix.

VIII. A QUARTER-PLANE PROCESSOR. The estimate $\hat{x}_{k,\ell|p,q}$ is defined as the estimate of $x_{k,\ell}$ obtained by fitting an approximating function in the region R(p,q) where

$$R(p,q) \triangleq \{(\xi,\eta)|\ \xi_1 \leq \xi \leq \xi_p\ ,\ \text{and}\ \eta_1 \leq \eta \leq \eta_q\} \qquad (26)$$

For $(p,q) = (k,\ell)$, $\hat{x}_{k,\ell|k,\ell}$ becomes a filtered estimate. For $p \geq k$ and $q \geq \ell$, except for $p = k$ and $q = \ell$, $\hat{x}_{k,\ell|p,q}$ becomes a smoothed estimate of $x_{k,\ell}$. In our formulation, $\hat{x}_{k,\ell|p,q}$ would be obtained by minimizing the objective function $J(p,q)$ in the region R(p,q):

$$J(p,q) = \sum_{j=1}^{q}\sum_{i=1}^{p} [(Hx_{i,j}-m_{i,j})^T R_{i,j}^{-1}(Hx_{i,j}-m_{i,j})]$$

$$+ [\sum_{j=1}^{q-1}\sum_{i=1}^{p-1} (x_{i,j}^T,\ x_{i+1,j}^T,\ x_{i,j+1}^T,\ x_{i+1,j+1}^T)$$

$$\cdot\ C(x_{i,j}^T,\ x_{i+1,j}^T,\ x_{i,j+1}^T,\ x_{i+1,j+1}^T)^T] \qquad (27)$$

where $H = (1,0,0,0)$.

In a quarter-plane processor, the filtered estimate $\hat{x}_{k,\ell|k,\ell}$ is obtained by using the previous estimates of $\hat{x}_{k-1,\ell-1}$, $\hat{x}_{k-1,\ell}$, and $\hat{x}_{k,\ell-1}$, and the measurement $m_{k,\ell}$. In the next iteration, the filtered estimate $\hat{x}_{k+1,\ell|k+1}$ is obtained by using the previous estimates of $\hat{x}_{k,\ell-1}$, $\hat{x}_{k,\ell}$, and $\hat{x}_{k+1,\ell-1}$ and the measurement $m_{k+1,\ell}$. After estimating all the states in the $\ell$th column the recursive processor moves to the next column, and estimates $\hat{x}_{k,\ell+1|k,\ell+1}$,

$k = 2,\ldots,N$, and so on. First, we will discuss a filtering procedure for $\hat{x}_{k,2|k,2}$, $k = 2, 3,\ldots,N$. Then, this procedure is extended for the filtered estimates $\hat{x}_{\ell,\ell|k,\ell}$ for $\ell = 2, 3,\ldots,M$.

By definition, the filtered estimate $\hat{x}_{k,2|k,2}$ is the estimate of $x_{k,2}$ obtained from an approximating function which minimizes the objective function in Eq. (27) for $(p,q) = (k,2)$. For minimization, we take the gradient of $J(k,2)$ with respect to $\underline{x}_1$ and $\underline{x}_2$, and set it to zero

$$\nabla_{\begin{pmatrix}\underline{x}_1\\ \underline{x}_2\end{pmatrix}} J(k,2) = 0$$

where

$$\underline{x}_j = [x_{1,j}^T,\ x_{2,j}^T,\ldots,x_{k,j}^T]^T \tag{28}$$

We can obtain a final recursive estimation equation as:

$$\begin{pmatrix}\hat{x}_{k,1|k,2}\\ \hat{x}_{k,2|k,2}\end{pmatrix} = E_{k,2}\begin{pmatrix}P_{k,1}^{-1}\,\hat{x}_{k,1}\\ H^T R_{k,2}^{-1}\,m_{k,2}\end{pmatrix} + F_{k,2}\begin{pmatrix}\hat{x}_{k-1,1|k-1,2}\\ \hat{x}_{k-1,2|k-1,2}\end{pmatrix} \tag{29}$$

For notation used in the above equation, see reference [17]. With reference to the final estimation equation above, it is noted that for the filtered estimate $\hat{x}_{k,2|k,2}$ the scheme uses the previous estimates $\hat{x}_{k,1}$, $\hat{x}_{k-1,1|k-1,2}$, and $\hat{x}_{k-1,2|k-1,2}$, and the measurement $m_{k,2}$. Here, it should be emphasized that the scheme uses the smoothed estimate $\hat{x}_{k-1,1|k-1,2}$ instead of $\hat{x}_{k-1,1}$. Thus, when the filtered estimate $\hat{x}_{k,2|k,2}$ is computed, it is needed to update the estimate $\hat{x}_{k,1}$ to $\hat{x}_{k,1|k,2}$ for use in the next iteration. Now, by using the recursive estimation equation in Eq. (28) and the pseudo error covariance equation in Reference [17], we can compute $\hat{x}_{k,2|k,2}$, $k = 2,\ldots,N$ recursively.

The resultant recursive filtering equation for $\hat{x}_{k,3|k,3}$ becomes similar to the one in Eq. (28). Also, the smoothing equation for $\hat{x}^*_{i,3/i+1,3}$ becomes similar to the one for $\hat{x}^*_{i,2/i+1,2}$. After all the estimates in the third column, $\hat{x}^*_{1,3|k,3}$ and $\hat{x}^*_{k,3|k+1,3}$ for $k = 1, 2,\ldots,N$ are obtained, the smoothed estimates $\hat{x}^*_{k,3|k+1,3}$, $k = 1,\ldots,N-1$, will be used for the estimates in the fourth column, $\hat{x}^*_{k,4|k,4}$ and $\hat{x}^*_{k,4|k+1,4}$, and so on.

The approximation method described above is one of the simplest ones. We can employ more elaborate approximation methods at the cost of more complicated computations. By now, we have introduced a recursive quarter-plane processor which computes the filtered estimate $\hat{x}^*_{k,\ell|k,\ell}$ as an approximation to $\hat{x}_{k,\ell|k,\ell}$. From the definition of the estimate $\hat{x}_{k,\ell|p,q}$, the value of $x_{k,\ell}$ which minimizes J is $\hat{x}_{k,\ell|N,M}$. Here, we note that $\hat{x}^*_{k,\ell|k,\ell}$ has its support in the region $R_{(k,\ell)}$, while the nonrecursive solution $\hat{x}^*_{k,\ell|N,M}$ has its support in the region $R_{(N,M)}$. Thus if we desire to have a better approximation to $\hat{x}_{k,\ell|N,M}$, it is needed to develop a smoothing algorithm which computes $\hat{x}_{k,\ell|k+d_1,\ell+d_2}$ where $d_1,d_2 \geq 1$, $k+d_1 \leq N$ and $\ell+d_2 \leq M$.

The smoothed estimate $\hat{x}_{k,\ell|k+d_1,\ell+d_2}$ is defined as the estimate of $x_{k,\ell}$ obtained by fitting an approximating function in the region $R_{(k+d_1,\ell+d_2)}$. It has been derived that $\hat{x}_{k+d_1,\ell+d_2}$ can be reasonably approximated by $\hat{x}_{k+d_1,\ell+d_2}$ where $\hat{x}_{k+d_1,\ell+d_2}$ is obtained by fitting an approximating function to a smaller region. The derivation procedure for the above approximation is similar to that of filtering discussed previously, and is omitted for conciseness.

## IX. FURTHER NAVIGATION PROBLEMS.

A. Terrain Slopes and Range Slopes. With reference to Figure 1, terrain in-path and cross-path slopes are defined as the two orthogonal slopes $dz/d\rho$ and $dz/\rho d\theta$ in a cylindrical coordinate system. During the past investigations, the terrain slopes were found to be appropriate measures for evaluating a terrain. A direct approach for estimating the terrain slopes would be to fit a smoothing spline to the measurement data in cylindrical coordinates. However, there is a major difficulty in this approach. Even though the two independent variables $\beta_i$ and $\theta_j$ for the rangefinder are changing with constant increments $\Delta\beta$ and $\Delta\theta$, respectively, the independent variable $\rho_{ij}$ in a cylindrical coordinate changes irregularly. The recursive smoothing algorithm in the previous subsection requires that the data points be located at the corners of rectangular grids of the two independent variables. Since the two independent variables $\rho_i$ and $\theta_j$ in a cylindrical coordinate system do not form rectangular grids, the smoothing algorithm cannot be applied directly. By noting that the positioning angles $\beta_i$ and $\theta_j$ are changing in regular fashion, it is proposed to obtain the smoothed estimates of the range slopes $dr/d\beta$ and $dr/d\theta$ defined in spherical coordinates. Then, these estimates are transformed to the terrain slopes. In applying the smoothing algorithm to terrain slope estimation, one point to be mentioned is that the basic philosophy of the smoothing spline approach is to suppress the noise elements by fitting a smooth approximating function to a noise corrupted data set. Thus, when the function to be approximated has

sharp changes in its values or derivatives, the smoothing algorithm will produce errors in the results by smoothing out these actual sharp changes. From the viewpoint of terrain slope estimation, such changes occur at the edges of a boulder, a crater, or a ridge on the terrain. Thus, it is proposed to detect these edges by using the rapid estimation scheme. Then, for the area which is free of discrete edges, the two-dimensional smoothing algorithm is utilized to estimate the slopes. The terrain slopes are estimated in the order: (1) discrete edges are detected by using the rapid estimation scheme; (2) for the area which is free of discrete edges, a two-dimensional smoothing algorithm is utilized to estimate the range slopes; (3) estimated range slopes are transformed into terrain slopes.

B. Estimated Terrain In-Path Slope. The simulation of terrain with hills and valleys is given in Figure 2. The estimated terrain in-path slopes [18] are displayed in terms of a slope map, Figure 2. Characters A,...,G represent a particular range of the terrain in-path slopes increasing from A to G, at the corresponding location. U represents undefined slopes. In Figure 3, we note circular slope regions on the faces of sinusoidal hills and valleys. Also, along a radial direction, the estimated slopes are changing slowly from one region of slopes to another. The large empty spaces are due to the hidden regions at the back of boulders or hills where laser rays could not reach. The undefined gradient represented by 'U' occurs when the recursive algorithm cannot be applied due to sharp changes in ranges between adjacent measurement data. The estimated in-path terrain slope maps are used for the evaluation of the terrain in front of the mobile robot vehicle.

C. Terrain Cross-Path Slopes. In discussing the terrain cross-path slopes [19], the data can be conveniently processed to generate smoothed in-path and cross-path range slopes recursively in spherical coordinate system due to the regularity of the elevation and azimuth angles. When we proceed to calculate the true terrain slopes on the base plane, the regularity of the data points are completely destroyed. For a fixed elevation angle $\beta$, the horizontal projection of the range data are not located at a fixed distance from the rover. It is desired to calculate the cross-path slope at point $(\beta_i,\theta_j)$. However, in general, $\rho_{i,j} \neq \rho_{i,j+1}$. Thus, the true cross-path slope is not along an arc connecting points $\rho_{i,j}$ and $\rho_{i,j+1}$.

Our algorithm to calculate the terrain cross-path slope can then be summarized as follows:

1. Obtain the range measurements.
2. Use the smoothing algorithm to calculate the range cross-path slope.
3. Obtain the terrain cross-path slope and its variance.

D. Evaluation of Terrain Variables. As mentioned in the introduction, in the evaluation of terrain variables, we only use the slopes at the spine and track points. The reason is two-fold. First, only a minor part of this path selection scheme need the data of elevation. Second, if we adopt the elevation estimates as our input data, we will get larger errors in the calculation of in-path and tilt slope terrain variables.

The terrain variables and the variances together with the corresponding explanations are listed below [20].

1. In-Path Terrain Variables The in-path slope terrain variable gives the average of the in-path slopes for the four vehicle wheels at each section. This variable is a measure of the risk in the forward direction.

2. Tilt Slope. Tilt slope terrain variable is used to estimate excessive cross-path slopes which may cause the vehicle to tip over.

3. Obstruction Height. The obstruction height is calculated for six different locations at each discrete section of the terrain. The maximum value is then chosen as representative of this whole section.

In deriving the formula for a typical obstruction height, we use a third order polynomial to approximate the terrain elevation in each location. By differentiating this polynomial with respect to the distance, we get an expression for the slope. With the known data of slopes at the three points substituted into this expression, we can determine the coefficients of the polynomial. Using this polynomial, we can then find the obstruction height in this direction.

4. Wheel Deviation. The wheel deviation variable describes the offset of any of the four wheels from a plane. Wheels on any three track points define a plane. For each combination of three wheels touching the terrain, the deviation of the fourth wheel with respect to this plane is defined as the wheel deviation.

A set of the measurement data is obtained by the described scanning scheme. The range measurement data are processed by gradient estimation scheme to evaluate in-path and cross-path slopes at the data points. Since the slopes are estimated in the spherical coordinate system, it is needed to transform the range slope in spherical coordinate system to terrain slope in cylindrical coordinate. The in-path and cross-path slopes and their covariances at the spline and track points along the corridors are evaluated by applying two-dimensional interpolation scheme over the estimated slopes at the data points. The terrain variables at a discrete section along each corridor are computed by using estimated slopes at the spline and track points. Since the terrain variable estimates have uncertainty, the present method increases the reliability by considering standard deviation as well as their mean values.

X. CONCLUSION. By taking an algebraic approach, a recursive smoothing algorithm was developed as an approximation to nonrecursive spline smoothing. Compared to the recursive smoothing algorithm suggested by Weinert, the smoothing algorithm in this paper is simpler in that the scheme is in a discrete form. Simulation result shows that the performance of the recursive smoothing algorithm is comparable to that of its nonrecursive counterpart. In addition, the computational complexity with recursive smoothing algorithm is much less than its nonrecursive one. Also, recursive smoothing by splines can be implemented on-line. By taking an algebraic approach, a two-dimensional recursive smoothing algorithm was developed as an approximation to a

nonrecursive smoothing spline technique. While the amount of computation required for a nonrecursive algorithm increases rapidly with the size of the two-dimensional data, the amount of computation for this smoothing algorithm increases only linearly.

## REFERENCES


1. Schafer, R. W. and Rabiner, L. R., "A Digital Signal Processing Approach to Interpolation," Proc. IEEE, Vol. 61, No. 6, June 1973.

2. Ahlberg, J. H., Nilson, E. N., and Walsh, J. L., The Theory of Splines and Their Application, Academic Press, Inc., 1967.

3. Prenter, P. M., Splines and Variational Methods, John Wiley & Sons, Inc., 1975.

4. Greville, T. N. E., Theory and Applications of Spline Functions, Academic Press, New York, 1969.

5. Hou, H. S. and Andrews, H. C., "Cubic Splines for Image Interpolation and Digital Filtering," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-26, December 1978.

6. Reinsch, C. H., "Smoothing by Spline Functions," Numer. Math., Vol. 10, 1967, pp. 177-183.

7. Hou, H. S. and Andrews, H. C., "Least Squares Image Restoration Using Spline Basis Functions," IEEE Trans. on Computers, Vol. C-26, September 1977, pp. 856-873.

8. Weinert, H. L. and Sidhu, G. S., "A Stochastic Framework for Recursive Computation of Spline Functions: Part I - Interpolating Splines," IEEE Trans. on Information Theory, Vol. IT-24, No. 1, January 1978.

9. Weinert, H. L., Byrd, R. H., and Sidhu, G. S., "A Stochastic Framework for Recursive Computation of Spline Functions: Part II - Smoothing Splines," Journal of Optimization Theory and Applications, Vol. 30, No. 2, February 1980.

10. Nahi, N. E., "Role of Recursive Estimation in Statistical Image Enhancement," Proc. IEEE, Vol. 60, pp. 872-877.

11. Habibi, A., "Two-dimensional Bayesian Estimate of Images," Proc. IEEE, Vol. 60, July 1972, pp. 878-883.

12. Strintzis, M. G., "Comments on Two-Dimensional Bayesian Estimates of Images," Proc. IEEE, Vol. 64, August 1976, pp. 1255-1257.

13. Ekstrom, M. P. and Woods, J. W., "Two-Dimensional Spectral Factorization with Application in Recursive Digital Filtering," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 2, April 1976.

14. Woods, J. W. and Radewan, C. H., "Kalman Filtering in Two-Dimension," IEEE Trans. Information Theory, Vol. IT-23, July 1977, pp. 473-482.

15. Marfredo, P. do Carmo, Differential Geometry of Curves and Surfaces, Prentice-Hall, Inc., 1976.

16. Den Hartog, J. P., Advanced Strength of Materials, McGraw-Hill, Inc., 1952.

17. Kim, C. S., "One-Dimensional and Two-Dimensional Recursive Digital Filters Using Spline Functions," Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, NY, August 1980.

18. Kim, C. S. and Shen, C. N., "Estimating Planetary Terrain Slopes From Range Measurements Using a Two-Dimensional Spline Smoothing Technique," Proceedings of the 8th Triannual World Congress, International Federation of Automatic Control, Kyoto, Japan, August 1981.

19. Shen, C. N. and Shen, Poueau, "The Estimation of Terrain Cross-Path Slopes," Proceedings of the 11th Annual Pittsburgh International Conference on Modeling and Simulation, University of Pittsburgh, May 1980.

20. Shen, Poueau and Shen, C. N., "Modified Vector Space Algorithms Applied to Path Selection Scheme," Proceedings of the 11th Annual Pittsburgh International Conference on Modeling and Simulation, University of Pittsburgh, May 1980.

## APPENDIX A

## EVALUATION OF SMOOTHING INTEGRALS

From Eqs. (2a) through (2e) of the text, a piecewise cubic Hermite polynomial in the section $[\xi_{i-1},\xi_i]$ is represented as:

$$s_{i-1,i}(\xi) = \begin{bmatrix} \phi_{i,1}(\xi) \\ \psi_{i,1}(\xi) \\ \phi_{i,0}(\xi) \\ \psi_{i,0}(\xi) \end{bmatrix}^T \cdot \begin{bmatrix} s(\xi_i) \\ s'(\xi_i) \\ s(\xi_{i-1}) \\ s'(\xi_{i-1}) \end{bmatrix} \tag{A-1}$$

where $s(\xi_{i-1})$, $s'(\xi_{i-1})$, $s(\xi_i)$, and $s'(\xi_i)$ are the function values and first derivatives at the nodes i-1, and i. We make the change of variable such as

$$\mu = \xi - \xi_{i-1} \tag{A-2}$$

This change of variables does not affect the value of the smoothing integral and results in a simpler computation. The smoothing integral in the interval $[\xi_{i-1},\xi_i]$ is

$$I_{i-1,j} = \int_{\xi^+_{i-1}}^{\xi^-_i} ||s''_{i-1,j}(\xi)||^2 \, d\xi = \int_{0^+}^{\Delta^-} ||s''_{i-1,j}(\mu)||^2 d\mu \tag{A-3}$$

where

$$\Delta = \mu_i - \mu_{i-1} = \xi_i - \xi_{i-1}$$

Using Eq. (A-2) and Eq. (2a) of the text, Eq. (A-1) becomes

$$s_{i-1,i}(\mu) = [\phi_{i,1}(\mu)\psi_{i,1}(\mu)\ \phi_{i,0}(\mu)\ \psi_{i,0}(\mu)][x_i^T\ x_{i-1}^T]^T \tag{A-4}$$

Thus, the norm square of the second derivative is written as:

$$||s''_{i-1,i}(\mu)||^2 = [x_i^T\ x_{i-1}^T]^T \cdot$$

$$\begin{bmatrix} \phi''_{i,1}(\mu)\phi''_{i,1}(\mu), & \phi''_{i,1}(\mu)\psi''_{i,1}(\mu), & \phi''_{i,1}(\mu)\phi''_{i,0}(\mu), & \phi''_{i,1}(\mu)\psi''_{i,0}(\mu) \\ \psi''_{i,1}(\mu)\phi''_{i,1}(\mu), & \psi''_{i,1}(\mu)\psi''_{i,1}(\mu), & \psi''_{i,1}(\mu)\phi''_{i,0}(\mu), & \psi''_{i,1}(\mu)\psi''_{i,0}(\mu) \\ \phi''_{i,0}(\mu)\phi''_{i,1}(\mu), & \phi''_{i,0}(\mu)\psi''_{i,1}(\mu), & \phi''_{i,0}(\mu)\phi''_{i,0}(\mu), & \phi''_{i,0}(\mu)\psi''_{i,0}(\mu) \\ \psi''_{i,0}(\mu)\phi''_{i,1}(\mu), & \psi''_{i,0}(\mu)\psi''_{i,1}(\mu), & \psi''_{i,0}(\mu)\phi''_{i,0}(\mu), & \psi''_{i,0}(\mu)\psi''_{i,0}(\mu) \end{bmatrix} \cdot$$

$$\begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix} = \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix}^T \cdot \begin{bmatrix} k_{i-1,i}(\mu) \end{bmatrix} \cdot \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix} \tag{A-5}$$

By utilizing Eq. (A-5), Eq. (A-3) becomes:

$$I_{i-1,i} = \int_{0^+}^{\Delta^-} ||s''_{i-1,i}(\mu)||^2 d\mu = \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix}^T \cdot \begin{bmatrix} \int_{0^+}^{\Delta^-} K_{i-1,i}(\mu) d\mu \end{bmatrix} \cdot \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix} \tag{A-6}$$

Thus, smoothing integral is obtained as:

$$I_{i-1,1} = \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix}^T \cdot \begin{bmatrix} 12\Delta^{-3} & -6\Delta^{-2} & -12\Delta^{-3} & -6\Delta^{-2} \\ -6\Delta^{-2} & 4\Delta^{-1} & 6\Delta^{-2} & 2\Delta^{-1} \\ -12\Delta^{-3} & 6\Delta^{-2} & 12\Delta^{-3} & 6\Delta^{-2} \\ -6\Delta^{-2} & 2\Delta^{-1} & 6\Delta^{-2} & 4\Delta^{-1} \end{bmatrix} \cdot \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix} \tag{A-7}$$

By defining $B^{-1}$ and $A^*$ as below:

$$B^{-1} = \begin{bmatrix} 12\Delta^{-3} & -6\Delta^{-2} \\ -6\Delta^{-2} & 4\Delta^{-1} \end{bmatrix} \qquad A^* = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \tag{A-8}$$

Equation (A-7) is rewritten in the following form as:

$$I_{i-1,i} = \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix}^T \cdot \begin{bmatrix} B^{-1} & -B^{-1}A^* \\ -A^{*T}B^{-1} & A^{*T}B^{-1}A^* \end{bmatrix} \cdot \begin{bmatrix} x_i \\ x_{i-1} \end{bmatrix} \tag{A-9}$$

$$= \begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix}^T \cdot \begin{bmatrix} A^{*T}B^{-1}A^* & -A^{*T}B^{-1} \\ -B^{-1}A^* & B^{-1} \end{bmatrix} \cdot \begin{bmatrix} x_{i-1} \\ x_i \end{bmatrix} \tag{A-10}$$

$$= (x_i - A^*x_{i-1})^T B^{-1}(x_i - A^*x_{i-1}) \tag{A-11}$$

which is Eq. (3) in the text.

## APPENDIX B

In this Appendix, we show that a piecewise bicubic Hermite polynomial which minimizes the objective function in Eq. (B-1) becomes a bicubic spline.

$$J = J_E + \rho J_s \tag{B-1}$$

where

$$J_E = \sum_{j=1}^{M} \sum_{i=1}^{N} [s(\xi_i,\eta_j) - m_{i,j}]^T R_{ij}^{-1} [s(\xi_i,\eta_j) - m_{i,j}] \tag{B-2}$$

and

$$J_s = \int_{\eta_1}^{\eta_M} \int_{\xi_1}^{\xi_N} \left(\frac{\partial^4}{\partial\xi^2\partial\eta^2} s(\xi,\eta)\right)^2 d\xi d\eta \tag{B-3}$$

Let a set S be a collection of all piecewise bicubic Hermite polynomials. Also, we define a set U as a collection of all piecewise bicubic Hermite polynomials which satisfy constraints set D in Eq. (B-4).

$$\begin{aligned}
&s(\xi_i,\eta_j) = c(i,j) && , \ i = 1,2,\ldots,N \text{ and } j = 1,2,\ldots,M \\
&\partial s(\xi_i,\eta_j)/\partial\xi = c_\xi(i,j) && , \ j = 1,2,\ldots,M \text{ and } i = 1,N \\
&\partial s(\xi_i,\eta_j)/\partial\eta = c_\eta(i,j) && , \ i = 1,2,\ldots,N \text{ and } j = 1,M \\
&\partial^2 s(\xi_i,\eta_j)/\partial\xi\partial\eta = c_{\xi,\eta}(i,J) && , \ i = 1,N \text{ and } j = 1,M
\end{aligned} \tag{B-4}$$

Then the minimizing problem in Eq. (A-1) is rewritten as:

$$\underset{s(\xi,\eta)\in S}{\text{Min}} \ J = \underset{s(\xi,\eta)\in S}{\text{Min}} \ [J_E + \rho J_s] = \underset{D}{\text{Min}} \ [J_E + \rho \underset{s(\xi,\eta)\in U}{\text{Min}} \ J_s] \tag{B-5}$$

In the paper by DeBoor [21], it is noted that there exists a unique bicubic spline $g(\xi,\eta)$ in the set U. Also, by using a standard technique to derive the minimum norm property [2] of a bicubic spline, it can be shown that:

$$J_s = \int_{\eta_1}^{\eta_M} \int_{\xi_1}^{\xi_N} \left(\frac{\partial^4 g(\xi,\eta)}{\partial\xi^2\partial\eta^2}\right)^2 d\xi d\eta \tag{B-6}$$

Since the bicubic spline $g(\xi,\eta)$ is unique, we have the following Lemma:

Lemma 1: A bicubic Hermite polynomial $s(\xi,\eta)\in U$ which minimizes the smoothing integral $J_s$, becomes a bicubic spline $g(\xi,\eta)$.

With reference to Eq. (B-5) and Lemma 1, we conclude that a piecewise bicubic Hermite polynomial $s(\xi,\eta)$, which minimizes Eq. (B-5), becomes a cubic spline.

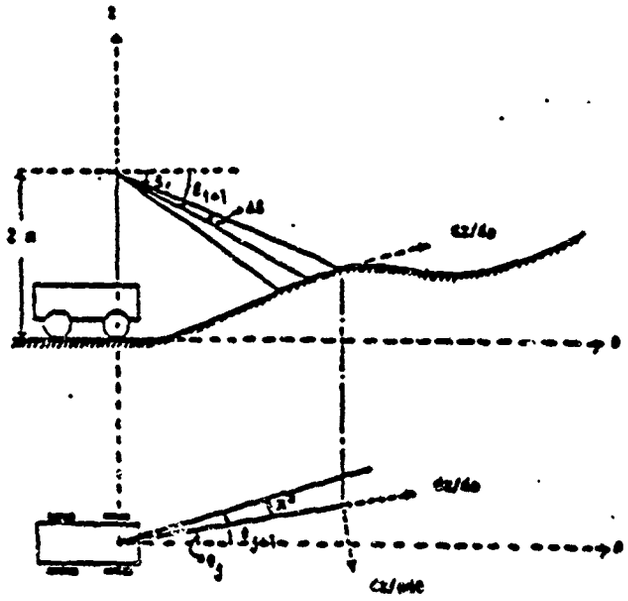


Fig. 1 Top and side views of a rangefinder



Fig. 2 A slope map in the x-y plane for the inpath terrain slopes

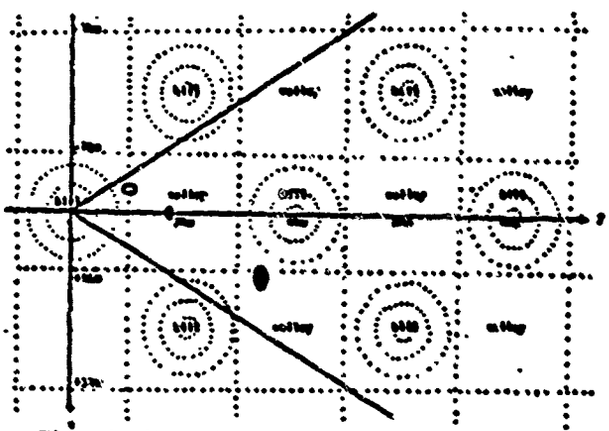


Fig. 3 Top view of the terrain model used for simulation.

# ON PHASE TRANSITIONS WITH INTERFACIAL ENERGY

Morton E. Gurtin*
Department of Mathematics
Carnegie-Mellon University
Pittsburgh, PA 15213

Consider a fluid which has free energy $\psi(\rho)$ a prescribed function of density $\rho$, and which occupies a fixed container $\Omega$, with $\Omega$ a bounded region in $\mathbb{R}^3$. If we neglect all other contributions to the free energy, then the total energy $E_0(\rho)$ corresponding to a density distribution $\rho(x)$, $x \in \Omega$, is

$$E_0(\rho) = \int_\Omega \psi(\rho(x))dx \quad . \tag{1.1}$$

If the fluid in $\Omega$ has mass $m$, and fluid is neither added to - nor removed from - $\Omega$, then the allowable density distributions must be consistent with the constraint

$$\int_\Omega \rho(x)dx = m \quad . \tag{1.2}$$

Following Gibbs, we postulate that the stable configurations of the fluid are those which minimize (1.1) subject to (1.2). Thus we are led to the problem:

$$P_0 \begin{cases} \underline{\text{minimize the energy}}\ E_0(\rho)\ \underline{\text{over}} \\ \underline{\text{all sufficiently regular fields}}\ \rho \\ \underline{\text{that satisfy the constraint}}\ (1.2) \quad . \end{cases}$$

Here we shall be concerned with situations in which $\psi(\rho)$ is nonconvex, of a form capable of supporting two phases (Figure 1). In this instance the solution $\rho_0$ is most easily described in terms of the Maxwell parameters $\alpha_0$, $\beta_0$ and $\mu_0$ defined by the conditions

---

Present address: Mathematics Research Center, University of Wisconsin-Madison, Madison, WI 53705.

$$\psi(\beta_0) - \psi(\alpha_0) = \mu_0(\beta_0 - \alpha_0) \quad ,$$

$$\mu_0 = \psi'(\alpha_0) = \psi'(\beta_0) \quad . \tag{1.3}$$

The line through $(\alpha_0, \psi(\alpha_0))$ and $(\beta_0, \psi(\beta_0))$ forms the convex envelope of the free energy between $\alpha_0$ and $\beta_0$ (Figure 1).

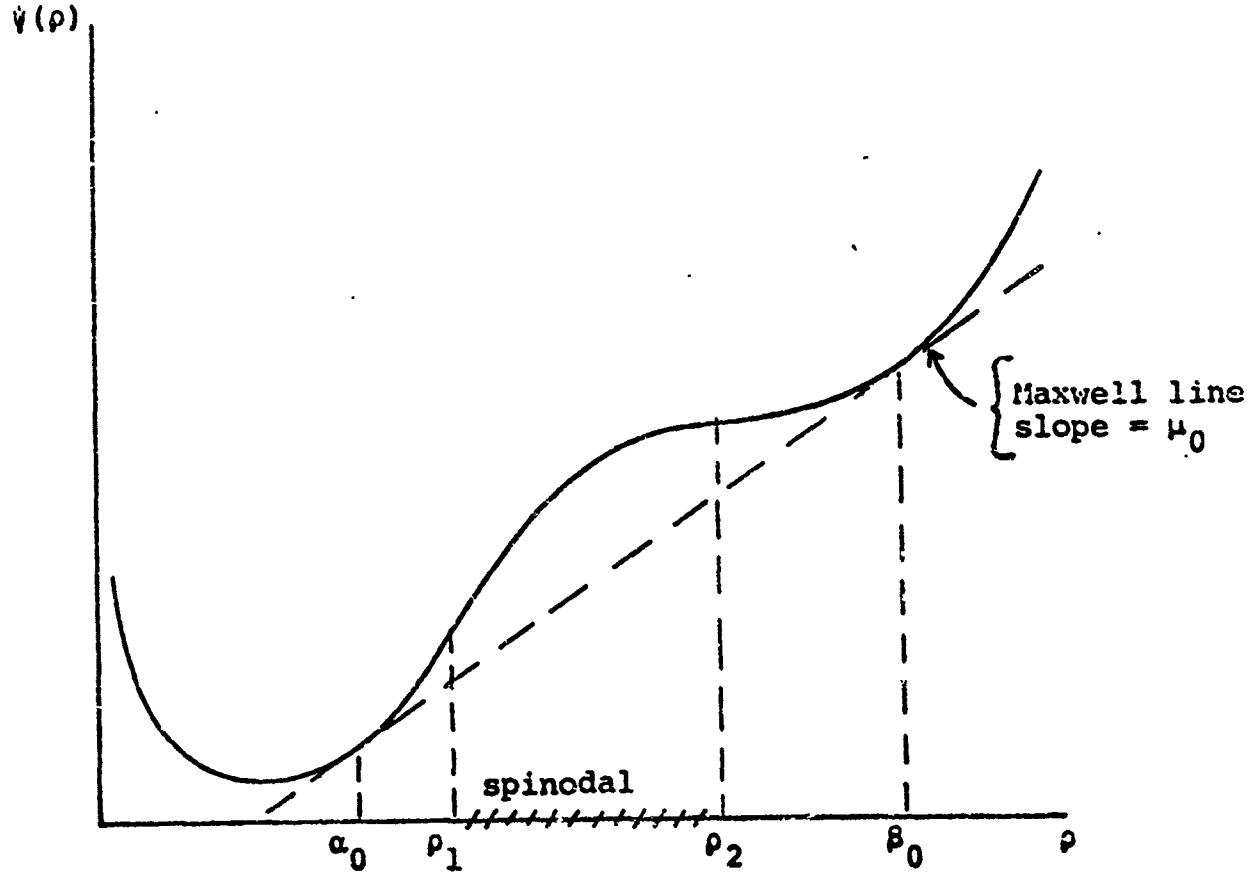



Figure 1. Free energy $\psi(\rho)$.

It is not difficult to show that if we restrict our attention to continuous $\rho$, then $P_0$ has no solution for $\alpha_0 < m < \beta_0$. If we take piecewise continuous $\rho$, and assume that the length scale is chosen with

$$\text{volume}(\Omega) = 1 \quad ,$$

then the solution of $P_0$ is as follows:

(i) for $m \leq \alpha_0$ or $m \geq \beta_0$ there is exactly one solution, the constant field

$$\rho(x) \equiv m \ ;$$

(ii) for $m \in (\alpha_0, \beta_0)$ <u>any</u> field $\rho$ of the form

$$\rho(x) = \begin{cases} \alpha_0, & x \in \Omega_1 \\ \beta_0, & x \in \Omega_2 \end{cases} ,$$

$$\Omega = \Omega_1 \cup \Omega_2, \qquad \Omega_1, \Omega_2 \subset \Omega$$

$$\text{volume}(\Omega_1) = \frac{\beta_0 - m}{\beta_0 - \alpha_0} \ ,$$

$$\text{volume}(\Omega_2) = \frac{m - \alpha_0}{\beta_0 - \alpha_0}$$

is a solution, and all solutions have this form.

Thus for each $m \in (\alpha_0, \beta_0)$ there are infinitely many solutions, some corresponding to sets $\Omega_i$ which are quite wild. This drastic loss of uniqueness occurs because interfaces - jumps in density - are allowed to form without a corresponding increase in energy.

A theory which attempts to overcome this difficulty was developed independently by van der Waals [1] and Cahn and Hilliard [2] and is based on an energy of the form

$$\int_\Omega [\psi(\rho(x)) + \varepsilon |\nabla \rho(x)|^2] dx \ .$$

Thus jumps in density are not allowed, but rapid changes are, and such changes are penalized in energy by the presence of the term $|\nabla \rho|^2$.

One problem with this theory is its difficulty, especially in space-dimensions larger than one. A second problem is that, because solutions are smooth, it is not a simple matter to locate - or even define - an interfacial zone between phases.

An alternative theory for problems of this type was developed in [3] and differs from its predecessors in two respects:

(1) Jumps in density are allowed, but are accompanied by an interfacial energy.

(2) The density distribution is not allowed to lie in the spinodal region $[\rho_1,\rho_2]$. It is assumed that $\psi(\rho)$ is defined and locally convex on the intervals $(0,\rho_1)$ and $(\rho_2,\infty)$, with nothing said about the behavior of $\psi$ on $[\rho_1,\rho_2]$.

For fields $\rho$ consistent with (2) we define complementary subsets

$$\Omega_1(\rho) = \{x \in \Omega: \ \rho(x) < \rho_1\} \ , \qquad \Omega_2(\rho) = \{x \in \Omega: \ \rho(x) > \rho_2\} \ , \tag{1.4}$$

$\Omega_i(\rho)$ being the region in which the fluid is in phase $i$. The surface

$$\Gamma(\rho) = \partial\Omega_1(\rho) \cap \partial\Omega_2(\rho)$$

then represents the <u>interface</u> between phases, and we endow $\Gamma(\rho)$ with <u>interfacial energy</u>

$$\sigma I(\rho) \ ,$$

where

(i) $I(\rho)$ is the area of $\Gamma(\rho)$ ,

(ii) $\sigma$ (assumed small) is the interfacial energy per unit area.

We are therefore led to an energy of the form

$$E(\rho) = \int_\Omega \psi(\rho(x))dx + \sigma I(\rho) \ ,$$

and to the following problem:

$P$ { <u>minimize the energy</u> $E(\rho)$ <u>over all sufficiently regular fields</u> $\rho$ <u>which have range outside</u> $[\rho_1,\rho_2]$ <u>and which satisfy the constraint</u> (1.2) .

We establish, in [3], an existence theorem for problem $P$ and show that solutions $\rho(x)$ are piecewise constant with constant chemical potential $\mu = \psi'(\rho(x))$, thereby reducing $P$ to finding the global minimum of the energy expressed as a function of $\mu$. We show further that, when the solution is considered a function of the parameter $m$, with $\mu_m$ the corresponding chemical potential and $E_m$ the associated minimal energy,

$$\mu_m = \frac{dE_m}{dm} .$$

We also prove that solutions $\rho$ of $P$ have <u>minimal interface</u>; more precisely, we show that $\Gamma(\rho)$ has minimal area when compared to all other interfaces

$$\tilde{\Gamma} = \partial\tilde{\Omega}_1 \cap \partial\tilde{\Omega}_2$$

with

$$\text{volume}(\tilde{\Omega}_i) = \text{volume}(\Omega_i(\rho)) .$$

Also contained in [3] is the general solution of $P$ in $\mathbb{R}^1$, as well as the solution for $\Omega$ the unit square in $\mathbb{R}^2$ with $\psi(\rho)$ piecewise quadratic.

<u>References</u>


[1] van der Waals, J. D., The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density (in Dutch), Verhandel. Konink. Akad. Weten. Amsterdam (Sect. 1) Vol. 1, No. 8 (1893).

[2] Cahn, J. W. and J. E. Hilliard, Free energy of a nonuniform system. I. Interfacial free energy, J. Chem. Phys. 28, 258-267 (1958).

[3] Gurtin, M. E., On a theory of phase transitions with interfacial energy, Arch. Rational Mech. Anal. To appear.


---


Sponsored by the United States Army under Contract Nos. DAAG-29-82-K-0002 and DAAG29-80-C-0041.

# A COMPUTATIONAL METHOD FOR FIELD DETECTION OF UNKNOWN SUBSTANCES

EDWARD W. ROSS
Aero-Mechanical Engineering Laboratory
US Army Natick Research & Development Center
Natick, Massachusetts 01760

ABSTRACT. This report is about computational methods in the detection of unknown substances by means of their spectra. The method of least-squares is used to test whether an unknown spectrum is one of a previously-stored set of spectra of compounds known to be of interest. The requirements are that this be done with as small and simple a program as possible and still get reasonable certainty of detection. The programs are first described, then tested by simulation. The results suggest that the procedure is feasible although further improvement is probably desirable.

1. INTRODUCTION. This report is about computational methods in the detection of unknown substances by means of their spectra. The intent of the work is to devise a simple, rapid system for this purpose. Conceptually, the system might consist of a mass spectrometer which produces spectra of air samples and feeds them to a computer that compares them with a reference set of spectra. The computer then decides whether the spectrum is that of a substance of interest (that is, a member of the reference set) or not.

An enormous amount of work has been done in the last twenty years on methods for identification of unknown chemical compounds. Many different experimental tools are available (mass spectrometry, gas chromatography, infrared spectroscopy, nuclear-magnetic resonance, etc.), most of which produce a spectrum of some kind. The spectrum is compared to a library of spectra, using a variety of mathematical procedures, loosely termed "pattern recognition" or "cluster analysis". Frequently, the final decision is made by a highly skilled scientist examining a number of sophisticated, visual representations of the unknown spectrum and the reference spectra. A typical example of the computational effort involved is ARTHUR, a collection of programs for chemical data manipulation [1].

While this and similar large statistical collections are of great value in a laboratory setting, our concerns here are different in important respects. Primarily, our procedure must be simple; it cannot rely on a skilled observer, nor can we employ massive computing power and elaborate visual displays. Presumably, modest computational capability will be available. We must expect substantial noise in the spectrum since strict laboratory procedures cannot be followed. To some extent these burdens are offset by the fact that we are interested only in a specific, small set of substances.

PREVIOUS PAGE IS BLANK

Reproduced from best available copy.

In the following sections we shall describe the general features that we desire for this system and the mathematical and statistical questions implied by them. The system requires various computing programs, which are listed and described. Also, simulations of the system's operation are presented and discussed.

2. GENERAL ASPECTS. The requirements on the system imply that it must be simple but capable of dealing with noise. The limitation on computing power and the absence of skillful users almost surely rule out the methods of pattern-recognition and cluster analysis. We would like to use highly developed, efficient computer programs so as to reduce the calculational burdens. These considerations suggest that we adopt least-squares regression as the basic method for detection.

The main body of this report concerns questions that arise in carrying out the detection by least-squares. There are many variants of the least-squares method, many programs for carrying out the computations and (unfortunately) many pitfalls in understanding the results. We deal with these in the following sections but wish now to state the problem in the mathematical notation that will be used.

The mass spectrum of a substance is represented by the vector of spectral intensities at a suitably chosen set of mass ratios. For example, the unknown substance with intensities $y_1, y_2, \ldots, y_N$ at mass ratios indexed by 1, 2, ..., N, respectively, are viewed as the vector y,

$$y^T = [y_1, y_2, \ldots, y_N] = \text{transpose of } y$$

Each of the P reference substances is also represented by its vector of spectral intensities at the same mass ratios. The vector X is given by

$$X_j^T = [X_{1j}, X_{2j}, X_{3j}, \ldots, X_{Nj}] \quad j = 1, 2, \ldots, P.$$

In the least squares method, we assume

$$y = \sum_{j=1}^{P} X_j B_j + \varepsilon$$

where $B_j$ are coefficients to be estimated and $\varepsilon$ is assumed to be a vector of N independent, Gaussian random-variables with mean = 0 and variance = $\sigma^2$.

The coefficient vector,

$$B^T = [B_1, B_2, \ldots, B_p]$$

is estimated by minimizing $\varepsilon^T\varepsilon$, which leads to the formula for $\hat{B}$ (the estimate of B)

$$\hat{B} = (X^TX)^{-1}X^Ty,$$

Reproduced from best available copy.

where the matrix X has N rows, P columns and is given in terms of its column vectors, $X_j$, by

$$X = [X_1, X_2, \ldots, X_p].$$

The quantity $\sigma^2$ is estimated by

$$S^2 = \varepsilon^T\varepsilon/(N-P)$$

and the covariance matrix of B is estimated by

$$\text{COV}(B) = S^2 (X^TX)^{-1}$$

Finally, the estimated Student-t of the j-th coefficient is

$$t_j = \hat{B}_j/[\text{Var}(\hat{B})_j]^{1/2} = \hat{B}_j/S[(X^TX)^{-1}_{jj}]^{1/2}$$

If the mass spectrometer presents us with a spectrum of an unknown substance, represented by the vector y, the algorithm consists of calculating the $t_j$ and comparing them with a value $t_c \sim 2$. If all $t_j \leq t_c$, there is little reason to think that the unknown is any of the reference substances. If $t_j \geq t_c$, there is good reason to think that the unknown contains a substantial fraction of the j-th reference substance.

We shall discuss this procedure in the next section.

3. DESCRIPTION OF THE PROCEDURE. The procedure sketched in the previous section for detecting the reference substances can be carried out in many different ways. There are several perils that attend the least-squares method, and many different methods have been developed to deal with them.

Perhaps the most interesting aspect of the present detection problem is the possibility that most of the perils can be eliminated by careful organization of the computations. In particular, most of the difficulties with the least-squares method arise out of poor conditioning of the matrix X, that is, some of the columns of X are nearly linearly dependent on other columns. Sometimes the columns are badly-scaled, and this also can cause trouble. The point is that none of these difficulties involve the vector y directly. The matrix X, the set of reference vectors, is known long before the system goes into operation. It seems, therefore, that we can study and simplify the matrix X under laboratory conditions. If this is done well, it should be possible to avoid most of the difficulties that might afflict the computations done after the unknown vector, y, is received.

This implies that we should divide the computational algorithm so that as much as possible is done "in laboratory", before its actual use. The results are stored in the "final computer", and only the final steps, the parts that involve the unknown, y, are done in the final computation. In particular, all exploratory work having to do with the conditioning of the matrix X should be done in the laboratory.

Reproduced from best available copy.

Once we have adopted the viewpoint that the computations should be divided into two stages, a preliminary stage and a final stage, we want to use existing algorithms that are similarly divided into phases. The LINPACK algorithms are so arranged and are at least as efficient and parsimonious of storage as any available programs for doing least-squares, [2]. For the present, then, we decide to base our algorithm on the LINPACK subroutines (in FORTRAN) for doing least-squares. It is possible that subroutines could be written, specifically designed for the present situation (for example, a very well-conditioned X), that would do better than the LINPACK routines in some respects. However, the improvement is not likely to be great, and so we limit ourselves to these reliable programs.

The LINPACK collection furnishes three procedures for doing least-squares problems. They are:

(a) Cholesky Decomposition of $X^TX$,

(b). QR Decomposition of X, and

(c) Singular-Value Decomposition of X.

All three provide methods for transforming the matrix X into a form that is convenient to solve. This will be done in the laboratory. The first two also give subroutines that solve for B when y is given. No such program companion to (c) is given, though it would be easy to write one. The singular-value decomposition is computationally expensive (even for laboratory circumstances), and its advantages become evident only when dealing with a poorly-conditioned matrix. We shall use it in the exploratory phase of our work, when we are studying the condition of a set of reference vectors, but we do not use it as part of the scheme once a well-conditioned set of reference vectors has been found.

For this situation we have to choose between (a) Cholesky Decompostion and (b) QR-Decompostion. The LINPACK manual equivocates on this question, so parallel programs were written that used both methods. There was no significant difference between them in time-tests, but the Cholesky Decompostion algorithms used about 10% fewer words of instruction than the QR algorithms for the same amount of data. Accordingly we decided to use Cholesky Decomposition as our algorithm. The machine version of the final program for analyzing a (30 x 13) matrix required in instructions and data 9 00 36 bit words on a UNIVAC 1106, of which 6300 words were FORTRAN overhead.

The computer programs LAB and FINAL are listed in Appendix A and described by their documentation. The LAB program reads in the matrix X of spectral intensities of the reference set of substances. The symmetric matrix

$$A = X^TX$$

is formed and factored into the product

$$A = R^tR$$

where R is an upper triangular matrix. The LINPACK routine SPPFA finds the matrix R. Also the matrix $A^{-1}$ is found by subroutine SPPDI, and the vector V with

$$V_j = |A^{-1}_{jj}/(NR-NC)|^{1/2}$$

is calculated. The arrays X, R and V are written in file 7 to be used by the FINAL program.

The FINAL program reads the contents of file 7 and the vector spectrum of of the unknown, y. The vector $W = X^T y$ is calculated and the linear system

$$AB \quad X^T X B = X^T y = W$$

solved by LINPACK routine SPPSL. Then the residual vector

$$\varepsilon = y - XB$$

and its sum of squares,

$$S^2 = \varepsilon^T \varepsilon = \sum_{i=1}^{N} \varepsilon_i^2$$

are found. Finally, the t-values associated with $B_j$ are written out if $t_j > t_c$. The FINAL program has a loop for simulating randomness effects. It reads in a basic, unknown vector, U, and generates NR sample y-vectors, obtained by adding to U a vector of Gaussian random numbers with mean = 0 and standard deviation DSD.

These programs are used later in studying the effects on detection of altering various properties of the reference set or unknown.

4. TESTING THE PROCEDURE IN THE LABORATORY. A trial set of data was furnished, perhaps typical of the spectra that the procedure must treat. It comprises the spectra of 13 substances, each spectrum being defined by 30 peak intensities, hence is represented by a matrix with 30 rows and 13 columns. The matrix is shown in Table 1. We now proceed to analyze it, that is, we do the laboratory part of the statistical treatment.

The first step is to rearrange the data matrix so the columns are normalized to have length 10,000, and the rows are ordered according to their median entries. That is, the median of each row is found and the rows exchanged so that the first row is the one with the largest median, the second row has the next largest median, etc. Table 2 shows this rearrangement of the array, which we now designate the matrix X; this table also includes a trailing column (not part of X), giving for each row its index in the original data-matrix. The matrix X is the object of the study that follows.

TABLE 1: ORIGINAL DATA-MATRIX

| ROW | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 133 | 304 | 200 | 260 | 614 | 178 | 387 | 883 | 442 | 73 | 181 | 306 | 487 |
| 2 | 62 | 103 | 63 | 69 | 82 | 51 | 27 | 31 | 34 | 23 | 23 | 71 | 55 |
| 3 | 1051 | 2077 | 1486 | 1685 | 1682 | 1713 | 1382 | 1088 | 843 | 1112 | 1157 | 1489 | 1504 |
| 4 | 123 | 148 | 112 | 110 | 204 | 138 | 78 | 83 | 54 | 70 | 123 | 144 | 53 |
| 5 | 543 | 430 | 1004 | 731 | 736 | 684 | 237 | 241 | 238 | 567 | 366 | 756 | 861 |
| 6 | 33 | 0 | 11 | 20 | 79 | 43 | 150 | 54 | 28 | 29 | 54 | 25 | 64 |
| 7 | 19 | 286 | 0 | 0 | 56 | 14 | 105 | 32 | 34 | 46 | 20 | 23 | 64 |
| 8 | 26 | 20 | 84 | 0 | 27 | 5 | 35 | 30 | 3 | 0 | 23 | 0 | 0 |
| 9 | 20 | 81 | 52 | 8 | 38 | 25 | 17 | 16 | 24 | 12 | 9 | 21 | 23 |
| 10 | 18 | 64 | 46 | 15 | 36 | 33 | 10 | 18 | 16 | 10 | 13 | 18 | 17 |
| 11 | 446 | 738 | 335 | 299 | 769 | 387 | 697 | 805 | 563 | 169 | 566 | 551 | 861 |
| 12 | 9 | 26 | 30 | 35 | 81 | 23 | 57 | 46 | 41 | 22 | 22 | 45 | 54 |
| 13 | 1263 | 2245 | 1320 | 1418 | 1111 | 1322 | 1822 | 1749 | 1315 | 888 | 1591 | 1512 | 1670 |
| 14 | 130 | 114 | 481 | 313 | 270 | 174 | 194 | 606 | 700 | 296 | 192 | 189 | 337 |
| 15 | 4 | 0 | 0 | 9 | 8 | 0 | 0 | 0 | 16 | 133 | 35 | 0 | 45 |
| 16 | 225 | 554 | 225 | 77 | 46 | 123 | 164 | 118 | 123 | 94 | 140 | 126 | 174 |
| 17 | 54 | 27 | 24 | 24 | 54 | 62 | 44 | 46 | 35 | 21 | 33 | 30 | 39 |
| 18 | 33 | 27 | 14 | 15 | 92 | 18 | 38 | 33 | 20 | 48 | 17 | 44 | 32 |
| 19 | 67 | 0 | 0 | 26 | 64 | 7 | 22 | 16 | 13 | 85 | 7 | 12 | 9 |
| 20 | 116 | 51 | 72 | 0 | 90 | 52 | 219 | 345 | 529 | 0 | 112 | 42 | 387 |
| 21 | 24 | 19 | 8 | 18 | 34 | 11 | 33 | 14 | 7 | 21 | 10 | 30 | 11 |
| 22 | 6 | 0 | 7 | 6 | 7 | 4 | 0 | 119 | 6 | 0 | 108 | 8 | 28 |
| 23 | 568 | 229 | 604 | 410 | 293 | 326 | 187 | 642 | 1495 | 233 | 230 | 383 | 717 |
| 24 | 36 | 492 | 90 | 17 | 178 | 48 | 157 | 189 | 173 | 66 | 182 | 83 | 241 |
| 25 | 92 | 19 | 148 | 62 | 152 | 189 | 145 | 79 | 383 | 108 | 96 | 82 | 54 |
| 26 | 122 | 365 | 865 | 111 | 159 | 176 | 98 | 827 | 286 | 54 | 134 | 200 | 173 |
| 27 | 1654 | 73 | 921 | 832 | 1089 | 962 | 1649 | 314 | 482 | 941 | 1333 | 468 | 918 |
| 28 | 177 | 0 | 70 | 89 | 102 | 100 | 138 | 127 | 84 | 117 | 107 | 863 | 46 |
| 29 | 367 | 369 | 811 | 157 | 386 | 881 | 165 | 362 | 104 | 181 | 286 | 378 | 184 |
| 30 | 8345 | 485 | 1547 | 2459 | 1148 | 2474 | 1711 | 1501 | 1772 | 4658 | 2664 | 2300 | 1249 |

TABLE 2: NORMALIZED AND RE-ARRANGED DATA-MATRIX, X

| ROW | A | B | C | D | E | F | G | H | I | J | K | L | M | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6841 | 1381 | 5043 | 7445 | 3999 | 6951 | 5461 | 5165 | 5703 | 0277 | 7242 | 6979 | 4057 | 30 |
| 2 | 3684 | 6669 | 4303 | 3692 | 3870 | 3714 | 3900 | 6015 | 4232 | 1769 | 4325 | 4415 | 5453 | 13 |
| 3 | 3066 | 6170 | 4844 | 4388 | 5650 | 4813 | 4166 | 3538 | 2713 | 2215 | 3145 | 4173 | 4885 | 3 |
| 4 | 4533 | 0217 | 3003 | 2166 | 3793 | 2703 | 5263 | 1081 | 1561 | 1874 | 3624 | 1367 | 2082 | 27 |
| 5 | 1301 | 2182 | 1255 | 0779 | 2644 | 0919 | 2225 | 2770 | 1812 | 0317 | 1522 | 1609 | 2796 | 11 |
| 6 | 1584 | 1277 | 3667 | 1903 | 2564 | 1922 | 0756 | 0829 | 0747 | 1129 | 0995 | 2208 | 0848 | 6 |
| 7 | 1657 | 0680 | 1969 | 1068 | 1021 | 0913 | 0597 | 2209 | 4811 | 0464 | 0650 | 1118 | 2329 | 23 |
| 8 | 0379 | 0339 | 1568 | 0815 | 0941 | 0489 | 0619 | 2086 | 2253 | 0590 | 0448 | 0552 | 1005 | 14 |
| 9 | 0388 | 0903 | 0652 | 0677 | 1790 | 0500 | 1171 | 0974 | 1423 | 0145 | 0519 | 0894 | 1614 | 1 |
| 10 | 1071 | 1096 | 0688 | 0409 | 1345 | 0621 | 0627 | 1246 | 0335 | 0360 | 0777 | 1098 | 0403 | 20 |
| 11 | 0356 | 1084 | 0864 | 0289 | 0554 | 0494 | 0313 | 0781 | 0727 | 0108 | 0364 | 0584 | 0562 | 26 |
| 12 | 0656 | 1646 | 0734 | 0201 | 0160 | 0346 | 0523 | 0385 | 0386 | 0187 | 0381 | 0368 | 0565 | 16 |
| 13 | 0338 | 0152 | 0235 | 0000 | 0314 | 0146 | 0699 | 1187 | 1793 | 0000 | 0304 | 0123 | 1257 | 28 |
| 14 | 0186 | 1194 | 0293 | 0044 | 0592 | 0135 | 0501 | 0650 | 0557 | 0131 | 0495 | 0242 | 0783 | 24 |
| 15 | 0268 | 0056 | 0482 | 0135 | 0529 | 0531 | 0463 | 0272 | 1233 | 0211 | 0258 | 0239 | 0175 | 25 |
| 16 | 0516 | 0000 | 0228 | 0232 | 0355 | 0506 | 0421 | 0437 | 0270 | 0233 | 0291 | 0768 | 0149 | 29 |
| 17 | 0359 | 0440 | 0365 | 0286 | 0711 | 0388 | 0249 | 0286 | 0174 | 0139 | 0334 | 0421 | 0172 | 4 |
| 18 | 0181 | 0306 | 0205 | 0188 | 0286 | 0143 | 0086 | 0107 | 0109 | 0046 | 0063 | 0207 | 0178 | 2 |
| 19 | 0055 | 0850 | 0000 | 0000 | 0195 | 0039 | 0335 | 0110 | 0109 | 0080 | 0054 | 0067 | 0208 | 7 |
| 20 | 0096 | 0000 | 0036 | 0052 | 0275 | 0121 | 0479 | 0186 | 0090 | 0058 | 0147 | 0073 | 0208 | 6 |
| 21 | 0626 | 0977 | 0098 | 0091 | 0282 | 0065 | 0182 | 0158 | 0132 | 0058 | 0076 | 0131 | 0175 | 12 |
| 22 | 0158 | 0580 | 0078 | 0062 | 0188 | 0174 | 0140 | 0158 | 0113 | 0042 | 0090 | 0114 | 0127 | 17 |
| 23 | 0096 | 0080 | 0046 | 0039 | 0320 | 0051 | 0121 | 0114 | 0064 | 0096 | 0046 | 0128 | 0104 | 18 |
| 24 | 0058 | 0241 | 0170 | 0021 | 0132 | 0070 | 0054 | 0055 | 0077 | 0024 | 0024 | 0061 | 0075 | 9 |
| 25 | 0195 | 0000 | 0000 | 0068 | 0223 | 0020 | 0070 | 0065 | 0042 | 0169 | 0019 | 0035 | 0029 | 19 |
| 26 | 0053 | 0190 | 0150 | 0039 | 0122 | 0093 | 0032 | 0062 | 0051 | 0020 | 0035 | 0053 | 0055 | 10 |
| 27 | 0678 | 0059 | 0274 | 0000 | 0094 | 0014 | 0112 | 0275 | 0018 | 0000 | 0063 | 0000 | 0000 | 8 |
| 28 | 0019 | 0000 | 0023 | 0016 | 0024 | 0011 | 0000 | 0418 | 0018 | 0000 | 0294 | 0023 | 0001 | 22 |
| 29 | 0012 | 0000 | 0000 | 0023 | 0028 | 0000 | 0000 | 0000 | 0051 | 0265 | 0095 | 0000 | 0146 | 15 |
| 30 | 0078 | 0056 | 0026 | 0047 | 0118 | 0031 | 0105 | 0048 | 0023 | 0042 | 0027 | 0008 | 0036 | 21 |

Our preliminary investigation is concerned mainly with the condition of the matrix X. Also we are interested in the following notion. A glance at Table 2, matrix X, shows that the lower rows have very small entries compared with the first few rows. Since there is some random error in the matrix elements in any case, it is doubtful that these rows are contributing much information to the identification of the precise substances. Perhaps little discerning-power would be lost if these rows were deleted. However, we have to be careful about doing this for reasons explained in the next paragraph.

There are two levels of identification that concern us in this project. Primarily we are interested in the question "Does the unknown spectrum contain a significantly non-zero concentration of any of the reference substances?" Only secondarily are we interested in which specific substance is present and its concentration. (This is somewhat different from the relative importance usually accorded these two questions in a laboratory setting and is one way in which this problem differs from chemical identification in the laboratory.) The importance of the rows with small entries is slightly different with respect to these two questions. The small entries may not have much effect on the determination of the specific substances but may have a telling impact in deciding whether an unknown substance belongs to the set or not. For the unknown could differ from a member of the set by having a high peak where all members of the reference set have low intensity. If this peak is deleted, we could mistakenly identify the unknown as a member of the set, that is, we risk setting off a false alarm. We must, therefore, be very cautious about deleting rows of X.

There are many ways of assessing the condition of the matrix X. They range from very simple procedures, such as finding the correlation matrix for the columns of X up to doing the singular value decomposition of the matrix. We hope to find and perhaps remove any near dependencies (linear relationships), among the columns of X. The most complete information about the presence of these dependencies is furnished by the singular-values of X, the positive square-roots of the eigenvalues of $A = X^T X$. In particular, the presence of dependencies is indicated by high values of the ratio of the largest to the smallest singular-value.

A convenient procedure for investigating the dependencies among the columns of X is called V-NULL and is part of the collection called ROSEPACK. It causes a singular-value decomposition to be done. Estimates of the singular-values and their ratio to the largest singular-value are printed out. The user is then asked to name the index of the smallest non-zero singular-value. This choice implies a decision that the rank of the matrix is, say, $\rho$. Then there will exist $N-\rho$ dependent columns and $\rho$ independent columns, and the program prints out estimates of the coefficients in the relations between the dependent and independent columns.

This program was run, using the matrix X as input, with results shown in Table 3. The ratio of the largest to smallest singular-value is $1/0.00935 \approx 107$, which is a bit large for comfort though not indicative of major trouble.

Table 3. (Singular Values/Largest Singular Value) for Matrix X.

| INDEX | FULL MATRIX | 19 ROWS | 16 ROWS |
|---|---|---|---|
| 1 | 1.0000 | 1.0000 | 1.0000 |
| 2 | .2663 | .2663 | .2659 |
| 3 | .1649 | .1648 | .1647 |
| 4 | .1397 | .1394 | .1395 |
| 5 | .0992 | .0989 | .0984 |
| 6 | .0561 | .0552 | .0552 |
| 7 | .0528 | .0516 | .0509 |
| 8 | .0332 | .0322 | .0314 |
| 9 | .0258 | .0251 | .0231 |
| 10 | .0200 | .0189 | .0186 |
| 11 | .0149 | .0140 | .0137 |
| 12 | .0136 | .0126 | .0119 |
| 13 | .0094 | .0082 | .0074 |

In particular, a stem-and-leaf display of the logarithm of the ratio of the largest singular-value to each of the others, Table 4, shows no striking gaps. Hence there is no obvious choice of rank. A tentative selection, $\rho = 12$, was made, and the program calculated that column 4 is dependent and the dependency relation is:

$$C_4 = -0.148C_2 + 0.117C_3 - 0.006C_1 + 0.039C_5 + 0.857C_6 - 0.570C_7 + 0.078C_8 -$$

$$0.235C_9 + 0.358C_{10} + 0.126C_{11} - 0.251C_{12} + 0.631C_{13}$$

We conclude from this that the matrix X is not extremely close to being singular. There is no strong relationship among the columns of X, but the strongest of the somewhat weak dependencies is between $C_4$ and the columns having large coefficients, namely $C_6$, $C_{13}$, and $C_7$. The relation between $C_4$ and $C_6$ is depicted as particularly close in this treatment, and scrutiny of X confirms that these columns are similar.

At this point we might choose to remove $C_4$ (or $C_6$) from X if we thought the relation between them was close enough to adversely affect the results. Rather than doing that, however, we use V-NULL to determine what is the effect on the condition of X when we delete rows with small entries, the lower rows of X. The results of deleting the last 11 and 14 rows of X are also shown in Table 3. The singular values are not changed much, which suggest that omitting, say, the last 10 rows of X would not greatly weaken our ability to distinguish the columns of X. However, we prefer to retain these rows for reasons mentioned earlier.

In this section, we have discussed procedures for detecting poor conditioning of the matrix X. The treatment has not been exhaustive by any means. For example, we have not used the idea of the variance-decompostions proportion for the sake of brevity, [3]. In general, we shall do everything necessary in a laboratory computation to assure that X is well-conditioned before it is used in the final computation.

5. TESTING THE FIELD PROCEDURE. In this section we describe the results of tests of the program for doing the final computations. These tests are in the form of Monte-Carlo simulations in which the unknown vector, y, is proportional to one of the columns of X with added noise. A few tests are also done in which y is purely random noise; its only resemblance to the reference set is a chance one.

The computational procedure starts with the matrix X, hopefully well-conditioned, and applies the LAB program to create the file that will be read by the FINAL program. The FINAL program reads this file, then reads the unknown vector, U, generates NR samples of y by adding noise, and analyzes each. The random noise is independent Gaussian with standard deviation (SD). Tests were made with SD = 100, 200, 300 and 500. It is thought that SD = 300 corresponds roughly to a noise level of 10% because

TABLE 4: DISTRIBUTION OF $\log_{10}$ (SUR)

SUR = (LARGEST SINGULAR VALUE/EACH SINGULAR VALUE)

STEM-AND-LEAF DISPLAY OF LOG(SUR)
LEAF DIGIT UNIT = .1000
1 2 REPRESENTS 1.2

```
  1   +0*   0
  1   +0T
  2   +0F   5
  3   +0S   7
  4   +0.   8
  5    1*   0
(2)    1T   22
  6    1F   45
  4    1S   7
  3    1.   88
  1    2*   0
```

2(SD)$\sim$(median of largest values in each column)/10, so that these levels bracket the noise we expect to encounter in the field.

The Monte-Carlo simulations for the first condition were done by taking each column of X in turn as the basic unknown vector, U, then generating NR = 20 or 40 samples of y. In this way 260 or 520 trials were made at each noise level. The results are stated in terms of three integers,

NND = number of trials in which nothing was detected

NED = number of trials in which the true column was detected

NWD = number of trials in which a column was wrongly detected,

where detection of a column means that the t-value for the regression coefficient of that column was (a) larger than 2 and (b) larger than the t-values for the other columns.

The results are shown in Table 5 for SD = 200, 300, and 500. Trials were run for SD = 100, but the results are not shown because they were perfect; all 260 trials resulted in exactly correct detection. In looking at these results we must bear in mind that we would like the system to yield primarily a small number for NND, because in these trials the unknown is always a member of the reference set. If NND is large it means that the system often fails to detect the presence of a reference substance. Table V shows that the overall ratio of NND to the number of trials is:

for SSD = 200, $r_{ND}$ = NND/260 = 8/260 = 0.031

for SSD = 300, $r_{ND}$ = NND/520 = 58/520 = 0.112

for SSD = 500, $r_{ND}$ = NND/260 = 60/260 = 0.231

These results suggest that the system detects the presence of some reference substance (when one is present) in about 89% of the trials at this noise level. If the instrumentation could be improved so that SSD = 200 (about a 6% noise level), the estimated detection probability rises to 97%.

The table also furnishes information about the less crucial determination of the exact substance. We see that the overall probability of this is estimated to be:

for SSD = 200, $r_{ED}$ = 236/260 = 0.908

for SSD = 300, $r_{ED}$ = 386/520 = 0.742

for SSD = 500, $r_{ED}$ = 128/260 = 0.492.

These are less satisfactory than the previous results, but also less important.

Table 5. Simulation Results for Matrix X.

| Index of Column | SD = 200 | | | SD = 300 | | | SD = 500 | | |
|---|---|---|---|---|---|---|---|---|---|
| | NED | NND | NWD | NED | NND | NWD | NED | NND | NWD |
| 1 | 20 | 0 | 0 | 34 | 2 | 4 | 9 | 6 | 5 |
| 2 | 20 | 0 | 0 | 40 | 0 | 0 | 17 | 0 | 3 |
| 3 | 20 | 0 | 0 | 39 | 0 | 1 | 16 | 1 | 3 |
| 4 | 13 | 2 | 5 | 12 | 16 | 12 | 2 | 8 | 10 |
| 5 | 20 | 0 | 0 | 39 | 0 | 1 | 12 | 5 | 3 |
| 6 | 13 | 3 | 4 | 19 | 10 | 11 | 2 | 8 | 10 |
| 7 | 17 | 0 | 3 | 24 | 7 | 9 | 8 | 3 | 9 |
| 8 | 20 | 0 | 0 | 37 | 1 | 2 | 16 | 1 | 3 |
| 9 | 20 | 0 | 0 | 40 | 0 | 0 | 14 | 2 | 4 |
| 10 | 19 | 0 | 1 | 38 | 0 | 2 | 11 | 5 | 4 |
| 11 | 16 | 3 | 1 | 21 | 6 | 13 | 5 | 9 | 6 |
| 12 | 20 | 0 | 0 | 26 | 6 | 8 | 7 | 7 | 6 |
| 13 | 18 | 0 | 2 | 17 | 10 | 13 | 9 | 5 | 5 |
| | 236 | 9 | 16 | 386 | 58 | 76 | 128 | 60 | 72 |

The table also tells us quite a lot about the specific columns that are well or poorly detected. We see that for SD = 300 columns 2, 3, 5, 9, and 10 are very distinctive. When they are present, something is always detected and almost always the correct column is found. Columns 1 and 8 are less well-defined but stand out fairly clearly. Contrarily when column 4 is present, the system has a lot of trouble in detecting anything, and when something is found, it is often identified incorrectly. To a lesser extent, this is also true of columns 6, 7, 11, 12, and 13. These results agree fairly well with the implications of the V-NULL analysis described in the previous section. For SD = 200 and 500 these distinctions are less clear than at SD = 300 but coarsely similar.

The effect of deleting the last ten rows of the matrix X was investigated by a similar Monte-Carlo procedure. The results are not shown in detail, but the overall probabilities of no detection and exact detection are:

for SD = 200, $r_{ND} = 0.077, r_{ED} = 0.862$

for SD = 300, $r_{ND} = 0.165, r_{ED} = 0.677$

for SD = 500, $r_{ND} = 0.315, r_{ED} = 0.431$

There is a statistically significant loss of discrimination (at the 95% level) in these results, as compared with those using the full matrix.

The V-NULL analysis suggests that the most nearly dependent column of the matrix X is column 4. It is natural, then, to investigate the effect on the detection-process of removing column 4 from the matrix. Table 6 shows the results for this simulation carried out for SD = 300 and compares them with the results for the complete matrix. The estimated detection probabilities are:

| | Complete X | X without col. 4 |
|---|---|---|
| $r_{ED}$ | 0.742 | 0.831 |
| $r_{ND}$ | 0.112 | 0.071 |
| $r_{WD}$ | 0.146 | 0.098 |

We see that deleting column 4 has a generally favorable effect on detection probability, although part of the improvement is due to the fact that we did not use column 4 as an unknown in these simulations. Examination of Table 6 shows that most of the improvement was caused by the elimination of column 4 from consideration and better detection of columns 6, 7, and 13. This agrees well with the results of the V-NULL analysis described in the previous section. However, we notice also that column 11 is detected less well than before. The reason for this is not clear. In a real life situation this would lead us to do further analysis of the null-space of X, but we shall not pursue it here.

Table 6. Comparison of Simulation Results With and Without Column 4.

| | Complete X | | | X Without Column 4 | | |
|---|---|---|---|---|---|---|
| | NED | NND | NWD | NED | NND | NWD |
| 1 | 34 | 2 | 4 | 35 | 2 | 3 |
| 2 | 40 | 0 | 0 | 40 | 0 | 0 |
| 3 | 39 | 0 | 1 | 40 | 0 | 0 |
| 4 | 12 | 16 | 12 | - | - | - |
| 5 | 39 | 0 | 1 | 39 | 0 | 1 |
| 6 | 19 | 10 | 11 | 27 | 6 | 7 |
| 7 | 24 | 7 | 9 | 26 | 4 | 10 |
| 8 | 37 | 1 | 2 | 39 | 0 | 1 |
| 9 | 40 | 0 | 0 | 40 | 0 | 0 |
| 10 | 38 | 0 | 2 | 37 | 0 | 3 |
| 11 | 21 | 6 | 13 | 18 | 12 | 10 |
| 12 | 26 | 6 | 8 | 28 | 5 | 7 |
| 13 | 17 | 10 | 13 | 30 | 5 | 5 |
| | 386 | 58 | 76 | 399 | 34 | 47 |

In all the cases described so far, the unknown vector, y, was assumed to consist of one column of X plus noise. A few simulations were also run in which y was a linear combination of two columns of X with added noise. The results were much worse than with one column. The set of simulations was far from complete, but the probability of detecting nothing when two columns are present appears to be about 0.25, which is quite poor.

It still remains to investigate the propensity of this system for giving false alarms. For this purpose it would be desirable to have a set of mass spectra typical of substances that might be present at the same time as those of the reference set. Members of this set would then be used as the y-vectors, and we could see whether the system misidentifies any of them. Unfortunately such a set is not available. A much less desirable procedure is to generate random vectors from some plausible distribution and see what fraction of them is identified as one of the reference set. Since we are using a critical t-value of 2, we expect that about 5% of the cases will be identified as members of the reference set if the underlying distribution is close to Gaussian. A few preliminary trials of this nature were carried out, with the expected results, but it is difficult to see what could be learned from a large simulation since relatively little is known about the distribution of the mass spectra in natural environments.

6. CONCLUSIONS AND DISCUSSION. The principal conclusion that can be drawn from this effort is that the proposed scheme appears workable, provided the real life situation is reasonably close to what is assumed in the simulations. It is likely that with care in the conditioning of the reference matrix, we can achieve successful detection in more than 90% of the trials. The probability of false alarms is less well determined, especially under conditions that might arise in military operations.

The following comments and suggestions appear relevant:

(a) It would be very desirable to obtain spectra of sample volatiles occurring under natural conditions. We must have this or equivalent information in order to see whether the system will produce false alarms under operating conditions.

(b) Further thought should be given to the problem of improving the condition X. The procedure exemplified in section 4, simply deleting the most dependent column is widely practiced but sometimes unreliable. On the other hand, the procedures usually suggested as replacements (see Belsley, Kuh, and Welsch[3]), ridge regression and mixed or Bayesian estimation, involve appending questionable information, beyond that available from the spectra, and are, therefore, also open to criticism.

(c) The final program used here is quite small and is probably within the capacity of most minicomputers. However, it is in FORTRAN and perhaps will need reprogramming into BASIC or some assembly language for actual use.

(d) We have not conducted any formal search for outlying or highly influential rows in the data matrix. We have omitted this aspect of the study because it is reasonably clear from the results already found that for our purposes none of the rows are excessively influential. For if, say, one row was very influential, the B values would fluctuate wildly from one trial to the next, according to the randomly changing y-entry in that row. Our principal concern is detecting the j for which $B_j$ is largest and deciding whether that $t_j > 2$; this primary conclusion is quite reproducible from one trial to the next. It is possible that other estimates are influenced mostly by one row, but there is little evidence that the primary conclusion is so affected. Nevertheless, it might be interesting to examine the hat matrix

$$H = X (X^T X)^{-1} X^T$$

for this problem. Because of the initial ordering of the rows of X, it seems plausible that each row would be more influential than its successor, but this has to be checked.

(e) In conclusion, it is a good idea to re-emphasize the factors that make this problem different from a routine application of least-squares regression. They are

(1) limited computing power and data accuracy,

(2) a modest number of reference vectors, and

(3) a very specific objective, namely, to see whether the unknown vector contains a statistically significant amount of any of the reference vectors. Although we would like to detect which vectors are present, that is much less important than detecting whether any are found.

## REFERENCES


1. Kowalski, B.R. Chemometrics: Theory and Application, Washington, DC, American Chemical Soc. (1977), pp. 14-53.

2. Dongarra, J.J., et. al., LINPACK Users Guide, Philadelphia, Society of Industrial and Applied Mathematics (1979).

3. Belsley, D.A., Kuh, E., and Welsch, R.E., Regression Diagnostics, New York, John Wiley & Sons, (1980).

# APPENDIX A

## PROGRAM LAB

```
      C***    THIS IS THE IN-LABORATORY OR PRELIMINARY
      C*** PART OF THE ALGORITHM FOR CHEMICAL-DETECTION
      C*** BY USE OF CHOLESKY-DECOMPOSITION. THE SPECTRAL-
      C*** INTENSITY MATRIX, X, IS READ IN AND SYMMETRIC
      C*** MATRIX A GENERATED. THIS IS THEN FACTORED AND
      C*** THE INVERSE FOUND. X,A AND THE DIAGONAL OF THE
      C*** INVERSE ARE WRITTEN IN FILE 7, WHENCE THEY
      C*** ARE READ BY THE FIELD PROGRAM CF.
            REAL X(30,13),A(105),U(13)
        100 FORMAT( )
        101 FORMAT(13F5.0)
        200 FORMAT(6E16.7)
        201 FORMAT(2I5)
      C*** READ NR,NC AND THE MATRIX X. THEN WRITE X
      C*** IN FILE 7.
            READ(5,100) NR,NC
            WRITE(7,201) NR,NC
            DO 4 I=1,NR
            READ(5,101) (X(I,J),J=1,NC)
          4 WRITE(7,200) (X(I,J),J=1,NC)
      C*** CALCULATE THE UPPER-TRIANGLE OF THE SYMMETRIC
      C*** MATRIX  A  =  (X-TRANSPOSE)*X   AND STORE
      C*** IN VECTOR REPRESENTATION.
            K=0
            DO 20 J=1,NC
            DO 10 I=1,J
            K=K+1
            A(K)=SDOT(NR,X(1,I),1,X(1,J),1)
         10 CONTINUE
         20 CONTINUE
      C*** USE LINPACK SUBROUTINE SPPFA TO DO CHOLESKY-
      C*** FACTORIZATION OF SYMMETRIC MATRIX A. WRITE THE
      C*** RESULTING MATRIX A IN FILE 7.
            CALL SPPFA(A,NC,INFO)
            IF(INFO.GT.0) WRITE(6,201) INFO
            WRITE(7,200) (A(KK),KK=1,K)
      C*** CALCULATE THE DIAGONAL TERMS OF THE INVERSE OF
      C*** SYMMETRIC MATRIX  A AND DIVIDE BY THE DEGREES
      C*** OF FREEDOM TO GET THE VECTOR U. THEN WRITE
      C*** SQRT(U) IN FILE 7. LINPACK SUBROUTINE SPPDI IS
      C*** USED TO FIND THE INVERSE OF A.
            CALL SPPDI(A,NC,D,1)
            C=NR-NC
            JJ=0
            DO 1 J=1,NC
            JJ=J+JJ
          1 U(J)=SQRT(A(JJ)/C)
            WRITE(7,200) (U(J),J=1,NC)
            END
```

PROGRAM FINAL

```
      C      THIS IS THE FINAL PART OF THE PROCEDURE FOR
      C  IDENTIFYING UNKNOWN SPECTRA. IT READS IN THE UNKNOWN
      C  SPECTRUM AND DOES A LINEAR REGRESSION VIA CHOLESKY
      C  DECOMPOSITION TO ESTIMATE THE REGRESSION COEFFICIENTS AND
      C  THEIR T-VALUES. LARGE T-VALUES CORRESPOND TO THE
      C  MOST PROBABLE KNOWN SUBSTANCES IN THE LIST OF THOSE
      C  PREVIOUSLY ACQUIRED FROM THE 'LAB' PROGRAM, CH . AS MUCH
      C  AS POSSIBLE OF THE WORK HAS BEEN DONE IN THE LAB
      C  PROGRAM.
      C
            PARAMETER LDR=30, LDC=13, LDP=105
            REAL X(LDR,LDC),V(LDC),A(LDP),U(LDR)
            REAL B(LDC),Y(LDR)
            INTEGER IR(LDC)
            DOUBLE PRECISION DSD
      C**** READ IN DATA CREATED BY THE 'LAB' PROGRAM, CH.
            READ (7,90) NR,NC
            DO 10  I=1,NR
         10    READ (7,120) (X(I,J),J=1,NC)
            K=NC*(NC+1)/2
            READ (7,120) (A(KK),KK=1,K)
            READ (7,120) (V(J),J=1,NC)
      C**** READ IN DATA FOR UNKNOWN SPECTRAL INTENSITIES.
            DO 20  I=1,NR
         20    READ (5,120) U(I)
            READ (5,110) NREP,SD,DSD
      C**** START OF LOOP FOR SIMULATING RANDOMNESS EFFECTS
            DO 80  L=1,NREP
      C**** INTRODUCE RANDOM GAUSSIAN NOISE .
               DO 30  I=1,NR
                  Y(I)=U(I)+SD*GGNQF(DSD)
```

```
32          30        IF (Y(I).LT.0.0)  Y(I)=0.0
33       C**** FIND REGRESSION COEFFICIENT ESTIMATES, B(J) .
34                 DO 40  J=1,NC
35          40        B(J)=SDOT(NR,X(1,J),1,Y,1)
36                 CALL SPPSL (A,NC,B)
37       C**** FIND RESIDUALS AND T-VALUES OF COEFF. ESTIMATES.
38       C**** ARRAY Y(I) IS USED FOR BOTH RESIDUALS AND T-VALUES.
39                 DO 58  I=1,NR
40          50        Y(I)=SDOT(NC,X(I,1),LDR,B,1)-Y(I)
41                 S=SNRM2(NR,Y,1)
42                 DO 60  J=1,NC
43                    IR(J)=J
44                    SE=V(J)*S
45          60        Y(J)=B(J)/SE
46       C**** SORT ON T-VALUES AND WRITE OUT FOR T > 1.0 .
47                 CALL USRTR (Y,NC,IR)
48                 WRITE (6,90) L
49                 DO 70  J=NC,1,-1
50                    K=IR(J)
51                    IF (Y(J).LT. 1.0)  GO TO 80
52          70        WRITE (6,100) K,B(K),Y(J)
53          80     CONTINUE
54       C
55          90 FORMAT (2I5)
56         100 FORMAT (I3,2F7.4)
57         110 FORMAT ()
58         120 FORMAT (5E15.7)
59       C
60             END
```

# MATHEMATICAL FOUNDATIONS FOR ROBOTICS

John E. Hopcroft
Department of Computer Science
Cornell University
Ithaca, N.Y. 14853

ABSTRACT. A major component of robotics is concerned with the representation and manipulation of physical objects. The design of computer algorithms to interact with these representations and to carry out task planning raises many mathematical and computational issues that are not well understood. This talk will explore some of these issues and outline areas where mathematical development is needed to support a robotics effort.

I. INTRODUCTION. Task level planning is an important aspect of robotics. What is needed is the ability to communicate a task to a robot at the level of "insert peg in hole" and have the detailed instructions such as open gripper, move to location x, rotate gripper, etc. be automatically generated by a computer. To do this requires solving many problems. First, one must be able to automatically deduce a grip position that will not interfere with the desired task. Then one must develop an approach strategy that will insure gripping even in the presense of some uncertainty of position. Finally one must be able to plan motion.

Each of these tasks requires the ability to represent physical objects in a computer in such a manner that they can easily be reasoned about. Although researchers in computer graphics have developed computer models of surface representations and efficient ray tracing algorithms, these techniques are not suited to representing the physical objects themselves for several reasons. A surface is easily represented by bicubic patches. However, a bicubic patch is an 18 degree surface and thus the intersection is a degree 36 curve. This

PREVIOUS PAGE IS BLANK

precludes symbolic or algebraic calculations of intersections. Furthermore, a circle does not have a polynomial parameterization and hence a cylinder cannot be represented exactly by bicubic patches. This prevents reasoning about a cylinder rotating in a circular bearing. One alternative to the bicubic patch is the Steiner surface with the canonical equation

$$x^2y^2 + x^2z^2 + y^2z^2 - 2xyzw = 0$$

This surface is of fourth degree so that algebraic techniques are still possible. Furthermore, all quadric surfaces are special cases of it.

Another difficulty with simply geometric and topological models of solids is that solid objects have an internal structure that is important. If one is designing an object by editing an already existing solid model of an object, it is important that changes to the object are reasonable. Thus if an object has two holes aligned on an axis, moving one of the holes should move the other. To design reasonable editors requires that a theory be developed to provide some insight into this structure.

Finally, to reason about objects and transformations of objects requires some mathematical framework. We will illustrate the type of theory that is needed by one example. It involves motion of objects in contact. In order to state the theorem we first provide a simple definition of an object and of a motion.

For our purposes an object is a parameterized homeomorphic mapping from a canonical form of the object to a region of three space. If parameters specify the position and orientation of the object, then to instantiate a copy of the object one supplies the x,y,z coordinates and the orientation. However, one can instantiate more than just the location and orientation. There could, for example, be parameters for length, width and height allowing one to

instantiate size. More generally even the shape can be parameterized allowing one to instantiate shape.

A motion is simply a path in parameter space. In this setting a motion may be a translation, rotation, growth or continuous deformation. This general approach allows us to deal with non rigid objects, objects that can change shape.

The type of theorem that can be proved is the following:

Theorem: Given an initial and a final configuration, I and F, of two objects in which the objects are in contact, if there is any motion from configuration I to F, then there is a motion in which the objects stay in contact at all time.

This work is part of a program to develop a theory of solids, how to represent, manipulate and reason about them. This theory, sterophenomenology, is essential to the development of advanced computer automated design and manufacture.

REFERENCES

1. Hopcroft, J.E. and G. Wilfong, "On the Motion of Objects in Contact", Department of Computer Science, Technical Report TR84-602, Cornell University, Ithaca, NY, May 1984.

# DYNAMIC RESPONSE IN AN ELASTIC-PLASTIC PROJECTILE DUE TO NORMAL IMPACT

P. C. T. Chen, J. E. Flaherty, and J. D. Vasilakis
U.S. Army Armament, Munitions, and Chemical Command
Armament Research and Development Center
Large Caliber Weapon Systems Laboratory
Benet Weapons Laboratory
Watervliet, NY 12189

ABSTRACT. A numerical study of the dynamic response of an elastic-plastic projectile due to normal impact has been made using the finite element structural response code ADINA. The projectile is a finite length circular cylindrical bar striking a rigid target. First, three (central-difference, Newmark, Wilson) direct integration schemes were used for uniaxial stress wave problems in a linear-hardening material and the results were compared with an exact analytical solution in order to evaluate accuracy and stability. Then, additional numerical results for perfectly-plastic materials were obtained in order to show the effect of strain-hardening. Finally, some results for a multi-linear material model based on two-dimensional elements are presented in order to show the lateral effect.

I. INTRODUCTION. The propagation of elastic-plastic waves in long rods has been treated extensively in the literature [1-3] since the pioneering works of Donell, Karman, Taylor, and Rakhmatulin. The study of plastic wave propagation is important because it attempts to explain the response of materials to intense dynamic loading and serves also as a basis for determining dynamic material properties.

Analytical solutions can be obtained for only a few idealized situations; hence, many impact studies have been performed using numerical methods. Many computer codes using either finite-element or finite-difference approaches have been developed. The computer simulation of impact phenomena in solids is still quite involved and it depends critically on the impact velocity. For high velocity impact and penetration problems, a good review was given by Zukas [3]. For low velocity contact-impact problems, many structural response codes were reviewed by Noor [4].

In the present paper, a numerical study of the dynamic response of an elastic-plastic projectile due to normal impact is made using the finite element structural response code ADINA [5]. The projectile is a finite length cylindrical bar made of a high strength steel. The bar is long and travels with velocity V = 75 m/s before it strikes a rigid target. First, three direct integration schemes are used for uniaxial stress wave problem in a linear-hardening material and the results are compared with an exact analytical solution in order to evaluate the accuracy and stability. Then, additional numerical results for perfectly-plastic materials are obtained in order to show the effect of strain-hardening for a multi-linear material model. Finally, some results based on two-dimensional elements are presented in order to show the lateral effect.



PREVIOUS PAGE IS BLANK

II. ANALYTICAL SOLUTION. The problem of the normal impact of a rod against a rigid target has been considered by many authors. Various schemes were used for different kinds of initial conditions and various material properties. For a linear work-hardening material due to sudden impact, an analytical solution for the uniaxial stress wave problem is available [1] and it is presented here for comparison with the corresponding ADINA results. Thus, consider a bar of length L and diameter D that is moving with velocity V in the negative Z direction. At time 0 the bar strikes a rigid wall Z = 0 (Figure 1a). Guided by the stress-strain curve for a high strength steel supplied to us [6], the following material data will be used

$$E = 208 \text{ GPa}, \qquad \rho = 0.783 \text{ g/cc}, \qquad \nu = 0.293$$
$$\sigma_y = 1.3 \text{ GPa}, \qquad E_p = 4 \text{ GPa}, \tag{1}$$

where E, $\rho$, $\nu$, $\sigma_y$, and $E_p$ are Young's modulus, density, Poisson's ratio, yield stress, and plastic modulus, respectively. The elastic and plastic one-dimensional wave speeds will then be $c_e = \sqrt{E/\rho} = 5154$ m/s and $c_p = \sqrt{E_p/\rho} = 715$ m/s, respectively. The velocity $V_y$ corresponding to the yield stress $\sigma_y$ of the material is

$$V_y = \sigma_y/(\rho c_e) = c_e \varepsilon_y = 32.21 \text{ m/s} \tag{2}$$

Both elastic and plastic waves will be generated if the impact velocity $V > V_y$. Let us consider the case V = 75 m/s, L = 1.1 m, D = 0.1 m. After impact, two shock wave fronts delimit three distinct regions in the bar (Figures 1b and 1c). The analytical solutions for the particle velocity, strain, and stress in these three regions are

$$V_0 = -V = -75 \text{ m/s}, \qquad \varepsilon_0 = 0, \qquad \sigma_0 = 0,$$
$$V_1 = -V + V_y = -42.79 \text{ m/s}, \qquad \varepsilon_1 = -\varepsilon_y = -0.625\%, \qquad \sigma_0 = -\sigma_y = -1.3 \text{ GPa},$$
$$V_2 = 0, \qquad \varepsilon_2 = -\varepsilon_y + (V_y - V)/c_p = -6.610\%$$
$$\sigma_2 = -\sigma_y + E_p(\varepsilon_2 + \varepsilon_y) = -1.539 \text{ GPa} . \tag{3}$$

After time $t = L/c_e = 213$ μs, the elastic wave front is reflected from the free end. Behind this front (Figure 1d), $\sigma_3 = \varepsilon_3 = 0$, $V_3 = 2V_y - V$. At time $t_S = 2L/(c_e + c_p) = 374.85$ μs, the wave fronts of the plastic and of the elastic unloading waves meet at the section S. The stress and velocity are continuous but the strain is discontinuous across this section S, a non-propagable discontinuity surface. Since the inequality

$$2V_y < V < \left(1 + \frac{2c_e}{c_e + c_p}\right) V_y = 88.79 \text{ m/s} \tag{4}$$

is satisfied, the plastic wave stops at S and elastic waves again propagate from S in both directions (Figure 1e). The analytical solutions in regions 4 and 5 are

$$V_4 = V_5 = \frac{1}{2}\left(3 - \frac{c_e}{c_p}\right)(V_y - V) + V = 13.79 \text{ m/s}$$

$$\varepsilon_4 = \frac{1}{2}(1+c_p/c_e)(\varepsilon_y - V/c_e) = -0.473\%$$

$$\varepsilon_5 = -\left(2\frac{c_e}{c_p} + 1 - \frac{c_p}{c_e}\right)(\varepsilon_y - V/c_e) = -6.342\%$$

$$\sigma_4 = \sigma_5 = -0.983 \text{ GPa} \tag{5}$$

Figures 1b through 1e show the locations of the wave fronts at time t = 100, 200, 300, 400 μs, respectively. The analysis can be continued until the contact between the bar and the target ceases at $t_c = 4L/(c_e+c_p) = 749.7$ μs. More detailed information about the analytical solution can be found in the book by Cristescu [1].

III. ADINA SOLUTION. The ADINA code, developed by K. J. Bathe, is a general purpose finite element program for Automatic Dynamic Incremental Nonlinear Analysis [5]. In nonlinear analysis the incremental finite element equations of motion used are, in implicit time integration,

$$\underline{M}\,{}^{t+\Delta t}\ddot{\underline{U}} + \underline{C}\,{}^{t+\Delta t}\dot{\underline{U}} + {}^{t}\underline{K}\,\underline{U} = {}^{t+\Delta t}\underline{R} - {}^{t}\underline{F} \tag{6}$$

and in explicit time integration,

$$\underline{M}\,{}^{t}\ddot{\underline{U}} + \underline{C}\,{}^{t}\dot{\underline{U}} + {}^{t}\underline{K}\,\underline{U} = {}^{t}\underline{R} - {}^{t}\underline{F} \tag{7}$$

where $\underline{M}$, $\underline{C}$, ${}^{t}\underline{K}$, ${}^{t}\underline{R}$, ${}^{t+\Delta t}\underline{R}$, ${}^{t}\underline{F}$ are constant mass matrix, constant damping matrix, tangent stiffness matrix at time t, external load vector applied at time t, t+Δt, nodal point force vector equivalent to the element stresses at time t, respectively, and $\underline{U}$ is the vector of nodal point displacement increments from time t to time t+Δt, i.e., $\underline{U} = {}^{t+\Delta t}\underline{U} - {}^{t}\underline{U}$. The solution of Eq. (6) yields, in general, an approximate displacement increment $\underline{U}$. To improve the solution accuracy and in some cases to prevent the development of numerical instabilities, it may be necessary to use equilibrium iteration in each or preselected time steps.

In ADINA, the central difference method is employed for explicit time integration and either the Newmark method or Wilson method are employed for implicit time integration. The integration schemes [7] are given by:

$${}^{t}\ddot{U} = \frac{1}{(\Delta t)^2}\left({}^{t+\Delta t}U - 2{}^{t}U + {}^{t-\Delta t}U\right) \tag{8}$$

$${}^{t}\dot{U} = \frac{1}{2\Delta t}\left({}^{t+\Delta t}U - {}^{t-\Delta t}U\right) \tag{9}$$

for the central difference method

$$\dot{t}^{t+\Delta t}U = \dot{t}U + [(1-\delta)\ddot{t}U + \delta\ddot{t}^{t+\Delta t}U]\Delta t \tag{10}$$

$$t^{t+\Delta t}U = tU + \dot{t}U\Delta t + [(\frac{1}{2} - \alpha)\ddot{t}U + \alpha\ddot{t}^{t+\Delta t}U](\Delta t)^2 \tag{11}$$

for the Newmark method, and

$$\ddot{t}^{t+\Delta t}U = \ddot{t}U + (\frac{\tau}{\theta\Delta t})(\ddot{t}^{t+\theta\Delta t}U - \ddot{t}U) \tag{12}$$

$$\theta \geq 1, \quad 0 \leq \tau \leq \theta\Delta t$$

for the Wilson method.

The Wilson and Newmark methods are unconditionally stable if $\theta \geq 1.37$ or $\alpha \geq 1/4(1/2+\delta)$, $\delta \geq 1/2$. In our numerical study, we have chosen $\theta = 1.4$, $\alpha = 1/4$, $\delta = 1/2$. In using the central difference method, the time step, $\Delta t$, has to satisfy the Courant condition

$$\Delta t = \Delta t_{cr} = K\Delta\ell/c \quad \text{or} \quad \frac{2}{\omega} \tag{13}$$

where $\Delta\ell$ is the minimum mesh size, $\omega$ is the maximum natural frequency, c is the local sound speed and $K \leq 1$.

IV. NUMERICAL COMPARISON. Consider a bar with the following geometrical and material data: L = 1.1 m, D = 0.1 m, E = 208 GPa, $\rho$ = 0.783 g/cc, $\sigma_y$ = 1.3 GPa, $E_p$ = 4 GPa, subjected to two values of impact velocity: V = 25 m/s or 75 m/s. Since the velocity corresponding to the yield stress $\sigma_y$ of the material is $V_y$ = 32.2 m/s, the impact is elastic for the first case and elastic-plastic for the second case. Analytical solutions are known for both cases. We use 100 one-dimensional truss elements to simulate this uniaxial stress wave problem. In order to satisfy the stability criterion for explicit integration by the central difference method, we have chosen the time step $\Delta t$ = 2μs which is less than the critical time step

$$\Delta t_{cr} = \Delta Z/c_e = 2.[illegible] \text{ s}$$

We use three integration schemes with the same time step. The computations are all stable and the numerical results for the axial stress and velocity when V = 25 m/s are shown in Figures 2 through 5. Figure 2 shows results for the particle velocity and stress along the rod at t = 100, 200, 300, 400, 500 μs when the Wilson method was used. Figure 3 shows the similar results for the Newmark method. The numerical results based on these two methods are less accurate when compared with the results based on the central difference method. Figure 4 shows a comparison of the results for the particle velocity along the length of the bar based on the central difference method, the Newmark method, and the analytical solution at t = 100, 200, 300, 400 μs. A similar comparison between the central difference method, and the analytic solution for the axial stress along the bar is shown in Figure 5. It should be noted that the computations are performed under the assumption that the rod and target remain in contact after impact while the theoretical interval of contact is $t_c = 2L/c_e$ = 426 μs.

Calculations were also performed for elastic-plastic impact with V = 75 m/s using the three integration schemes with the same mesh size and time step. A comparison of numerical results for the axial stress and velocity using the central difference and Newmark methods is shown in Figure 6 at t = 200 μs. The solid and dotted curves represent the results based on the central difference method and Newmark method, respectively. Similar results are shown in Figure 7 at t = 400 μs. As can be seen from these two figures, the central difference method gives more accurate results than the Newmark method. The numerical results based on the Wilson method are compared with those based on the central difference method in Figure 8. This also demonstrates that the numerical results based on the central difference method is more accurate. We may then conclude from this study on elastic as well as elastic-plastic impact that the central difference method will give more accurate results than the other two integration schemes.

V. HARDENING AND LATERAL EFFECTS. After reaching the above conclusion, we use the central difference method for the rest of this study. In order to show hardening effects, we obtained numerical results for displacement, velocity, strain, and stress in a long rod of an elastic-perfectly-plastic material. Figures 9 and 10 show the results of the particle velocity and stress for a linear work-hardening as well as a perfectly-plastic material at t = 300 and 400 μs, respectively. It can be seen from these comparsions that the effect of strain-hardening on the particle velocity and stress is quite significant even though the plastic modulus $E_p$ = 4 GPa is small when compared with the elastic modulus E = 208 GPa.

In order to study lateral effects, we have used two-dimensional four-node quadrilateral ring elements to obtain numerical results. We choose the same mesh size $\Delta r = \Delta z = 0.011$ m and use 50 elements along the length of the bar with L/D = 25. In this arrangement, the new length of the bar is only a half of the original. We have used the same time step $\Delta t$ = 2 μs as in the one-dimensional truss elements. This time step yields stable computations for one-dimensional truss elements but not for two-dimensional quadrilateral ring elements. This seems due to the lateral effect such as the Poisson's ratio. Including the effect of Poisson's ratio ($\nu = 0.293$), the speed of longitudinal elastic wave is $c_d = [(E/\rho)((1-\nu)/(1+\nu))/(1-2\nu)]^{1/2}$ = 5923 m/s, which reduces to $c_e = (E/\rho)^{1/2}$ = 5154 m/s in case of $\nu = 0$. Guided by the stability criterion for linear elastic problems, we shall choose $\Delta t < \Delta t_{cr} = \Delta Z/c_d$ = 0.011/5923 sec = 1.857 μs. For this reason, we have carried the computations for 250 time steps using $\Delta t$ = 1 μs. The total number of time steps is the same for both the one and two-dimensional problems. The length of the bar and time increment for the two-dimensional case are only half of the one-dimensional case. For the two-dimensional case, we have carried out the computations for impact velocities of V = 25 m/s and 75 m/s. The results for the elastic impact (V = 25 m/s) are not shown here. For elastic-plastic impact (V = 75 m/s), we have carried out the computations for a bilinear as well as a multi-linear material model. Seven points are used to represent the stress-strain curve, i.e., ($\sigma$ in GPa, $\varepsilon$ in %) = (1.3, 0.63), (1.355, 1.03), (1.38, 1.83), (1.394, 2.63), (1.415, 4.23), (1.43, 5.83), (1.45, 8.63). The numerical results for the case of a multi-linear material are shown in Figures 11 through 14. Figure 11 shows the effective stress and axial velocity along the length of the bar at the end of 50 and 100 time steps. Curves 1 and 2

represent the results at t = 50 and 100 μs, respectively. Figure 12 shows the axial stresses along the length of the bar. We have 2x2 stations to carry out the numerical integration in the spatial direction in each element. The results for the stresses are calculated at four integration stations. As can be expected from a long rod (L/D = 25), the differences in the results along the stations near the centerline or outside are very small. Figures 13 and 14 show the similar results for the effective stress, particle velocity, and axial stress along the length of the bar at t = 150 and 200 μs. As can be seen in these two figures, the axial stress in the plastic zone shows bigger oscillations than corresponding effective stress. Comparing the results shown in Figures 11 through 14 with the one-dimensional results shown in Figures 6 through 8, we conclude that the transition near the wave front is not as steep as the one-dimensional case and dispersion behind the wave front can be observed. Some of the oscillations are real due to lateral effects such as radial inertia, radial shear, etc., but some are due to numerical errors such as truncation error in the finite element system, approximations in time integration schemes, numerical integration in the spatial directions, etc. It is difficult to identify how much of the oscillations are real and how much are due to numerical error. We hope to develop a numerical model to minimize the numerical error for this purpose while trying to improve the theoretical model.

VI CONCLUSION. Based on our numerical study for uniaxial stress wave problem in a linear-hardening material, the central difference method gives more accurate results than the Wilson and Newmark methods. The effect of hardening is significant and lateral effects due to radial motion needs further study. We plan to improve the theoretical model and to develop a numerical scheme. We hope to compare our numerical results with experiments involving normal impact of cylindrical rods.

## REFERENCES


1. Cristescu, N., Dynamic Plasticity, John Wiley & Sons, Inc., New York, 1967.

2. Nowacki, W. K., Stress Waves in Non-Elastic Solids, Pergamon Press, Oxford, 1978.

3. Zukas, J. A., Nicholas, T., Swift, H. F., Greszczuk, L. B., and Curran, D. R., Impact Dynamics, John Wiley & Sons, Inc., New York, 1982.

4. Noor, A. K., "Survey of Computer Programs for Solution of Nonlinear Structural and Solid Mechanics Problems," Computer and Structures, Vol. 13, 1981, pp. 425-465.

5. Bathe, K. J., "ADINA Users' Manual," Report AR-81-1, ADINA Engineering, Inc., Watertown, MA, 1981.

6. Wright, T. W., Private Communication, 1982.

7. Bathe, K. J., Numerical Methods in Finite Element Analysis, Prentice-Hall, Inc., New Jersey, 1976.

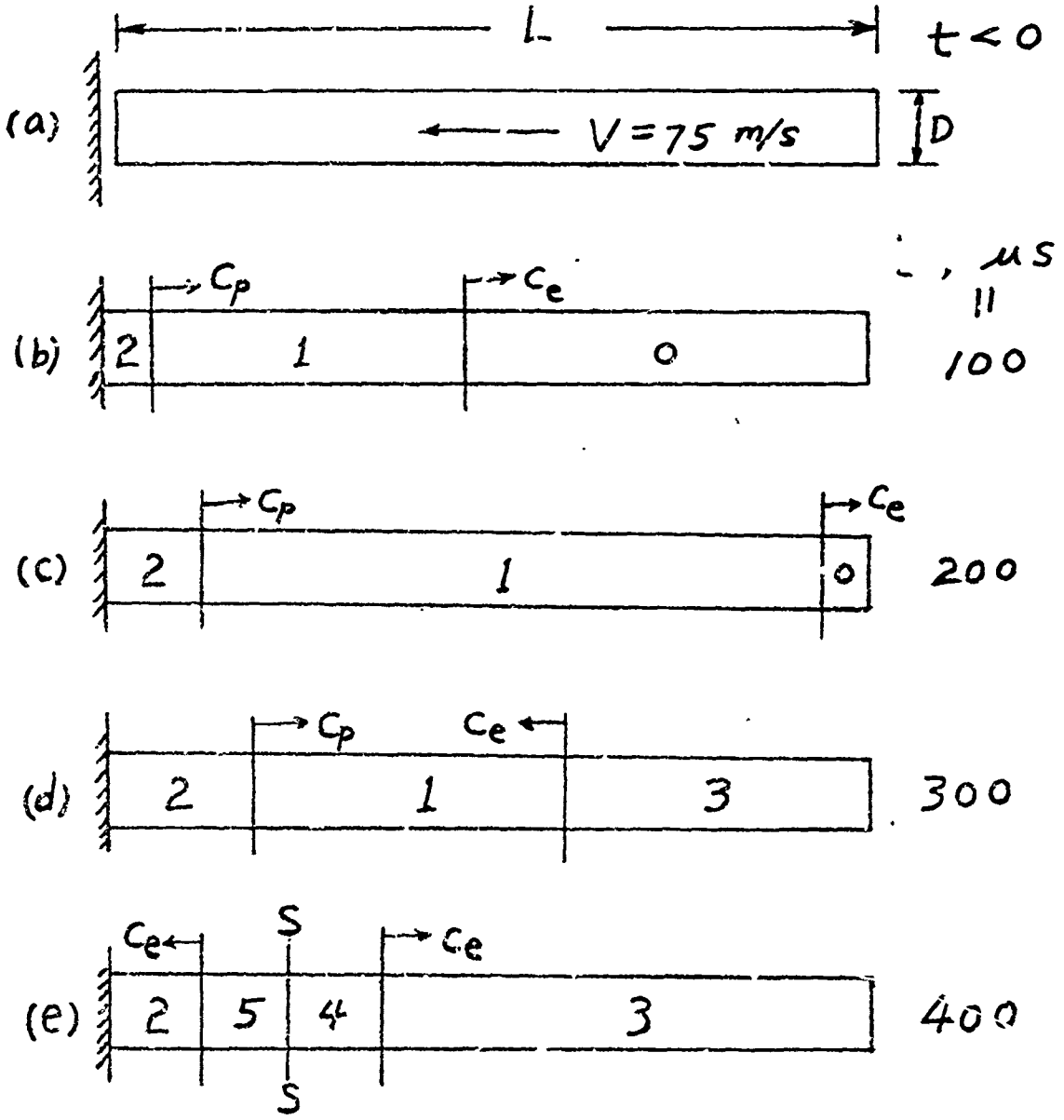
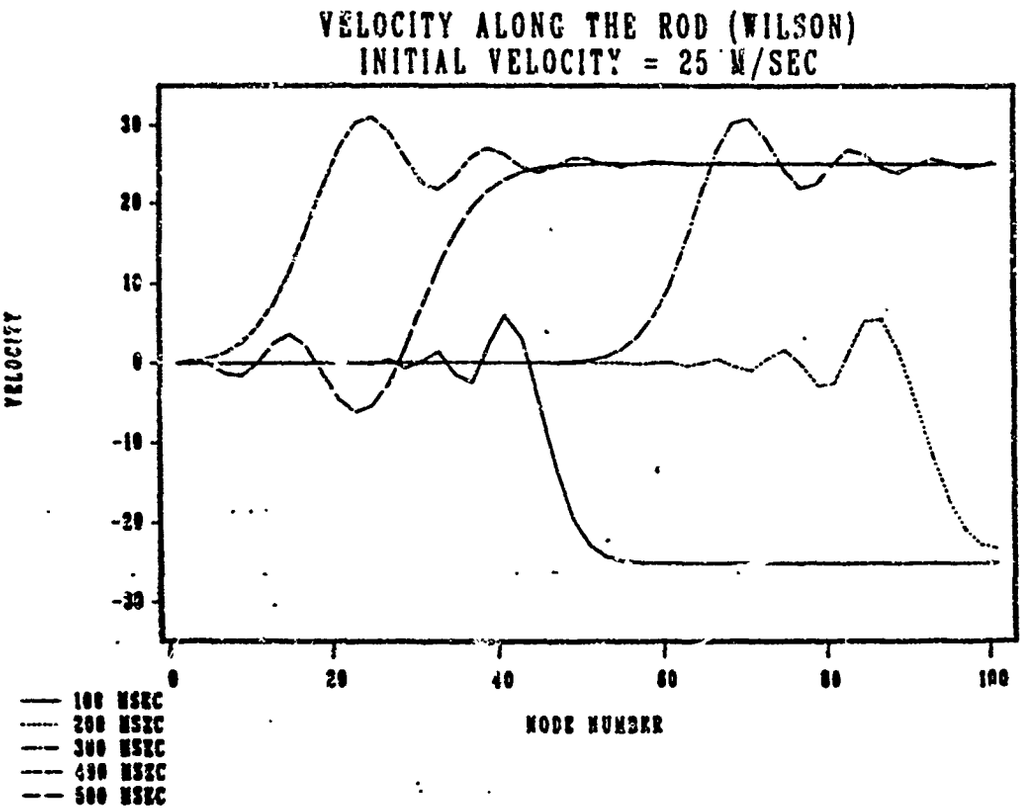



Figure 1. Analytical solution to a projectile striking a rigid target.



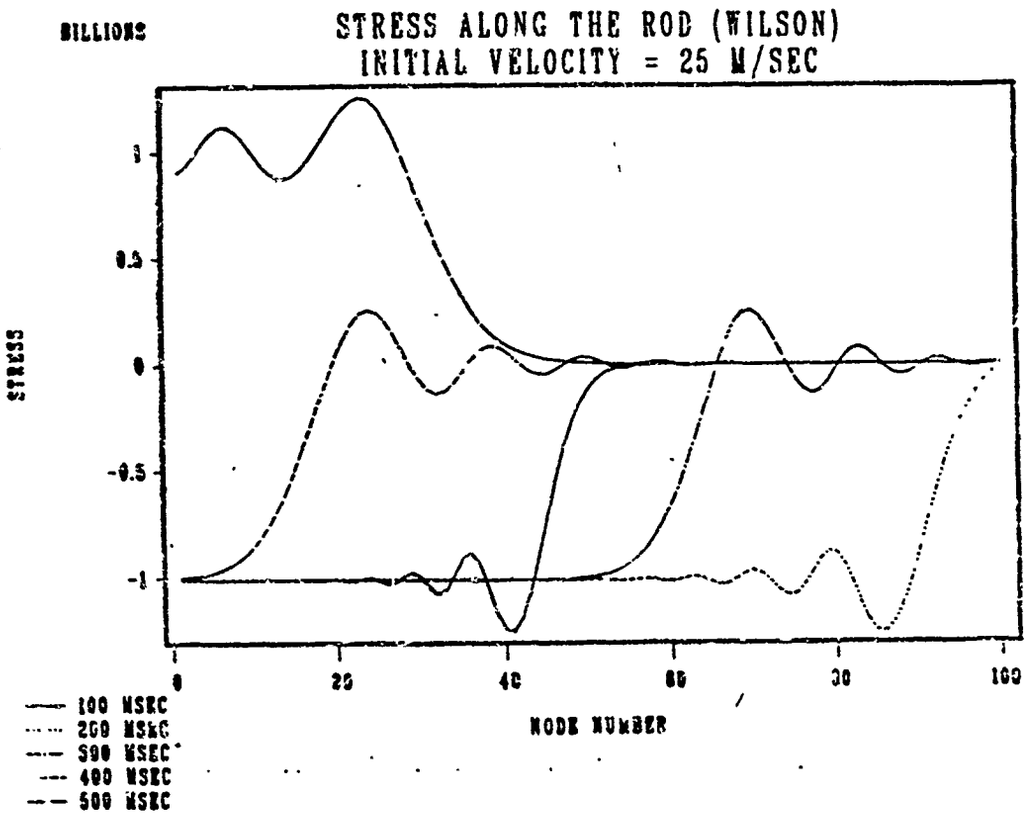
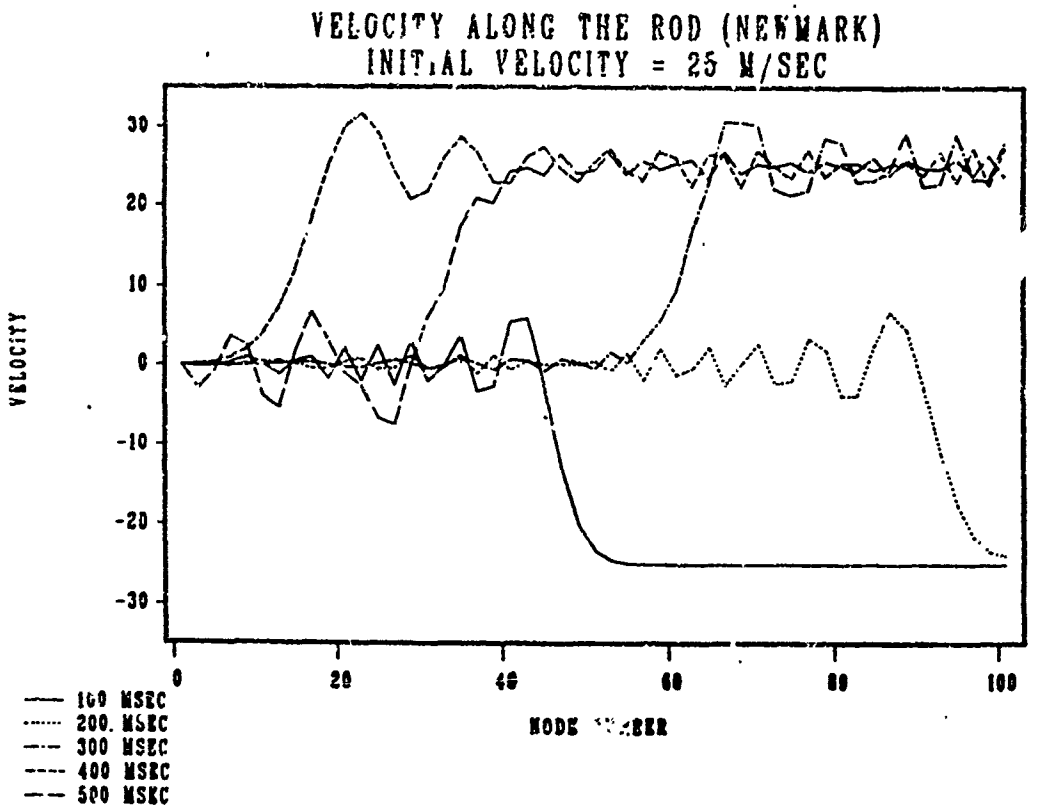




Figure 2. Axial velocity and stress based on the Wilson method (V = 25 m/s).



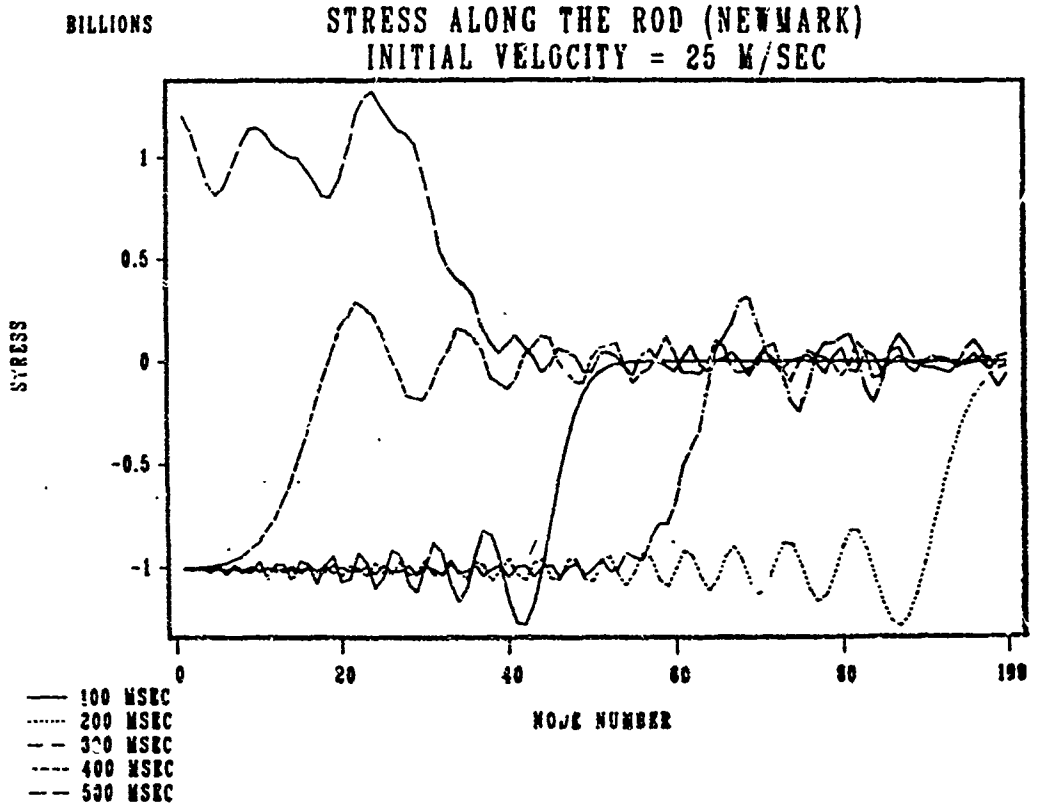




Figure 3. Axial velocity and stress based on the Newmark method (V = 25 m/s).

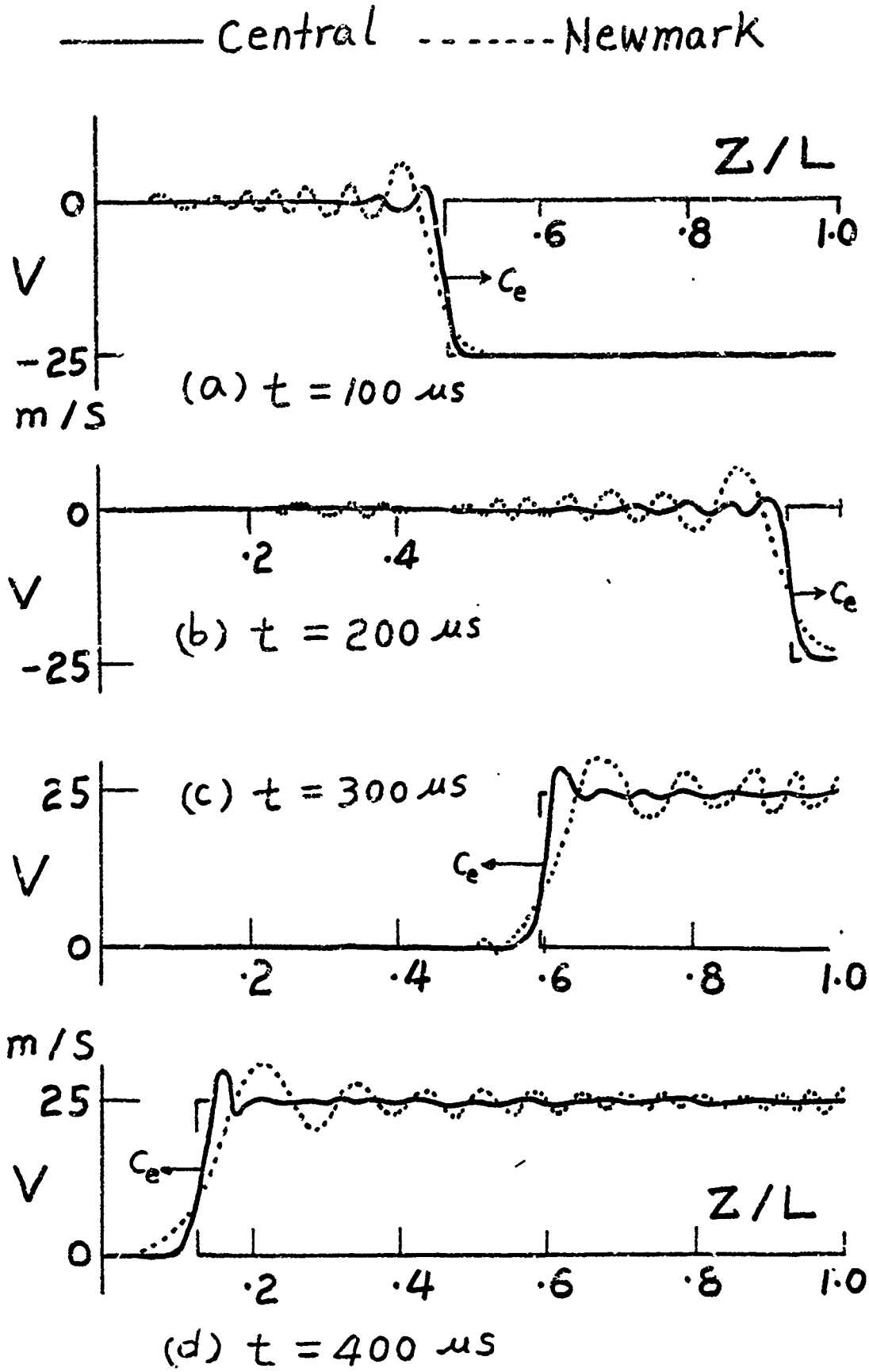



Figure 4. A comparison of the central difference method, the Newmark method, and the analytical solution for the particle velocity (V = 25 m/s).

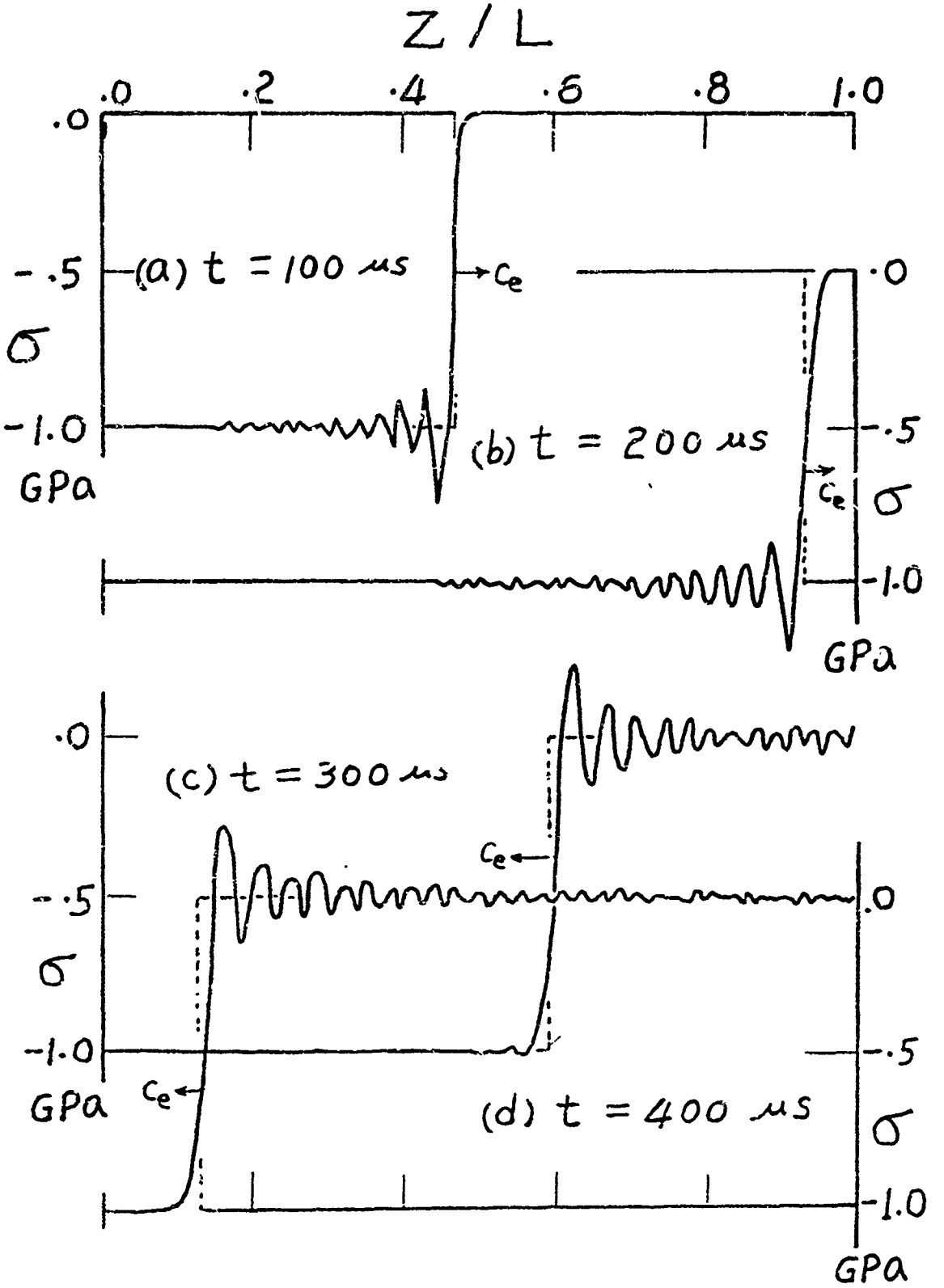



Figure 5. A comparison of the central difference method and the analytical solution for the axial stress (V = 25 m/s).

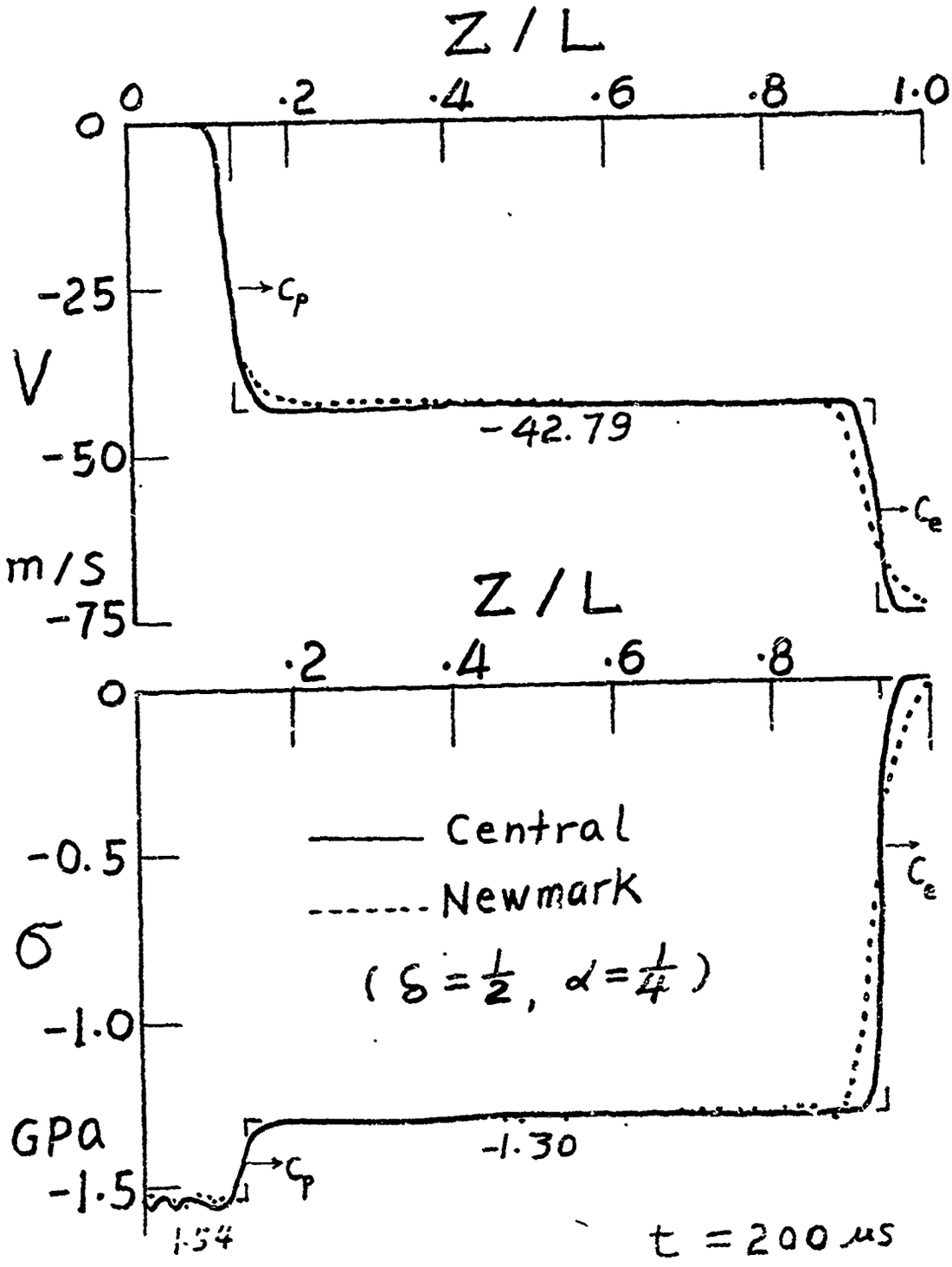



Figure 6. A comparison of the central difference method and the Newmark method for V and σ at t = 200 μs (V = 75 m/s).

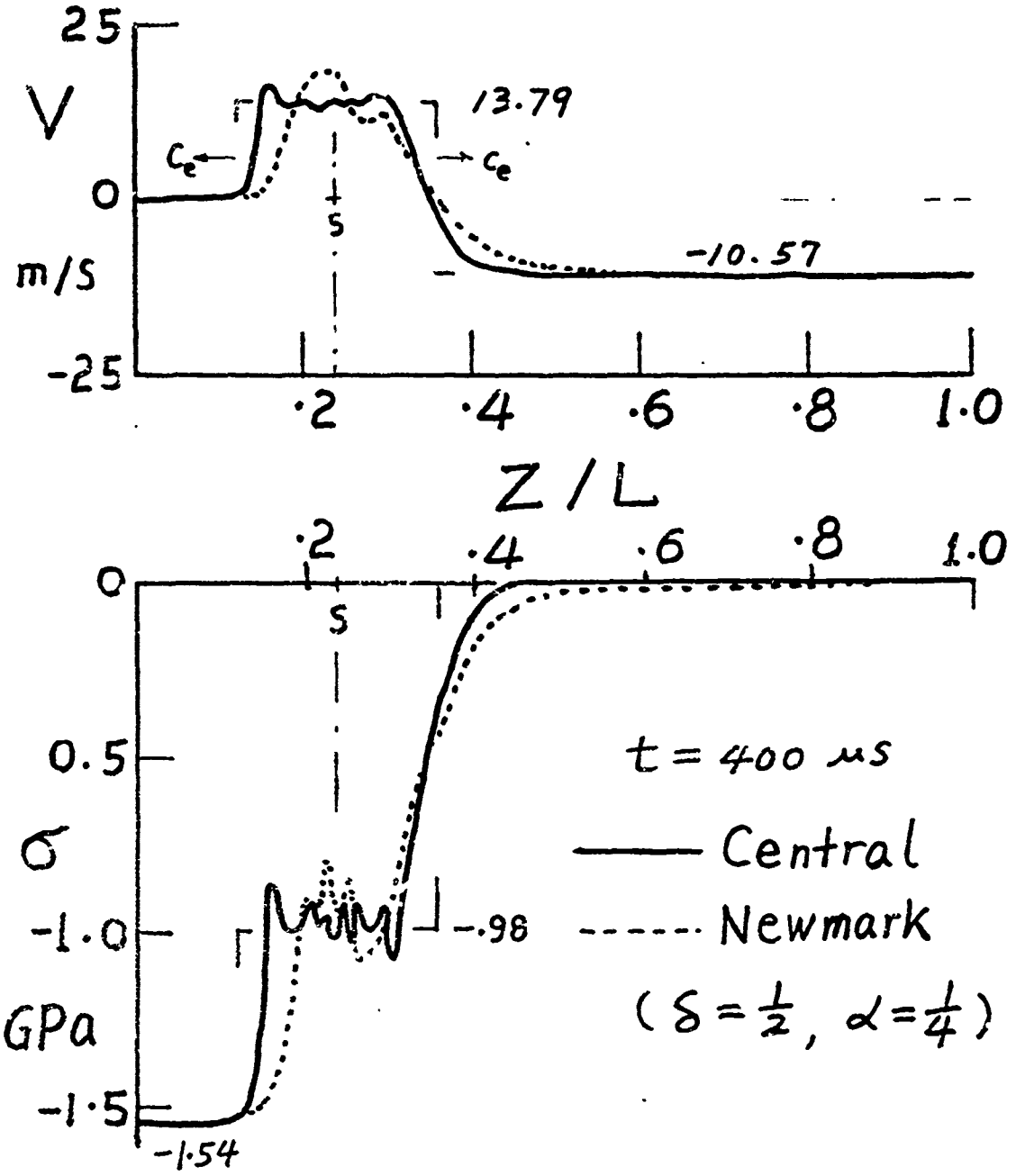



Figure 7. A comparison of the central difference method and the Newmark method for V and σ at t = 400 μs (V = 75 m/s).

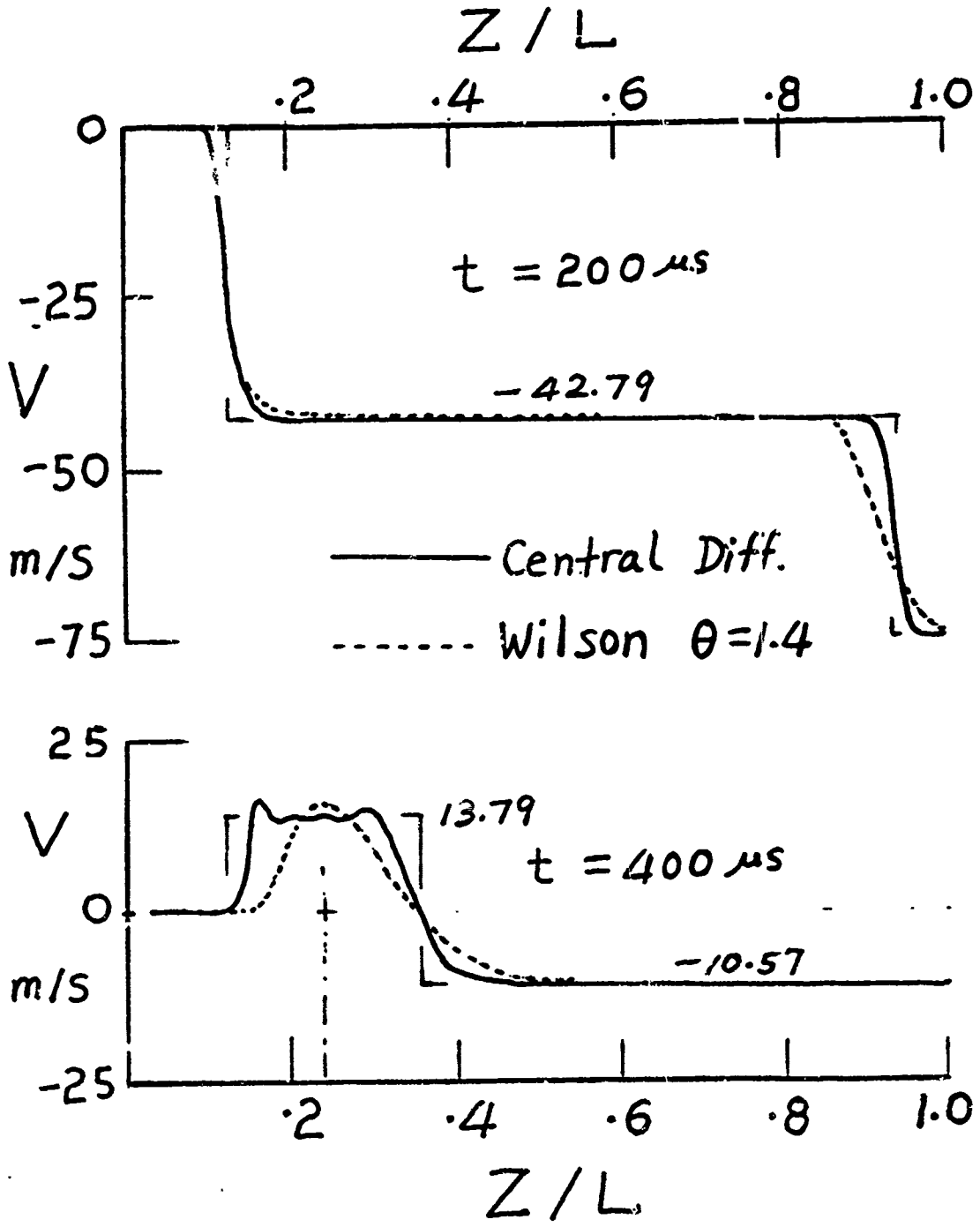



Figure 8. A comparison of the central difference method and the Wilson method for V at t = 200 and 400 μs (V = 75 m/s).

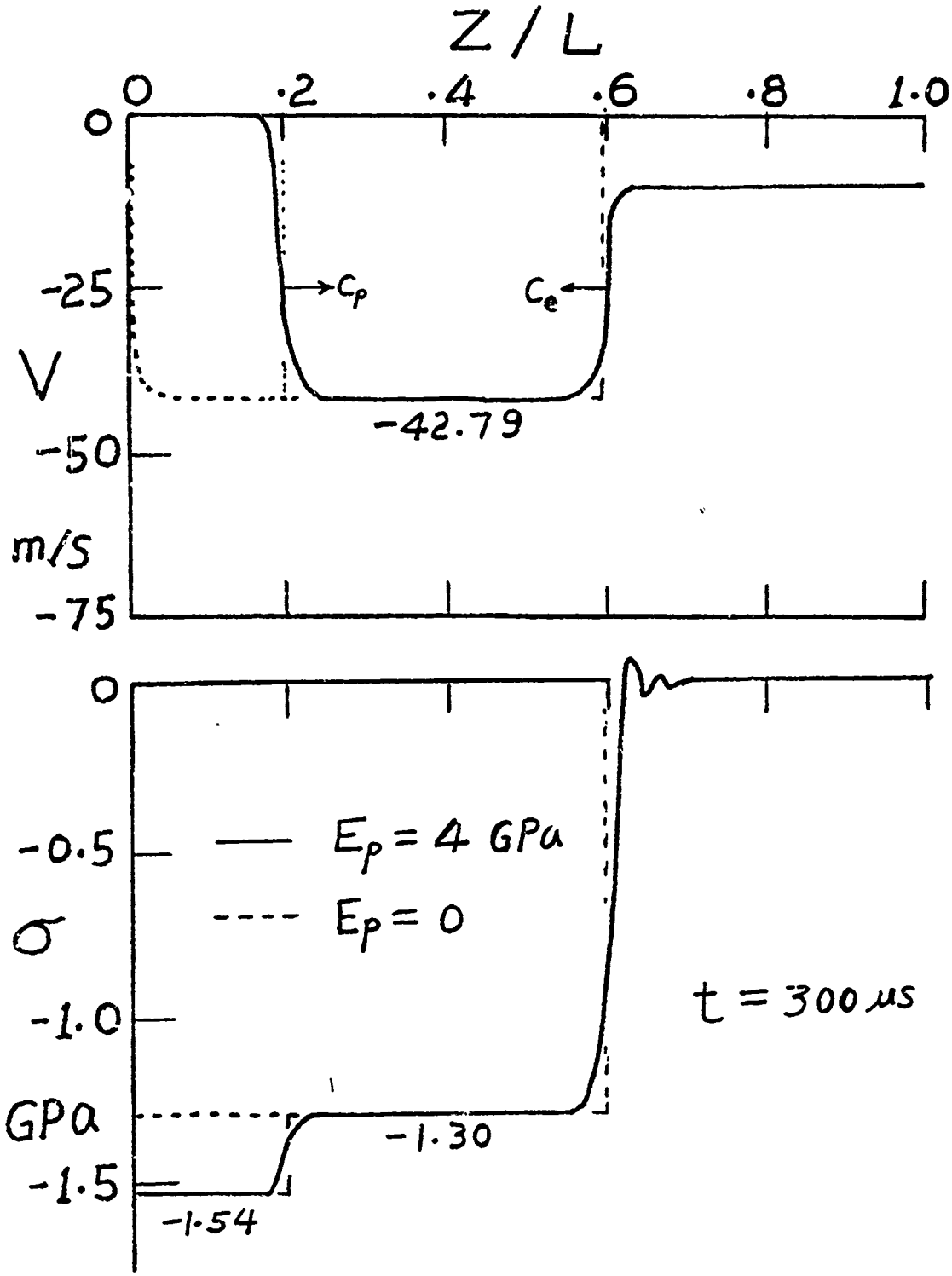



Figure 9. The effect of hardening on V and σ at t = 300 μs.

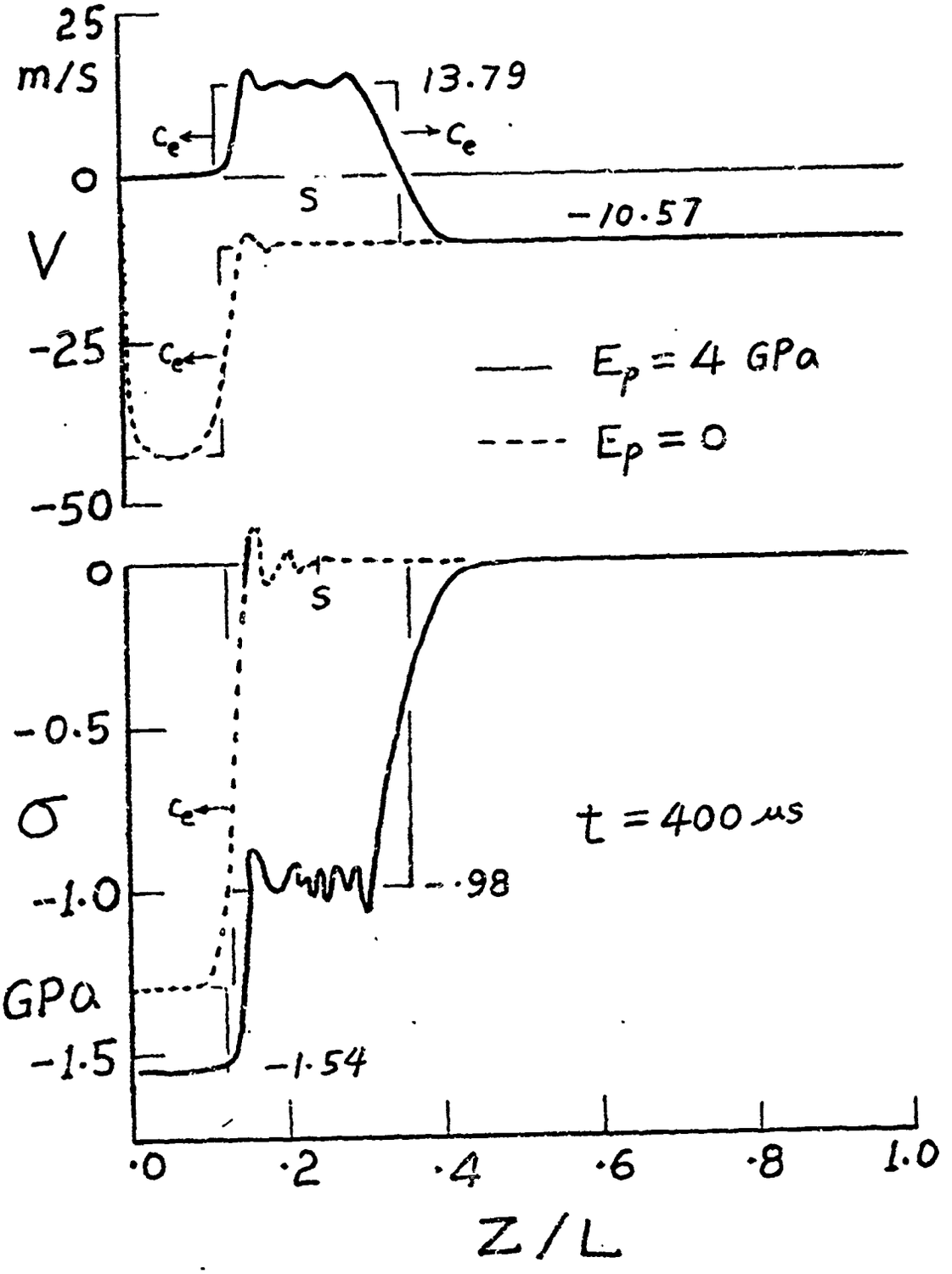



Figure 10. The effect of hardening on V and σ at t = 400 μs.

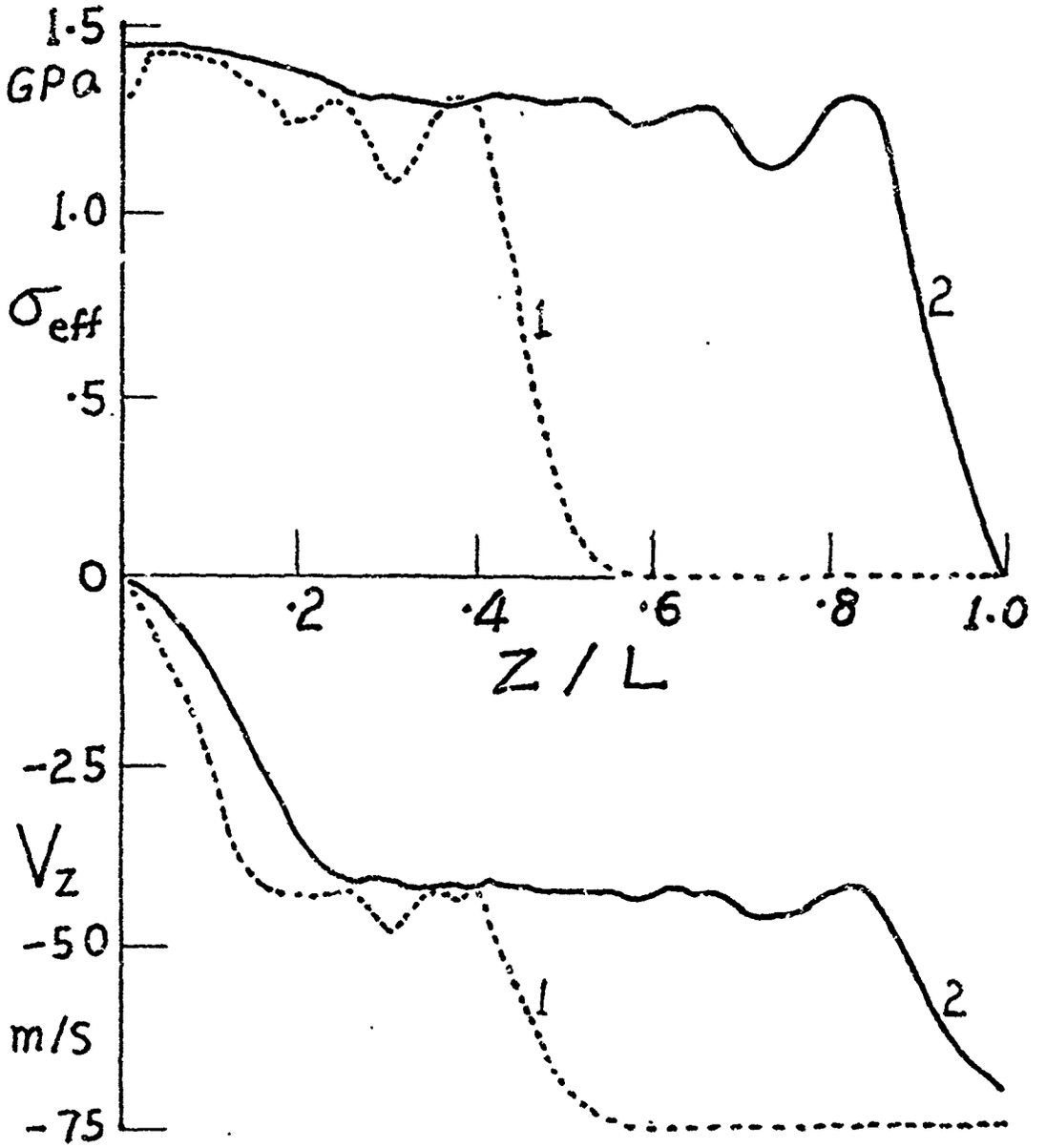
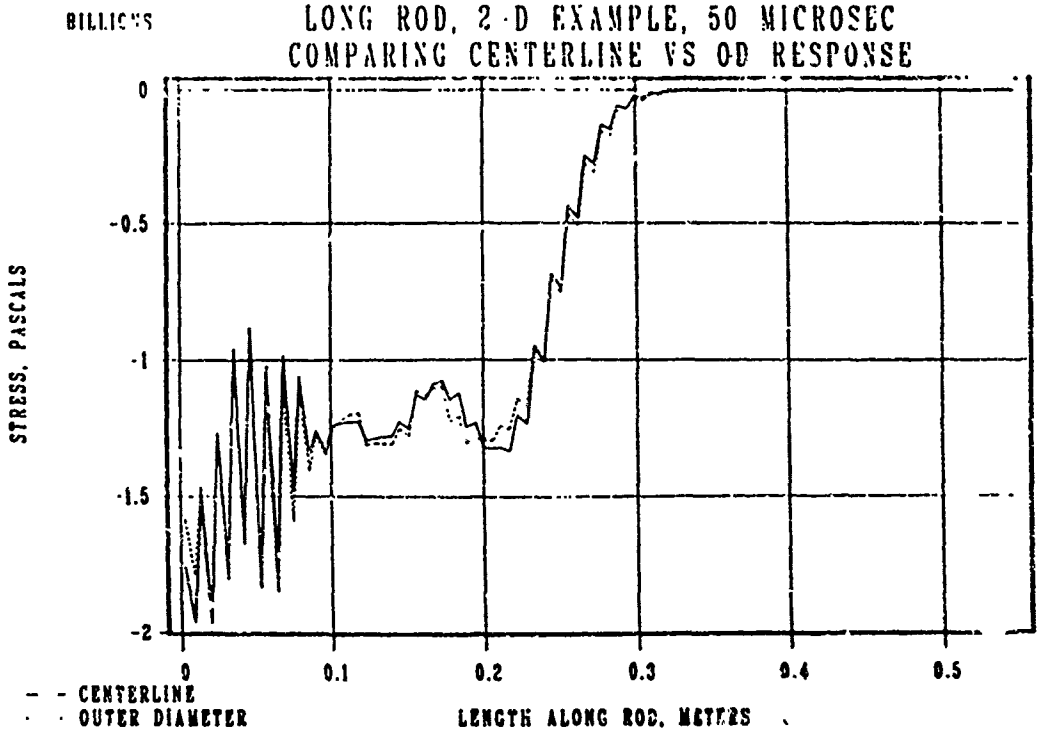



Figure 11. The effective stress and axial velocity based on 2-D elements at t = 50 and 100 μs.



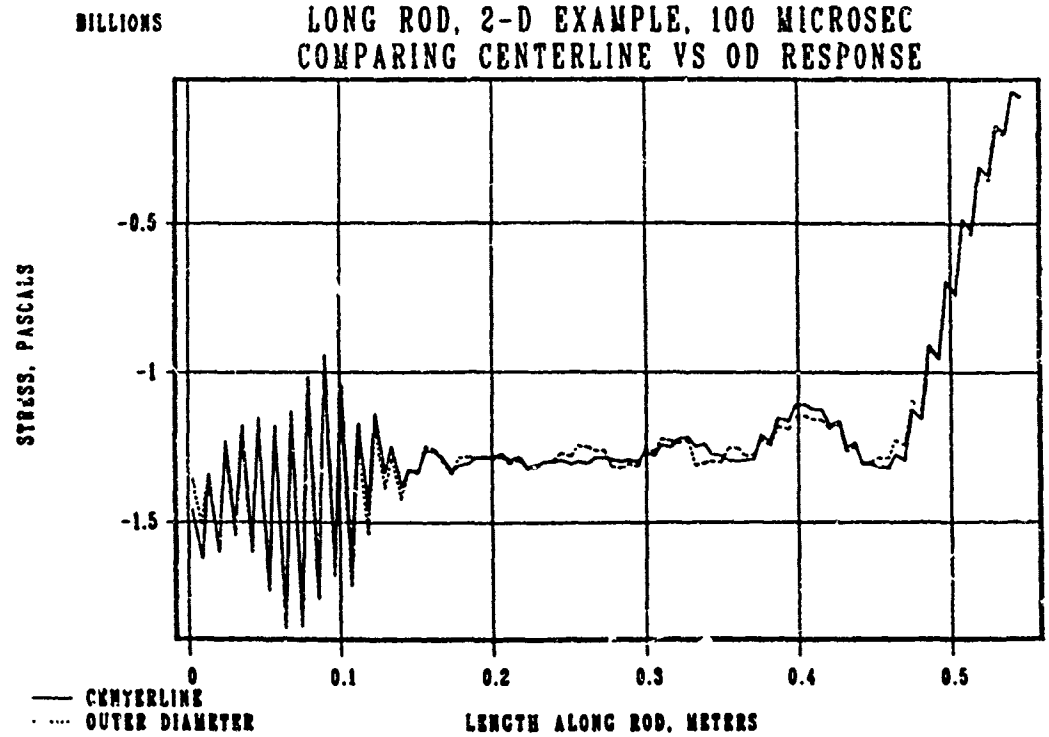




Figure 12. Axial stress ($\sigma_z$) based on 2-D elements at $t$ = 50 and 100 μs.

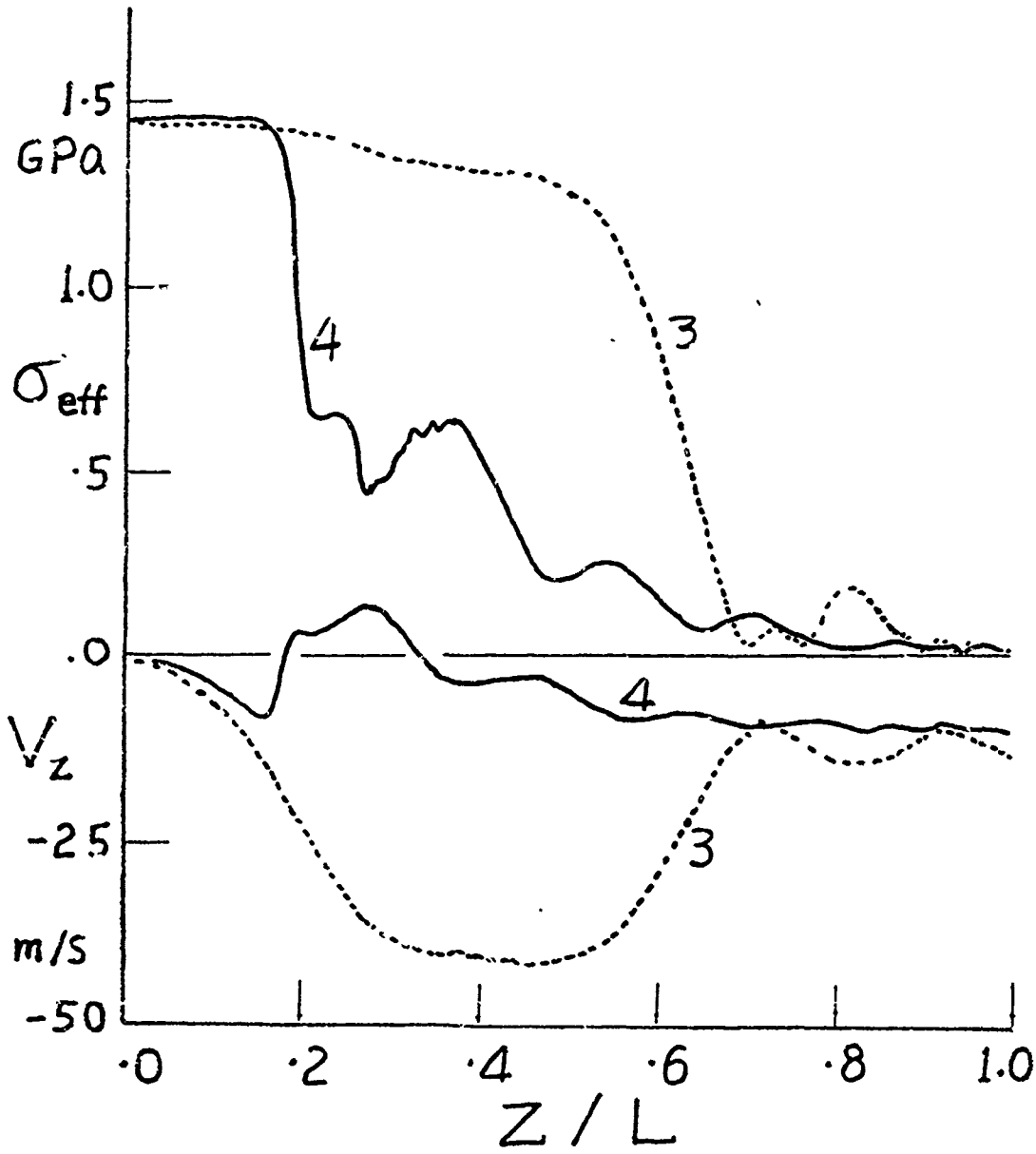
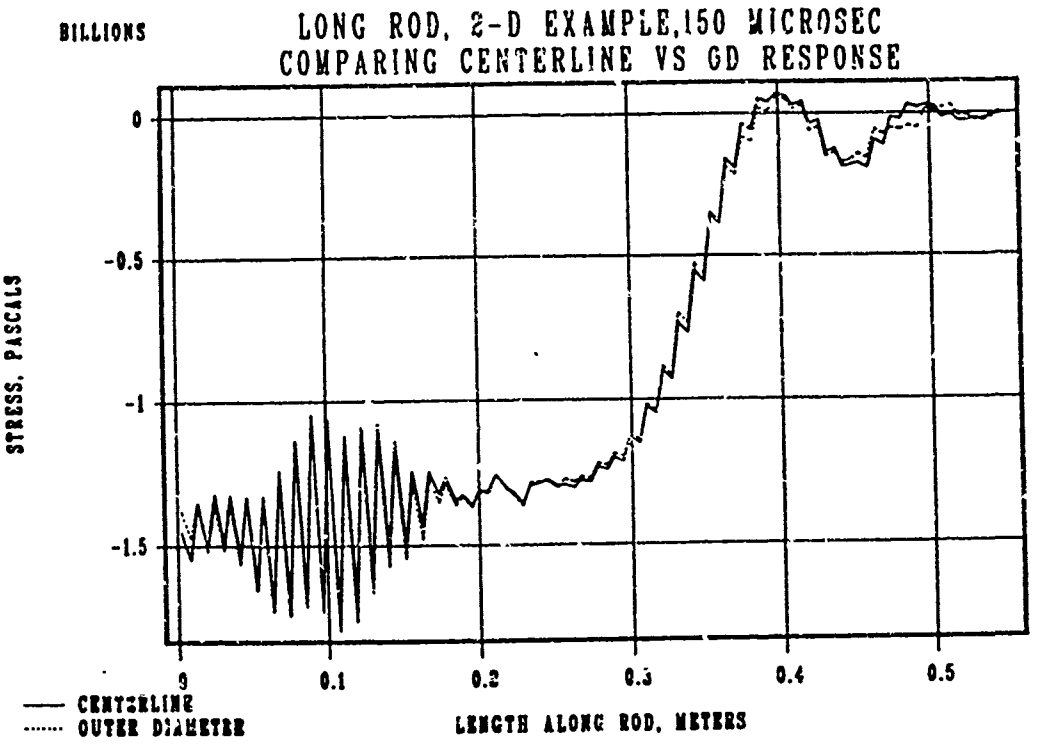



Figure 13. The effective stress and axial velocity based on 2-D elements at t = 150 and 200 us.



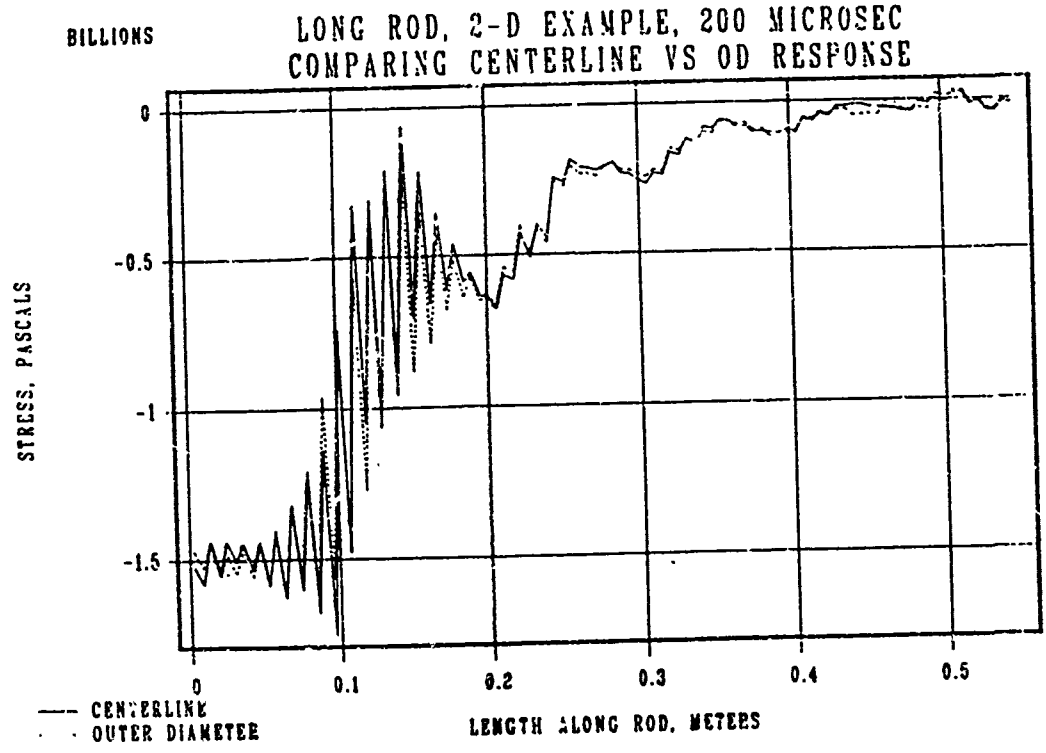




Figure 14.. Axial stress ($\sigma_z$) based on 2-D elements at t = 150 and 200 μs.

# ON THE ASYMPOTIC ANALYSIS OF TRAVELLING SHOCKS AND PHASE BOUNDARIES IN ELASTIC BARS

Thomas J. Pence
Mathematics Research Center
University of Wisconsin-Madison
610 Walnut Street
Madison, WI 53705

I. INTRODUCTION. This investigation is concerned with the propagation of shocks and phase boundaries in elastic solids. Attention is restricted to one-dimensional motions; for simplicity, imagine a bar in which transverse displacements are absent. The problem is introduced in section II, where it is reviewed how change of phase phenomena can be modelled by means of a nonmonotonic stress-strain law. These laws have been studied previously in [1], [2], [3]; a relevant experimental study is [4]. This section also treats the simple wave that develops whenever a nonzero load $\sigma_\infty$ is suddenly applied to the end of the bar. This simple wave would be expected to mirror in some fashion the ultimate state of affairs whenever the bar is gradually loaded at one end to the level $\sigma_\infty$ provided waves are not subsequently reflected back from the opposite end. Issues involved in such an asymptotic study are discussed in the third section. Section IV addresses special considerations for materials in which the stress-strain law is piecewise linear. Then, in section V, we carry out an asymptotic analysis for an example problem involving such a material.

II. FORMULATION OF THE PROBLEM AND THE RIEMANN SOLUTION FOR IMPULSIVE LOADING. Consider a homogeneous, semi-infinite elastic bar which occupies $x \geq 0$ in a reference configuration. Pure longitudinal motion is governed by the momentum equation

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial \sigma}{\partial x}, \tag{2.1}$$

where $u = u(x,t)$ is the longitudinal displacement of the bar and $\sigma$ is the stress along the axis of the bar. The density in the reference state is taken to be one. Let

$$\varepsilon = \frac{\partial u}{\partial x}, \qquad v = \frac{\partial u}{\partial t} \tag{2.2}$$

denote respectively the strain and velocity in the bar. For elastic materials, the stress at time $t$ at a location which was originally at position $x$ is completely determined by the value of $\varepsilon(x,t)$ by means of the constitutive relation $\sigma = \hat{\sigma}(\varepsilon)$. The sound speed of the material can be identified by expressing (2.1) in characteristic form and is found to be $\sqrt{\hat{\sigma}'(\varepsilon)}$. Here the $'$ symbol is the usual shorthand notation for derivative. We shall focus attention on a hypothetical material for which the stress-strain law $\hat{\sigma}(\varepsilon)$ is given by the smooth curve in Fig. 1. Here

Sponsored by the U.S. Army under Contract No. DAAG29-80-C-0041.
This material is based upon work supported by the National Science Foundation under Grant No. MCS-8210950.

$\sigma(0) = 0$ and we shall restrict attention to positive values of $\sigma$ and $\varepsilon$. This entails no loss in generality since compressive motions can be treated by considering a corresponding extensional problem for a material with a stress-strain curve found by reflecting the original curve through the origin. Unlike the state of affairs in gas dynamics, this technique gives correct results even in the presence of shocks [5].

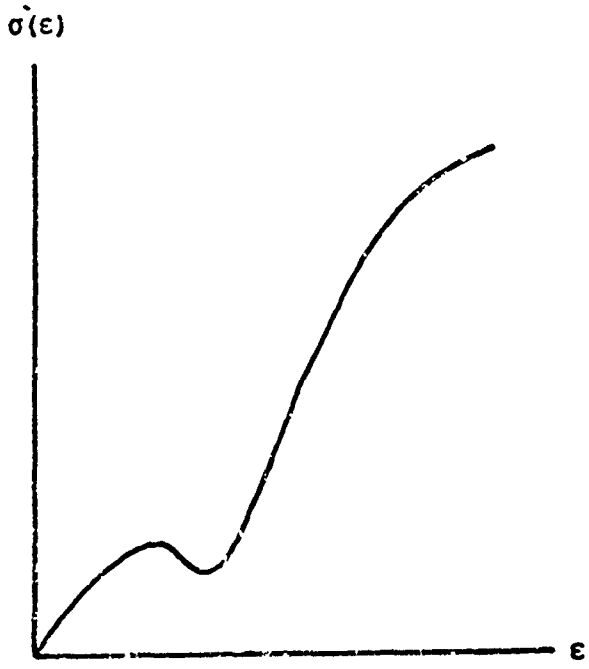



Fig. 1. Stress-strain curve

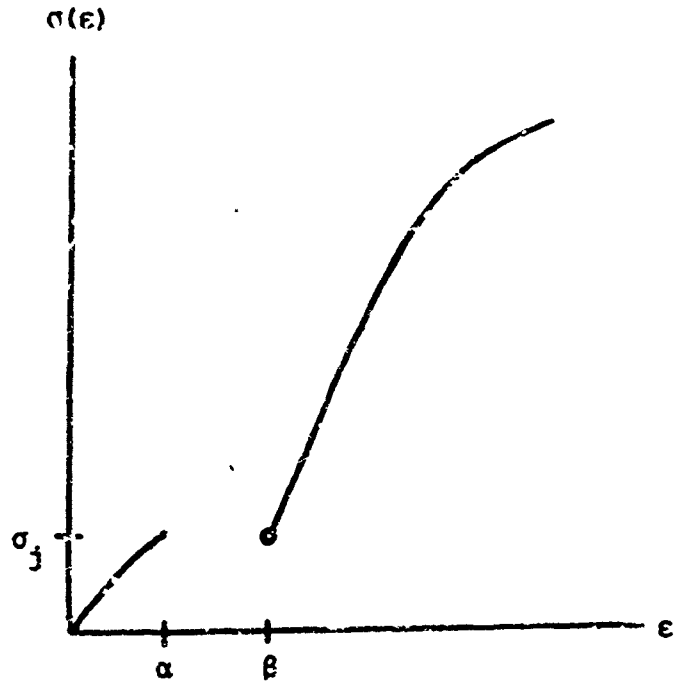



Fig. 2. Restricted stress-strain curve indicating the jump associated with a change of phase

The descending portion of the curve in Fig. 1 is its most conspicuous feature. In an equilibrium setting, strains associated with this portion are found to be unstable [1]. In the nonequilibrium case the sound speed is imaginary and the equation of motion is elliptic at these strains. These difficulties can be overcome by precluding these strains. This is accomplished in a natural fashion by considering an inverse to the $\sigma(\varepsilon)$ relation, namely strain as a function of stress. Since the $\sigma(\varepsilon)$ relation is not monotonic, there are innumerable such inverse functions. We take the particular inverse that leads to the clipped or restricted curve as shown in Fig. 2. The separated strain intervals are now associated with different material phases. Here the value of the transition stress $\sigma_j$ is assumed to be a property of the material (see [4]). Several other methods for dealing with non-monotonic constitutive laws have also been studied (see [6], [7], [8]).

If a load $\sigma_0(t) > 0$ is applied to the end of the bar beginning at time $t = 0$, (2.1) is to be solved subject to

$$\sigma\big(\varepsilon(0,t)\big) = \sigma_0(t), \qquad t > 0, \tag{2.3}$$

$$\varepsilon(x,0) = 0, \qquad v(x,0) = 0, \qquad x > 0 \tag{2.4}$$

The condition (2.4) indicates that the bar is taken to be initially undeformed and at rest. The condition (2.3) can also be written

$$\varepsilon(0,t) = \varepsilon_0(t) , \qquad t > 0 , \tag{2.5}$$

where $\varepsilon_0(t)$ is the strain associated with the stress $\sigma_0(t)$ by means of the curve in Fig. 2, thus

$$\sigma\big(\varepsilon_0(t)\big) = \sigma_0(t) . \tag{2.6}$$

In the event $\sigma_0(t)$ exceeds the value $\sigma_j$, $\varepsilon_0(t)$ will be discontinuous and the strain field $\varepsilon(x,t)$ will necessarily include a discontinuity front associated with a change of phase. In addition, other shock discontinuities of a more familiar kind may arise from the intersection of characteristics associated with different sounds speeds of the nonlinear $\sigma(\varepsilon)$ relation. Across any such discontinuity front, say $x = s(t)$, the jump in field quantities are to be restricted by the shock conditions

$$\frac{ds}{dt}\,\big|\varepsilon\big| + \big|v\big| = 0 , \qquad \frac{ds}{dt}\,\big|v\big| + \big|\sigma\big| = 0 . \tag{2.7}$$

In general the problem given by (2.1) - (2.7) does not admit simple analytical solutions.

An important practical problem associated with this system is that of impulsive loading. By this is meant the situation where $\sigma_0(t)$ is given by $\sigma_0(t) \equiv \sigma_\infty$. In this case the problem, which is given on the quadrant $x > 0$, $t > 0$, is a variant of what is known as the Riemann Problem [2]. The solution follows from the absence of both length and time scales in the initial and boundary conditions. It is given by

$$\varepsilon(x,t) = \tilde{\varepsilon}(\lambda) , \qquad v(x,t) = \tilde{v}(\lambda) \tag{2.8}$$

where $\lambda = x/t$ ; $\tilde{\varepsilon}(\lambda)$ can be found by a construction which is outlined in the following paragraph. Other treatments of similar problems may be found in [2].

Let $\hat{\sigma}(\varepsilon)$ be the upper convex envelope on the interval $[0,\varepsilon_\infty]$ of the clipped curve $\sigma(\varepsilon)$ of Fig. 2. Here $\varepsilon_\infty$ is the root of $\sigma(\varepsilon_\infty) = \sigma_\infty$. The curve $\hat{\sigma}(\varepsilon)$ will in general consist of a number of line segments connecting portions of the clipped curve. The curve $\hat{\sigma}(\varepsilon)$ inherits the differentiability of the original restricted curve $\sigma(\varepsilon)$ and so will be smooth at all values of strain in the interval $0 < \varepsilon < \varepsilon_\infty$ with the possible exception of $\varepsilon = \alpha$. At the value $\varepsilon = \alpha$, the derivative $\hat{\sigma}'(\varepsilon)$ may - or may not - be discontinuous (see Fig. 3.) In the event $\hat{\sigma}'(\varepsilon)$ is discontinuous at $\varepsilon = \alpha$, $\hat{\sigma}'(\alpha)$ is to be regarded as given by the interval $\big[\hat{\sigma}'(\alpha+), \hat{\sigma}'(\alpha-)\big]$. In this way the graph of $\hat{\sigma}'(\varepsilon)$ is a monotonically decreasing curve on $0 < \varepsilon < \varepsilon_\infty$. Hence the equation

$$\hat{\sigma}'(\tilde{\varepsilon}) = \lambda^2 \tag{2.9}$$

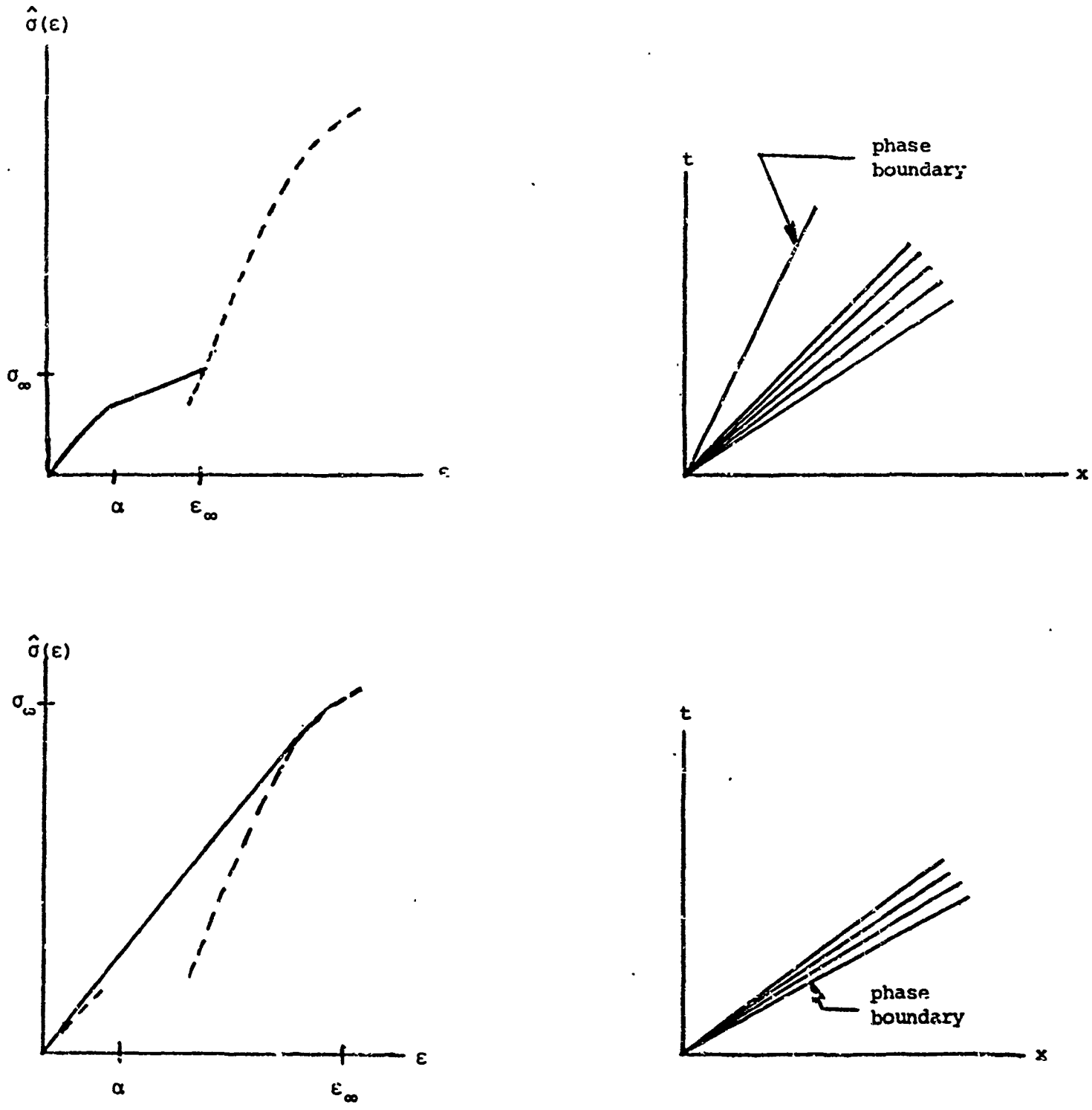



Fig. 3. Upper convex envelopes and corresponding simple waves for two values of $\sigma_\infty$. Here $\hat{\sigma}'(\varepsilon)$ is discontinuous at $\varepsilon = \alpha$ in the example on top; no such discontinuity in $\hat{\sigma}'(\varepsilon)$ occurs in the bottom example.

has solutions $\tilde{\varepsilon}$ for each value of $\lambda$ on the interval $[\sqrt{\hat{\sigma}'(\varepsilon_\infty)},\ \sqrt{\hat{\sigma}'(0)}]$. Whenever $\hat{\sigma}(\varepsilon)$ contains line segments, (2.9) will have an interval of solutions in $\tilde{\varepsilon}$ for certain values of $\lambda$, say $\lambda = \lambda_1 \ldots \lambda_n$. At all other values of $\lambda$, (2.9) uniquely determines a function $\tilde{\varepsilon}(\lambda)$. Next, $\tilde{\varepsilon}(\lambda)$ is extended to $0 \leq \lambda \leq \infty$, $\lambda \neq \lambda_i$, by defining $\varepsilon(\tilde{\lambda}) = 0$ for $\lambda > \sqrt{\hat{\sigma}'(0)} \equiv \lambda_0$ and $\hat{\varepsilon}(\lambda) = \varepsilon_\infty$ for $\lambda < \sqrt{\hat{\sigma}'(\varepsilon_\infty)}$.

This construction orders the sound speeds of the rays $x = \lambda t$ in such a manner that values of strain associated with higher sound speeds are located further down the bar compared to those strain values with lower sound speeds. In the process it naturally positions shock and phase boundary curves $x = \lambda_i t$, as well as ensuring the correct values of $\varepsilon$ on the $x$ and $t$ axes. By virtue of (2.1), (2.7), the function $\tilde{v}(\lambda)$ is found from $\tilde{\varepsilon}(\lambda)$ through

$$\tilde{v}(\lambda) = \begin{cases} 0 & \lambda > \lambda_0 \\ \tilde{v}(\lambda_i^-) - \displaystyle\int_{\varepsilon(\lambda_i^-)}^{\tilde{\varepsilon}(\lambda)} \sqrt{\sigma'(s)}\, ds & \lambda_{i+1} \leq \lambda \leq \lambda_i \end{cases} \tag{2.10}$$

$$(i = 1 \ldots n,\ \lambda_{n+1} \equiv 0)$$

where

$$v(\lambda_i^-) = v(\lambda_i^+) + \lambda_i\left[\varepsilon(\lambda_i^+) - \varepsilon(\lambda_i^-)\right] .$$

The solution (2.8) of the impulsive load problem thus consists of a partitioning of the quadrant $x > 0$, $t > 0$ into sectors by rays through the origin. Across these rays the solution may be either smooth or discontinuous. In the sectors, the strain and velocity fields are either constant or arise from solutions of (2.9). In both cases it can be shown that at least one of the two Riemann invariants associated with (2.1) remains constant. This being the case, the solution is said to be a simple wave.

It is worth mentioning that if at some time $t_1 > 0$ the bar is subsequently impulsively unloaded back to $\sigma = 0$, the wave pattern which results (before any interactions with the loading waves which at $t_1$ are further down the bar) can be found by an analogous construction involving lower convex envelopes of $\sigma(\varepsilon)$.

III. NONIMPULSIVE LOADING. Whenever the load $\sigma_0(t)$ is not simply a positive constant, the solution of the system (2.1) - (2.7) is not a simple wave similarity solution like that of the previous section. Instead one must take account of both families of characteristics associated with (2.1). Suppose, however, that the applied load eventually attains or merely approaches the final value $\sigma_\infty$. In this event, one expects the corresponding simple wave solution with $\sigma_0(t) \equiv \sigma_\infty$ to give the ultimate number of shocks

and phase boundaries, and to yield the correct limiting order of the waves in the bar as $t$ tends to infinity. The spacing of these waves will depend on the manner in which $\sigma_0(t) \to \sigma_\infty$, nevertheless the value of the dynamical fields in the limit $\lambda = x/t$ fixed, $t \to \infty$ is given by the simple wave solution.

The approach to the asymptotic state may be studied by decomposing the displacement $u$ into the corresponding simple wave solution, and a correction to be denoted by $\hat{u}$. By virtue of (2.8), (2.2) the former is expressed $u_0(\lambda)t$; the latter $\hat{u}$ is assumed to be $o(t)$ as $t \to \infty$, $\lambda$ fixed.

Thus

$$u = u_0(\lambda)t + \hat{u}(\lambda,t) , \tag{3.1}$$

while the governing equation (2.1) becomes

$$\frac{2\lambda}{t^2} u_\lambda + \frac{\lambda^2}{t^2} u_{\lambda\lambda} - \frac{2\lambda}{t} u_{\lambda t} + u_{tt} - \sigma'\left(\frac{1}{t} u_\lambda\right) \frac{1}{t^2} u_{\lambda\lambda} = 0 . \tag{3.2}$$

Here subscripts of $\lambda$ and $t$ denote partial differentiation. Entering (3.2) with (3.1) one obtains to leading and second order

$$\begin{aligned} 0 = & \left\{\frac{1}{t}\left[\lambda^2 - \sigma'(u_0')\right]u_0''\right\} \\ & + \left\{\frac{1}{t^2}\left[\lambda^2 - \sigma'(u_0')\right]\hat{u}_{\lambda\lambda} - \frac{2\lambda}{t}\hat{u}_{\lambda t} + \hat{u}_{tt} - \frac{1}{t^2}\left[2\lambda - \frac{d}{d\lambda}\sigma'(u_0')\right]\right\} \\ & + \text{terms higher than second order.} \end{aligned} \tag{3.3}$$

In arriving at the above result it is necessary to make use of the expansion

$$\sigma'\left(\frac{1}{t} u_\lambda\right) \sim \sigma'(u_0') + \sigma''(u_0') \frac{1}{t} \hat{u}_\lambda .$$

The expression in the last parenthesis of (3.3) dominates all succeeding expressions for large times and so must independently vanish. Thus one draws

$$u_0(\lambda) = c_1\lambda + c_2 \tag{3.4}$$

for constants $c_1$, $c_2$, or

$$u_0(\lambda) = \int^\lambda \sigma'^{-1}(s^2)ds \tag{3.5}$$

where the superscript $^{-1}$ denotes functional inverse. It is easily verified that (3.4) gives rise to the constant strain and velocity fields of the impulsive load problem, while (3.5) corresponds to solutions of (2.9). The Riemann solution discussed in the previous section is the unique way in which it is possible to assemble these solutions (3.4) and (3.5) in a manner which satisfies (2.4), (2.7), $\sigma_0(t) \equiv \sigma_\infty$ and also jumps over the specified interval $(\alpha,\beta)$.

The nature of the approach to this simple wave is governed by the expression in the second parenthesis of (3.3). Requiring this expression to vanish leads to

$$-2\lambda\, \hat{u}_{\lambda t} + t\, \hat{u}_{tt} = 0 \tag{3.6}$$

in the case of (3.5), while for (3.4) it is easiest to revert back to independent variables $x$ and $t$ whereupon one finds that the equation can be written

$$\hat{u}_{tt} = \sigma'(c_1)\hat{u}_{xx} .$$

The latter has the familiar solution

$$\hat{u} = A(x - \sqrt{\sigma'(c_1)}\, t) + B(x + \sqrt{\sigma'(c_1)}\, t) , \tag{3.7}$$

for any functions $A(\cdot)$ and $B(\cdot)$. The former (3.7) is also easily solved

$$\hat{u} = \frac{1}{\sqrt{\lambda}}\, \bar{C}(\sqrt{\lambda}\, t) + \bar{D}(\lambda) = t\, C(xt) + D(x/t) \tag{3.8}$$

for any functions $C(\cdot)$ and $D(\cdot)$. Unfortunately the boundary condition (2.3) cannot be used directly for determining the free functions which have emerged from this treatment. Instead one anticipates appropriate conditions to arise from a more penetrating analysis of various short- and intermediate-time solutions to the problem, each of which is appropriate in a different region of the $(x,t)$-quadrant. In general one expects to continue the fields between these as yet unknown regions through a detailed matching layer analysis. Rather than pursuing this program, a method appropriate to certain special materials will be introduced.

IV. PIECEWISE LINEAR STRESS RESPONSE We turn now to a class of model materials in which the stress-strain curve consists of a number of linear segments. Such models are often associated with the theory of plasticity, nevertheless here we shall assume that loading and unloading follow the same stress-strain curve.

The fields arising from an impulsive load can be found by the procedure discussed in section 2. As before, each line segment in the upper convex envelope $\hat{\sigma}(\varepsilon)$ of $\sigma(\varepsilon)$ is associated with a front across which the strain suffers a discontinuity. For the materials now under consideration, $\hat{\sigma}(\varepsilon)$ will consist of nothing but line segments. It is convenient to distinguish among three types of discontinuity fronts by drawing a distinction between the line segments comprising $\hat{\sigma}(\varepsilon)$. We shall say

a phase boundary is a front which is associated with a line segment which spans a clipped portion of the original curve $\sigma(\varepsilon)$,

a contact discontinuity is a front which is associated with a line segment which coincides with the original curve $\sigma(\varepsilon)$, and

a shock is a front which is associated with a line segment which neither coincides with the original curve, nor spans a clipped portion of the curve.

The solution of the impulsive load problem will consist of partitioning of the $(x,t)$-quadrant into sectors of constant strain and velocity. These sectors are separated by either shocks, phase boundaries or contact discontinuities.

We now inquire as to what is the relation between this simple wave solution and the fields which would be produced if the bar were gradually loaded to the level $\sigma_\infty$? As before, the simple wave solution will yield the limiting order of waves in the bar as $t$ tends to infinity. It will also produce the correct final number and ordering of shocks and phase boundaries. The contact discontinuities, however, are not associated with true curves of discontinuity in the field variables. Instead they indicate nondispersive wave packets across which the dynamical fields gradually change.

The approach to this simple wave can be examined by exploiting the linearity of the material in the different strain intervals. The difficulty lies in determining the regions in the $(x,t)$-plane in which the strain takes values in the individual intervals. The boundaries of these curves must be either curves of constant strain or curves across which the strain jumps between values from different intervals. In either case these conditions lead to functional equations when the strains are expressed in terms of D'Alembert's solution to the linear wave equation.

In order to illustrate this, we consider the material of Fig. 5. The particular form of this stress-strain curve is motivated by its similarity to the material previously introduced in Fig. 1. We shall suppose that $\lim_{t\to\infty} \sigma_0(t) = \sigma_\infty$ where $\sigma_\infty$ is as indicated in Fig. 5. The corresponding impulsive load solution is also given in Fig. 5. This solution consists of a phase boundary and a contact discontinuity separating regions in which the dynamical fields are constant. Suppose further that

$$\sigma_0(0) = 0 , \quad \sigma_0'(t) > 0 .$$

We shall let $x = s(t)$ denote the phase boundary, while $x = q(t)$ shall denote the curve upon which the strain has value $\mu$, thus

$$\varepsilon(q(t),t) = \mu . \tag{4.1}$$

Let $A_0$, $A_1$, $A_2$ denote the regions in the $(x,t)$-plane in which the strain lies in the respective intervals $(0,\alpha)$, $(\beta,\mu)$, $(\mu,\infty)$. In each region $A_i$ the strain obeys the equation $\dfrac{\partial^2\varepsilon}{\partial t^2} = c_i^2 \dfrac{\partial^2\varepsilon}{\partial x^2}$ which implies that

$$\varepsilon(x,t) = f_i(x + c_i t) + g_i(-x + c_i t) \tag{4.2}$$

for as yet undetermined functions $f$ and $g$. The velocity in $A_i$ must then

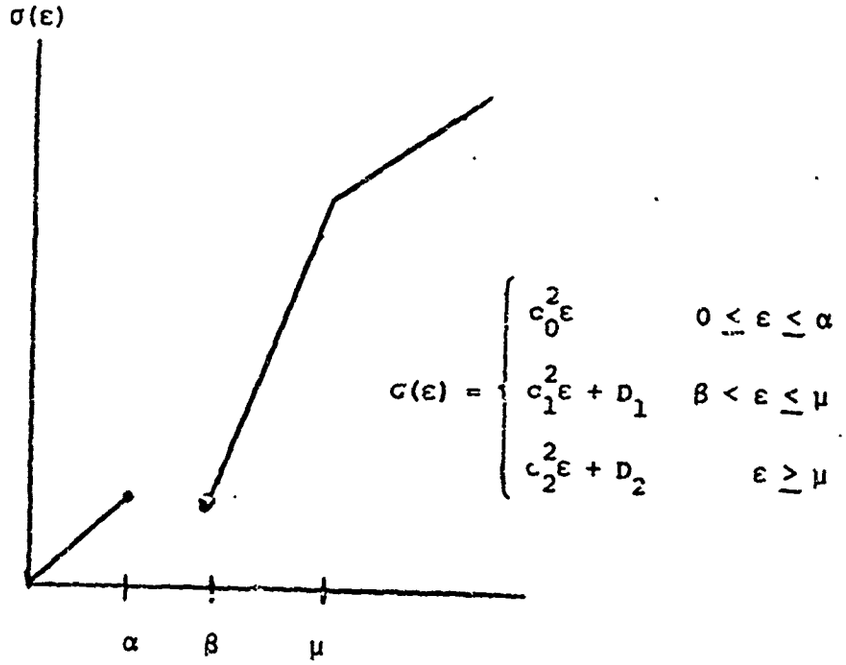



Fig. 4. Model material with piecewise linear stress response

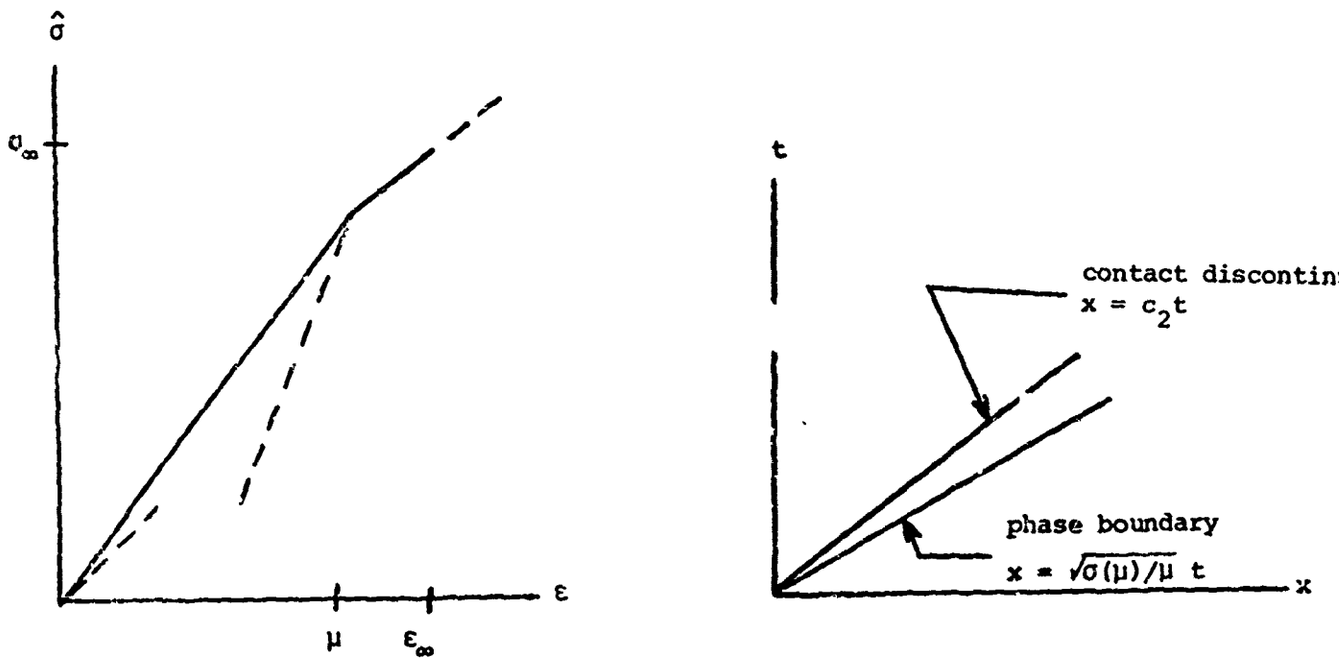



Fig. 5. Upper convex envelope $\hat{\sigma}(\varepsilon)$ and corresponding simple wave for the material in Fig. 4 when $\varepsilon_\infty > \mu$

be $v(x,t) = c_i f_i(x + c_i t) - c_i g_i(-x + c_i t)$. The $f_i$ and $g_i$, along with $s(t)$ and $q(t)$, are eight unknown functions, each of a single argument. They must satisfy eight conditions given by: the two initial conditions (2.4), the boundary condition (2.3), the two shock conditions (2.7), two conditions stemming from (4.1), and a condition that the velocity is continuous across $x = q(t)$. The last three conditions in the above list give rise to functional equations since $q(t)$ appears as an argument of $f_1$, $f_2$, $g_1$, $g_2$. Moreover, since $\dot{s}(t)$ appears in (2.7), the condition on $x = s(t)$ furnishes a pair of functional differential equations. In the expressions which follow, we shall employ parenthesis solely to indicate the argument of a function.

From the initial conditions (2.4) it can be shown that the Riemann invariant associated with the characteristics $\frac{dx}{dt} = -c_0$ is identically zero on $A_0$. Thus $f_0(z) \equiv 0$ which, by virtue of (4.2), (2.5), yields

$$g_0(z) = \begin{cases} \varepsilon_0(z/c_0) & z > 0 , \\ 0 & z < 0 . \end{cases}$$

The other six unknown functions do not admit simple solution representations. In what follows we shall restrict attention to large times. The analysis will be shortened considerably by taking advantage of the simple wave solution depicted in Fig. 5. This is not necessary; the asymptotic simple wave could be deduced from the analysis. Such a program, however, requires the consideration - too lengthy to be included here - of numerous possible cases.

V. LARGE-TIME DYNAMICAL FIELDS The boundary condition (2.5) can be incorporated directly into the functions $f_2$, $g_2$. In place of the functions $f_1$, $g_1$, $f_2$, $g_2$ we shall instead employ $f$, $g$, $h$ through the expressions

$$\varepsilon(x,t) = \begin{cases} f(x + c_1 t) + g(-x + c_1 t) & \text{in } A_1 \\ h(x + c_2 t) - h(-x + c_2 t) + \varepsilon_0(t - x/c_2) & \text{in } A_2 \end{cases} \tag{5.1}$$

$$v(x,t) = \begin{cases} c_1 f(x + c_1 t) - c_1 g(-x + c_1 t) & \text{in } A_1 \\ c_2 h(x + c_2 t) + c_2 h(-x + c_2 t) - c_2 \varepsilon_0(t - x/c_2) & \text{in } A_2 \end{cases} \tag{5.2}$$

The five unknowns $f$, $g$, $h$, $s$, $q$ are to be determined from: two conditions holding on $x = q(t)$ which stem from (4.1), a condition expressing the continuity of velocity on $x = q(t)$, and the two shock conditions (2.7) which hold on $x = s(t)$. The first two of the conditions holding on $x = q(t)$ become

$$f(q(t) + c_1 t) + g(-q(t) + c_1 t) = \mu , \tag{5.3}$$

$$h(q(t) + c_2 t) - h(-q(t) + c_2 t) + \varepsilon_0\left(t - \frac{1}{c_2} q(t)\right) = \mu . \tag{5.4}$$

With the aid of these two equations, the continuity of velocity condition may be written as

$$(5.5) \qquad f(q(t) + c_1 t) = \frac{c_2}{c_1} h(q(t) + c_2 t) + \frac{1}{2}\left(1 - \frac{c_2}{c_1}\right)\mu .$$

The two shock conditions can be manipulated into the form

$$(5.6) \quad [c_1 + \dot{s}(t)] f(s(t) + c_1 t) + \frac{D_1}{2c_1} = 0 , \qquad t \text{ sufficiently large,}$$

$$(5.7) \quad [c_1 - \dot{s}(t)] g(-s(\ ) + c_1 t) + \frac{D_1}{2c_1} = 0 , \qquad t \text{ sufficiently large .}$$

The phrase "t sufficiently large" indicates that these are the equations which hold once the phase boundary $x = s(t)$ has become the leading disturbance. This eventuality follows from the simple wave solution, moreover the simple wave solution also indicates that $\dot{s}(t) \to \sqrt{\sigma(\mu)/\mu}$ and $\dot{q}(t) \to c_2$.

Assume that the wave and front speeds obey the ordering

$$(5.8) \qquad c_1 > \dot{s}(t) > c_2 > \dot{q}(t) > 0$$

for $t$ sufficiently large. One may conclude directly from (5.3) - (5.7) that

$$f(z) \to f_\infty , \quad g(z) \to g_\infty , \quad h(z) \to h_\infty \qquad \text{as} \quad z \to \infty ,$$

$$\dot{s}(t) \to \alpha , \quad \dot{q}(t) \to \beta \qquad \text{as} \quad t \to \infty .$$

The values for $f_\infty$, $g_\infty$, and $\alpha$ are obtained from a consideration of the equations (5.3), (5.6), (5.7) for large times. As $t \to \infty$ these equations become respectively

$$f_\infty + g_\infty = \mu , \quad [c_1 + \alpha] f_\infty + \frac{D_1}{2c_1} = 0 , \quad [c_1 - \alpha] g_\infty + \frac{D_1}{2c_1} = 0 ,$$

which give

$$(5.9) \quad \alpha = \sqrt{c_1^2 + \frac{D_1}{\mu}} = \sqrt{\frac{\sigma(\mu)}{\mu}} , \quad f_\infty = \frac{-D_1}{2c_1[c_1 + \alpha]} , \quad g_\infty = \frac{-D_1}{2c_1[c_1 - \alpha]} .$$

The values

$$(5.10) \qquad h_\infty = \frac{c_1}{c_2} f_\infty - \frac{1}{2}\left[\frac{c_1}{c_2} - 1\right]\mu = \frac{1}{2}\mu\left[1 - \frac{\alpha}{c_2}\right] , \quad \beta = c_2 ,$$

follow from (5.4), (5.5). Since $q(t) \to c_2 t$ and $s(t) \to \sqrt{\frac{\sigma(\mu)}{\mu}}\, t$ as $t \to \infty$, it is immediate that $A_0$, $A_1$, $A_2$ tend toward the sectors of the simple wave solution displayed in Fig. 5. Upon entering the values of $f_\infty$, $g_\infty$, $h_\infty$ into (5.1), (5.2), the corresponding sector values for strain and velocity are recovered.

In order to study the approach to these values let

$$
\begin{aligned}
s(t) &= \sqrt{\frac{\sigma(\mu)}{\mu}}\, t + s_1(t) \\
q(t) &= c_2 t + q_1(t) \\
f(z) &= f_\infty + f_1(z) \\
g(z) &= g_\infty + g_1(z) \\
h(z) &= h_\infty + h_1(z)
\end{aligned}
\tag{5.11}
$$

for new functions $s_1$, $q_1$, $f_1$, $g_1$, $h_1$. Upon substituting from these expressions into (5.3) - (5.7) and considering a balance of the second order terms, one obtains asymptotic expressions for these functions. This procedure applied to (5.6) yields

$$
\begin{aligned}
0 &= \left[c_1 + \alpha + \dot{s}_1(t)\right]\left\{f_\infty + f_1\left([c_1 + \alpha]t + s_1(t)\right)\right\} + \frac{D_1}{2c_1} \\
&= f_\infty \dot{s}_1(t) + \left[c_1 + \alpha + \dot{s}_1(t)\right] f_1\left([c_1 + \alpha]t + s_1(t)\right) \\
&\sim \dot{s}_1(t) f_\infty + [c_1 + \alpha] f_1\left([c_1 + \alpha]t\right) ,
\end{aligned}
$$

which in turn implies

$$
\dot{s}_1(t) \sim -\frac{[c_1 + \alpha]}{f_\infty} f_1\left([c_1 + \alpha]t\right) . \tag{5.12}
$$

Similarly (5.7) leads to

$$
\dot{s}_1(t) \sim -\frac{[c_1 - \alpha]}{g_\infty} g_1\left([c_1 - \alpha]t\right) , \tag{5.13}
$$

while (5.3) leads to

$$
f_1\left([c_1 + c_2]t\right) \sim - g_1\left([c_1 - c_2]t\right) . \tag{5.14}
$$

Eliminating $g_1$ and $s_1$ between (5.12) (5.13) (5.14) reveals that $f_1$ obeys

$$f_1\left(\left[\frac{c_1 + c_2}{c_1 - c_2}\right] z\right) \sim \left[\frac{c_1 + \alpha}{c_1 - \alpha}\right]^2 f_1\left(\left[\frac{c_1 + \alpha}{c_1 - \alpha}\right] z\right), \quad \text{as} \quad z \to \infty .$$

This condition will be written

$$f_1(k_2 z) \sim k_1 f_1(z) \quad \text{as} \quad z \to \infty \tag{5.15}$$

where

$$k_1 = \left[\frac{c_1 - \alpha}{c_1 + \alpha}\right]^2 , \quad k_2 = \frac{[c_1 + \alpha][c_1 - c_2]}{[c_1 - \alpha][c_1 + c_2]} . \tag{5.16}$$

It follows from (5.8) that

$$k_2 > 1 > k_1 > 0 . \tag{5.17}$$

The asymptotic expression (5.15) indicates that

$$f_1(z) \sim Az^n , \quad n = \ln k_1 / \ln k_2 . \tag{5.18}$$

where A is undetermined. In light of (5.17), it follows that $n < 0$ so that $f_1(z)$ is indeed dominated by $f_\infty$ whenever $z \to \infty$, as was assumed in the development leading to (5.15). Moreover the following argument demonstrates that

$$n < -2 . \tag{5.19}$$

Proof: (5.18) implies $(k_1^{1/2} k_2)^n = k_1^{1 + \frac{1}{2} n}$, which, in conjunction with $0 < k_1^{1/2} k_2 = \frac{c_1 - c_2}{c_1 + c_2} < 1$ and $n < 0$ yields $k_1^{1 + \frac{1}{2} n} > 1$. This last result, along with (5.17), yields $1 + \frac{1}{2} n < 0$, which is (5.19). /

One finds from (5.18), (5.12), (5.13) that

$$g_1(z) \sim k_1^{-\left(1 + \frac{n}{2}\right)} Az^n , \quad \dot{s}_1(t) \sim - \frac{[c_1 + \alpha]^{n+1}}{f_\infty} At^n . \tag{5.20}$$

On account of (5.20) and (5.19) one has

$$s_1(t) \sim s_0 - \frac{[c_1 + \alpha]^{n+1}}{(n+1) f_\infty} At^{n+1} , \tag{5.21}$$

where $s_0$ is undetermined.

The asymptotic behavior of the corrections $h_1(z)$ and $q_1(t)$ are found from the remaining equations (5.4), (5.5). Entering (5.5) with (5.11), (5.10) one obtains

$$\frac{c_2}{c_1} h_1\big(2c_2t + q_1(t)\big) - f_1\big([c_1 + c_2]t + q_1(t)\big) = 0$$

from which it follows that

$$h_1(z) \sim \frac{c_1}{c_2} \Big[\frac{c_1 + c_2}{2c_2}\Big]^n Az^n \ . \tag{5.22}$$

Equation (5.4) is the most delicate in the analysis. The exact equation yields

$$\mu = h_\infty + h_1\big(2c_2t + q_1(t)\big) - h\big(-q_1(t)\big) + \varepsilon_0\big(-q_1(t)/c_2\big) \ . \tag{5.23}$$

In this equation $h_1\big(2c_2t + q_1(t)\big) = o(1)$ as $t \to \infty$. It is, however, not appropriate to expand $h\big(-q_1(t)\big)$ by means of (5.11) unless $q_1(t) \to -\infty$. Assume for the moment that $q_1(t) \to -\infty$; then $\varepsilon_0\big(-q_1(t)/c_2\big) \to \varepsilon_\infty$, $h\big(-q_1(t)\big) \to h_\infty$ so that (5.23) gives up $\varepsilon_\infty = \mu$. Since $\varepsilon_\infty > \mu$ the assumption is false. Instead (5.23) demands that $q_1(t) \to q_0 > -\infty$ with $\mu = h_\infty - h(-q_0) + \varepsilon_0(-q_0/c_2)$. In the event that $h(z)$ admits a Taylor expansion at $z = -q_0$, then (5.22), (5.23) yields

$$\Big\{\dot h(-q_0) + \frac{1}{c_2}\dot\varepsilon_0(-q_0/c_2)\Big\}\big[-q(t) + q_0\big] \sim h_1(2c_2t) \sim [c_1/c_2][c_1 + c_2]^n At^n \ .$$

This in turn implies that $q(t) \sim q_0 + O(t^n)$ whenever the coefficient in parenthesis is not zero. Collecting the results from this section we have

$$f(z) = f_\infty + Az^n + o(z^n)$$

$$g(z) = g_\infty + Ak_1^{-(1 + \frac{n}{2})} z^n + o(z^n)$$

$$s(t) = \alpha t + s_0 - \frac{[c_1 + \alpha]^{n+1}}{(n+1)f_\infty} At^{n+1} + o(t^{n+1}) \tag{5.24}$$

$$h(z) = h_\infty + A\frac{c_1[c_1 + c_2]^n}{2^n c_2^{n+1}} z^n + o(z^n)$$

$$q(t) = c_2t + q_0 + O(t^n) \ .$$

The constants $f_\infty$, $h_\infty$, $g_\infty$, $\alpha$, $k_1$, $n$ are given in (5.9), (5.10), (5.16), (5.18). The constants $A$, $q_0$, $s_0$ are as yet unspecified. Indeed, in order to assign values to $A$, $q_0$, $s_0$ one may surmise that it is necessary to take account of the complete loading history $\sigma_0(t)$. Consider, for example, the constant $s_0$. The phase boundary $x = s(t)$ asymptotically approaches the line $x = \alpha t + s_0$. This line issues from the $t$ axis at time $t = -s_0/\alpha$, so that $s_0$ defines a time scale for the problem. The only source of such a time scale lies in the applied load $\sigma_0(t)$.

The dynamical fields are obtained by substituting from (5.24) into (5.1), (5.2). This yields

$$\varepsilon(x,t) \sim \varepsilon_\infty + \frac{Ac_1[c_1 + c_2]^n}{2^n c_2^{n+1}} \left[ \left[c_2 + \frac{x}{t}\right]^n - \left[c_2 - \frac{x}{t}\right]^n \right] t^n ,$$

$$v(x,t) \sim -\alpha\mu - c_2[\varepsilon_\infty - \mu] + \frac{Ac_1[c_1 + c_2]}{2^n c_2^{n+1}} \left[ \left[c_2 + \frac{x}{t}\right]^n - \left[c_2 - \frac{x}{t}\right]^n \right] t^n$$

in region $A_2$. In region $A_1$ one obtains

$$\varepsilon(x,t) \sim \mu + A \left[ \left[c_1 + \frac{x}{t}\right]^n + \left[\frac{c_1 + \alpha}{c_1 - \alpha}\right]^{n+2} \left[c_1 - \frac{x}{t}\right]^n \right] t^n ,$$

$$v(x,t) \sim -\alpha\mu + Ac_1 \left[ \left[c_1 + \frac{x}{t}\right]^n - \left[\frac{c_1 + \alpha}{c_1 - \alpha}\right]^{n+2} \left[c_1 - \frac{x}{t}\right]^n \right] t^n .$$

Although the value of $A$ is not found from this analysis, the exponent $n$ governing the rate of approach to the simple wave is determined from both the stress-strain behavior of the material and the ultimate level of the applied load by means of (5.18).

## REFERENCES


1. J. L. Ericksen, Equilibrium of Bars, J. Elasticity, 5 (1975) 3-4, p. 191.

2. R. D. James, The Propagation of Phase Boundaries in Elastic Bars, Arch. Rational Mech. Anal., 73 (1980), p. 125.

3. T. J. Pence, On the Emergence and Propagation of a Phase Boundary in an Elastic Bar with a Suddenly Applied End Load, to appear J. Elasticity.

4. D. E. Grady, R. E. Hollenbach and K. W. Schuler, Compression Wave Studies in Calcite Rock, J. Grophys. Res., 83 (1978), p. 2839.

5. R. Courant and K. O. Friedrichs, Supersonic Flow and Shock Waves, Interscience, N.Y. 1956.

6. J. Serrin, Phase Transitions and Interfacial Layers for van der Waals Fluids, in "Proceedings of SAFA IV Conference, Recent Methods in Nonlinear Analysis and Applications, Naples, March 21-28, 1980" (A. Canfora, S. Rionero, C. Sbordone, C. Trombetti, Eds.)

7. M. Slemrod, Admissibility Criteria for Propagating Phase Boundaries in a van der Waals Fluid, Arch. Rational Mech. Anal, 81 (1983), p. 301.

8. R. Hagan and M. Slemrod, The Viscosity - Capillarity Criterion for Shocks and Phase Transitions, Arch. Rational Mech. Anal, 83 (1983) p. 333.

# EIGENFUNCTIONS AT A SINGULAR POINT IN TRANSVERSELY ISOTROPIC MATERIALS UNDER AXISYMMETRIC DEFORMATIONS

T. C. T. Ting and Yijing Jin
Department of Civil Engineering, Mechanics and Metallurgy
University of Illinois at Chicago
P.O. Box 4348, Chicago, IL 60680

S. C. Chou
U.S. Army Materials and Mechanics Research Center
Watertown, MA 02172

ABSTRACT. When a two-dimensional elastic body which contains a notch or a crack is under a plane stress or plane strain deformation, the asymptotic solution of the stress near the apex of the notch or crack is simply a series of eigenfunctions of the form $\rho^{\delta}f(\psi,\delta)$ in which $(\rho,\psi)$ is the polar coordinate with origin at the apex and $\delta$ is the eigenvalue. If the body is a three-dimensional elastic solid which contains axisymmetric notches or cracks and subjected to an axisymmetric deformation, the eigenfunction associated with an eigenvalue contains not only the $\rho^{\delta}$ term, but also the $\rho^{\delta+1}$, $\rho^{\delta+2}$... terms. Therefore, the second and higher order terms of the asymptotic solution are not simply the second and subsequent eigenfunctions. We present the eigenfunctions for transversely isotropic materials under an axisymmetric deformation. The degenerate case in which the eigenvalues $p_1$ and $p_2$ of the elasticity constants are identical is also considered. The latter includes the isotropic materials under axisymmetric deformations.

1 INTRODUCTION. It is well-known that the stress distribution near the apex of an isotropic elastic wedge or a notch under a plane-stress or plane-strain deformation can be expressed in terms of a series of eigenfunctions of the form $\rho^{\delta}f(\psi,\delta)$ where $\rho$ is the radial distance from the apex and $f$ is a function of the polar angle $\psi$ and the eigenvalue $\delta$, [1,2]. For given wedge angle and homogeneous boundary conditions at the sides of the wedge, there are in general infinitely many eigenvalues $\delta$ and the associated eigenfunctions $\rho^{\delta}f(\psi,\delta)$. Particularly important in applications is when one or more of the $\delta$'s is negative and the stress is singular at the apex. For instance, if the specimen shown in Fig. 1 represents a two-dimensional body under an external loading, $\delta = -1/2$ at the crack tip Q. At points R, N and M, it can be shown [3] that there are two negative $\delta$'s. Hence the stress is singular at points Q, R, N and M. In solving stress distribution in the entire specimen numerically by a finite element scheme, one may use regular finite elements everywhere except at the singular points Q, R, N and M. At these singular points, a special element is used in which the singular nature of the stress is given by the analytical expression $\rho^{\delta}f(\psi,\delta)$. It may be sufficient to consider only the first term (or terms) for which $\delta$ is negative in the special element. In many cases, however, more terms including those associated with positive $\delta$ are required [4,5].

In this paper, we consider the case in which the specimen shown in Fig. 1 is a cross section of an axisymmetic body under an axisymmetric deformation. The material is assumed to be transversely isotropic with the z-axis being the axis of symmetry. The associated problem for isotropic materials was investigated by Delale and Erdogan [6]. However, their objectives are different from ours and hence their series solution is different from the one presented here. It is seen that the eigenfunction associated with an eigenvalue $\delta$ no longer contains a single term $\rho^{\delta} f(\psi,\delta)$. It also has the terms $\rho^{\delta+1} f_1(\psi,\delta)$, $\rho^{\delta+2} f_2(\psi,\delta)$... Therefore, the inclusion of the second and higher order terms in the special element is not simply the inclusion of the eigenfunctions associated with the subsequent smallest eigenvalues $\delta$. A similar situation occurs in wedges with curved sides under a two-dimensional deformation [7].

2. MATHEMATICAL FORMULATION. We choose a cylindrical coordinate system $(r,\theta,z)$ in the transversely isotropic medium such that the z-axis is the axis of material symmetry. Let $(u_r,u_\theta,u_z)$ be the displacement components. We assume that the deformation is axisymmetric in which $u_\theta = 0$ while $u_r$ and $u_z$ are functions of $r$ and $z$ only. Introducing the displacement potential $\phi(r,z)$ from which $u_r$ and $u_z$ are given by [8-10]

$$u_r = \frac{\partial \phi}{\partial r}, \qquad u_z = m \frac{\partial \phi}{\partial z}, \tag{1}$$

where $m$ is a constant to be determined, the stresses have the expressions

$$\sigma_r = c_{11} \frac{\partial^2 \phi}{\partial r^2} + c_{12} \frac{\partial \phi}{r \partial r} + c_{13} m \frac{\partial^2 \phi}{\partial z^2} \tag{2}$$

$$\sigma_\theta = c_{12} \frac{\partial^2 \phi}{\partial r^2} + c_{11} \frac{\partial \phi}{r \partial r} + c_{13} m \frac{\partial^2 \phi}{\partial z^2}$$

$$\sigma_z = c_{13} \frac{\partial^2 \phi}{\partial r^2} + c_{13} \frac{\partial \phi}{r \partial r} + c_{33} m \frac{\partial^2 \phi}{\partial z^2}$$

$$\sigma_{rz} = c_{44}(1+m) \frac{\partial^2 \phi}{\partial r \partial z}$$

in which $c_{ij}$ are the elasticity constants. The equations of equilibrium are satisfied if

$$\frac{\partial^2 \phi}{\partial r^2} + \frac{\partial \phi}{r \partial r} - \frac{1}{p^2} \frac{\partial^2 \phi}{\partial z^2} = 0 \tag{3}$$

where

$$p^2 = \frac{-c_{11}}{c_{13}m + (1+m)c_{44}} = \frac{c_{13} + (1+m)c_{44}}{-c_{33}m} \tag{4a}$$

or

$$-m = \frac{c_{11} + c_{44}p^2}{(c_{13} + c_{44})p^2} = \frac{c_{13} + c_{44}}{c_{44} + c_{33}p^2} \tag{4b}$$

The second equality of Eqs. (4a) and (4b) yield, respectively,

$$m^2 - 2\left[\frac{c_{11}c_{33} - c_{13}^2}{2c_{44}(c_{13} + c_{44})} - 1\right]m + 1 = 0 \tag{5a}$$

$$p^4 + 2\left[\frac{c_{11}c_{33} - c_{13}^2 - 2c_{13}c_{44}}{2c_{33}c_{44}}\right]p^2 + \frac{c_{11}}{c_{33}} = 0 \tag{5b}$$

It can be shown [11] that $p$ cannot be real if the strain energy is positive definite. We therefore have two pairs of complex conjugates for $p$ denoted by $p_1$, $p_2$, $\bar{p}_1$ and $\bar{p}_2$ where an overbar denotes the complex conjugate. The associated $m$ are denoted by $m_1$, $m_2$, $\bar{m}_1$ and $\bar{m}_2$. We see from Eq. (5a) that

$$m_1 m_2 = 1 \tag{5c}$$

Since Eq. (5b) is a quadric in $p^2$ with real coefficients, if $p_1$ is purely imaginary so is $p_2$. In this case $m_1$, $m_2$ are real and $m_1 = \bar{m}_1$, $m_2 = \bar{m}_2$. If $p_1$ and $p_2$ are not purely imaginary, we may choose

$$p_1 = u + iv\,, \qquad p_2 = -u + iv = -\bar{p}_1 \tag{6}$$

where $u$, $v$ are real. In this case $m_1$, $m_2$ are complex and $m_1 = \bar{m}_2$. In view of the fact that Eq. (3) is linear in $\phi$, the general solution for $\phi$ is obtained by superimposing $\phi$'s associated with $p = p_1$, $p_2$, $\bar{p}_1$ and $\bar{p}_2$. We will assume that $p_1 \neq p_2$. The degenerate case $p_1 = p_2$ will be discussed in Section 4.

3. EIGENFUNCTIONS FOR SMALL $\rho$. Let $(r,z) = (a,0)$ be a singular point which may be the apex of a wedge, a notch or a crack. In this paper, we consider the cases in which $a \neq 0$. Using the singular point as the origin, we define (Fig. 1)

$$x = r - a = \rho\cos\psi, \qquad z = \rho\sin\psi\,. \tag{7}$$

We assume that $\psi = \alpha$ and $\psi = \alpha'$ are the free surfaces. To find the eigenfunction for $\phi$ which is valid for small $\rho$, we rewrite Eq. (3) as

$$\frac{\partial^2\phi}{\partial x^2} - \frac{1}{p^2}\frac{\partial^2\phi}{\partial z^2} = -\frac{1}{a+x}\frac{\partial\phi}{\partial x} = -\frac{1}{a}\frac{\partial\phi}{\partial x}\sum_{s=0}^{\infty}\left(\frac{-x}{a}\right)^s \tag{8}$$

Let

$$\phi = \phi^{(0)} - \frac{1}{a}\phi^{(1)} + \frac{1}{a^2}\phi^{(2)} - \ldots = \sum_{k=0}^{\infty}\left(\frac{-1}{a}\right)^k \phi^{(k)} , \tag{9a}$$

$$\phi^{(k)} = \sum_{t=0}^{k} A_t^{(k)} x^t Z^{\delta+k-t+2}/(\delta+k-t+2)(\delta+k-t+1) , \tag{9b}$$

$$Z = x + pz \tag{10}$$

where $\delta$ is the eigenvalue and $A_t^{(k)}$ are constants to be determined. Using Eq. (7), we have

$$x^t Z^{\delta+k-t+2} = \rho^{\delta+k+2}(\cos\psi)^t \zeta^{\delta+k-t+2} \tag{11}$$

$$\zeta = \cos\psi + p\sin\psi \tag{12}$$

Hence $\phi^{(k)}$ is of order $\rho^{\delta+k+2}$. By substituting Eqs. (9) into (8) and equating the coefficients of $x^t Z^{\delta+k-t+2}$, it can be shown that

$$A_k^{(k)} = \frac{2k-1}{2k} A_{k-1}^{(k-1)} , \qquad (k \geq 1) \tag{13a}$$

$$A_t^{(k)} = \frac{2t-1}{2t} A_{t-1}^{(k-1)} + \frac{1}{2(\delta+k-t)}\left[tA_t^{(k-1)} - (t+1)A_{t+1}^{(k)}\right] , \tag{13b}$$

$$(t = k-1, k-2 \ldots 1 \; ; \; k \geq 2)$$

Hence the only unknowns are $A_o^{(k)}$, $(k = 0,1,2\ldots)$ which are determined from the boundary conditions at $\psi = a$ and $a'$.

We will let the solutions given by Eqs. (9-13) apply to $p = p_1$. For $p = p_2$, $\bar{p}_1$ and $\bar{p}_2$, we will use the same expressions except that $A_t^{(k)}$ is replaced by $B_t^{(k)}$, $C_t^{(k)}$ and $D_t^{(k)}$, respectively. Thus the general solution for $\phi^{(k)}$ is

$$\phi^{(k)} = \sum_{t=0}^{k}\left\{A_t^{(k)} x^t Z_1^{\delta+k-t+2} + B_t^{(k)} x^t Z_2^{\delta+k-t+2} + C_t^{(k)} x^t \bar{Z}_1^{\delta+k-t+2} + D_t^{(k)} x^t \bar{Z}_2^{\delta+k-t+2}\right\}\Big/(\delta+k-t+2)(\delta+k-t+1) \tag{14}$$

$$Z_s = x + p_s z \ , \qquad (s = 1,2) \tag{15}$$

4. DEGENERATE CASE: $m_1 = m_2 = 1$. When $p_1 = p_2$, $p$ must be purely imaginary. This follows from Eq. (5b) and the fact that $p$ cannot be real. By Eqs. (4b) and (5c), we have $m_1 = m_2 = 1$. We cannot have $m_1 = m_2 = -1$ because this would make $p$ real.

In a degenerate case $p_1 = p_2$, the terms associated with $B_t^{(k)}$ and $D_t^{(k)}$ are identical, respectively, to the terms associated with $A_t^{(k)}$ and $C_t^{(k)}$. We therefore need a new solution for $B_t^{(k)}$ and $D_t^{(k)}$. This can be accomplished by replacing the coefficients of $B_t^{(k)}$ and $D_t^{(k)}$ by their derivatives with respect to $p_2$ and $\bar{p}_2$ [12,13]. Thus, for instance, Eq. (14) becomes

$$\phi^{(k)} = \sum_{t=0}^{k} \left\{ A_t^{(k)} x t Z^{\delta+k-t+2} + C_t^{(k)} x t \bar{Z}^{\delta+k-t+2} \right\} \Big/ (\delta+k-t+2)(\delta+k-t+1) \tag{16}$$

$$+ \sum_{t=0}^{k} \left\{ B_t^{(k)} z x t Z^{\delta+k-t+1} + D_t^{(k)} z x t \bar{Z}^{\delta+k-t+1} \right\} \Big/ (\delta+k-t+1)$$

where we have omitted the subscripts 1 and 2 for $Z$ and $\bar{Z}$.

5. ASYMPTOTIC SOLUTION FOR SMALL $\rho$. When boundary conditions at $\psi = \alpha$ and $\alpha'$ are imposed, one obtains a system of homogeneous equations for the eigenvalue $\delta$ and the associated eigenvectors $q_o^{(0)}$ whose elements are $A_o^{(0)}$, $B_o^{(0)}$, $C_o^{(0)}$ and $D_o^{(0)}$. Thus $q_o^{(0)}$ is determined uniquely within an arbitrary constant $\xi$. One also obtains systems of linear but non-homogeneous equations for $A_t^{(k)}$, $B_t^{(k)}$, $C_t^{(k)}$ and $D_t^{(k)}$ in terms of $q_o^{(0)}$. Thus the eigenfunction associated with an eigenvalue $\delta$ contains only one arbitrary constant $\xi$. Let $\delta = \delta_1, \delta_2, \ldots$ be the eigenvalues arranged in the order of increasing magnitude and $\xi = \xi_1, \xi_2, \ldots$ be the associated arbitrary constants. Using Eqs. (9a), (11) and (14), the asymptotic solution for $\phi$ for small $\rho$ can be written as

$$\phi = \xi_1 \left\{ \rho^{\delta_1} f_{1,0}(\psi,\delta_1) + \rho^{\delta_1+1} f_{1,1}(\psi,\delta_1) + \rho^{\delta_1+2} f_{1,2}(\psi,\delta_1) + \ldots \right\} \tag{17}$$

$$+ \xi_2 \left\{ \rho^{\delta_2} f_{2,0}(\psi,\delta_2) + \rho^{\delta_2+1} f_{2,1}(\psi,\delta_2) + \rho^{\delta_2+2} f_{2,2}(\psi,\delta_2) + \ldots \right\}$$

$$+ \ldots$$

in which $f_{i,j}$ are known functions of $\psi$ and $\delta_i$. The first term asymptotic solution is given by the term $\rho^{\delta_1} f_{1,0}(\psi,\delta_1)$. The second term is $\rho^{\delta_1+1} f_{1,1}(\psi,\delta_1)$ if $\delta_1+1 < \delta_2$ and $\rho^{\delta_2} f_{2,0}(\psi,\delta_2)$ if $\delta_1+1 > \delta_2$. Notice that the eigenfunction associated with an eigenvalue $\delta$ has infinitely many terms.

Thus the second and higher order terms of asymptotic solution are not simply the second and subsequent eigenfunctions.

## REFERENCES


[1] Knein, M., 'Zur Theorie des Druckversuchs,' Zeit. Ang. Math. Mech., Vol. 6, 1926, 414-416.

[2] Williams, M. L., 'Stress Singularities Resulting from Various Boundary Conditions in Angular Corners of Plates in Extension,' J. Appl. Mech., Vol. 19, 1952, 526-528.

[3] Ting, T. C. T., 'The Wedge Subjected to Tractions: A Paradox Remined,' J. Elasticity. To Appear.

[4] Tong, P., Pian, T. H. H. and Lasry, S. J., 'A Hybrid-Element Approach to Crack Problems in Plane Elasticity,' Int. J. Numerical Meth. in Eng., Vol. 7, 1973, 297-308.

[5] Lin, K. Y. and Mar, J. W., 'Finite Element Analysis of Stress Intensity Factors for Cracks at a Bi-Material Interface,' Int. J. Fracture, Vol. 12, 1976, 521-531.

[6] Delale, F. and Erdogan, F., 'The Axisymmetric Elasticity Problem for a Laminated Plate Containing a Circular Hole,' Lehigh University Report, July 1981.

[7] Ting, T. C. T. ,'Asymptotic Solution Near the Apex of an Elastic Wedge with Curved Boundaries,' Q. Appl. Math. To appear.

[8] Elliott, H. A., 'Three-Dimensional Stress Distributions in Hexagonal Aeolotropic Crystals,' Proc. Cambridge Phil. Soc., Vol. 44, 1948, 522-533.

[9] Green, A. E. and Zerna, W., Theoretical Elasticity, Oxford University Press, Oxford, 1954.

[10] Kassir, M. K. and Sih, G. C., 'Three-Dimensional Crack Problems,' Mechanics of Fracture, Vol. 2, Noordhoff Int. Pub., 1975, 336-342.

[11] Eshelby, J. D., Read, W. T. and Shockley, W., 'Anisotropic Elasticity with Applications to Dislocation Theory,' Act. Met., Vol. 1, 1953, 251-259.

[12] Ting, T. C. T. and Chou, S. C., 'Edge Singularities in Anisotropic Composites,' Int. J. Solids Structures, Vol. 17, 1981, 1057-1068.

[13] Ting, T. C. T., 'Effects of Change of Reference Coordinates on the Stress Analyses of Anisotropic Elastic Materials,' Int. J. Solids Structures, Vol. 18, 1982, 139-152.

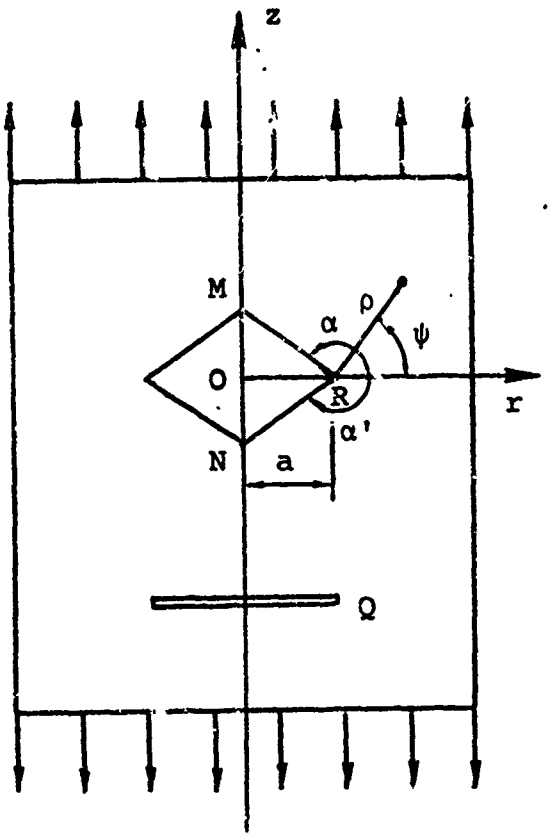



Fig. 1 Cross section of an axi-symmetric body which contains a crack and notches

# HIGHLY VISCOUS FLUID FLOW IN A SPINNING AND NUTATING CYLINDER

Thorwald Herbert
Department of Engineering Science and Mechanics
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT. Spin-stabilized projectiles with liquid payloads can experience a severe flight instability characterized by a rapid yaw angle growth and a simultaneous loss in spin rate. Laboratory experiments and field tests have shown that this instability originates from the internal fluid motion in the range of high viscosity. Evaluation of the experimental data and analysis of the equations for the fluid motion in a spinning and nutating cylinder suggest a theoretical approach in three major steps: (1) analysis of the steady viscous flow in an infinitely long cylinder, (2) hydrodynamic stability analysis of this basic flow, and (3) analysis of the end effects. The basic flow has been found in analytical form. At low Reynolds number, this flow agrees well with computational results for the center section of a cylinder of aspect ratio 4.3. The despin moment caused by this flow largely agrees with experimental data for a wide range of Reynolds numbers. Current work aims at the stability of this flow.

1. INTRODUCTION. It is well-known that spin-stabilized shells carrying liquid payloads can suffer dynamical instability. For cylindrical cavities and low viscosity of the liquid, the instability due to basically inviscid inertial waves can be predicted by the Stewartson-Wedemeyer theory [1,2]. This theory rests on the boundary-layer approach and is, therefore, restricted to the range of sufficiently large Reynolds numbers. The instability of certain shells like the XM 761 [3,4], however, escaped such a prediction and is also distinguished in character owing to the rapid loss in spin rate. Experiments with a full-scale liquid cylinder [5] and subsequent field tests [6] established that this new flight instability is most pronounced for liquid-fills of very high viscosity.

We conduct a theoretical analysis of this problem in order to support the ongoing experiments and to independently obtain insight into the anatomy of the flow phenomena. The initial steps of this analysis are reported elsewhere [7]: evaluation of the experimental data base, dimensional analysis, scaling aspects, governing equations, and discussion of various simplifying assumptions. Two observations in this earlier work led to the building-block approach discussed in the following. First, if the despin (negative roll) moments [5] and void observations [8] are correlated with the Reynolds number Re, at least three regions can be distinguished. At low Re, the despin moment increases proportional to Re, and the void in an incompletely filled cylinder is parallel to the spin axis. This suggests a simple fluid motion that is essentially independent of the axial coordinate, except in the neighborhood of the end walls. In a middle range of Re, the despin moment assumes a maximum, and a wavy distortion of the void seems to indicate a cellular structure of the fluid motion. This cellular motion can, in principle, originate from hydrodynamic instability of the basic flow with respect to axially periodic disturbances. At still higher Reynolds numbers, the despin moment decreases with increasing Re in a manner not clearly defined by the few available data

points. The void observations indicate, however, that the motion ultimately becomes turbulent.

The second observation was the appearance of the nutation rate and angle as a small parameter in the equations for the deviation from solid-body rotation. The forcing term due to nutation can be considered small enough for linearization of the equations in the situations of practical and theoretical interest.

In the following, we describe the development of a simple system of equations for the basic flow. Analytical solutions are given for the flow field and for the liquid moments. A comparison is made with computer simulations of the flow [9] and with experimental data [5]. The properties of inertial modes at low Reynolds numbers and the possibility of instability due to primary resonance is discussed.

2. GOVERNING EQUATIONS. We consider the motion of a fluid of density $\rho$ and viscosity $\mu$ in a cylinder of radius a and length 2c that rotates with the spin rate $\omega$ about its axis of symmetry, the z-axis. We consider the motion with respect to the nutating coordinate system x,y,z. This system is obtained from the inertial system X,Y,Z by a rotation with the nutation angle $\theta$ about the axis Y=y. Therefore, x is in the Z,z-plane, and this plane rotates about the Z-axis with the nutation rate $\Omega$. The two axes of rotation intersect in the center of mass of the cylinder, as shown in Fig. 1. In contrast to the experimental procedures [5], we consider $\omega>0$, $\Omega$, and $0\le\theta\le\pi/2$ as constant. The fluid motion is governed by the Navier-Stokes equations written in the nutating coordinate system:

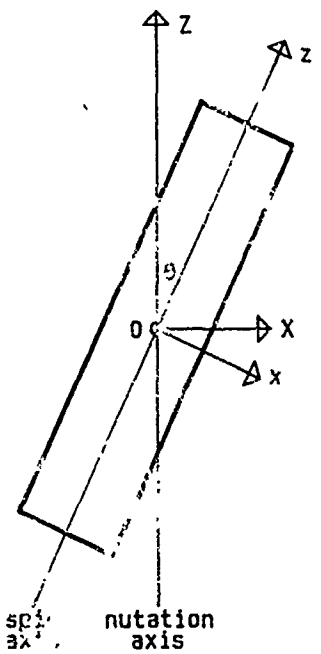



Figure 1. Definition sketch.

$$\rho\left[\frac{D\underset{\sim}{V}_n}{Dt} + 2\underset{\sim}{\Omega}\times\underset{\sim}{V}_n + \underset{\sim}{\Omega}\times(\underset{\sim}{\Omega}\times\underset{\sim}{r})\right] = -\nabla P_n + \mu\nabla^2\underset{\sim}{V}_n \qquad (1)$$

$$\nabla\cdot\underset{\sim}{V}_n = 0.$$

$\underset{\sim}{V}_n$ is the velocity measured in the nutating frame, $P_n$ the pressure, and $\underset{\sim}{r}$ the position vector. The body force due to gravity has been disregarded. Equations (1) are subject to the no-slip and no-penetration conditions at the cylinder walls.

It is convenient [7] to split the velocity and pressure fields according to

$$\underset{\sim}{V}_n = \underset{\sim}{V}_s + \underset{\sim}{V}_d, \quad P_n = P_s + P_d$$

where $\underset{\sim}{V}_s, P_s$ describe the state of pure solid body rotation, whereas $\underset{\sim}{V}_d, P_d$ represent the deviation from solid body rotation. The advantage of this isolated view on the deviation is obvious: $\underset{\sim}{V}_d$ and the reduced pressure $P_d$ are responsible for the observed flight instability. A glance at the equations shows that $V_d \equiv 0$ and $P_d \equiv 0$ if either one of the following conditions is satisfied: $\omega=0$, $\Omega=0$, $\theta=0$ or $\mu\to\infty$ (solid fill).

The equations for $\underset{\sim}{V}_d$, $P_d$ are then written in terms of nondimensional quantities $\underset{\sim}{v}_d$, $p_d$. We use $a$, $\omega$, and $\rho$ for scaling length, time, and mass. Note that this choice is ambiguous [7] and excludes the case $\omega=0$ (which lacks practical interest). The problem then depends on four nondimensional parameters:

| | |
|---|---|
| $\lambda = c/a$ | aspect ratio |
| $\sigma = \sin\theta$ | |
| $\tau = \Omega/\omega$ | frequency |
| $Re = \rho\omega a^2/\mu$ | Reynolds number. |

The aspect ratio enters the solution only through the boundary conditions. The boundary conditions on $\underset{\sim}{v}_d$ are homogeneous.

In cylindrical coordinates $r,\phi,z$, the equations for the nondimensional deviation velocity $\underset{\sim}{v}_d = (v_r, v_\phi, v_z)$ and pressure $p_d$ take the form

$$\frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial v_\phi}{\partial \phi} + \frac{\partial v_z}{\partial z} = 0 \tag{2a}$$

$$D'v_r - \frac{v_\phi^2}{r} - 2(1+\tau_z)\,v_\phi + 2\tau_\phi v_z = -\frac{\partial p_d}{\partial r} + \frac{1}{Re}\left[D''v_r - \frac{v_r}{r^2} - \frac{2}{r^2}\frac{\partial v_\phi}{\partial \phi}\right] \tag{2b}$$

$$D'v_\phi + \frac{v_r v_\phi}{r} + 2(1+\tau_z)\,v_r - 2\tau_r v_z = -\frac{1}{r}\frac{\partial p_d}{\partial \phi} + \frac{1}{Re}\left[D''v_\phi - \frac{v_\phi}{r^2} + \frac{2}{r^2}\frac{\partial v_r}{\partial \phi}\right] \tag{2c}$$

$$D'v_z + 2\tau_r v_\phi - 2\tau_\phi v_r = -\frac{\partial p_d}{\partial z} - 2r\tau_r + \frac{1}{Re}D''v_z \tag{2d}$$

where

$$D' = \frac{\partial}{\partial t} + \frac{\partial}{\partial \phi} + v_r\frac{\partial}{\partial r} + \frac{v_\phi}{r}\frac{\partial}{\partial \phi} + v_z\frac{\partial}{\partial z}$$

$$D'' = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \phi^2} + \frac{\partial^2}{\partial z^2}$$

and

$$\tau_r = -\varepsilon\cos\phi,\ \tau_\phi = \varepsilon\sin\phi,\ \tau_z = \tau\cos\theta,\ \varepsilon = \tau\sin\theta \tag{3}$$

The primary effect of nutation is contained in the $\phi$-periodic force term $-2r\tau_r = 2\varepsilon r\cos\phi$ in the z-momentum equation (2d). If this term vanishes throughout, $\varepsilon=0$, equations (2) support a trivial solution $\underset{\sim}{v}_d=0$, $p_d=0$. For sufficiently small $\varepsilon\neq0$, it is obvious that the deviation velocity is of order $O(\varepsilon)$. In the situations of practical interest, $\varepsilon = (\Omega/\omega)\sin\theta$ turns out to be a rather small parameter. Even a conservative estimate with $\Omega \leq 500$ rpm, $\omega \geq 3000$ rpm, and $\theta \leq 20^0$ provides values of $\varepsilon \leq 0.057$. Consequently, it seems well justified to linearize the equations in $\varepsilon$. This linearization imposes no restriction on the Reynolds number.

3. THE BASIC FLOW. The system of equations after linearization is still quite difficult to solve. Any serious attempt to satisfy all boundary conditions leads directly to a purely computational approach. Use of the boundary-layer approximation would simplify the task but seems inappropriate in the interesting range of low Reynolds numbers. Recalling that the flow in a relatively long cylinder (aspect ratio $\lambda=4.3$) at low Re exhibits little axial variation over much of the cylinder length [7], we have relaxed the boundary conditions at the end walls. As a first step, we seek for a steady flow in an infinitely long cylinder.

At closer analysis, the z-independent force term in eq. (2d) can only be balanced by a purely axial deviation velocity. It is consistent with the linearized equations to assume a solution in the form

$$\underset{\sim}{v}_d = (0,\ 0,\ v_z),\ p_d = 0 \tag{4}$$

and moreover,

$$v_z = v_z(r,\phi) = 2\varepsilon\left[f(r)\cos\phi + g(r)\sin\phi\right] \tag{5}$$

Substituting (4),(5) into the linearized equations provides

$$f'' + \frac{1}{r}f' - \frac{1}{r^2}f - \mathrm{Re}\,g = -\mathrm{Re}\,r \tag{6a}$$

$$g'' + \frac{1}{r}g' - \frac{1}{r^2}g + \mathrm{Re}\,f = 0 \tag{6b}$$

$$f=0,\ g=0 \quad \text{at} \quad r=1 \tag{6c}$$

$$f,\ g \quad \text{finite} \quad \text{at} \quad r=0 \tag{6d}$$

The primes denote d/dr. For $\mathrm{Re}\to0$, the solution of these equations can be found in the form of series

$$f = \frac{\mathrm{Re}}{8}(r - r^3) - \frac{\mathrm{Re}^3}{9216}(7r - 12r^3 + 6r^5 - r^7) + O(\mathrm{Re}^5) \tag{7a}$$

$$g = \frac{Re^2}{192}(2r - 3r^3 + r^5) + O(Re^4). \qquad (7b)$$

With higher terms included, these series converge for $Re \leq 12$. In the limit $Re \to \infty$, one obtains

$$f \to 0, \quad g \to r \quad \text{as} \quad Re \to \infty. \qquad (8)$$

Owing to the loss of the highest derivatives, however, this solution cannot satisfy the boundary conditions (6c) and is valid only outside the thin boundary layers near the wall at r=1. Even without any knowledge of the solution in the intermediate range, the different character of the basic flow at low and high Reynolds numbers is evident. At low Re, the component f in the z,x-plane $\phi=0$ is dominating. At high Re, f is negligible in the core of the cylinder while g in the z,y-plane $\phi=90^0$ is dominating.

In earlier work [10], we have applied a spectral collocation method for numerical solution of eqs. (6). Here, we derive an analytical solution by introducing the complex function F=g+if. Eqs. (6) can then be written in the form

$$r^2F'' + rF' - (1 + iRe\, r^2)F = - iRe\, r^3 \qquad (9a)$$

$$F = 0 \quad \text{at} \quad r = 1 \qquad (9b)$$

$$F \text{ finite} \quad \text{at} \quad r = 0 \qquad (9c)$$

A particular solution of the inhomogeneous equation (9a) is $F_0=r$, whereas the homogeneous part of (9a) is the equation for the modified Bessel functions $I_1(qr)$ and $K_1(qr)$ of the complex argument qr with $q = \sqrt{Re/2}\,(1 + i)r$. For (9c), $K_1(qr)$ cannot contribute to the solution. Finally, (9b) provides

$$F(r) = g + if = r - I_1(qr)/I_1(q). \qquad (10)$$

Expressing the solution in terms of Kelvin functions of real argument is of little advantage for the numerical evaluation. The solution is valid for arbitrary Re but may be unstable as Re exceeds some critical value. It is straightforward to derive the approximations (7) from the ascending series for $I_1$ (and to explain the convergence problem for larger Re). The asymptotic expansion for large arguments provides

$$F \sim r - \sqrt{r}\, e^{q(r-1)}. \qquad (11)$$

This expression agrees to within 1% with (8) provided that $r \leq 1 - \delta$. The boundary layer thickness $\delta$ can be obtained from the transcendental equation

$$\delta = \sqrt{2/Re}\,[4.605 - \tfrac{3}{2}\ln(1 - \delta)], \qquad (12)$$

e.g., $\delta=0.223$ for Re=1000. The characteristic changes in the flow structure with increasing Re, in particular the shift of the velocity maximum from $\phi\approx0$ at Re=2 to $\phi\approx90^0$ at Re=200 are shown in Fig. 2.

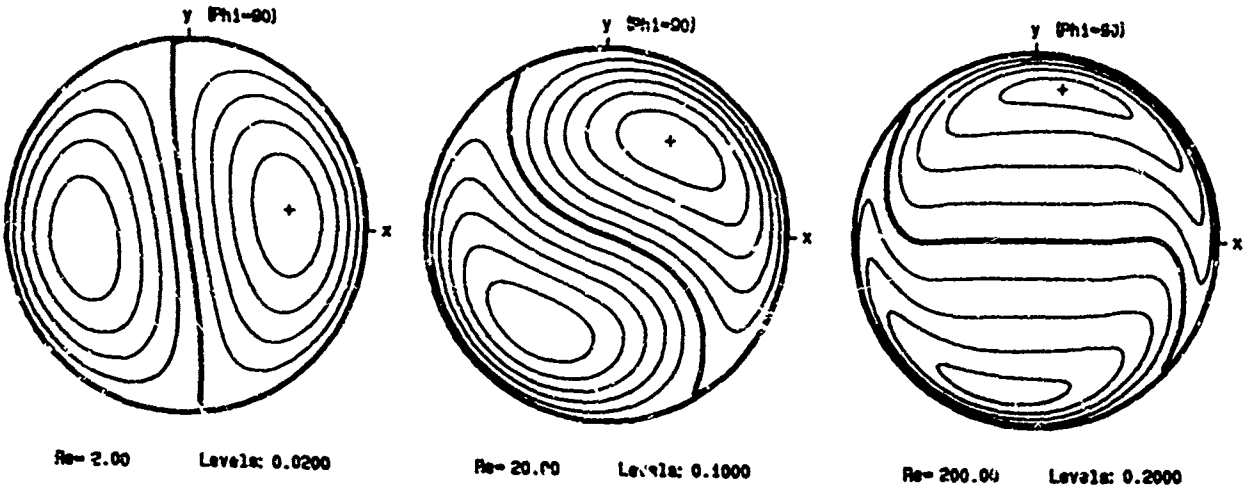



Figure 2. Contour lines of equal axial velocity, $v_z/(2\varepsilon)$ = const., for Re=2, 20, and 200. The difference between levels is 0.02, 0.1, and 0.2, respectively. The + marks the velocity maximum.

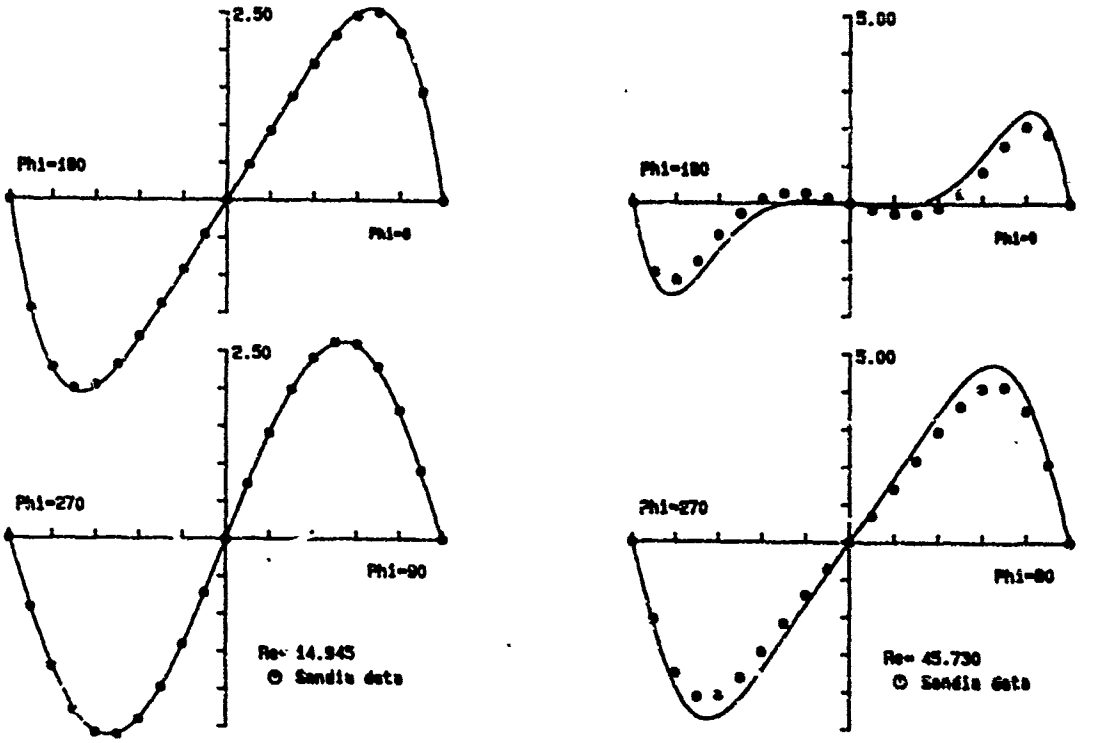



Figure 3. Radial distribution of the axial velocity (in fps) at z=0 for Re=14.9 and Re=45.7. The symbols show the solution to the full Navier-Stokes equations.

Fig. 3 compares the dimensional velocity distributions obtained from (5), (10) with computational results for the center cross-section (z=0) of the cylinder.* The agreement for Re=14.9 is considered representative for the range of lower Reynolds numbers. The numerical simulation provides very small components $v_r$, $v_\vartheta$ and hence verifies our estimates. At the higher value Re=45.7, a systematic deviation between the two results seems to be due to a superposed cellular motion that is not yet incorporated into our analysis. It is encouraging, however, that the simple theory of the basic flow yields results in essential agreement with the computational solution of the full Navier Stokes equations for a finite-length cylinder.

4. MOMENTS. With the deviation velocity $\underset{\sim}{V}_d = (0,0,\omega a v_z)$ and $v_z$ given, the moments on a finite-length section of the cylinder can be calculated. We consider a control volume R (surface S) formed by the solid cylindrical wall and liquid surfaces at both ends. Conservation of angular momentum requires

$$\underset{\sim}{M} + \Sigma\,(\underset{\sim}{r} \times \underset{\sim}{F}_d) = \frac{\partial}{\partial t}\iiint_R (\underset{\sim}{r} \times \underset{\sim}{V}_d)\rho dR + \iiint_R [\underset{\sim}{r} \times (2\underset{\sim}{\Omega} \times \underset{\sim}{V}_d)]\rho dR$$
$$+ \iint_S (\underset{\sim}{r} \times \underset{\sim}{V}_d)\rho(\underset{\sim}{V}_d \cdot \underset{\sim}{n})dS + \iint_S (\underset{\sim}{r} \times \underset{\sim}{V}_s)\rho(\underset{\sim}{V}_d \cdot \underset{\sim}{n})dS \qquad (13)$$

where $\underset{\sim}{n}$ is the outer unit normal. On the left-hand side, M is the resultant torque on the control volume. The second term accounts for the moments due to the shear force $\underset{\sim}{F}_d$ and vanishes owing to the solid sidewall and cancellation of the contributions from both ends**. On the right-hand side, the first term vanishes for steady $\underset{\sim}{V}_d$. The second term originates from Coriolis forces in the nutating system. The third term vanishes since $\underset{\sim}{V}_d$ has only an axial component. The last term then provides the net rate of angular momentum flux through the control surface.

Substitution of $\underset{\sim}{V}_d$ leads to the following expressions for the cartesian components of $\underset{\sim}{M}$:

$$M_x = m_\ell(2\Omega a\sin\theta)(\omega a)\, m_x,\quad m_x = -\int_0^1 r^2 f dr \qquad (14a)$$

$$M_y = m_\ell(2\Omega a\sin\theta)(\omega a)\, m_y,\quad m_y = -\int_0^1 r^2 g dr \qquad (14b)$$

$$M_z = m_\ell(2\Omega a\sin\theta)^2\, m_z,\quad m_z = \int_0^1 r^2 f dr = -m_x \qquad (14c)$$

where $m_\ell$ is the liquid mass in the cylinder. In this form, the components $M_x$, $M_y$ represent the net rate of angular momentum flux through the liquid endwalls, whereas the roll moment $M_z$ is solely due to Coriolis forces. A close

*The data were kindly provided by Dr. H. Vaughn, Sandia National Laboratories.

**Improper account of the sidewall conditions introduced an incorrect factor of two in earlier results for $M_x$, $M_y$ [10].

relation between roll moment $M_z$ and yaw moment $M_x$ has also been found by Murphy [11] for the range of high Reynolds numbers.

A different interpretation can be derived using the differential equation (9a), integrating by parts, applying (9b), and separating real and imaginary part:

$$m_z = -m_x = \int_0^1 r^2 f dr = -\frac{g'(1)}{Re} \tag{15a}$$

$$m_y = -\int_0^1 r^2 g dr = -\frac{f'(1)}{Re} - \frac{1}{4}. \tag{15b}$$

In this form, the moments are directly related to the shear forces at the cylindrical sidewall, r=1. Since $f'(1)<0$, $g'(1)<0$, the roll moment $M_z$ is always positive (even for $\Omega<0$), while $M_x$ is negative for $\Omega<0$ and changes sign with $\Omega$. For small Re, the series (7) provide the approximations

$$m_z \approx \frac{Re}{96},\ m_y \approx -\frac{Re^2}{1536} \tag{16}$$

that can be used for quick estimates up to $Re \leq 10$. The linear increase of $m_z$ and $M_z$ with Re is consistent with the experimental data. From the analytical solution (10), we obtain

$$F'(1) = g'(1) + if'(1) = 2 - qI_0(q)/I_1(q). \tag{17}$$

Substitution into (15) provides the variation of $m_z$, $m_y$ with the Reynolds number shown in Fig. 4. The coefficient $m_z$ assumes a pronounced maximum at

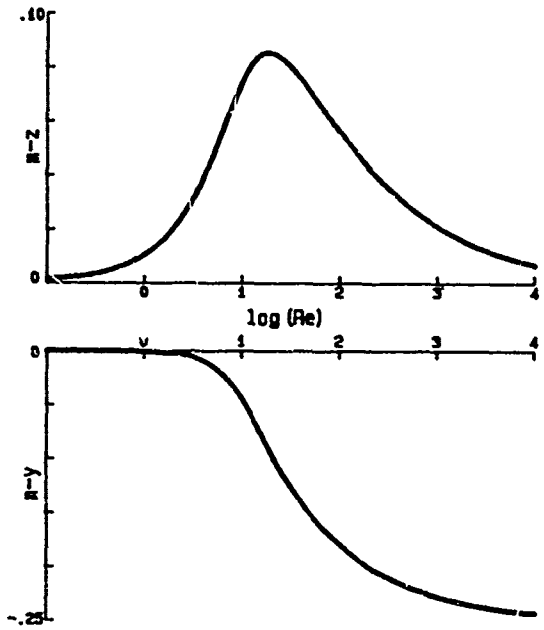



Fig. 4. The nondimensional coefficients $m_z$, $m_y$ in eq. (14) versus the Reynolds number, Re.

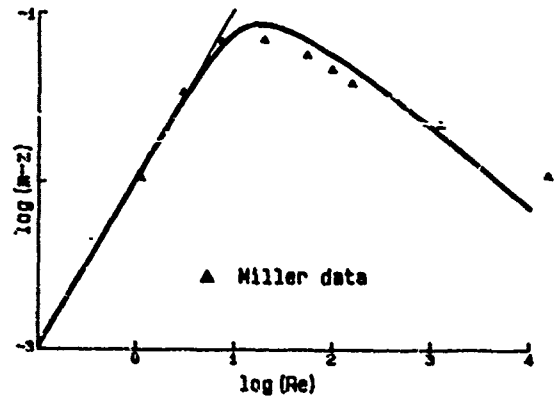



Fig. 5. Comparison of the theoretical result for $m_z$ with experimental data [5]. The straight line shows the asymptotic law $m_z \approx Re/96$.

Re≈19. The coefficient $m_y$ is negative and reaches an asymptotic value of $m_y \to -1/4$ as $Re \to \infty$. Hence, for $\Omega>0$, $M_y$ reduces the pitch moment due to the solid body rotation. This result is consistent with computations [9]. A comparison of theoretical and experimental [5] results for $m_z$ is given in Fig. 5 on a double logarithmic scale. The initial spin rate $\omega$=4000 rpm was used for reducing the experimental data. For Re<10, the experimental points match the analytical result as well as the asymptotic law $m_z \approx Re/96$. The occurrence of a maximum of $m_z$ is found to be a property of the basic flow. Only the systematic deviation for higher Reynolds numbers Re<200 may be attributed to a cellular motion. The two data points for $Re>10^3$ probably indicate turbulent flow.

The basic flow, hence, can be considered a first but essential step toward understanding and predicting the gross features of the fluid motion in a spinning and nutating cylinder. Some observations, however, such as the virtual independence of the despin moment on the spin rate require further analysis, especially of the end effects. The occurrence of a cellular motion may be due to hydrodynamic instability of the basic flow.

5. STABILITY ANALYSIS. The stability analysis is currently conducted and only a brief outline is given here. We superpose to the steady flow $v_n=(o,r,v_z)$, $p_n$ disturbances $v'=(u,v,w)$, p sufficiently small for linearization. Substitution into eqs.~(2) and neglect of products between disturbances and terms of order $O(\epsilon^2)$ provides the following stability equations:

$$\left\{\frac{\partial u}{\partial t} + \frac{\partial u}{\partial \phi} - 2(1+\tau_z)v + \frac{\partial p}{\partial r}\right\} - \frac{1}{Re}\left\{D''u - \frac{u}{r^2} - \frac{2}{r^2}\frac{\partial v}{\partial \phi}\right\} + \left\{v_z\frac{\partial u}{\partial z} + 2\tau_\phi w\right\} = 0 \tag{18a}$$

$$\left\{\frac{\partial v}{\partial t} + \frac{\partial v}{\partial \phi} + 2(1+\tau_z)u + \frac{1}{r}\frac{\partial p}{\partial \phi}\right\} - \frac{1}{Re}\left\{D''v - \frac{v}{r^2} + \frac{2}{r^2}\frac{\partial u}{\partial \phi}\right\} + \left\{v_z\frac{\partial v}{\partial z} - 2\tau_r w\right\} = 0 \tag{18b}$$

$$\left\{\frac{\partial w}{\partial t} + \frac{\partial w}{\partial \phi} + \frac{\partial p}{\partial z}\right\} - \frac{1}{Re}\{D''w\} + \left\{\left(\frac{\partial v_z}{\partial r} - 2\tau_\phi\right)u + \left(\frac{1}{r}\frac{\partial v_z}{\partial \phi} + 2\tau_r\right)v + v_z\frac{\partial w}{\partial z}\right\} = 0 \tag{18c}$$

$$\frac{1}{r}\frac{\partial}{\partial r}(ru) + \frac{1}{r}\frac{\partial v}{\partial \phi} + \frac{\partial w}{\partial z} = 0 \tag{18d}$$

Three groups of terms have been separated by braces in eqs. (18a-c). The first group, if set to zero, represents the equations for inviscid inertial modes $\sim\exp(im\phi + i\alpha z + st)$, where m is the (integer) azimuthal, $\alpha$ the axial wavenumber, and $s = s_r + is_i$ provides the amplification rate $s_r$ (=0) and frequency $s_i$. Usually, an equation for the pressure is used for obtaining the analytical solution. We have derived an alternative system in terms of u,v and applied the spectral method to be used for more general cases in order to check the numerical results against the exact values.

The second group of terms multiplied with 1/Re represents the viscous correction to the inertial modes. By eliminating w and p, a system of ordinary differential equations for u and v has been derived. Due to the higher order of this system, all boundary conditions can be satisfied. Programs have been developed for calculating spectra of (complex) eigenvalues, for tracing single eigenvalues as function of Re, $\alpha$, and for obtaining the eigenfunctions. At high Re, the results follow the trends predicted by asymptotic theories. Our analysis, however, also covers the range of low Reynolds numbers, where the inertial modes suffer rapid decay ($s_r<0$).

The most interesting aspect of the stability equations (18) is the third group of terms. The coefficients in this group, $v_z$, $\tau_r$, and $\tau_\phi$, are (1) of order $O(\varepsilon)$ and (2) periodic in $\phi$. The periodicity in $\phi$ leads to a coupling of the mode equations for m and m±1, and may cause primary resonance between inertial modes. In view of viscous damping, this resonance is likely to occur as $\varepsilon$ exceeds a critical value that decreases as Re increases. The analysis of this parametric instability is currently in preparation.

## ACKNOWLEDGMENT

The open cooperation and sharing of data with Miles C. Miller (CRDC) and Harold R. Vaughn (Sandia Laboratories) are greatly appreciated. This work is supported by the Army Research Office under Contract DAAG29-82-K-0129 and by the Army AMCCOM under Contract DAAK11-83-K-0011.

## REFERENCES

[1] Stewartson, K. 1959 "On the Stability of a Spinning Top Containing Liquid," Journal of Fluid Mechanics, Vol. 5, Part 4, pp. 577-592.

[2] Wedemeyer, E. H. 1966 "Viscous Corrections to Stewartson's Stability Criterion," Ballistic Research Laboratory, Report 1325.

[3] D'Amico, W. P. 1977 "Field Tests of the XM761: First Diagnostic Test," Ballistic Research Laboratory, Memorandum Report 2792.

[4] D'Amico, W. P. 1978 "Field Tests of the XM761: Second Diagnostic Test," Ballistic Research Laboratory, Memorandum Report ARBRL-MR-02806.

[5] Miller, M. C. 1982 "Flight Instabilities of Spinning Projectiles Having Nonrigid Payloads," Journal of Guidance, Control, and Dynamics, Vol. 5, pp. 151-157.

[6] D'Amico, W. P. & Miller, M. C. 1979 "Flight Instability Produced by a Rapidly Spinning, Highly Viscous Liquid," Journal of Spacecraft and Rockets, Vol. 16, pp. 62-64.

[7] Herbert, Th. 1982 "Fluid Motion in a Rotating and Nutating Cylinder - Part I," Report prepared under the Scientific Services Program.

[8] Miller, M. C. 1981 "Void Characteristics of a Liquid Filled Cylinder Undergoing Spinning and Coning Motion," Journal of Spacecraft and Rockets, Vol. 18, 286-288.

[9] Vaughn, H. R., Oberkampf, W. L. & Wolfe, W. P. 1983 "Numerical Solution for a Spinning Nutating Fluid-Filled Cylinder," Sandia Report SAND 83-1789.

[10] Herbert, Th. 1983 "The Flow of Highly Viscous Fluid in a Spinning and Nutating Cylinder," Proceedings of the 1983 Scientific Conference on Chemical Defense Research, Aberdeen Proving Ground, Md.

[11] Murphy, C. M. 1984 "A Relationship between Liquid Roll Moment and Liquid Side Moment," Ballistic Research Laboratory Memorandum Report ARBRL-MR-03347.

# Computing Sets

Charles R. Leake

USA Concepts Analysis Agency
ATTN: CSCA-RQR
8120 Woodmont Avenue
Bethesda, MD 20814-2797

Abstract. In large linear programing problems, rounding off type errors can cause problems. These errors stem from the computational set which is used by a computer. Consider $\epsilon > 0$, sets F and R where R is the set of rational numbers and $F \subset R$, there exists $\alpha$, $\beta$, $\gamma$, and $\delta \in F$ with $\alpha \neq 0$, $\beta \neq 0$ and $\gamma \neq -\delta$ such that whenever $\alpha \cdot \beta < \epsilon$ or $\gamma + \delta < \epsilon$, then

$$\begin{aligned} \alpha \cdot \beta &= 0 \\ \gamma + \delta &= 0. \end{aligned}$$

It is shown that this set is commutative, contains additive and multiplicative inverses, is nonassociative, nondistributive with zero divisors.

Two other sets which are quadratic extensions of the field $\Gamma$, $C_N$ and $C_J$ that are nonassociative algebras with zero divisors are developed with an example of how they can be used to solve a differential equation of the form

$$P = \frac{dS_i}{dt} = \sum J_p K_p$$

where P, $J_p$ and $X_p$ are vectors of any dimension. $C_N$ is used to solve the case where $J_p K_p = K_p J_p$ and $C_i$ the more general noncommutative case. Finally, the field properties of F, $C_N$ and $C_J$ are compared with some implications for software as a result of the loss of field properties.

1. The concept of a computational set F which permits round-off errors. $F \subset R$ where R is the set of rational numbers. The arithmetic on F, for not necessarily zero elements $\alpha, \beta, \gamma, \delta, \lambda$ and $\omega$ with $\gamma \neq -\delta$ and $\epsilon > 0$, is

$$(1) \quad \alpha + \beta = \begin{cases} \lambda \in F, \text{ if } |\lambda| > \epsilon \\ 0 \in F, \text{ if } |\lambda| \leq \epsilon \end{cases}$$

$$(2) \quad \gamma \cdot \delta = \begin{cases} \omega \in F, \text{ if } |\omega| > \epsilon \\ 0 \in F, \text{ if } |\omega| \leq \epsilon \end{cases}$$

An additional property of this set is the existence of a whole number N such that $N > N + 1$.

As a consequence of the definition of the set arithmetic as well as the properties of the rational numbers, the set is commutative and closed under the operations of addition (+) and multiplication (·). The set also contains

PREVIOUS PAGE IS BLANK

multiplicative and additive inverses. However, not all elements have multiplicative inverses. For example, the number 3 does not have an inverse because the reciprocal of 3 is some truncated version of 1/3. For this example let us consider 1/3 = .33, then 3 x .33 = .99≠1.

Additionally, as shown by the following two examples (3) and (4) additive inverses are not unique. Let $\alpha$ = .011, $\beta$ = -.012, $\delta$ = -.013 and $\varepsilon$ = .01

(3) $\alpha+\beta = -.001$

$= 0$

because $|-.001| < \varepsilon$, and

(4) $\alpha+\delta = 0$

for the same reason. Thus additive inverses exist but are not unique. Because of this characteristic, the associative property for the set F does not always prevail.

The set F has zero divisors. Consider $\alpha$, $\beta$ and $\varepsilon$ defined as above, then

(5) $\alpha\cdot\beta = -0.000132.$

Since $|-.000132| < \varepsilon$, $\alpha\cdot\beta = 0$. Moreover, for $\alpha$, $\beta$ and $\varepsilon$ defined as above and $\delta = 1000$

(6) $(\alpha\cdot\beta)\cdot\delta \neq \alpha\cdot(\beta\cdot\delta)$

and

(7) $\delta\cdot(\alpha+\beta) \neq \delta\cdot\alpha + \delta\cdot\beta$

Thus F is nonassociative and nondistributive.

A consequence of the properties of F is that software using assumed properties can give misleading results. For example in FORTRAN although it is often assumed to be correct, statements (8) and (9) are not always the same.

(8) SUMX = SUMX + C*X

and

(9) SUMX = SUMX + X

SUMX = C*SUMX

where C is a constant.

2. The concept of a quadratic extension $\Gamma$ of a field K of characteristic ≠2. $\Gamma$ is a set with a product xy and sum x + y defined on it with a subset $\overline{X}$ = K as well as an automorphism $Z \rightarrow \overline{Z}$. For each $Z \in \Gamma$ there is a $T = Z + \overline{Z}$ and an $N = Z\overline{Z}$, for T, N ∈X. For each $Z \in \Gamma$, we have

(10) $Z^2 - TZ + N = 0$

It can be assumed that $\Gamma \neq X$. In this case there exists at least one element $v \in X$. These elements constitute the set V. It can be shown that

$$(11)\quad \Gamma = X \oplus V.$$

The image of the automorphism $\bar{Z}$ is called the conjugate of Z. T is its trace, N the norm, X is the set of invariant elements and V the set of skew-conjugate elements.

Considerable work has been done on sets $\Gamma$. In [1] $\Gamma$ is characterized for the real, complex, quaternion and Cayley number systems. In [2] $\Gamma$ is generalized to a field K where there is an element i such that $i^2 - \beta i - \alpha = 0$. The concept is then extended to the case where K is a commutative ring with unit that admits an involutorial automorphism in [3] and in [4] the geometry of the Z-plane of a quadratic extension $\Gamma$ of a field K is discussed. In [5] and [6] there are examples of when $\Gamma$ is a nonassociative algebra.

Essentially quadratic extensions belong to a class of algebras commonly known as Clifford numbers. Van der Waerden in [7] discusses a class of these numbers in his section on hypercomplex numbers.

3. The concept of the commutative algebra $C_J$ over a field K of characteristics $\neq 2$. $C_J$ is an algebra of order $J \geq 1$ where $1, e_2, e_3, \ldots, e_J$ is a basis for $C_J$. In the case of $J = 1$, $C_J = K$. Using the operations defined on K, $C_J$ has the following sum and products defined on it for all $a, b \in C_J$ and $\alpha, a_i, b_i \in K$

$$(12)\quad a + b = \sum_i (a_i + b_i)e_i,$$

$$(13)\quad a = \alpha \sum_i a_i e_i,$$

$$(14)\quad ab = (a_1 b_1 - \sum_{i=2}^{J} a_i b_i)1 + \sum_{i=2}^{J} (a_1 b_i + a_i b_1)e_i.$$

$$(15)\quad \text{The automorphism } a \rightarrow \bar{a} = a_1 1 - \sum_{i=2}^{J} a_i e_i.$$

The unit or 1 in $\Gamma$ is

$$(16)\quad 1 = 1 + \sum_{i=2}^{J} o e_i.$$

The trace T and norm N are

$$(17)\quad T_a = \bar{a} + a \quad \text{and} \quad (18)\quad N_a = a\bar{a}.$$

$T_a$ and $N_a \in K$ and $ab = ba$ for all $a, b \in C_J$. When the characteristic of K is 0, K is real. The set $C_J$ has inverses

$$(19)\quad a^{-1} = \frac{\bar{a}}{N_a}$$

$C_J$ has some interesting properties. For $n = 2$, $C_J \simeq \Phi \subseteq \mathbb{C}$, the complex numbers, for appropriate restrictions on K. For $n > 2$, $C_J$ is nonassociative, nonalternative and contains zero divisors. Moreover,

$$(20)\quad N_{ab} \neq N_a N_b$$

$$(21)\quad N_a^2 \neq (N_a)^2$$

in general. When $K = \Phi \subset \mathbb{C}$ which contains at least one element $v = -\bar{v}$, (19) does not always hold true. For $K \subseteq \mathbb{C}$, (10) holds. Hence, this class c algebras $C_J$ are quadratic extensions of the field K.

4. An application of the class of algebras $C_J$ to the physical sciences. In the physical sciences such equations as

$$(22)\quad P = \frac{dS_i}{dt} = \sum J_p K_p$$

are frequently encountered where P, $J_p$ and $X_p$ can be vectors of any dimension $n = 1,2,\ldots,m$. If $P, J_p, X_p$ $C_J$ with $J = n$, then (22) has the representation.

$$(23)\quad \begin{cases} \sum_{P=1}^{L} \left(j_{p1}\, x_{p1} - \sum_{1=2}^{n} j_{pi}\, x_{pi}\right) = P_1 \\ \sum_{P=1}^{L} \sum_{i=2}^{n} \left(j_{p1}\, x_{pi} + j_{pi}\, x_{pi}\right) = P_i \end{cases}$$

and for $N_p$

$$(24)\quad \begin{cases} N_p = \sum_i P_i^2 = \sum_p \sum_i P_i\, j_{pi}\, x_{pi} \\ \quad = \sum_P \left(\sum_i j_{pi}^2\, x_{pi}^2 + \sum_{k \neq p} \sum_i j_{pi}\, x_{pi}\, j_{ki}\, x_{ji}\right) \end{cases}$$

$$(25)\quad \begin{cases} \sum_P \sum_{i=2}^{n} \left(P_1 j_{pi} x_{pi} - P_i j_p x_{p1}\right) = 0 \\ \sum_P \sum_{k \neq p} \sum_{i=2}^{n} \left(j_{p1} x_{p1} j_{ki} x_{ki} - j_{pi} x_{pi} j_{k1} x_{k1}\right) = 0 \end{cases}$$

and where $p, j, x \in K$.

(23), (24) and (25) hold regardless of whether K is the set of real numbers or the set of complex numbers. In this example it should be remembered that $C_J$ is commutative. The noncommutative case will be treated in the next two sections. The last section will address the case that $J_p$ is a matrix or linear transformation.

5. The concept of the noncommutative algebra $C_n$ over a field K of characteristic $\neq 2$. $C_n$ is an algebra of order $n \geq 1$ defined as in section 2 with the one modification.

$$
(14')\quad \begin{cases} ab = (a_1b_1 - \sum_{i=2}^{m} a_ib_i)1 + (a_1b_2 + a_2b_1 + a_3b_4 + a_4b_3)e_2 \\ + (a_1b_3 - a_2b_4 + a_3b_1 - a_4b_2)e_3 \\ + \sum_{i=4}^{n} (a_1b_i + a_2b_3 - a_3b_2 + a_ib_1)e_i. \end{cases}
$$

$C_n$ has all the properties discussed in section 2 with the exceptions that $C_n$ is no longer generally commutative and for $n = 4$ there exists a $C_n \simeq Q$, the set of quaternions.

6. An application of the class of algebras $C_n$ to the physical sciences. In the case that the elements P, $J_P$ and $X_P$ are noncommutative, we have as a solution to (22)

$$
(23')\quad \begin{cases} \sum_P (j_{p1}x_{pi} - \sum_{i=2}^{n} j_{pi}\, x_{pi}) = P_1 \\ \sum_P (j_{pi}x_{p2} + j_{p2}x_{p1} + j_{p3}x_{p4} + j_{p4}x_{p3}) = P_2 \\ \sum_P (j_{p1}x_{p3} - j_{p2}x_{p4} + j_{p3}x_{p1} - j_{p4}x_{p2}) = P_3 \\ \sum_P [j_{p2}x_{p3} - j_{p3}x_{p2} + \sum_{i=4}^{n} (j_{p1}x_{pi} + j_{pi}x_{p1})] = P_i \end{cases}
$$

$N_P$ is again given by (24) or is invariant and (25) is given below by (25').

$$
(25')\quad \begin{cases} \sum_P \sum_{k\neq P} (j_{p1}x_{p1}j_{k2}x_{k2} - j_{p2}x_{p2}j_{k1}x_{k1} - j_{p3}x_{p3}j_{k4}x_{k4} - j_{p4}x_{p4}j_{k3}x_{k3}) = 0 \\ \sum_P \sum_{k\neq P} (j_{p1}x_{p1}j_{k3}x_{k3} + j_{p2}x_{p2}j_{k4}x_{k4} - j_{p3}x_{p3}j_{k1}x_{k1} + j_{p4}x_{p4}j_{k2}x_{k2}) = 0 \\ \sum_P \sum_{k\neq P} [(j_{p2}x_{p2}j_{k3}x_{k3} + j_{p3}x_{p3}j_{k2}x_{k2} + \sum_{i=4}^{n} (j_{p1}x_{p1}j_{ki}x_{ki} \\ \qquad - j_{pi}x_{pi}j_{k1}x_{k1})] = 0 \\ \sum_P (P_1j_{p2}x_{p2} - P_2j_{p1}x_{p1} - P_3j_{p4}x_{p4} - P_4j_{p3}x_{p3}) = 0 \\ \sum_P (P_1j_{p3}x_{p3} + P_2j_{p4}x_{p4} - P_3j_{p1}x_{p1} + P_4j_{p2}x_{p2}) = 0 \\ \sum_P [-P_2j_{p3}x_{p3} + P_3j_{p2}x_{p2} + \sum_{i=4}^{n} (P_ij_{pi}x_{pi} - P_ij_{p1}x_{p1})] = 0 \end{cases}
$$

Moreover, it can be shown that the equations given by (23), (23'), (24), 5) and (25') hold when K is an associative ring with unit. In the event that K is the real numbers(24)$\geq$ 0.

7. An application of $C_1$ and $C_n$ in the case that the $J_p$'s are linear transformations. For $J_p$ a linear transformation and P,$X_p$ $C_J$ the following representations hold for $N_p$.

$$(25)\quad S_1 = \sum_i P_i^2 = \sum_P \sum_k \sum_i P_i j_{pki} x_{pi}$$

$$(27)\quad S_2 = \sum_P \sum_{k=2}^{n} [P_1 \sum_i j_{pki} x_{pi} - P_k \sum_i j_{p1i} x_{pi}]$$

$$(28)\quad S_3 = \sum_P \sum_k (\sum_i j_{pki} x_{pi})^2$$

$$(29)\quad S_4 = \sum_P \sum_{l\neq P} \sum_k [(\sum_j j_{pki} x_{pi})(\sum_i j_{lri} x_{pi})]$$

$$(30)\quad S_5 = \sum_P \sum_{q\neq P} \sum_{k=2}^{n} [(\sum_i j_{p1i} x_{pi})(\sum_i j_{qki} x_{qi}) - (\sum_i j_{pki} x_{pi})(\sum_i j_{q1i} x_{qi})]$$

with

$$(31)\quad S_1 = S_3 + S_4 \text{ and}$$

$$(32)\quad S_2 = S_5 = 0$$

For $J_p$ a linear transformation and P, $X_p$ $C_n$ the following representations hold for $N_p$.

$$(33)\quad S_6 = \sum_P \sum_{Q\neq P} [(\sum_i j_{p1i} x_{pi})(\sum_i j_{q2i} x_{qi}) - (\sum_i j_{p2i} x_{pi})(\sum_i j_{q1i} x_{qi})$$
$$- (\sum_i j_{p3i} x_{pi})(\sum_i j_{q4i} x_{li}) - (\sum_i j_{p4i} x_{pi})(\sum_i j_{q3i} x_{qi})]$$

$$(34)\quad S_7 = \sum_P \sum_{q\neq P} [(\sum_i j_{p1i} x_{pi})(\sum_i j_{q3i} x_{qi}) + (\sum_i j_{p2i} x_{pi})(\sum_i j_{q4i} x_{qi})$$
$$- (\sum_i j_{p3i} x_{pi})(\sum_i j_{q1i} x_{qi}) + (\sum_i j_{p4i} x_{pi})(\sum_i j_{q2i} x_{qi})]$$

$$(35)\quad S_8 = \sum_P \sum_{q\neq P} [-(\sum_i j_{p2i} x_{pi})(\sum_i j_{q3i} x_{qi}) + (\sum_i j_{p3i} x_{pi})(\sum_i j_{q2i} x_{qi})$$
$$+ \sum_{k=4}^{n} [(\sum_i j_{p1i} x_{pi})(\sum_i j_{qki} x_{qi}) - (\sum_i j_{pki} x_{pi})(\sum_i j_{q1i} x_{qi})]]$$

$$(36)\quad S_9 = \sum_P [P_1 \sum_i j_{p2i} x_{pi} - P_2 \sum_i j_{p1i} x_{pi} - P_3 \sum_i j_{p4i} x_{pi} - P_4 \sum_i j_{p2i} x_{pi}]$$

$$(37)\quad S_{10} = \sum_{P}\left[P_1\sum_{i} jp_{3i}x_{pi} + P_2\sum_{i} jp_{4i}x_{pi} - P_3\sum_{i}^{n} jp_{1i}x_{pi} + P_4\sum_{i} jp_{2i}x_{pi}\right]$$

$$(38)\quad S_{11} = \sum_{P}\left[-P_2\sum_{i} jp_{3i}x_{pi} + P_3\sum_{i} jp_{2i}x_{pi} + \sum_{k=4}^{n}\left(P_1\sum_{i} jp_{ki}x_{pi} - P_i\sum_{i} jp_{1i}x_{pi}\right)\right]$$

with (31) as for the commutative case and

$$(39)\quad S_6 = S_7 = S_8 = S_9 = S_{10} = S_{11} = 0.$$

Thus it has been shown that $C_J$ and $C_n$ can be used to solve vector equations.

8. A comparison of the field properties of F, $C_J$ and $C_N$. As shown below in Table 1, F does not have all the properties that $C_J$ and $C_N$ have with respect to addition and multiplication inverses. It is interesting to note that there is not much interest in using sets such as $C_J$ and $C_N$. However, we use a set such

| | F | | $C_J$ | | $C_N$ | |
|---|---|---|---|---|---|---|
| | + | · | + | · | + | · |
| Closure | yes | yes | yes | yes | yes | yes |
| Associativity | no | no | yes | no | yes | no |
| Commutativity | yes | yes | yes | yes | yes | no |
| Identity | yes | yes | yes | yes | yes | yes |
| Inverse | not unique | no | yes | yes | yes | yes |
| Distributive | no | | no | | no | |

Table 1

as F in our computers and it has clearly fewer properties than $C_J$. Perhaps this will change with time when people become adjusted to the realization that the number sets which we use may not contain all the features that one might desire.

## Bibliography


1. Curtis, C. W., "The four and eight square problem and division algebras," Studies In Modern Algebra, vol. 2, Mathematical Association of America, 1963.

2. De Cisco, J., "Introduction to the theory of a quadratic extension of a field K," Universita e Politecnico di Torino Rendiconti del Seminario Matematico, vol. 17, 1957/58, pp. 223-251.

3. _____, "Some theorems concerning commutative rings with unit which admit involutorial automorphisms." Reale Accademia della Scienze di Torino. Atti Classe di Scienze, Fisiche, Mathematiche e Naturali, vol. 92, 1957/58, pp, 225-242.

4. _____, "The geometry of the Z-plane based on a quadratic extension of a field K," Universita e Politecnico di Torino Rendiconti del Seminario Matematico, vol. 18, 1958/59, pp. 91-119.

5. Jacobson, N., "Structure and representations of Jordan algebras," American Mathematical Society Colloquium Publications, vol. 39, American Mathematical Society, 1968.

6. Kleinfeld, E., "A characterization of the Cayley numbers," Studies In Modern Algebra, vol. 2, Mathematical Association of America, vol. 2, 1963.

7. van der Waerden, B. L., Modern Algebra, vol. 1, New York, Frederick Ungar Publishing Co., 1953.

# CALCULATION OF LOWER CONFIDENCE BOUNDS ON SYSTEM RELIABILITY

Joseph V. Michalowicz

Harry Diamond Laboratories, USA ERADCOM
Adelphi, MD 20783

ABSTRACT. A general methodology, based on algorithms developed by the Ad-Hoc Methodology Working Group on Nuclear Weapons Reliability Assessment, is described for evaluating 90% lower confidence bounds on system reliability for configurations of series/parallel circuits. General configurations of non-repeated and repeated components are examined and a method for unpooling data is discussed. A technique is derived for representing "m out of n" decision logic gates. The methodology is applied to an example of the type of a sophisticated weapon fuzing system. Maximum likelihood estimates of reliability and 90% lower confidence bounds are calculated for the system and critical components are identified.

1. INTRODUCTION. For critical and expensive weapon systems, such as nuclear projectiles, highly reliable subsystems are required to produce a high probability of successful system performance. Not only must the reliability of these integral subsystems be very high, but, since often relatively few of such weapon systems will be used to attack an enemy target, there must also be a high degree of confidence that such reliability will be achieved. This report describes a general methodology for calculating maximum-likelihood estimates of reliability as well as 90-percent lower confidence bounds on the system reliability for general systems representable as configurations of series/parallel circuits.

In testing these weapon systems, because of the scarcity and cost of some of the components, the tester must be quite selective in the number and type of subsystems to be included in field tests. An important byproduct of the methodology to be presented is that it evinces those components that are critical, in that they constrain the lower confidence bounds, and those that are not. Therefore, it would be highly cost-effective to schematize the system in the format of this methodology before testing has begun, so that the test director can effectively allocate his test resources to the critical components.

The next section discusses the methodology for calculating confidence bounds on circuit system reliability in a completely general way. It is hoped that this section will serve as a handy reference to the analyst who desires to make confidence-bound determinations for many types of circuit systems. For example, the methodology should be readily applicable to various kinds of sensors, radars, and missile guidance systems. In later sections, the methodology is applied to a system of the type of an actual weapon fuzing system. Based on simulated test data, 90-percent lower confidence bounds on system reliability are calculated and critical components are identified.

2.1 EVALUATION OF CONFIDENCE BOUNDS. By a 90-percent lower confidence bound on system reliability is meant a statistic computed from the test data with the property that there is at least a probability of 0.90 that this statistic is lower than the unknown system reliability. Under the assumption that tests on a component are binomial, that is, the tests are independent with constant failure probability, the 90-percent lower confidence bound, $LCB_{90}$, on component reliability is computed as follows when the test data indicate N tests with F failures:

$$LCB_{90} = B_{90}(N,F) = 1 - p \quad , \tag{1}$$

where p satisfies the binomial relationship

$$\sum_{i=F+1}^{N} \binom{N}{i} p^i (1 - p)^{N-i} = 0.90$$

or, equivalently,

$$\sum_{i=0}^{F} \binom{N}{i} p^i (1 - p)^{N-i} = 0.10 \quad .$$

These formulas assume that N and F are integers; in calculating equivalent components later, there will be a need for evaluating 90-percent lower confidence bounds when N and/or F are not integral. In this case, the following linear interpolation formula is useful:

$$\begin{aligned} B_{90}(N,F) \approx\ & (1 - N_D)[(1 - F_D)B_{90}(N_I,F_I) + F_D B_{90}(N_I,F_I + 1)] \\ & + N_D[(1 - F_D)B_{90}(N_I + 1,F_I) + F_D B_{90}(N_I + 1,F_I + 1)] \end{aligned} \tag{2}$$

where

$N_I = [N]$, the integer part of N,
$N_D = N - [N]$,
$F_I = [F]$, the integer part of F, and
$F_D = F - [F]$.

Tables [1] are available from which binomial confidence bounds can be read for N = 1 to 150. In the calculation of equivalent-system lower confidence bounds later in this paper, we shall frequently encounter very large values of N together with very small values of F. To obtain such lower confidence bounds, the Poisson approximation to the binomial is used when N > 150 and F < 10. As long as F is an integer, regardless of whether or not N is an integer, the Poisson estimate is given by

$$P_{90}(F) \approx \frac{1}{2}\chi^2_{0.90}(2F + 2) \quad , \tag{3}$$

where $\chi^2_{0.90}(2F + 2)$ denotes the 90th percentile of a chi-square distribution with $2F + 2$ degrees of freedom. Tables of the chi-square percentiles can be found in many statistics textbooks (see [2]). When F is not an integer, $P_{90}(F)$ may be calculated by linear interpolation:

$$P_{90}(F) \approx (1 - F_D)P_{90}(F_I) + F_D P_{90}(F_I + 1) \quad . \tag{4}$$

In either case, the 90-percent lower confidence bound is then estimated from the formula

$$LCB_{90} = B_{90}(N,F) = 1 - \frac{P_{90}(F)}{N} \quad . \tag{5}$$

Another useful formula for calculating component lower confidence bounds arises from the observation that, when F = 0 in equation (1), we have

$$[B_{90}(N,0)]^N = (1 - p)^N = 0.10$$

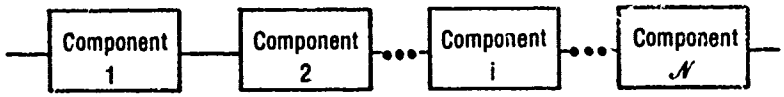
which leads to the exact solution

$$B_{90}(N,0) = (0.10)^{1/N} \quad . \tag{6}$$

It should be clear that all the preceding formulas can be readily extended to the computation of component lower confidence bounds at other than the 90-percent level.

The next several sections describe techniques for calculating lower confidence bounds for general series and parallel configurations of components. These procedures are taken from those recommended by a special Working Group chaired by the Army Materiel Systems Analysis Activity [3].

2.2 CALCULATION OF CONFIDENCE BOUNDS FOR A SERIES SYSTEM OF NONREPEATED COMPONENTS. The simplest case is a system whose configuration consists of a series arrangement of independent components, as exemplified in figure 1.




| | | | | |
|---|---|---|---|---|
| Test Data: | $N_1$ Tests | $N_2$ Tests | $N_i$ Tests | $N_{\mathcal{N}}$ Tests |
| | $S_1$ Successes | $S_2$ Successes | $S_i$ Successes | $S_{\mathcal{N}}$ Successes |
| | $F_1$ Failures | $F_2$ Failures | $F_i$ Failures | $F_{\mathcal{N}}$ Failures |

Figure 1. Series system of nonrepeated components.

None of the $\mathcal{N}$ components in this series are repeated; that is, all are independently functioning components which appear only once and have specific test data in terms of observed successes and failures. Note that $S_i + F_i = N_i$ for all values of i.

The lower confidence bound on the reliability of this series is obtained by reducing the combination to an equivalent component. This is done by means of the Lindstrom-Madden method [4] which calculates the maximum-likelihood estimate of the system reliability, $R_s$, by the formula

$$R_s = \prod_{i=1}^{\mathcal{N}} \frac{S_i}{N_i} \tag{7}$$

and takes the equivalent number of tests, N, for the system to be

$$N = \min_{1 \le i \le \mathcal{N}} N_i \quad . \tag{8}$$

The equivalent number of successes and failures of the system, S and F respectively, are then given by

$$S = NR_s \quad , \tag{9}$$

$$F = N(1 - R_s) \quad . \tag{10}$$

Thus the series combination is now represented by a single equivalent component with S successes and F failures out of N tests. The 90-percent lower confidence bound for the series combination can now be computed by the methods of section 2.1.

For example, consider the three components in series in figure 2.

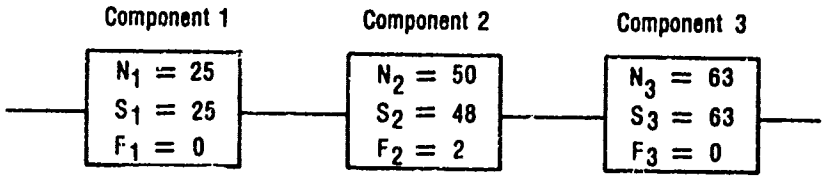



Figure 2. Example of series system.

The computational procedure gives the following:

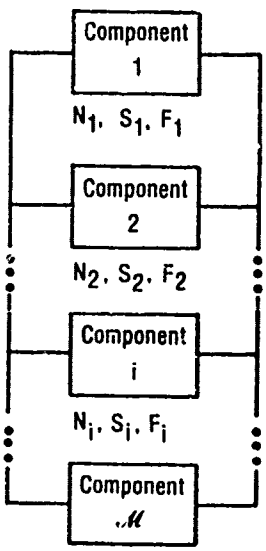
Step 1. $R_s = 1 \times \frac{48}{50} \times 1 = 0.96$

Step 2. $N = \min(25, 50, 63) = 25$

Step 3. $S = 0.96(25) = 24$
$F = 0.04(25) = 1$

Step 4. $LCB_{90} = B_{90}(25, 1) = 0.853$ (from eq (1) and table lookup)

2.3 CALCULATION OF CONFIDENCE BOUNDS FOR A PARALLEL SYSTEM OF NONREPEATED COMPONENTS. For a system configured as in figure 3 with independent, nonrepeated components in parallel, an equivalent single component is again derived. The equivalent number of tests, N, is computed from the equation




$N_i$ = Number of Tests
$S_i$ = Number of Successes
$F_i$ = Number of Failures

Figure 3. Parallel system of nonrepeated components.

$$N = \frac{1 - Q'}{Q' - Q} , \tag{11}$$

where

$$Q = \prod_{i=1}^{\mathcal{M}} \frac{F_i}{N_i} ,$$

$$Q' = \prod_{i=1}^{\mathcal{M}} \frac{F_i + 1}{N_i + 1} ,$$

and the maximum likelihood estimate of the system reliability is then given by:

$$R_s = 1 - Q . \tag{12}$$

The equivalent numbers of successes and failures are then derived:

$$S = NR_s , \tag{13}$$

$$F = NQ . \tag{14}$$

The 90-percent lower confidence bound for the system reliability can now be computed as that for the equivalent single component with F failures out of N tests.

An example of a parallel system is given in figure 4. The computational steps proceed as follows:

Step 1. $Q = 0 \times \frac{2}{20} \times \frac{1}{30} = 0$

Step 2. $Q' = \frac{1}{11} \times \frac{3}{21} \times \frac{2}{31} = 0.000838$

Step 3. $N = \frac{1 - Q'}{Q' - 0} = 1192.5$

Step 4. $R_s = 1$

Step 5. $S = 1192.5$

$F = 0$

Step 6. $LCB_{90} = B_{90}(1192.5, 0) = (0.10)^{1/1192.5} = 0.9981$ (from eq (6))

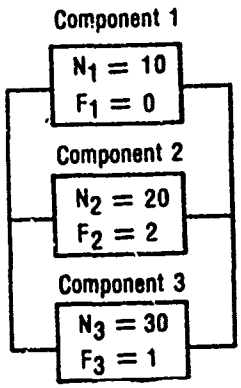



Figure 4. Example of parallel system.

2.4 CONSTRUCTION OF AN EQUIVALENT COMPONENT WITH SPECIFIED RELIABILITY AND CONFIDENCE BOUND. In reducing a complex combination of components to an equivalent single component, a sequence of substitutions may be required. It may occur, in the techniques to be developed in subsequent sections, that some reductions will calculate the maximum-likelihood estimate of reliability as well as the lower confidence bound for a subsystem without specifying the equivalent test data. Therefore, it will be useful in the sequel to have a technique for constructing the equivalent test data for a subsystem when given only the maximum-likelihood estimate of reliability, R, and the 90-percent lower confidence bound, $B_{90}$.

The technique for solving for the equivalent number of tests, N, and failures, F, given R and $B_{90}$, is actually just the solution of the following two equations in two unknowns:

$$R = 1 - \frac{F}{N} \quad ,$$

$$B_{90} = B_{90}(N,F) \quad .$$

However, the second of these equations cannot be solved explicitly, so an iterative approach to solution is used.

The iteration begins with an initial estimate of N, denoted by $N_1$, calculated from the formula

$$N_1 = \frac{\ln 0.10}{\ln B_{90}} \quad . \tag{15}$$

If the reliability estimate R is equal to 1, set $N = N_1$ and $F = 0$. If not, an estimate of F, denoted by $F_1$, is obtained from

$$F_1 = (1 - R)N_1 \quad , \tag{16}$$

and the confidence bound $B_{90}(N_1, F_1)$ is determined by the techniques in section 2.1. An adjustment factor given by

$$t = \frac{\ln B_{90}(N_1, F_1)}{\ln B_{90}} \tag{17}$$

is used to obtain the next estimate of N, denoted by $N_2$:

$$N_2 = tN_1 \quad . \tag{18}$$

If the adjustment factor is near enough to 1 (i.e., $|t - 1| < 0.01$), then use $N = N_2$ and $F = (1 - R)N_2$ as the equivalent test data. If not, $N_2$ is taken as the estimate of N and the above process (eq (16) through (18)) is repeated until the adjustment factor converges close enough to 1, resulting in the equivalent values of N and F. This procedure is illustrated in the next section.

2.5 CALCULATION OF CONFIDENCE BOUNDS FOR A SYSTEM CONSISTING OF ONLY A SINGLE COMPONENT REPEATED IN ANY CONFIGURATION. This section describes the methodology to be used for calculating confidence bounds for a system or subsystem which is a combination of series and/or parallel circuits composed solely of repetitions of the same component. More precisely, the components, although separate physical devices, are the same in the sense that they are of the same generic type and are described by the same test data.

Suppose the system to be analyzed is a series/parallel configuration consisting of repetitions of a component, C, characterized by test data indicating $F_C$ failures in $N_C$ tests. The maximum-likelihood estimate for the reliability of the component, C, is given by

$$R_C = 1 - \frac{F_C}{N_C} \quad . \tag{19}$$

Analysis of the system into its series and parallel branches of C components gives rise to a reliability estimate for the system which is a function of $R_c$:

$$R_s = f(R_c) \quad . \tag{20}$$

For example, if the configuration consisted of n components C in series, $f(R_c)$ would be $R_c^n$, whereas if the configuration were n components C in parallel, $f(R_c)$ would be $1 - (1 - R_c)^n$.

To calculate confidence bounds for the general series/parallel configuration of C components, the methodology begins by evaluating the 90-percent lower confidence bound for C:

$$LCB_c = B_{90}(N_c, F_c) \quad .$$

The 90-percent lower confidence bound for the system, $LCB_s$, is then calculated by means of the function in equation (20):

$$LCB_s = f(LCB_c) \quad . \tag{21}$$

Therefore, we have obtained the maximum-likelihood estimate of reliability (eq (20)) and the 90-percent lower confidence bound (eq (21)) for the system. Equivalent test data for the system (that is, $N_s$ and $F_s$) can now be calculated by the method of section 2.4.

For the special case where the configuration of the system is just a series arrangement of n repeats of C and where $F_c$ is small compared to $N_c$ (that is, $F_c \leq N_c/10$), two simple but accurate approximations for $N_s$ and $F_s$ are available. Both of these approximations are conservative in that they tend to underestimate $N_s$:

Approximation 1:
$$N_s \approx \frac{N_c}{n} \tag{22}$$

$$F_s = (1 - R_s)N_s \tag{23}$$

Approximation 2:
$$N_s \approx \frac{F_c}{1 - R_s} \tag{24}$$

$$F_s = (1 - R_s)N_s = F_c \tag{25}$$

Note that the second approximation cannot be used when $R_c = 1$ (or, equivalently, $F_c = 0$), but in this case the first approximation yields exactly the same values as the general method in section 2.4, since

$$N_s = N_1 = \frac{\ln 0.10}{\ln LCB_s} \qquad \text{from equation (15)}$$

$$= \frac{\ln 0.10}{\ln LCB_c^n}$$

$$= \frac{\ln 0.10}{n \ln LCB_c}$$

$$= \frac{\ln 0.10}{n \ln (0.10)^{1/N_c}} \qquad \text{from equation (6)}$$

$$= \frac{N_c}{n} \ .$$

These approximations are often useful in the reduction of a complex system with series subsystems to an equivalent system.

As an example, consider the series/parallel configuration in figure 5, where $N_c = 15$ and $F_c = 1$. Computations proceed as follows:

Step 1. $R_c = 1 - \frac{1}{15} = 0.93333$

Step 2. $R_s = [1 - (1 - R_c)^2][1 - (1 - R_c)^3] = 0.99526$

Step 3. $LCB_c = B_{90}(15, 1) = 0.7643$

Step 4. $LCB_s = [1 - (1 - LCB_c)^2][1 - (1 - LCB_c)^3] = 0.93206$

Step 5. The iterative method of section 2.4 with $R = 0.99526$ and $B_{90} = 0.93206$ then gives the following table, where the $B_{90}(N_1, F_1)$ values are obtained by the interpolation formula (2).

| Iteration | $N_1$ | $F_1$ | $B_{90}(N_1, F_1)$ | t | $N_2$ |
|---|---|---|---|---|---|
| 1 | 32.73 | 0.155 | 0.9247 | 1.112 | 36.42 |
| 2 | 36.42 | 0.173 | 0.9314 | 1.011 | 36.81 |
| 3 | 36.81 | 0.174 | 0.9320 | 0.9997 | 36.80 |

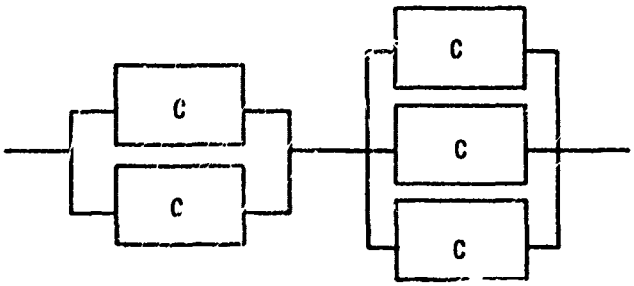



Figure 5. Series/parallel configuration.

Consequently, the equivalent test data for the system are given by

$$N_s = 36.80$$

$$F_s = (1 - R_s)N_s = 0.174 \quad .$$

2.6 CALCULATION OF CONFIDENCE BOUNDS FOR GENERAL CONFIGURATIONS--METHOD FOR UNPOOLING DATA. The techniques discussed so far permit the calculation of lower confidence bounds on system reliability for series/parallel systems of independent, nonrepeated components, as well as for systems which contain repeated component types, as long as each repeated component appears in only one subsystem. In order to handle configurations in which repeated components are distributed throughout several subsystems in combination with other repeated or nonrepeated components, a method will be described for unpooling the data for repeated components. This method divides the component test data into groups corresponding to the various subsystems in which the component appears, and then treats the component as distinct and independent within each subsystem. It has been found that such unpooling schemes provide somewhat conservative lower confidence bounds on reliability.

The basic idea behind the unpooling method is as follows. Suppose C is a component, with test data indicating $F_c$ failures in $N_c$ tests, which occurs in n subsystems, where the subsystems are chosen to each contain as many appearances of C as possible and still be analyzable by the techniques of sections 2.2 through 2.5. Thus each subsystem either contains just one appearance of C or, if it contains two or more appearances, that portion of the subsystem can be reduced to a configuration composed of repetitions of a single equivalent component. The component C will be relabeled as $C_1$, $C_2$, ..., $C_n$, respectively, for each of the n subsystems in which it appears. The test data for C is then allocated over the n subsystems in such a way as to keep the maximum-likelihood estimate of reliability for each $C_i$, i = 1, 2, ..., n, equal to that for C. That is, the constraints on the unpooling are

$$\sum_{i=1}^{n} F_{c_i} = F_c \quad .$$

$$\sum_{i=1}^{n} N_{c_i} = N_c \quad ,$$

$$\frac{F_{c_i}}{N_{c_i}} = \frac{F_c}{N_c} \quad , \qquad \text{for } i = 1, 2, \ldots, n \quad .$$

There are many ways of unpooling which satisfy these constraints. The method used here unpools according to the following scheme:

(1) Unpool equally in a series direction.
(2) Then unpool equally in a parallel direction.
(3) Then unpool equally in a series direction.
etc

This sequence is best illustrated by an example, as shown in figure 6. In this system the component C appears in four subsystems and has been relabeled accordingly. The first step of the unpooling would allocate $N_c/2$ and $F_c/2$ to $C_1$ and the other $N_c/2$ and $F_c/2$ to the parallel combination. Since there are two branches in parallel, the second step of the unpooling would divide in half the equivalent test data for the parallel combination, thus allocating $N_c/4$ and $F_c/4$ to $C_4$ and the other $N_c/4$ and $F_c/4$ to the series combination containing $C_2$ and $C_3$. In turn the third step of the unpooling allocates $N_c/8$ and $F_c/8$ to each of $C_2$ and $C_3$. In summary, the unpooled test data for each appearance of C would be as follows.

| Component | Test data | |
|---|---|---|
| | N | F |
| $C_1$ | $N_c/2$ | $F_c/2$ |
| $C_2$ | $N_c/8$ | $F_c/8$ |
| $C_3$ | $N_c/8$ | $F_c/8$ |
| $C_4$ | $N_c/4$ | $F_c/4$ |
| Total: | $N_c$ | $F_c$ |

After unpooling, each of the $C_i$'s is treated as a separate, independent component and the techniques in sections 2.2 through 2.5 are applied, as appropriate, to each of the subsystems.

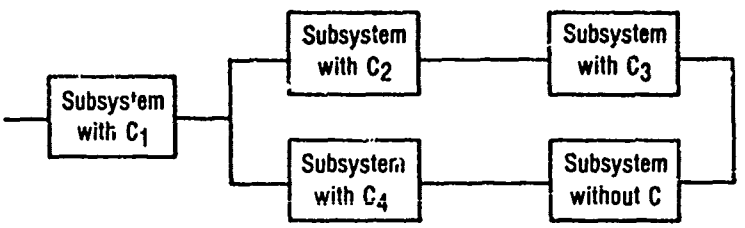



Figure 6. Example of unpooling scheme

2.7 REPRESENTATION OF "TWO OUT OF THREE" DECISION GATE. Synthesis of the techniques presented so far enables one to calculate lower confidence bounds on reliability for any series/parallel configuration. However, many of the sophisticated circuits of today contain other configurations, such as decision gates, to gain greater reliability and efficiency. A straightforward procedure will be formulated to handle decision gates by approximate equivalent combinations of series and parallel circuits.
The methodology will be illustrated for a "two out of three" decision gate; the extension to general "k out of $n$" decision logic gates should be clear.

First, observe that for a series combination of components $C_1$, $C_2$, ..., $C_K$, with component failure probabilities $Q_{C_1}$, $Q_{C_2}$, ..., $Q_{C_K}$, the failure probability of the combination, Q, is given by

$$Q = 1 - \left(1 - Q_{C_1}\right)\left(1 - Q_{C_2}\right) \cdots \left(1 - Q_{C_K}\right)$$
$$= Q_{C_1} + Q_{C_2} + \cdots + Q_{C_K} + \text{second and higher order terms} .$$

Mission reliability equations for modern weapon systems typically neglect the second and higher order terms and simply add together failure probabilities of components in series. On the other hand, if $C_1$, $C_2$, ..., $C_K$ were in parallel, the failure probability for the system would be, simply,

$$Q = Q_{C_1} Q_{C_2} \cdots Q_{C_K} .$$

For a decision gate configuration which requires success in two (or more) of the three branches (with each branch consisting of the same component C) for a YES vote, the probability of failure, Q, of the gate (i.e., a NO vote) is given by

$$Q = \text{probability that 2 or 3 branches fail}$$
$$= 3Q_C^2 + Q_C^3 ,$$

where $Q_C$ is the failure probability of the component C. In terms of failure probability, the decision gate is, therefore, approximately equivalent to the series/parallel combination shown in figure 7, which has a failure probability given by

$$Q = Q_C^2 + Q_C^2 + Q_C^2 + Q_C^3 + \text{fourth and higher order terms} .$$

Since the terms omitted by introducing this approximation are two orders less than those already typically neglected in the mission reliability equation, this series/parallel combination should afford a sufficiently accurate representation of the decision gate.

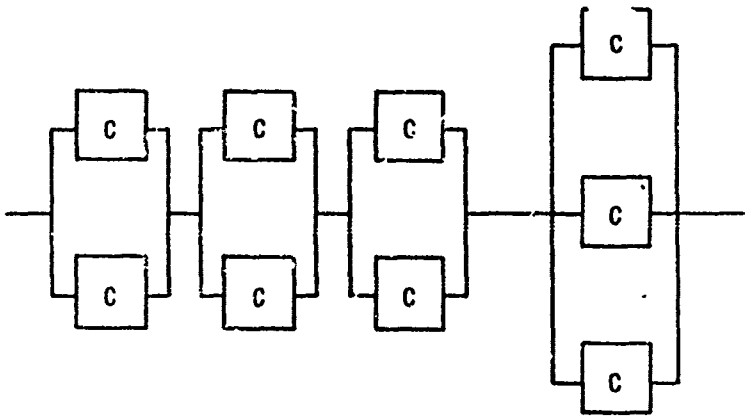



Figure 7. Decision gate approximate equivalent combination.

3. CASE STUDY - APPLICATION OF THE METHODOLOGY. The methodology described in the previous sections will now be applied to an example based on the actual fuzing system of a battlefield weapon. The system schematic, shown in figure 8, is at the same level of sophistication as the fuzing system. However, for purposes of keeping this report unclassified, a few modifications have been made to the actual schematic and simulated component test data is used. Note the "two out of three" decision gate equivalent in the upper right hand part of the system schematic in figure 8. The simulated component test data is displayed in table 1. For some of the components only a reliability value, R, is available, presumably based on a large number of tests by the manufacturer; such components are denoted by an asterisk in figure 8.

The lower-confidence-bound computation for this system will proceed through two reductions of the system, unpooling into subsystems and calculation of equivalent components, and then the calculation of the system lower confidence bound itself. In the process, components critical to the confidence-bound assessment will be evinced and pertinent observations made.

In the first reduction many of the series combinations which are repeated in a particular type of configuration throughout the system schematic are simplified. The computations are sketched in appendix A. Note that those components which have reliability estimates only are treated as having essentially an infinite number of trials; thus they do not affect the calculation of the equivalent component N (number of trials) but only the calculation of the equivalent component R (reliability). After the first reduction, the system schematic takes the form shown in figure 9.

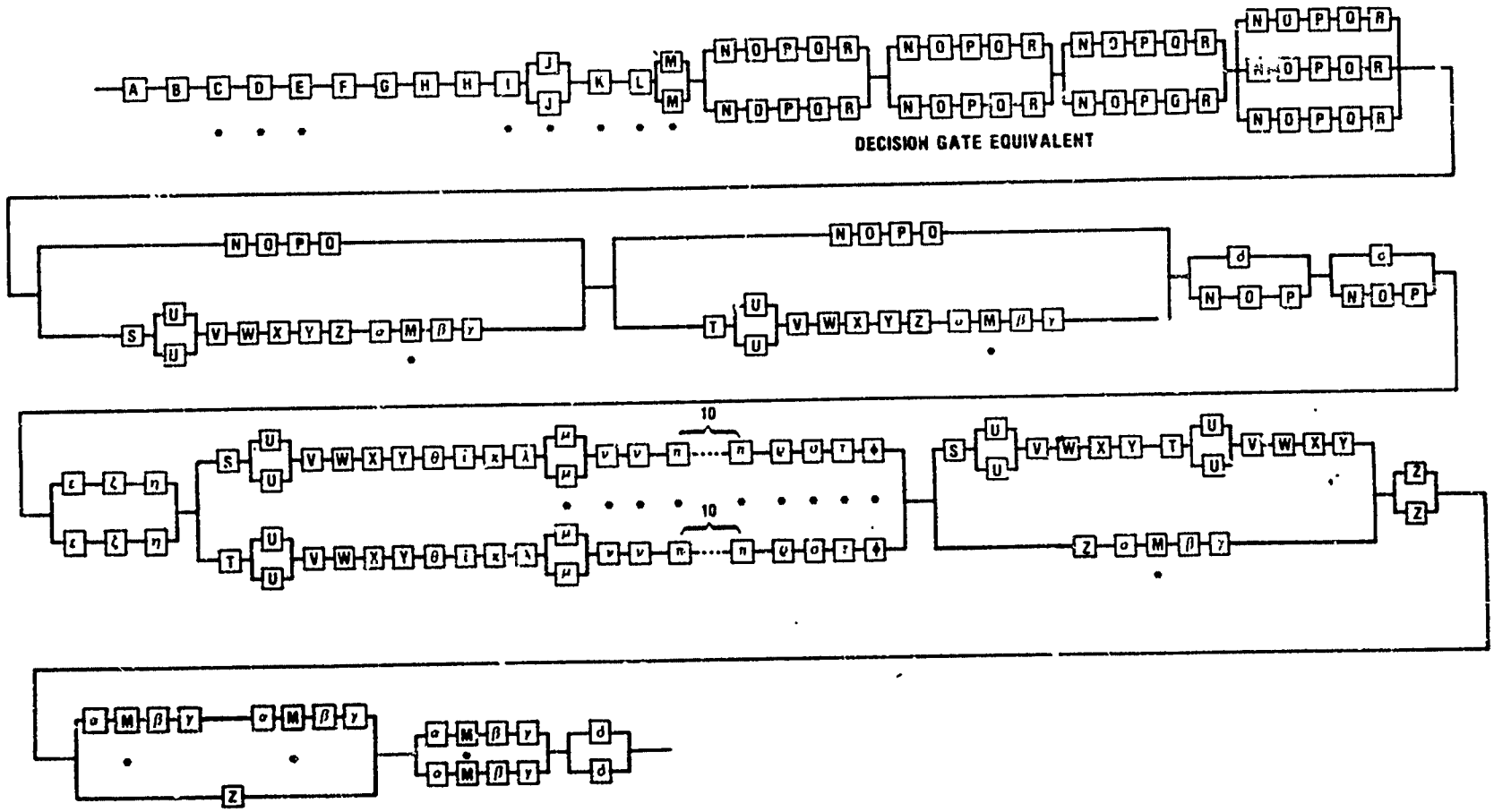



**FIGURE 8. SYSTEM SCHEMATIC**

TABLE 1. SIMULATED COMPONENT TEST DATA

| Component | Number of Trials | Number of Failures |
|---|---|---|
| A | 5000 | 0 |
| B | 201 | 0 |
| C | R = .995 | |
| D | R = .994 | |
| E | R = .9997 | |
| F | 192 | 0 |
| G | 192 | 0 |
| H | 401 | 0 |
| I | R = .9978 | |
| J | R = .97 | |
| K | R = .999 | |
| L | R = .999 | |
| M | R = .999 | |
| N | 573 | 0 |
| O | 573 | 0 |
| P | 573 | 0 |
| Q | 570 | 1 |
| R | 572 | 0 |
| S | 250 | 15 |
| T | 328 | 1 |
| U | 384[a] | 0 |
| V | 384 | 0 |
| W | 383 | 0 |
| X | 383 | 0 |
| Y | 384 | 1 |
| Z | 92 | 0 |
| α | 401 | 0 |
| β | 1260 | 0 |
| γ | 1260 | 0 |
| δ | 384 | 0 |
| ε | 382 | 0 |
| ζ | 382 | 3 |
| η | 381 | 0 |
| θ | 381 | 0 |
| ι | 399 | 0 |
| κ | 371 | 0 |
| λ | 375 | 0 |
| μ | R = .992 | |
| ν | R = .998 | |
| ξ | R = .9998 | |
| ρ | R = .9999 | |
| σ | R = .9995 | |
| τ | R = .9995 | |
| φ | R = .9982 | |

[a] For a parallel pair of U components

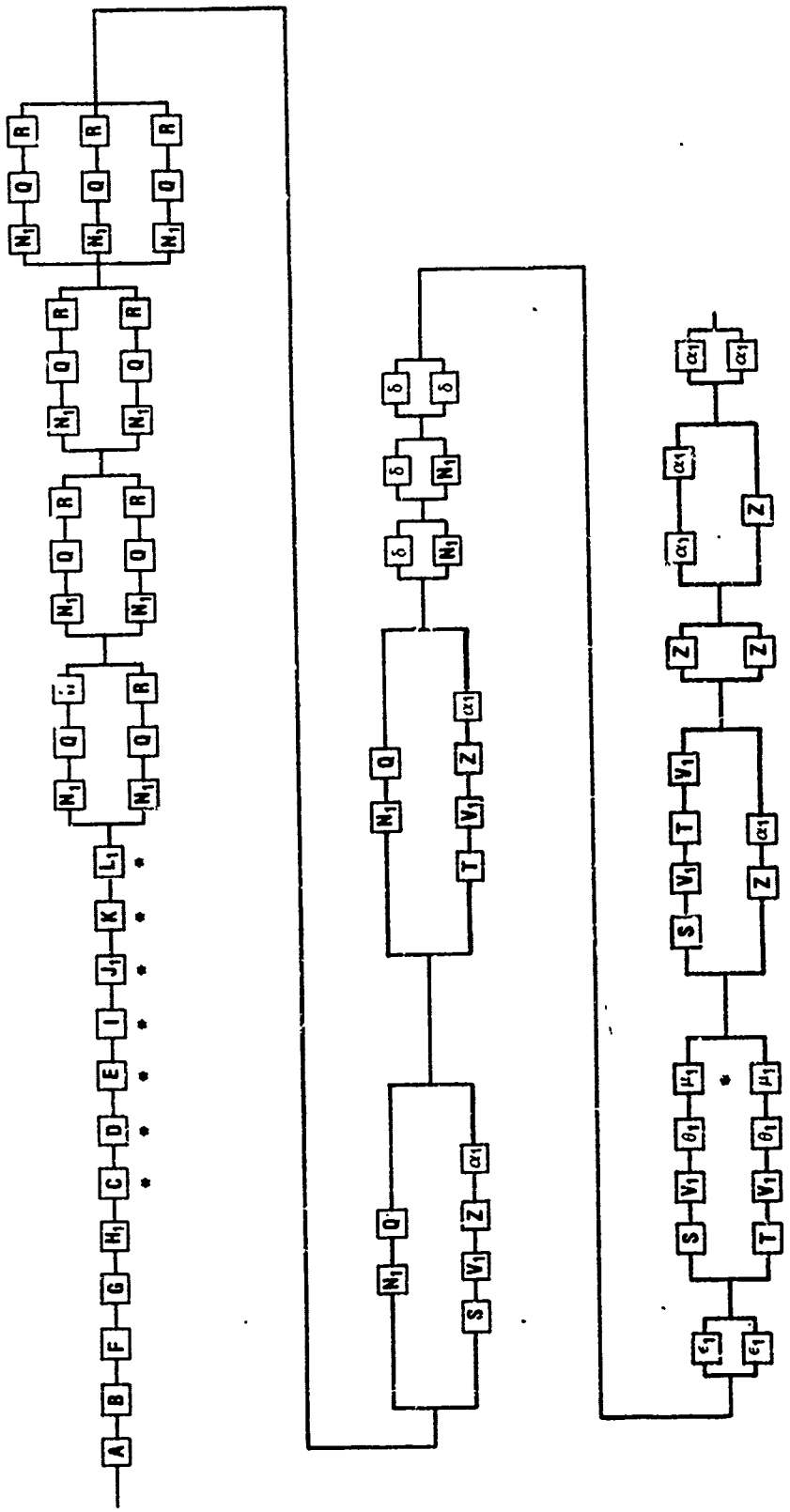


FIGURE 9. SYSTEM SCHEMATIC — AFTER FIRST REDUCTION

The primary function of the second reduction is to consolidate the long series of components at the beginning of the schematic in figure 9. After these substitutions are made, the reduced schematic assumes the tractable form shown in figure 10. Details of the reduction procedure are given in appendix B.

The reduced schematic, divided into subsystems as shown in figure 10, can now be treated by applying the methodology developed previously to each of the numbered subsystems and then determining the equivalent N and R for the overall series configuration of subsystems. However, the data must first be unpooled for components which appear in more than one subsystem. The components which appear in the reduced schematic, before unpooling, are listed in table 2 along with their equivalent test data. The equivalent test data after unpooling are shown in table 3. Note that those components which appear in more than one subsystem have had an extra subscript appended to indicate those repetitions. (For example, $V_{13}$ refers to the third distinct appearance of $V_1$, in the top branch of subsystem 6.)

The equivalent number of trials, N, and the maximum-likelihood reliability estimate, R, are computed, subsystem by subsystem, in appendix C and tabulated in table 4. Since the overall system configuration is now represented as a series combination of these subsystems, the maximum-likelihood estimate of the overall system's reliability is just the product of the subsystem reliabilities (R = 0.9824), and the equivalent number of trials is the minimum of those for the subsystems (N = 165). This minimum number (indicated by an asterisk in table 4) corresponds to the critical subsystem--that which delimits the equivalent number of trials. Note how only a few subsystems, and thus only a few components, may determine the calculation of the confidence bound. Examination of the critical subsystem 8 identifies the critical component of the overall system (i.e., that component for which additional test data could increase the equivalent number of trials for the overall system and hence improve the resulting lower confidence bound ), to be the Z component.

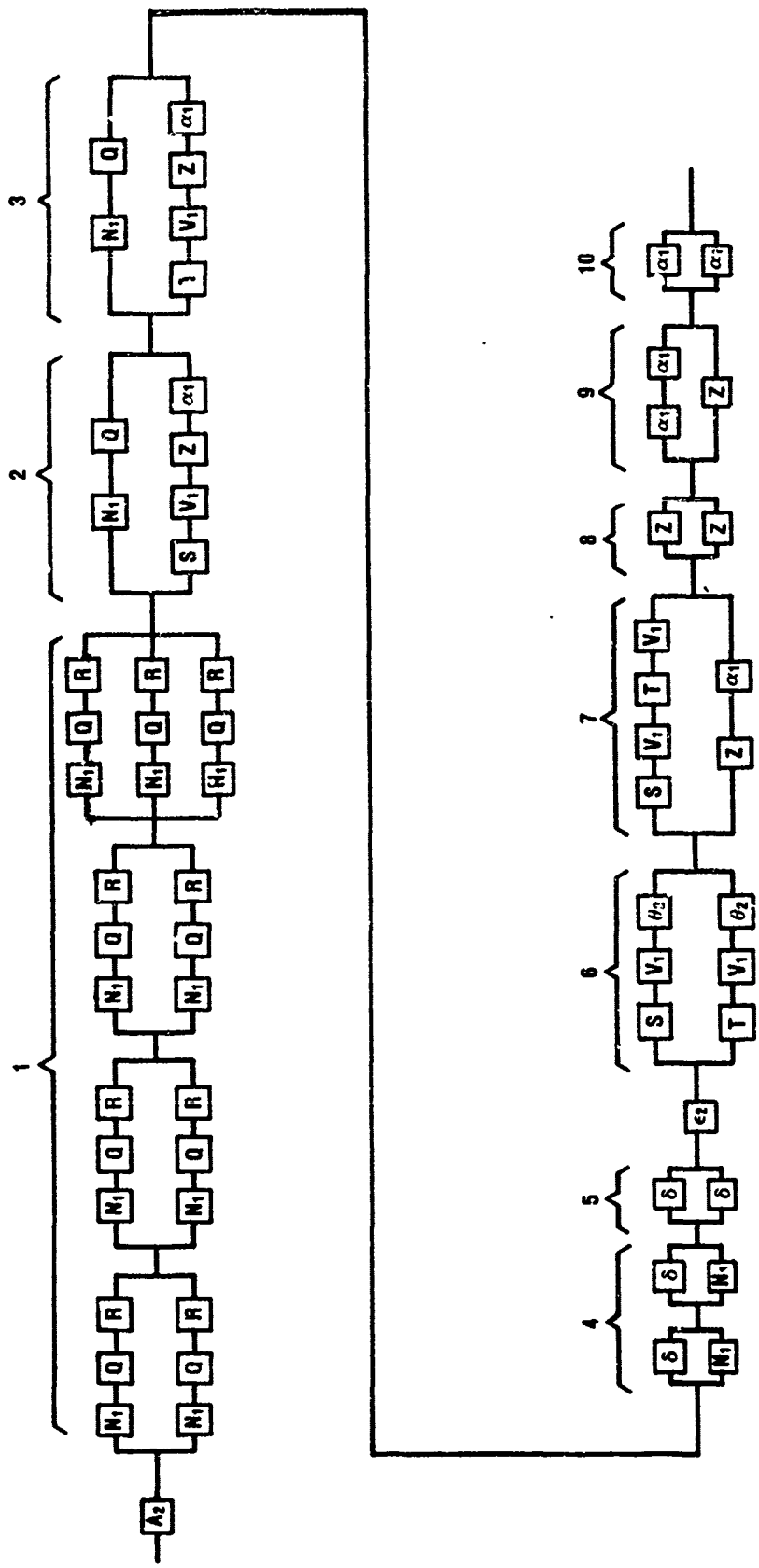


FIGURE 10. REDUCED SCHEMATIC

TABLE 2. EQUIVALENT TEST DATA — BEFORE UNPOOLING

| Component | Number of Trials | Number of Failures |
|---|---|---|
| $A_2$ | 192 | 3.130 |
| $N_1$ | 573 | 0 |
| Q | 570 | 1 |
| R | 572 | 0 |
| S | 250 | 15 |
| T | 328 | 1 |
| $V_1$ | 383 | 0.997 |
| Z | 92 | 0 |
| $\alpha_1$ | 401 | 0.401 |
| $\delta$ | 384 | 0 |
| $\epsilon_2$ | 10,961 | 0.680 |
| $\theta_2$ | 371 | 3.313 |

TABLE 3. EQUIVALENT TEST DATA — AFTER UNPOOLING

| Component | Number of Trials | Number of Failures |
|---|---|---|
| $A_2$ | 192 | 3.130 |
| $N_{11}$ | 143.25 | 0 |
| $N_{12}$ | 143.25 | 0 |
| $N_{13}$ | 143.25 | 0 |
| $N_{14}$ | 143.25 | 0 |
| $Q_1$ | 190 | 0.3333 |
| $Q_2$ | 190 | 0.3333 |
| $Q_3$ | 190 | 0.3333 |
| R | 572 | 0 |
| $S_1$ | 83.33 | 5 |
| $S_2$ | 83.33 | 5 |
| $S_3$ | 83.33 | 5 |
| $T_1$ | 109.33 | 0.3333 |
| $T_2$ | 109.33 | 0.3333 |
| $T_3$ | 109.33 | 0.3333 |
| $V_{11}$ | 95.75 | 0.24925 |
| $V_{12}$ | 95.75 | 0.24925 |
| $V_{13}$ | 47.875 | 0.124625 |
| $V_{14}$ | 47.875 | 0.124625 |
| $V_{15}$ | 95.75 | 0.24925 |
| $Z_1$ | 18.4 | 0 |
| $Z_2$ | 18.4 | 0 |
| $Z_3$ | 18.4 | 0 |
| $Z_4$ | 18.4 | 0 |
| $Z_5$ | 18.4 | 0 |

TABLE 3. EQUIVALENT TEST DATA -- AFTER UNPOOLING (CONT'D)

| Component | Number of Trials | Number of Failures |
|---|---|---|
| $\alpha_{11}$ | 80.2 | 0.0802 |
| $\alpha_{12}$ | 80.2 | 0.0802 |
| $\alpha_{13}$ | 80.2 | 0.0802 |
| $\alpha_{14}$ | 80.2 | 0.0802 |
| $\alpha_{15}$ | 80.2 | 0.0802 |
| $\delta_1$ | 192 | 0 |
| $\delta_2$ | 192 | 0 |
| $\epsilon_2$ | 10,961 | 0.680 |
| $\theta_{21}$ | 185.5 | 1.6565 |
| $\theta_{22}$ | 185.5 | 1.6565 |

TABLE 4. SUMMARY OF SUBSYSTEM DATA

| Subsystem | R | Equivalent N |
|---|---|---|
| $A_2$ | 0.98370 | 192 |
| $\epsilon_2$ | 0.999938 | 10,961 |
| 1 | 0.999991 | 2,158 |
| 2 | 0.9998891 | 1,166 |
| 3 | 0.9999884 | 2,040 |
| 4 | 1 | 13,919 |
| 5 | 1 | 15,989 |
| 6 | 0.9989715 | 478 |
| 7 | 0.9999323 | 222 |
| 8 | 1 | 165 * |
| 9 | 1 | 737 |
| 10 | 0.999999 | 2,586 |
| System | 0.9824 | 165 |

R = Maximum-likelihood estimate of reliability

N = Number of trials

The data are now in place to calculate the 90-percent lower confidence bound on the reliability of this example system. We have:

$$R = 0.9824$$

$$N = 165$$

$$F = (1-R)N = 2.904$$

The interpolation formula (2) and the Poisson estimate (3) give the 90-percent lower confidence bound:

$$\begin{aligned} LCB &= 0.096\ B_{90}(165,\ 2) + 0.904\ B_{90}(165,\ 3) \\ &= 0.096\left(1 - \frac{\frac{1}{2}\chi^2_{0.90}(6)}{165}\right) + 0.904\left(1 - \frac{\frac{1}{2}\chi^2_{0.90}(8)}{165}\right) \\ &= 0.096(0.96788) + 0.904(0.95939) \\ &= 0.9602 \end{aligned}$$

Note that R is a point estimate of the reliability of the system, whereas the lower confidence bound is a bound on the unknown actual system reliability, not on the point estimate.

In summary, the general methodology described in this report has been utilized to estimate the system reliability of a practical weapon system design, to obtain a 90-percent lower confidence bound on the system reliability, and to determine those system components which are prime candidates for further design tests.

## REFERENCES


1. Engineering Design Handbook, Tables of Cumulative Binomial Probabilities, AMCP706-109, HQ U.S. Army Materiel Command (June 1972).

2. G. Hahn and S. Shapiro, Statistical Models in Engineering, J. Wiley & Sons, New York (1968).

3. Handbook for the Calculation of Lower Statistical Confidence Bounds on System Reliability Assessment, Ad-Hoc Methodology Working Group on Nuclear Weapons Reliability Assessment (February 1980).

4. D. K. Lloyd and M. Lipow, Reliability: Management, Methods, and Mathematics, Prentice Hall, New Jersey (1962).

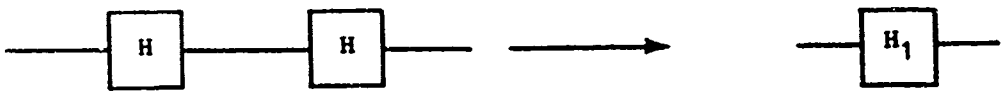
APPENDIX A

## FIRST REDUCTION

In each replacement of a subsystem, shown in figure 8 in the body of the report, by an equivalent component, both the original subsystem and the new equivalent component will be depicted. The methodology used for the reduction will be referred to by the appropriate section in the body of the report. The symbols N, F, R, and Q will be used throughout to denote number of tests, number of failures, maximum likelihood reliability estimate, and failure probability, respectively.

Original subsystem

Equivalent component




$N_H = 401$
$F_H = 0$
$R_H = 1$

(method of sect. 2.5, eq (22))

$N_{H_1} = 200.5$
$F_{H_1} = 0$

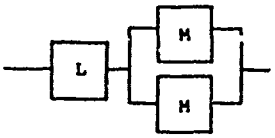



$Q_J = 0.03$

$Q_{J_1} = (0.03)^2 = 0.0009$

Original subsystem

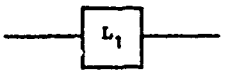
Equivalent component




$Q_L = 0.001$

$Q_M = 0.001$

$R_{L_1} = (0.999999)\ (0.999)$



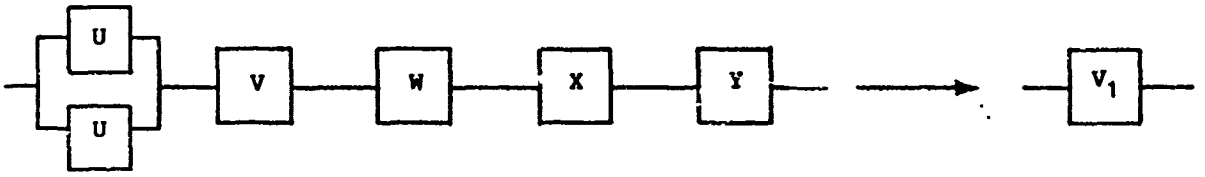

$Q_{L_1} = 0.001001$




$N_N = 573$ $N_O = 573$ $N_P = 573$ (method of sect. 2.2) $N_{N_1} = 573$

$F_N = 0$ $F_O = 0$ $F_P = 0$ $F_{N_1} = 0$

$R_N = 1$ $R_O = 1$ $R_P = 1$ $R_{N_1} = 1$




$N_{U2} = 384$ $N_V = 384$ $N_W = 383$ $N_X = 383$ $N_Y = 384$ (method of sect. 2.2) $N_{V_1} = 383$

$F_{U2} = 0$ $F_V = 0$ $F_W = 0$ $F_X = 0$ $F_Y = 1$ $F_{V_1} = 0.997$

$R_{V_1} = \frac{383}{384}$

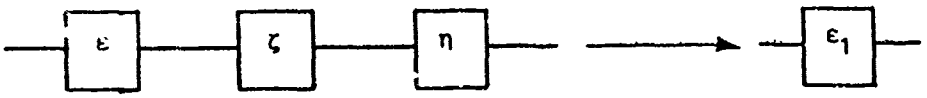
Original subsystem

Equivalent component



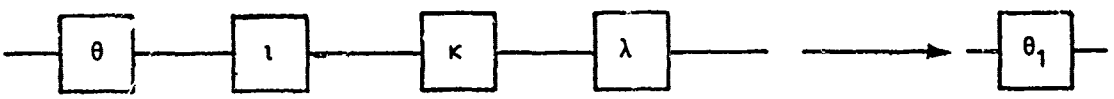

$N_\alpha = 401$ $R_M = 0.999$ $N_\beta = 1260$ $N_\gamma = 1260$ (method of sect. 2.2) $N_{\alpha_1} = 401$

$F_\alpha = 0$ $F_\beta = 0$ $F_\gamma = 0$ $R_{\alpha_1} = 0.999$ $F_{\alpha_1} = 0.401$




$N_\varepsilon = 382$ $N_\zeta = 382$ $N_\eta = 381$ (method of sect. 2.2) $N_{\varepsilon_1} = 381$

$F_\varepsilon = 0$ $F_\zeta = 3$ $F_\eta = 0$ $R_{\varepsilon_1} = \frac{379}{382}$ $F_{\varepsilon_1} = 2.992$



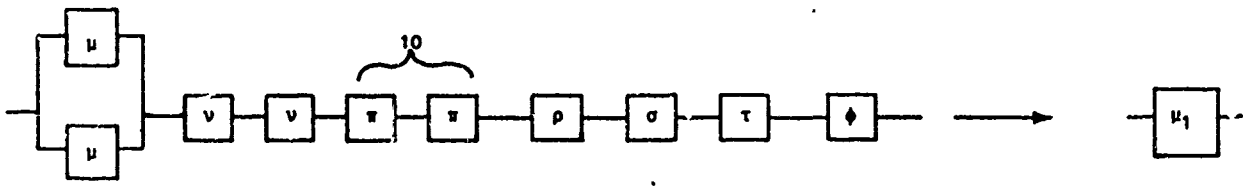

$N_\theta = 381$ $N_\iota = 399$ $N_\kappa = 371$ $N_\lambda = 375$ (method of sect. 2.2) $N_{\theta_1} = 371$

$F_\theta = 0$ $F_\iota = 0$ $F_\kappa = 0$ $F_\lambda = 0$ $R_{\theta_1} = 1$ $F_{\theta_1} = 0$

*This warhead component is repeated in other subsystems but, since it is being treated as having essentialy an infinite number of trials, it cannot affect calculation of the equivalent N and so it can be treated as independent, affecting only the calculation of R.*

Original subsystem

Equivalent component




$Q_\mu$ = 0.008 $Q_\nu$ = 0.002 $Q_\pi$ = 0.0002 $Q_\rho$ = 0.0001 $Q_\sigma$ = 0.0005 $Q_\tau$ = 0.0005 $Q_\phi$ = 0.0018

$R_{\mu_1} = (1 - 0.008)^2 \times 0.998^2 \times 0.9998^{10} \times 0.9999 \times 0.9995 \times 0.9995 \times 0.9982 = 0.99107$

$Q_{\mu_1}$ = 0.00893

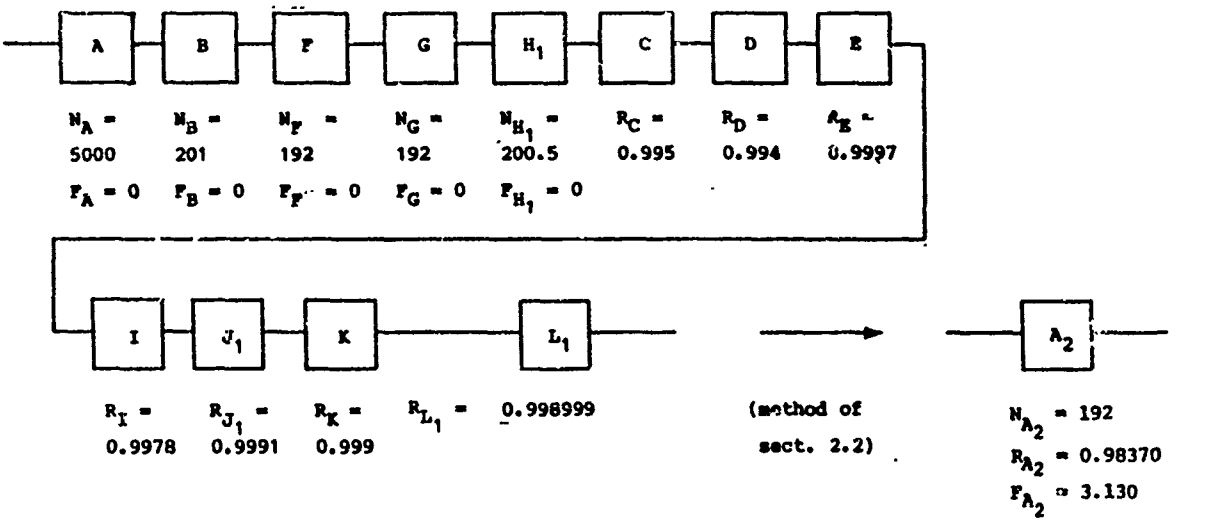
APPENDIX B

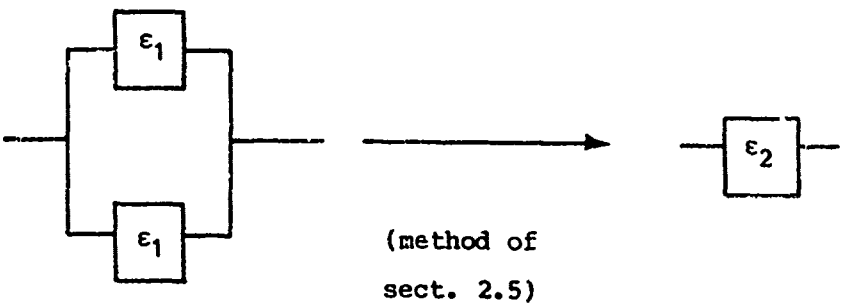
## SECOND REDUCTION

In each replacement of a subsystem, shown in figure 9 in the body of the report, by an equivalent component, both the original subsystem and the new equivalent component will be depicted. The methodology used for the reduction will be referred to by the appropriate section in the body of the report. The symbols N, F, R, and LCB will be used throughout to denote number of tests, number of failures, maximum likelihood reliability estimate, and 90-percent lower confidence bound, respectively.

Original subsystem

Equivalent component




$N_{\varepsilon_1}$ = 381
$F_{\varepsilon_1}$ = 2.992
$R_{\varepsilon_1}$ = 0.99215

$N_{\varepsilon_2}$ = 10,961
$F_{\varepsilon_2}$ = 0.680

$$LCB_{\varepsilon_1} = 0.98244 \text{ by the Poisson estimate (eq (5))}$$

$$R_{\varepsilon_2} = 1 - \left(1 - R_{\varepsilon_1}\right)^2 = 0.999938$$

$$LCB_{\varepsilon_2} = 1 - \left(1 - LCB_{\varepsilon_1}\right)^2 = 0.999692$$

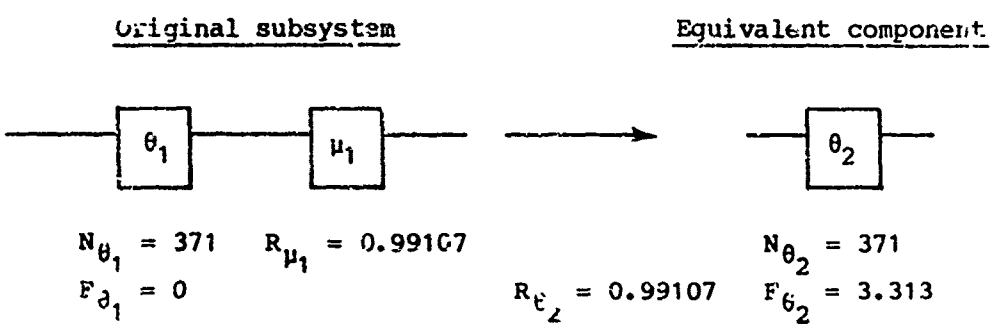
The iterative method of section 2.4 is used to find $N_{\varepsilon_2}$ and $F_{\varepsilon_2}$, producing the results in the following table.

| Iteration | $N_1$ | $F_1$ | $B_{90}\left(N_1, F_1\right)$ | t | $N_2$ |
|---|---|---|---|---|---|
| 1 | 7,475 | 0.4635 | 0.999593 | 1.32 | 9,878 |
| 2 | 9,878 | 0.6124 | 0.999668 | 1.078 | 10,648 |
| 3 | 10,648 | 0.6602 | 0.999685 | 1.023 | 10,890 |
| 4 | 10,890 | 0.6752 | 0.999690 | 1.006 | 10,961 |

From these results,

$$N_{\varepsilon_2} = 10,961 \quad ,$$

$$F_{\varepsilon_2} = \left(1 - R_{\varepsilon_2}\right)N_{\varepsilon_2} = 0.680 \quad .$$



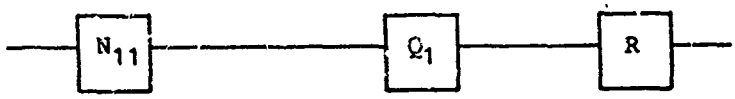
APPENDIX C

## ANALYSIS OF SUBSYSTEMS

The equivalent number of trials and the maximum-likelihood reliability estimate will be calculated for each of the 10 subsystems in the reduced schematic in figure 10 in the body of the report. The methodology used for each subsystem will be referred to by the appropriate section in the body of the report. The symbols N, F, R, and LCB will be used throughout to denote number of tests, number of failures, maximum likelihood reliability estimate, and 90-percent lower confidence bound, respectively.

### SUBSYSTEM 1

This subsystem is a series/parallel configuration consisting of repetitions of a single series combination:




$N_{N_{11}} = 143.25$ $N_{Q_1} = 190$ $N_R = 572$

$F_{N_{11}} = 0$ $F_{Q_1} = 0.3333$ $F_R = 0$

For this series,

N = 143.25
R = 0.99825
LCB = 0.98107 (by interpolation formula (2) in the body of the report

For subsystem 1 (using the method of sect. 2.5), we obtain

$$R_I = [1 - (1 - R)^2]^3[1 - (1 - R)^3] = 0.999991 ,$$

$$LCB_I = [1 - (1 - LCB)^2]^3[1 - (1 - LCB)^3] = 0.998918 ;$$

which leads to the results in the following table:

| Iteration | $N_1$ | $F_1$ | $B_{90}(N_1, F_1)$ | t | $N_2$ |
|---|---|---|---|---|---|
| 1 | 2127 | 0.0191 | 0.998905 | 1.010 | 2158 |

This gives the final data for subsystem 1:

$N_I = 2158$ ,

$R_I = 0.999991$ .

SUBSYSTEM 2

For the upper series:




$N_{N_{12}} = 143.25$ $\quad N_{Q_2} = 190$

$F_{N_{12}} = 0$ $\quad F_{Q_2} = 0.3333$

Equivalent N = 143.25

R = 0.99825

For the lower series:




$N_{S_1} = 83.33$ $\quad N_{V_{11}} = 95.75$ $\quad N_{Z_1} = 18.4$ $\quad N_{\alpha_{11}} = 80.2$

$F_{S_1} = 5$ $\quad F_{V_{11}} = 0.24925$ $\quad F_{Z_1} = 0$ $\quad F_{\alpha_{11}} = 0.0802$

Equivalent N = 18.4

R = 0.93662

For subsystem 2 (using the method of sect. 2.3), we obtain

$$Q = 0.00175 \times 0.06338 = 0.0001109 \ ,$$

$$Q' = 0.000968 \ ,$$

$$N_{II} = \frac{1 - Q'}{Q' - Q} = 1166 \ ,$$

$$R_{II} = 0.9998891 .$$

SUBSYSTEM 3

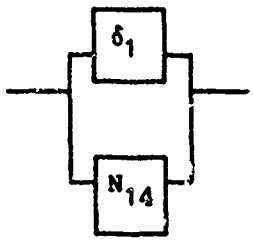
Subsystem 3 is the same as subsystem 2 except that S is replaced by T. A similar computation yields

$$N_{III} = 2040 \ ,$$

$$R_{III} = 0.9999884 \ .$$

SUBSYSTEM 4

For the parallel pair:




$N_{\delta_1} = 192$ $\quad N_{N_{14}} = 143.25$

$F_{\delta_1} = 0$ $\quad F_{N_{14}} = 0$

By the method of section 2.3, we obtain

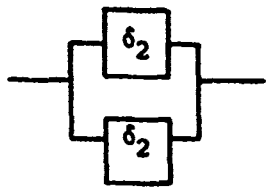
$$\text{Equivalent } N = 27838 \ ,$$
$$R = 1 \ .$$

Subsystem 4 is just this parallel pair repeated twice in series. By the approximation in equation (22), we have

$$N_{IV} = \frac{27,838}{2} = 13,919 \ ,$$
$$R_{IV} = 1 \ .$$

SUBSYSTEM 5

Subsystem 5 is just a single component repeated in parallel:




$N_{\delta_2} = 192$

$F_{\delta_2} = 0$

By the method of section 2.5 in the special case where R = 1, we have

$$LCB_{\delta_2} = 0.98799$$

$$R_V = 1 \quad ,$$

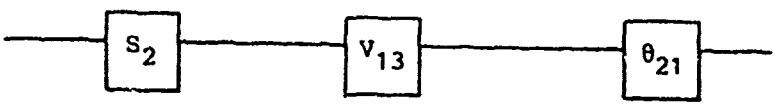
$$LCB_V = 1 - (0.01201)^2 = 0.999856 \quad ,$$

$$N_V = \frac{\ln 0.10}{\ln LCB_V} = 15,989 \quad ,$$

$$R_V = 1 \quad .$$

SUBSYSTEM 6

For the upper series:



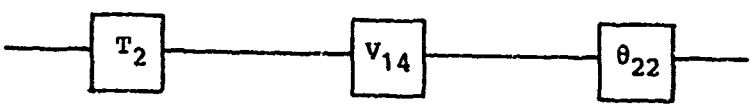

$N_{S_2} = 83.33$ $N_{V_{13}} = 47.875$ $N_{\theta_{21}} = 185.5$

$F_{S_2} = 5$ $F_{V_{13}} = 0.124625$ $F_{\theta_{21}} = 1.6565$

Equivalent N = 47.875

R = 0.92918

For the lower series:




$N_{T_2} = 109.33$ $N_{V_{14}} = 47.875$ $N_{\theta_{22}} = 185.5$

$F_{T_2} = 0.3333$ $F_{V_{14}} = 0.124625$ $F_{\theta_{22}} = 1.6565$

Equivalent N = 47.875

R = 0.98548

For subsystem 6 (using the method of sect. 2.3), we obtain

$$Q = 0.0010285 \ ,$$

$$Q' = 0.0031159 \ ,$$

$$N_{VI} = \frac{1 - Q'}{Q' - Q} = 478 \ ,$$

$$R_{VI} = 0.9989715 \ .$$

SUBSYSTEM 7

First the series repetition of $V_1$ is reduced:



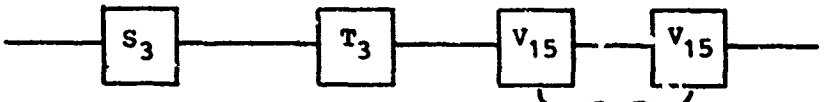

$$N_{V_{15}} = 95.75$$
$$F_{V_{15}} = 0.24925$$
$$R_{V_{15}} = 0.997397$$

By the approximation in equation (24) in the body of the report, we have

$$\text{Equivalent } N = \frac{F_{V_{15}}}{1 - R^2_{V_{15}}} = 47.9 \ ,$$

$$F = 0.24925 \ .$$

For the upper series:




$N_{S_3} = 83.33$ $N_{T_3} = 109.33$ $N = 47.9$

$F_{S_3} = 5$ $F_{T_3} = 0.3333$ $F = 0.24925$

Equivalent N = 47.9

R = 0.93226

For the lower series:




$N_{Z_3} = 18.4$ $\quad$ $N_{\alpha_{13}} = 80.2$

$F_{Z_3} = 0$ $\quad$ $F_{\alpha_{13}} = 0.0802$

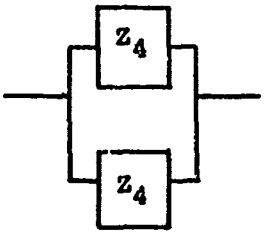
Equivalent $N = 18.4$

$R = 0.999$

For subsystem 7 (using the method of sect. 2.3), we obtain

$$Q = 0.00006774 \ ,$$
$$Q' = 0.0045571 \ ,$$
$$N_{VII} = 222 \ ,$$
$$R_{VII} = 0.9999323 \ .$$

SUBSYSTEM 8




$N_{Z_4} = 18.4$

$F_{Z_4} = 0$

By the method of section 2.5 with $R = 1$, we have

$$LCB_{Z_4} = 0.88230 \text{ by interpolation,}$$
$$R_{VIII} = 1 \ ,$$
$$LCB_{VIII} = 1 - \left(1 - LCB_{Z_4}\right)^2 = 0.98615,$$
$$N_{VIII} = 165 \ ,$$
$$R_{VIII} = 1 \ .$$

SUBSYSTEM 9

For the upper series:




$$N_{\alpha_{14}} = 80.2$$
$$F_{\alpha_{14}} = 0.0802$$
$$R_{\alpha_{14}} = 0.999$$

By the approximation in equation (24) in the body of the report, we have

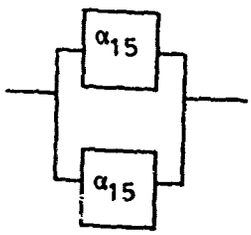
$$\text{Equivalent } N = \frac{F_{\alpha_{14}}}{1 - R^2_{\alpha_{14}}} = 40.1 \quad ,$$

$$F = 0.0802 \quad .$$

For subsystem 9 (using the method of sect. 2.3), we combine the upper series in parallel with $Z_5$, which has $N_{Z_5} = 18.4$ and $F_{Z_5} = 0$, and obtain

$$Q = 0 \quad ,$$
$$Q' = 0.0013548 \quad ,$$
$$N_{IX} = 737 \quad ,$$
$$R_{IX} = 1 \quad .$$

SUBSYSTEM 10




$$N_{\alpha_{15}} = 80.2$$

$$F_{\alpha_{15}} = 0.0802$$

By the method of section 2.5 we have

$$R_{\alpha_{15}} = 0.999$$

$$LCB_{\alpha_{15}} = 0.97011 \text{ by interpolation}$$

$$R_X = 1 - \left(1 - R_{\alpha_{15}}\right)^2 = 0.999999 \ ,$$

$$LCB_X = 1 - \left(1 - LCB_{\alpha_{15}}\right)^2 = 0.999107 \ .$$

The method of section 2.4 is then used to get equivalent test data, as follows.

| Iteration | $N_1$ | $F_1$ | $B_{90}(N_1, F_1)$ | t | $N_2$ |
|---|---|---|---|---|---|
| 1 | 2577 | 0.002577 | 0.999104 | 1.003 | 2586 |

These results lead to the following data for subsystem 10:

$$N_X = 2586 \ ,$$

$$R_X = 0.999999 \ .$$

# B-SPLINES ON NONUNIFORM TRIANGULATIONS

Charles K. Chui
Center for Approximation Theory
Department of Mathematics
Texas A&M University
College Station, Texas 77843

ABSTRACT. $C^1$ B-splines of lowest total degrees on triangulations of nonuniform rectangular partitions are considered. Several interesting properties are discovered for the type-one setting. In particular, there are B-splines, that is splines with minimum supports, whose supports are concave, and B-splines do not necessarily form a partition of unity.

1. INTRODUCTION. Consider a rectangular region $D = [a,b] \times [c,d]$ and let the lines $x - x_i = 0$ and $y - y_j = 0$, where $a = x_0 < \ldots < x_{m+1} = b$ and $c = y_0 < \ldots < y_{n+1} = d$, partition $D$ into $(m+1)(n+1)$ rectangular cells $D_{ij}$. By drawing in all the upward sloping diagonals of these rectangles, we obtain a unidiagonal (or type-1) triangulation of $D$, and by drawing in both diagonals to each $D_{ij}$ we have a crisscross (or type-2) triangulation. Of course, in the special case when $x_{i+1} - x_i = x_i - x_{i-1}$ and $y_{j+1} - y_j = y_j - y_{j-1}$ for $i = 1,\ldots,m$ and $j = 1,\ldots,n$, these triangulations of $D$ become 3-directional and 4-directional meshes respectively, and many interesting results in this special setting have been obtained recently (cf. [1], [2], [4], [7], and [8], for instance). This article is a continuation of [3] where the results on nonuniform crisscross triangulation obtained in [6] were reported. We take this opportunity to report two misprints in [3], namely: on page 879, $B_j$ should be $(y_j - y_{j-1})/(y_{j+1} - y_{j-1})$ and the identity on page 881 should read

$$\sum_{j=-1}^{n+1} \sum_{i=-1}^{m+1} (-1)^{i+j}(x_{i+1} - x_i)(y_{j+1} - y_j)B_{ij}(x,y) = 0.$$

Of course in the equally spaced setting, i.e. on a 4-directional mesh, this linear dependence relationship becomes

$$\sum_{j=-1}^{n+1} \sum_{i=-1}^{m+1} (-1)^{i+j}B_{ij}(x,y) = 0$$

and this special case has recently been generalized by Dahmen and Micchelli [8] to a more general regular grid partition.

PREVIOUS PAGE IS BLANK

One very nice property of the B-splines $B_{ij}$ on crisscross triangulations reported in [3] is that they are very flexible, in the sense that their graphs are continuous with respect to the lines $x - x_i = 0$ and $y - y_j = 0$ as shown in the pictures on pages 880 and 881 in [3]. This property is not enjoyed by the $C^1$ cubic splines on unidiagonal triangulations as observed in [5]. More results have now been found and it is the purpose of this article to report some of the unusual properties these smooth cubic splines on unidiagonal triangulations.

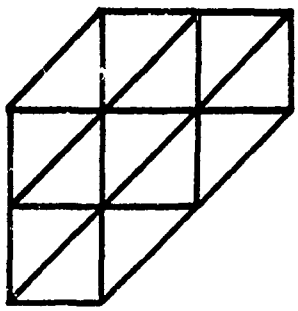


Fig. 1a

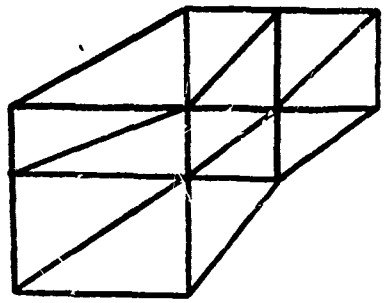


Fig 1b

2. SMOOTH CUBIC B-SPLINES. In this section we only consider unidiagonal triangulation for the rectangle D as described above. Inthe equally spaced setting, i.e. on a 3-directional mesh, there are two $C^1$ cubic B-splines where one is a 180°-rotation of the other. The support of one of them is given in Fig. 1a. In general, the grid configuration of this support would look like the one shown in Fig. 1b. Let us call this grid configuration $\Delta_{ij}$. It was observed in [5] that $\Delta_{ij}$ is the support of a nontrivial bivariate $C^1$ cubic spline if and only if

$$\frac{(x_{i+1} - x_i)^2}{(x_{i+2} - x_{i+1})(x_i - x_{i-1})} = \frac{(y_{j+1} - y_j)^2}{(y_{j+2} - y_{j+1})(y_j - y_{j-1})} = 1. \qquad (1)$$

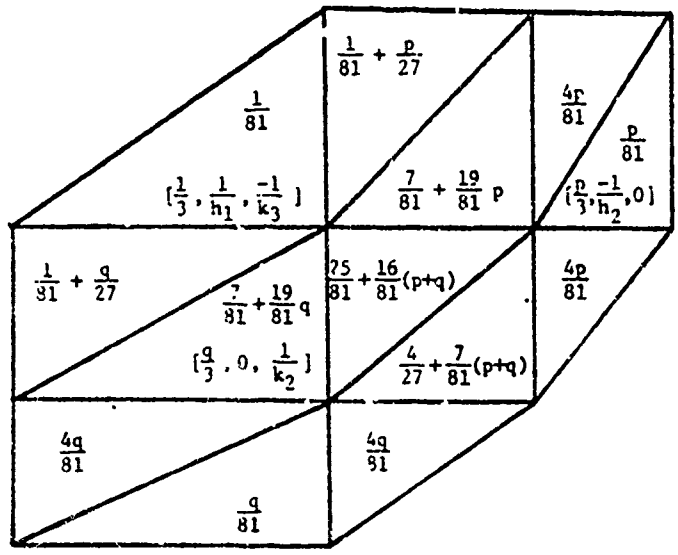



Fig. 2

It is clear that a condition equivalent to (1) is that $\{x_s - x_{s-1}\}$, $s = i, i+1, i+2$ and $\{y_t - y_{t-1}\}$, $t = j, j+1, j+2$, are geometric progressions. Suppose that this condition is satisfied and let

$$p = \frac{x_{i+1} - x_i}{x_i - x_{i-1}} = \frac{x_{i+2} - x_{i+1}}{x_{i+1} - x_i} \quad \text{and} \quad q = \frac{y_j - y_{j-1}}{y_{j+1} - y_j} = \frac{y_{j+1} - y_j}{y_{j+2} - y_{j+1}}.$$

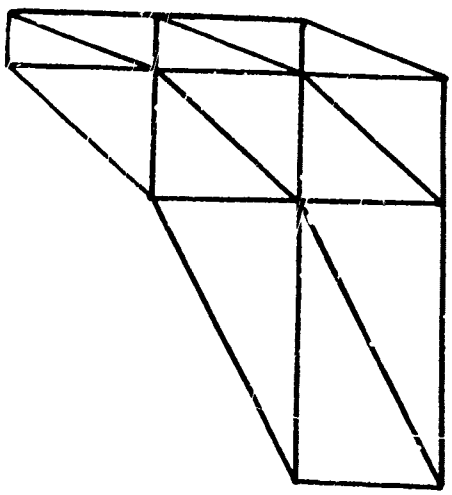


Fig 3a

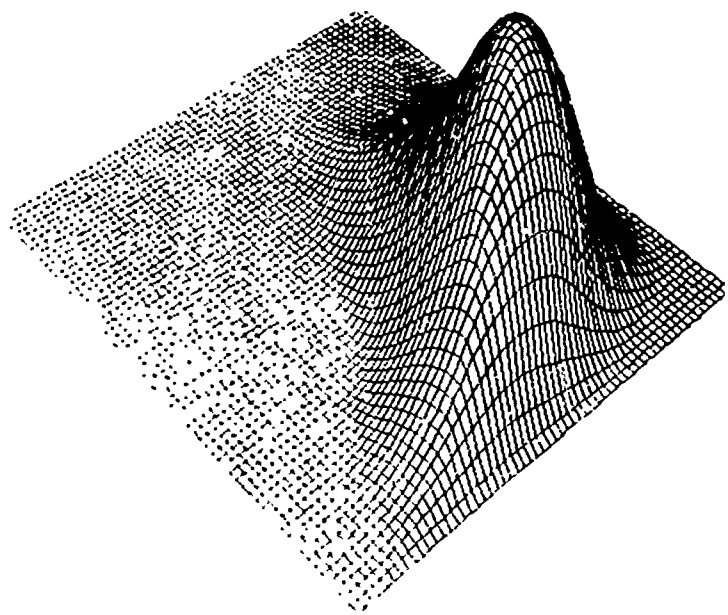


Fig 3b

It [6] the "unique" cubic B-spline supported by $\Delta_{ij}$ was constructed as shown in Fig. 2. We remark here that $\Delta_{ij}$ may happen to be concave. A simple example is given by setting $p = \frac{7}{3}$ and $q = 1$ as shown in Fig. 3a and a picture of the B-spline it supports is given in Fig 3b. Suppose for the time being that both the horizontal and vertical spacings of the entire original rectangular grid partition of $D$ are geometric progressions. It is perhaps surprising to note that these B-splines do not necessarily produce constants. More precisely, if one of the geometric ratios $p$ and $q$ is different from one, then the constant 1 is not a linear combination of these B-splines.

If the lines $x - x_i = 0$ and $y - y_j = 0$ are arbitrarily given, then the supports of the bivariate $C^1$ cubic B-splines, if they exist, have to increase. In fact the sizes and shapes of the B-splines depend very heavily on the sequences $\{x_{i+1} - x_i\}$ and $\{y_{j+1} - y_j\}$. These and other related results will be discussed in a forthcoming joint paper with my student G. Chen.

## REFERENCES


1. C. de Boor and K. Höllig, Bivariate box splines and smooth pp functions on a three direction mesh, J. Comp. Appl. Math. 9 (1983), 13-28.

2. C. de Boor and K. Höllig, Bivariate splines of minimal support on regular meshes, in manuscript.

3. C. K. Chui, Bivariate quadratic splines on crisscross triangulations, Trans. of First Army Conf. on Appl. Math. and Comp., 1983, pp. 877-882.

4. C. K. Chui and R. H. Wang, Spaces of bivariate cubic and quartic splines on type-1 triangulations, J. Math. Anal. and Appl. 101 (1984).

5. C. K. Chui and R. H. Wang, Bivariate B-splines on triangulated rectangles, in Approximation Theory IV, Ed. by C. K. Chui, L. L. Schumaker, and J. D. Ward, Academic Press, N.Y. 1983, pp. 413-418.

6. C. K. Chui and R. H. Wang, Concerning $C^1$ B-splines on triangulations of non-uniform rectangular partition, J. Approx. Theory and Its Appl., To appear.

7. W. Dahmen and C. A. Micchelli, Recent progress in multivariate splines, in Approximation Theory IV, Ed. by C. K. Chui, L. L. Schumaker, and J. D. Ward, Academic Press, N. Y. 1983, 27-121.

8. W. Dahmen and C. A. Micchelli, On the solution of certain systems of partial difference equations and linear dependence of translates of box splines, CAT Report #52, Texas A&M University, 1984.

# A MODEL FOR ASYNCHRONOUS DISTRIBUTED COMPUTATION*

Dimitri P. Bertsekas
Laboratory for Information and Decision Systems
Department of Electrical Engineering and Decision Systems
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT. We present an algorithmic model for distributed computation of fixed points whereby several processors participate simultaneously in the calculations while exchanging information via communication links. We place essentially no assumptions on the ordering of computation and communication between processors thereby allowing for completely uncoordinated execution. We find that even under these potentially chaotic circumstances it is possible to solve several important classes of problems including the calculation of fixed points of contraction and monotone mappings arising in linear and nonlinear systems of equations, shortest path problems, and dynamic programming.

1. INTRODUCTION. There is presently a great deal of interest in distributed implementations of various iterative algorithms whereby the computational load is shared by several processors while coordination is maintained by information exchange via communication links. In most of the work done in this area the starting point is some iterative algorithm which is guaranteed to converge to the correct solution under the usual circumstances of centralized computation in a single processor. The computational load of the typical iteration is then divided in some way between the available processors, and it is assumed that the processors exchange all necessary information regarding the outcomes of the current iteration before a new iteration can begin.

The mode of operation described above may be termed synchronous in the sense that each processor must complete its assigned portion of an iteration and communicate the results to every other processor before a new iteration can begin. This assumption certainly enhances the orderly operation of the algorithm and greatly simplifies the convergence anaysis. On the other hand synchronous distributed algorithms also have some obvious disadvantages such as the need for an algorithm initiation and iteration synchronization protocol. Furthermore the speed of computation is limited to that of the slowest processor. It is thus interesting to consider algorithms that can tolerate a more flexible ordering of computation and communication between processors. Such algorithms have so far found applications in computer communication networks like the ARPANET [1] where processor failures are common and it is quite complicated to maintain synchronization between the nodes of the entire network as they execute real-time twork functions such

as the routing algorithm.

Processor network environments for which weakly coordinated distributed computation seems particularly advantageous typically possess one or more of the following characteristics all of which involve occurance of some type of unpredictable event.

1) Computation nodes and communication links are subject to frequent and/or unexpected failures. (For example packet radio networks.)

2) Computation nodes have different and/or time varying speeds of execution. (For example each processor is assigned to a perhaps time varying number of tasks involving computation loads which are not fixed a priori.)

3) Computation at various nodes is event driven. (For example in data collection or sensor networks where the timing and ordering of measurements may not be predictable.)

It is possible to consider various degrees of coordination in different types of distributed algorithms. An interesting question is to determine the minimum degree of coordination needed in a given algorithm in order to obtain the correct solution. To this end we consider an extreme model of uncoordinated distributed algorithms whereby computation and communication are performed at each processor completely independently of the progress in other processors. It is perhaps surprising that even under these chaotic circumstances it is still possible to solve correctly important classes of fixed point problems. The complete analysis is given in [2] for broad classes of dynamic programming and in [3] for more general fixed point problems involving contraction and monotonicity assumptions. Further related work is [5] and [6].

## 2. A Model for Distributed Uncoordinated Fixed Point Algorithms

The fixed point problem considered in this paper is defined in terms of a set $X$, a class $F$ of functions mapping $X$ into the extended real line $[-\infty,+\infty]$, and a mapping $T$ which maps $F$ into itself. We wish to find an element $J^*$ of $F$ such that

$$J^* = T(J^*) \tag{1}$$

or equivalently

$$J^*(x) = T(J^*)(x), \quad \forall\, x \in X, \tag{2}$$

where J*(x) and T(J*)(x) denote the values of the functions J* and T(J*) respectively at the typical element x∈X. We will assur throughout that T has a unique fixed point J* within the set F.

We provide some examples:

Example 1: (Fixed points of mappings on $R^n$). Let X be the finite set

$$X = \{1,2,\ldots,n\},$$

and F be the set of all real-valued functions on X. Then F can be identified with the n-dimensional space $R^n$ in the sense that with each J∈F we can associate the n-dimensional vector (J(1), J(2),...,J(n)). Similarly T(J) can be identified with the n-dimensional vector (T(J)(1),...,T(J)(n)), so the fixed point problem (1) amounts to solving the system of n equations

$$J^* = T(J^*) \text{ or } J^*(i) = T(J^*)(i), \quad \forall\, i = 1,\ldots,n \qquad (3)$$

with the n unknowns J*(1),...,J*(n). It is also evident that any system of n (possibly nonlinear) equations with n unknowns can be formulated into a fixed point problem such as (3).

Example 2: (Shortest path problems). Let (**N**,**L**) be a directed graph where **N** = (1,2,...,n) denotes the set of nodes and **L** denotes the set of links. Let N(i) denote the downstream neighbors of node i, i.e., the set of nodes j for which (i,j) is a link. Assume that each link (i,j) is assigned a positive scalar $a_{ij}$ referred to as its length. Assume also that there is a directed path to node 1 from every other node. Then it is known ([4], p. 67) that the shortest path distances J*(i) to node 1 from all other nodes i solve uniquely the equations

$$J^*(i) = \min_{j \in N(i)} \{a_{ij} + J^*(j)\}, \quad i \neq 1 \tag{4a}$$

$$J^*(1) = 0 \tag{4b}$$

If we make the identifications $X = \{1,2,\ldots,n\}$, $F$: Set of all functions mapping $X$ into $[0,+\infty]$, and define $T(J)$ for all $J \in F$ by means of

$$T(J)(i) = \begin{cases} \min_{j \in N(i)} \{a_{ij} + J(j)\} & \text{if } i \neq 1 \\ 0 & \text{if } i = 1 \end{cases} \tag{5}$$

then we find that the fixed point problem (2) reduces to the shortest path problem.

The shortest path problem above is representative of a broad class of dynamic programming problems which can be viewed as special cases of the fixed point problem (2) and can be correctly solved by using the distributed algorithms of this paper (see [3]).

Our algorithmic model can be described in terms of a collection of n computation centers (or processors) referred to as nodes and denoted 1,1,...,n. The set $X$ is partitioned into n disjoint sets denoted $X_1,\ldots,X_n$, i.e.

$$X = \bigcup_{i=1}^{n} X_i, \quad X_i \cap X_j = \emptyset, \quad \text{if } i \neq j.$$

Each node i is assigned the responsibility of computing the values of the solution function $J^*$ [c.f. (1), (2)] at all $x \in X_i$.

At each time instant, node i can in be one of three possible states compute, transmit, or idle. In the compute state node i computes a new estimate of the values of the solution function $J^*$ for all $x \in X_i$. In the transmit state node i communicates the estimate obtained from the latest computation to one or more nodes j ($j \neq i$). In the idle state node i does nothing related to

the solution of the problem. It is assumed that a node can receive a transmission from other nodes simultaneously with computing or transmitting. We assume that computation and transmission for each node takes place in uninterrupted time intervals $[t_1,t_2]$ with $t_1 < t_2$, but do not exclude the possibility that a node may be simultaneously transmitting to more than one nodes nor do we assume that the transmission intervals to these nodes have the same origin and/or termination. We also make no assumptions on the length, timing and sequencing of computation and transmission intervals other than the following:

Assumption (A): There exists a positive scalar P such that, for every node i, every time interval of length P contains at least one computation interval for j and at least one transmission interval from i to each node $j \neq i$.

Each node i also has a buffer $B_{ij}$ for each $j \neq j$ where it stores the latest transmission from j, as well as a buffer $B_{ii}$ where it stores its own estimate of values of the solution function for all $x \in X_i$. The contents of each buffer $B_{ij}$ at time t are denoted $J^t_{ij}$. Thus $J^t_{ij}$ is, for every t, a function from $X_j$ into $[-\infty,\infty]$ and may be viewed as the estimate by node i of the restriction of the solution function J* on $X_j$ available at time t. The rules according to which the functions $J^t_{ij}$ are updated are as follows:

1) If $[t_1,t_2]$ is a transmission interval from node j to node i the contents $J^{t_1}_{jj}$ of the buffer $B_{jj}$ at time $t_1$ are transmitted and entered in the buffer $B_{ij}$ at time $t_2$, i.e.

$$J^{t_2}_{ij} = J^{t_1}_{jj}. \tag{6}$$

2) If $[t_1,t_2]$ is a computation interval for node i the contents of buffer $B_{ii}$ at time $t_2$ are replaced by the restriction of the function $T(J^t_i)$ on $X_i$ where, for all t, $J^t_i$ is defined by

$$J^t_i(x) = \begin{cases} J^t_{ii}(x) & \text{if } x \in X_i \\ J^t_{ij}(x) & \text{if } x \in X_j, \quad j \neq i. \end{cases} \tag{7}$$

In other words we have

$$J^{t_2}_{ii}(x) = T(J^{t_1}_i)(x), \quad \forall\, x \in X_i. \tag{8}$$

3) The contents of a buffer $B_{ii}$ can change only at the end of a computation interval for node i. The contents of a buffer $B_{ij}$, $i \neq i$ can change only at the end of a transmission interval from j to i.

Additional conditions under which there holds

$$\lim_{t\to\infty} J_i^t(x) = J^*(x), \quad \forall\, x\in X_i, \quad i = 1,\ldots,n \tag{9}$$

may be found in [2], [3]. An interesting aspect of results of this type is that they do not require that the initial processor buffer contents be identical and indeed these initial conditions can vary within a broad range. This means that for problems that are being solved continuously in real time it is not necessary to reset the initial conditions and resynchronize the algorithm each time the problem data changes. As a result the potential for tracking slow variations on the solution function is improved and algorithmic implementation is considerably simplified.

---

*This research was conducted at the M.I.T. Laboratory for Information and Decision Systems with partial support provided by the Defense Advanced Projects Agency under Contract No. ONR-N00014-75-C-1183 and Army Research Office under Contract No. DAAG29-84-K-0005.

References

[1] J. McQuillan, G. Falk, and I. Richer, "A Review of the Development and Performance of the ARPANET Routing Algorithm". IEEE Trans. on Communications, Vol. COM-26, 1978, pp. 1802-1811.

[2] D. P. Bertsekas, "Distributed Dynamic Programming", IEEE Trans. on Automatic Control, Vol. AC-27, 1982, pp. 610-616.

[3] D. P. Bertsekas, "Distributed Computation of Fixed Points. Math. Programming, Vol. 27, 1983, pp. 107-120.

[4] E. L. Lawler, Combinatorial Optimization: Networks and Matriods, Holt, Rinehart, and Winston, N.Y., 1976.

[5] D. P. Bertsekas, J. N. Tsitsiklis, and M. Athans, "Convergence Theories of Distributed Iterative Processess: A Survey", LIDS Report P-1342, M.I.T., Dec. 1983.

[6] J. N. Tsitsiklis, "Problems in Distributed Decision Making", Ph.D. Thesis, Dept. of Electrical Engineering and Computer Science, Mass. Institute of Technology.

Second Army Conference
on
Applied Mathematics and Computing

May 22-25, 1984

Rensselaer Polytechnic Institute
Troy, New York 12181

LIST OF PARTICIPANTS

Peter Alfeld
Mathematics Research Center
University of Wisconsin-Madison

Professor W. F. Ames
School of Mathematics
Georgia Institute of Technology

Dr. Charles M. Bowden
US Army Missile Command
Redstone Arsenal, Alabama

Harold J. Breaux
US Army Ballistics Research Lab
Aberdeen Proving Ground, Maryland

Paul H. Broome
US Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

Alfred S. Carasso
Center for Applied Mathematics
National Bureau of Standards
Washington, DC

Dr. Ben L. Carnes
US Army Eng. Waterways Exper. Sta
Vicksburg, MS

Dr. Garry Carofano
Benet Weapons Laboratory
Watervliet Arsenal

Dr. Alvars Celmins
US Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

Jagdish Chandra
US Army Research Office
Mathematical Sciences Division
Research Triangle Park, NC

Peter C. T. Chen
Benet Weapons Laboratory
Watervliet Arsenal

P. L. Chow
Department of Mathematics
Wayne State University

A. Brinton Cooper, III
U.S. Army Ballistic Research Lab.
Aberdeen Proving Ground, Maryland

Eugene Coppola
Benet Weapons Laboratory
Watervliet Arsenal

Donald Drew
Dept. of Mathematical Sciences
RPI

Walter O. Egerland
Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland

Dr. Paul D. Fedele
Research Division
Chemical Research & Devel. Center
Aberdeen Proving Ground, Maryland

Allan Finkel
IBM-T.J. Watson Research Center
Yorktown Heights, New York

Barry D. Fishburn
Armament Research & Devel. Center
US Army Armament, Munitions &
Chemical Command
Dover, New Jersey

Joseph E. Flaherty
Department of Mathematical Sciences
R.P.I.

Ferdinand Freudenstein
Department of Mechanical Engng.
Columbia University

Anthony Gabriele
Benet Weapons Laboratory
Watervliet Arsenal

Second Army Conference on Applied Mathematics and Computing

## LIST OF PARTICIPANTS

James Glimm
Courant Institute
NYU

Dennis Grady
Sandia National Labs
Albuquerque, New Mexico

Mort Gurtin
Mathematics Department
Carnegie Mellon University

Edward Haug
Center for Computer Aided Design
The University of Iowa

Dr. Kenneth Heaton
Defense Research Estab. Valcartier
Courcelette, Quebec, Canada

Dr. Rudi Heiser
Fraunhofer-Institut fur
Kurzzeitdynamik

Ernst-Mach-Institut
West Germany

Th. Herbert
Virginia Polytechnic Institute
and State University

Professor John E. Hopcroft
Computer Science Department
Cornell University

George W. Hoppe
National Guard Bureau
Army Comptroller Division
Pentagon, Washington, DC

John W. Hutchinson
Division of Applied Sciences
Harvard University

William Jackson
U.S. Army Tank-Automotive Command
Warren, MI

Daniel D. Joseph
University of Minnesota

H. T. Kung
Department of Computer Science
Carnegie-Mellon University

Charles R. Leake
US Army Concepts Analysis Agency
Bethesda, MD

P. LeTallec
Mathematics Research Center
University of Wisconsin

Sung P. Lin
Mathematics Department
Clarkson University

Geoffrey S. S. Ludford
Cornell University

Yan-Chow Ma
TRW Space & Technology Group
Redondo Beach, CA

Olvi L. Mangasarian
Mathematics Research Center
University of Wisconsin

Richard Meyer
Mathematics Research Center
University of Wisconsin

Toshio Mura
Department of Civil Engineering
Northwestern University

Alan Needleman
Brown University

Bart S. Ng
Department of Mathematical Sciences
RPI

John Nohel
Mathematics Research Center
University of Madison-Wisconsin

Kevin C. Nunan
IBM
T. J. Watson Research Center
Yorktown Heights, New York

Peter O'Hara
Benet Weapons Laboratory
Watervliet Arsenal

R. E. O'Malley, Jr.
Department of Mathematical Sciences
RPI

## LIST OF PARTICIPANTS

Thomas J. Pence
Mathematics Research Center
University of Wisconsin-Madison

George A. Pflegl
Benet Weapons Laboratory
Watervliet Arsenal

Colonel Jack Pollin
Department of Mathematics
United States Military Academy
West Point, New York

San-Li Pu
Benet Weapons Laboratory
Watervliet Arsenal

Ronald L. Racicot
Benet Weapons Laboratory
Watervliet Arsenal

Professor Louis B. Rall
Mathematics Research Center
University of Madison-Wisconsin

Louise A. Raphael
Howard University

Professor J. N. Reddy
Engineering Science & Mechanics
Virginia Polytechnic & State Uni

Harry L. Reed, Jr.
Ballistic Research Laboratory
System Engineering & Concepts
Aberdeen Proving Ground, MD

Michael Renardy
Mathematics Research Center
University of Wisconsin-Madison

Christian A. Ringhofer
Mathematics Research Center
University of Wisconsin-Madison

Dr. Edward W. Ross
US Army Natick R&D Center
Natick, MA

Edward Saibel
U.S. Army Research Office
Research Triangle Park, NC

James A. Schmitt
Ballistic Research Laboratory
Aberdeen Provind Ground, MD

Harvey Segur
ARAP
Princeton, NJ

T.-L. Sham
Dept. of Mechanical Engineering
RPI

C. N. Shen
Benet Weapons Laboratory
Watervliet Arsenal

George C. Sih
Lehigh University

Thomas E. Simkins
Benet Weapons Laboratory
Watervliet Arsenal

Royce Soanes
Benet Weapons Laboratory
Watervliet Arsenal

Ram P. Srivastav
Dept. of Applied Math. & Stat.
State Univ. of NY at Stony Brook

D. M. Swingle
Consultant
Las Cruces, NM

Iradj Tadjbakhsh
RPI

Alexander Tessler
Army Materials & Mechanics
Reseach Center
Watertown, MA

T. C. T. Ting
University of Illinois at Chicago

Dr. John D. Vasilakis
Benet Weapons Laboratory
Watervliet Arsenal

## LIST OF PARTICIPANTS

Roger A. Wehage
US Army Tank-Automotive Command
Warren, MI

Arthur Wouk
Army Research Office
Research Triangle Park, NC

Norman J. Zabusky
Department of Mathematics
University of Pittsburgh
Pittsburgh, PA